Online Spectrogram Inversion for Low-Latency Audio Source Separation

Paul Magron, Tuomas Virtanen

Online Spectrogram Inversion for Low-Latency Audio Source Separation

Paul Magron, Tuomas Virtanen, *Senior Member, IEEE*

***Abstract*—Audio source separation is usually achieved by estimating the short-time Fourier transform (STFT) magnitude of each source, and then applying a spectrogram inversion algorithm to retrieve time-domain signals. In particular, the multiple input spectrogram inversion (MISI) algorithm has been exploited successfully in several recent works. However, this algorithm suffers from two drawbacks, which we address in this paper. First, it has originally been introduced in a heuristic fashion: we propose here a rigorous optimization framework in which MISI is derived, thus proving the convergence of this algorithm. Besides, while MISI operates offline, we propose here an online version of MISI called oMISI, which is suitable for low-latency source separation, an important requirement for e.g., hearing aids applications. oMISI also allows one to use alternative phase initialization schemes exploiting the temporal structure of audio signals. Experiments conducted on a speech separation task show that oMISI performs as well as its offline counterpart, thus demonstrating its potential for real-time source separation.**

***Index Terms*—Audio source separation, low-latency, online spectrogram inversion, phase recovery, sinusoidal modeling.**

I. Introduction

Audio source separation [1] consists in extracting the underlying *sources* that add up to form an observable audio *mixture*. This task finds applications in many areas such as speech enhancement and recognition [2], musical signal processing [3] and hearing aid devices [4]. In particular, hearing aids require a very low processing latency, as significant discomfort can be experienced by listeners for delays exceeding 20 ms [5]. State-of-the-art 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) [6]. Recent works such as [7], [8] operate in the time domain directly, but TF approaches remain interesting since they make it possible to better exploit the structure of sound.

Applying a nonnegative mask to the input STFT results in assigning the mixture's phase to each isolated source. Even though this yields somewhat satisfactory results in practice, it has been pointed out [9] that when sources overlap in the TF domain, using the mixture's phase induces residual interference and artifacts in the estimates. With the advent of deep learning, magnitudes can nowadays be estimated with a high accuracy, which highlights the need for advanced phase recovery [10], [11], a problem hereafter termed *spectrogram inversion*. Consequently, several recent works have focused on phase recovery in DNN-based source separation, whether phase recovery algorithms are applied as a post-processing [12], [13] or integrated within end-to-end systems for time-domain separation [14], [15], [16], [17], [18].

Among the variety of phase recovery techniques, the multiple input spectrogram inversion (MISI) algorithm [19] is particularly popular. This iterative procedure consists in retrieving time-domain sources from their STFT magnitudes while respecting a mixing constraint: the estimates must add up to the original mixture. This algorithm has shown promising for audio source separation when combined with DNNs [12], [14]. However, the MISI algorithm has been introduced in a heuristic fashion, therefore there is currently no proof that it converges. Besides, while several recent works addressed the problem of low-latency magnitude estimation [20], [21], [22], the MISI algorithm operates offline, as it computes the whole STFT and its inverse at each iteration. This makes it impracticable for real-time applications such as hearing aids.

In this paper, we investigate and overcome the drawbacks of the MISI algorithm. First, we propose a rigorous optimization framework for spectrogram inversion. Using the auxiliary function method we derive a procedure that is equivalent to the MISI algorithm. To the best of our knowledge, this is the first proof of convergence of MISI. Second, we propose an online adaptation of MISI that is suitable for low-latency applications. This adaptation is based on approximating the STFT and its inverse in a causal fashion by only accounting for the past context and for an arbitrarily small number of future frames. It also allows us to exploit the temporal structure of the phase to use alternative phase initialization schemes. Here, we propose to use a sinusoidal phase [23] instead of the mixture's phase as initial estimate. We experimentally demonstrate the potential of this technique for a speech separation task.

The rest of this paper is structured as follows. Section II presents some mathematical notations. The MISI algorithm is derived in Section III and adapted online in Section IV. Section V presents the experimental results, and Section VI concludes the paper.

II. Mathematical notations

- $\mathbf{A}$ (capital, bold font): matrix.
- $\tilde{\mathbf{x}}$ (lower case, bold font, with tilde): time-domain signal, whose n-th sample is denoted $\tilde{\mathbf{x}}(n)$.
- $\mathbf{x}$ (lower case, bold font, without tilde): TF domain vector, whose m-th entry is denoted $\mathbf{x}(m)$.
- z (regular): scalar.
- $|.|$, $\angle(.)$: magnitude and complex angle, respectively.

P. Magron is with IRIT, Université de Toulouse, CNRS, France (paul.magron@irit.fr). T. Virtanen is with the Audio Research Group, Tampere University, Finland (tuomas.virtanen@tuni.fi). The work of P. Magron was conducted while he was with Tampere University and supported by the Academy of Finland, project no. 290190.

- $\mathbf{x}^\mathsf{T}$, $\mathbf{x}^\mathsf{H}$: transpose and Hermitian transpose, respectively.
- $\Re(.)$: real part function.
- $||.||$: Euclidean norm.
- $\odot$, $\dot{\div}$: element-wise matrix or vector multiplication, and division, respectively.

III. Algorithm derivation

In this section we derive MISI using the auxiliary function method, which proves the convergence of this algorithm.

A. Problem setting

Let us consider an instantaneous and linear mixing model:

$$\tilde{\mathbf{x}} = \sum_{j=1}^{J} \tilde{\mathbf{s}}_j, \tag{1}$$

where $\tilde{\mathbf{x}} \in \mathbb{R}^N$ is the mixture, $\tilde{\mathbf{s}}_j \in \mathbb{R}^N$ are the J sources, and N denotes the length of the time-domain signals in samples. The STFTs of the sources are denoted by $\mathbf{s}_j \in \mathbb{C}^M$ with $M = F \times T$, where F and T are the number of frequency channels and time frames respectively, and we have

$$\mathbf{s}_j = \mathbf{A}\tilde{\mathbf{s}}_j, \tag{2}$$

where the matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$ encodes the STFT operation.

Let us assume that some STFT magnitude estimates denoted $\mathbf{v}_j \in \mathbb{R}_+^M$ are available (e.g., estimated beforehand using a DNN). The goal of multiple-input spectrogram inversion is to estimate time domain source signals $\tilde{\mathbf{s}}_j$ given the STFT magnitude estimates $\mathbf{v}_j$. Since these magnitude estimates are usually not equal to the ground truth, one should allow the magnitudes of the estimated sources to deviate from those values. Thus, we consider the following objective function:

$$\psi(\tilde{\mathbf{s}}) = \sum_j |||\mathbf{A}\tilde{\mathbf{s}}_j| - \mathbf{v}_j||^2, \tag{3}$$

We treat the mixing constraint (1) as a hard constraint, leading to the following optimization problem:

$$\min_{\tilde{\mathbf{s}}} \psi(\tilde{\mathbf{s}}) \text{ subject to } \sum_j \tilde{\mathbf{s}}_j = \tilde{\mathbf{x}}. \tag{4}$$

Directly addressing the optimization problem (4) is however difficult since ψ is not differentiable with respect to $\tilde{\mathbf{s}}$. Therefore, we propose to use the auxiliary function method [24], which consists in finding a function $\psi^+(\tilde{\mathbf{s}}, \mathbf{y})$ such that

$$\psi(\tilde{\mathbf{s}}) = \min_{\mathbf{y}} \psi^+(\tilde{\mathbf{s}}, \mathbf{y}), \tag{5}$$

where $\mathbf{y}$ is a set of auxiliary parameters. It can be shown [24] that ψ is non-increasing when alternating the minimization of ψ^+ with respect to (w.r.t.) $\tilde{\mathbf{s}}$ and $\mathbf{y}$.

B. Auxiliary function

Let us first introduce a set of auxiliary parameters $\mathbf{y}_j$ such that $|\mathbf{y}_j| = \mathbf{v}_j$ and rewrite (3) as:

$$\psi(\tilde{\mathbf{s}}) = \sum_j |||\mathbf{A}\tilde{\mathbf{s}}_j| - |\mathbf{y}_j|||^2. \tag{6}$$

We then use the property

$$\forall (z, z') \in \mathbb{C}^2,\ ||z| - |z'|| \leq |z - z'|, \tag{7}$$

that arise from the triangle inequality, and where equality holds if and only if $\angle z = \angle z'$. This leads to $\psi(\tilde{\mathbf{s}}) \leq \psi^+(\tilde{\mathbf{s}}, \mathbf{y})$ with:

$$\psi^+(\tilde{\mathbf{s}}, \mathbf{y}) = \sum_j ||\mathbf{A}\tilde{\mathbf{s}}_j - \mathbf{y}_j||^2. \tag{8}$$

In order to minimize ψ^+ w.r.t. $\mathbf{y}$ under the constraint $|\mathbf{y}_j| = \mathbf{v}_j$, we introduce this constraint using the Lagrange multipliers method. We therefore aim at finding a saddle point for:

$$\psi^+(\tilde{\mathbf{s}}, \mathbf{y}) + \sum_{j,m} \boldsymbol{\lambda}_j(m)(|\mathbf{y}_j(m)|^2 - \mathbf{v}_j(m)^2), \tag{9}$$

where $\boldsymbol{\lambda}_j \in \mathbb{R}^M$ are the Lagrange multipliers. We set the partial derivative of (9) w.r.t $\mathbf{y}$ at 0, which leads to:

$$(1 + \boldsymbol{\lambda}_j(m))\mathbf{y}_j(m) = \mathbf{s}_j(m). \tag{10}$$

Using the constraint $|\mathbf{y}_j(m)| = \mathbf{v}_j(m)$, we have

$$|1 + \boldsymbol{\lambda}_j(m)| = \frac{|\mathbf{s}_j(m)|}{\mathbf{v}_j(m)}. \tag{11}$$

Finally, injecting (11) into (10) leads to the update for $\mathbf{y}_j$:

$$\mathbf{y}_j = \pm \frac{\mathbf{s}_j}{|\mathbf{s}_j|} \odot \mathbf{v}_j, \tag{12}$$

We consider the update that does not modify the phase of $\mathbf{s}_j$ (i.e., with a '+' sign in (12)), as it corresponds to the equality case of Eq. (7). Under such an update, $\psi(\tilde{\mathbf{s}}) = \psi^+(\tilde{\mathbf{s}}, \mathbf{y})$, which shows that ψ^+ is an auxiliary function for ψ.

C. Including the mixing constraint

Now, let us introduce the hard mixing constraint (1) within the auxiliary function ψ^+ by means of the Lagrange multipliers as in [25]. This results in finding a saddle point for:

$$\Psi(\tilde{\mathbf{s}}, \mathbf{y}, \tilde{\boldsymbol{\delta}}) = \sum_j ||\mathbf{A}\tilde{\mathbf{s}}_j - \mathbf{y}_j||^2 + 2\Re\left(\tilde{\boldsymbol{\delta}}^\mathsf{H}(\sum_j \tilde{\mathbf{s}}_j - \tilde{\mathbf{x}})\right), \tag{13}$$

where $\tilde{\boldsymbol{\delta}} \in \mathbb{C}^N$ is the vector of Lagrange multipliers. Setting the derivative of Ψ w.r.t. $\tilde{\mathbf{s}}_j$ at 0 yields:

$$2\mathbf{A}^\mathsf{H}\mathbf{A}\tilde{\mathbf{s}}_j - 2\mathbf{A}^\mathsf{H}\mathbf{y}_j + 2\tilde{\boldsymbol{\delta}} = 0. \tag{14}$$

Let us point out that the matrix $\mathbf{A}^\mathsf{H}$ encodes the inverse STFT [26]. Indeed, one can show [27] that if the synthesis window is equal to the analysis window up to a specific normalization constant [28], the STFT is an Hermitian operator. Assuming such analysis-synthesis windows are used, $\mathbf{A}^\mathsf{H}\mathbf{A}$ is the identity matrix. We define:

$$\tilde{\mathbf{y}}_j = \mathbf{A}^\mathsf{H}\mathbf{y}_j = \text{iSTFT}(\mathbf{y}_j), \tag{15}$$

and therefore Eq. (14) rewrites:

$$\tilde{\mathbf{s}}_j - \tilde{\mathbf{y}}_j + \tilde{\boldsymbol{\delta}} = 0. \tag{16}$$

Summing (16) over j and using the mixing constraint yields:

$$\tilde{\mathbf{x}} - \sum_j \tilde{\mathbf{y}}_j + J\tilde{\boldsymbol{\delta}} = 0. \tag{17}$$

Finally, solving for $\tilde{\boldsymbol{\delta}}$ and injecting it in Eq. (16) leads to:

$$\tilde{\mathbf{s}}_j = \tilde{\mathbf{y}}_j + \frac{1}{J}(\tilde{\mathbf{x}} - \sum_p \tilde{\mathbf{y}}_p). \tag{18}$$

Combining the updates given by Eq. (2), (12), (15), and (18) yields the MISI algorithm, as introduced in the original paper [19]; this derivation therefore proves its convergence.

IV. Online MISI

First, let us reshape the STFTs $\mathbf{s}_j$ onto matrix form as they are usually processed this way: $\mathbf{S}_j \in \mathbb{C}^{F\times T}$. We rewrite the MISI algorithm in the TF domain, as done in [14], [15]: $\forall j$,

$$\mathbf{Z}_j = \text{STFT}(\text{iSTFT}(\mathbf{S}_j)), \tag{19}$$

$$\mathbf{Y}_j = \frac{\mathbf{Z}_j}{|\mathbf{Z}_j|} \odot \mathbf{V}_j, \tag{20}$$

$$\mathbf{S}_j = \mathbf{Y}_j + \frac{1}{J}(\mathbf{X} - \sum_p \mathbf{Y}_p), \tag{21}$$

A. Problem setting

It is straightforward to implement (20) and (21) online, as these are performed bin-wise, but this is not the case of (19). Indeed, the inverse STFT of $\mathbf{S}_j$ is computed through the overlap-add (OLA) procedure as follows:

$$\tilde{\mathbf{s}}'_{j,t} = \text{iDFT}(\mathbf{S}_{j,t}) \odot \tilde{\mathbf{w}}, \tag{22}$$

$$\tilde{\mathbf{s}}_j(n) = \sum_{t=0}^{T-1} \tilde{\mathbf{s}}'_{j,t}(n - tl), \tag{23}$$

where $\mathbf{S}_{j,t}$ is the t-th column of $\mathbf{S}_j$, iDFT denotes the inverse discrete Fourier transform, $\tilde{\mathbf{w}}$ is a window of length N_w, and l is the hop size. For notation convenience, we consider that $\tilde{\mathbf{s}}'_{j,t}(n - tl) = 0$ if $n \notin \{0, ..., N_w - 1\}$.

Using OLA (23), a sample n is reconstructed by accounting for all frames whose iDFT has a temporal support that includes n, which is not suitable for online applications.

B. Online implementation

Let us assume that we are currently processing the frame indexed by t. We first decompose the OLA procedure (23) as:

$$\tilde{\mathbf{s}}_j(n) = \underbrace{\sum_{k=0}^{t-1} \tilde{\mathbf{s}}'_{j,k}(n - tl)}_{\tilde{\mathbf{s}}^{\text{past}}_{j,t}(n)} + \underbrace{\sum_{k=t}^{T-1} \tilde{\mathbf{s}}'_{j,k}(n - tl)}_{\tilde{\mathbf{s}}^{\text{fut}}_{j,t}(n)}, \tag{24}$$

where $\tilde{\mathbf{s}}^{\text{past}}_{j,t}$ (resp. $\tilde{\mathbf{s}}^{\text{fut}}_{j,t}$) contains the contributions of the previous (resp. current and future) frames. Drawing on prior work [29], [30], [31], we propose hereafter to approximate $\tilde{\mathbf{s}}^{\text{fut}}_{j,t}$ by using only the current frame and an arbitrarily small number $K \geq 0$ of future frames:

$$\tilde{\mathbf{s}}^{\text{fut}}_{j,t}(n) \approx \sum_{k=t}^{t+K} \tilde{\mathbf{s}}'_{j,k}(n - tl). \tag{25}$$

Even though this approach results in losing the contributions of some of the future time frames involved in the calculation of $\tilde{\mathbf{s}}_j(n)$ (*cf.* IV-A), it still enforces a form of coherence over time by accounting for the overlap with the previous time frames. It also preserves the coherence with the near-future frames which overlap with the current one. With such an approach, the algorithmic latency is reduced to $N_w + Kl$ samples. It is illustrated in Fig. 1. This procedure is called oMISI (for "online MISI") and summarized in Algorithm 1.

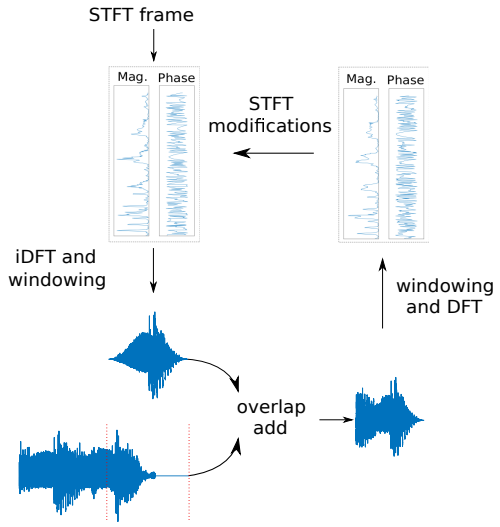

Fig. 1. Illustration of the proposed approach in one time frame with $K = 0$.

Algorithm 1: Online MISI

Inputs: Mixture $\mathbf{X} \in \mathbb{C}^{F\times T}$, magnitudes $\mathbf{V}_j \in \mathbb{C}^{F\times T}$, number of iterations N_i and future frames K.
Initialize the past segments: $\forall j$, $\tilde{\mathbf{s}}_j^{\text{past}} = 0$
for $t = 0$ **to** $T - 1 - K$ **do**
  Initialize the phase in the new frame $t + K$ (*cf.* IV-C)
  for $iter = 1$ **to** N_i **do**
    $\forall j$: Compute $\tilde{\mathbf{s}}_j^{fut}$ using (22) and (25)
    $\forall j$: $\mathbf{Z}_j = \text{STFT}(\tilde{\mathbf{s}}_j^{past} + \tilde{\mathbf{s}}_j^{fut})$
    $\forall j$: Update $\mathbf{S}_{j,t}, ..., \mathbf{S}_{j,t+K}$ using (20) and (21)
  end
  Compute, $\forall j$, $\tilde{\mathbf{s}}'_{j,t}$ from $\mathbf{S}_{j,t}$ using (22)
  Update the sources: $\forall j$, $\tilde{\mathbf{s}}_j(tl + n) = \tilde{\mathbf{s}}'_{j,t}(n) + \tilde{\mathbf{s}}_j^{past}(n)$ for $n < l$
  Update the past segments: $\forall j$, $\tilde{\mathbf{s}}_j^{\text{past}}(n) = \tilde{\mathbf{s}}'_{j,t}(n + l)$ for $n < N_w - l$ and 0 otherwise
end
Outputs: $\tilde{\mathbf{s}}_j$

C. Phase initialization

MISI is usually initialized by assigning the mixture's phase to each source. However, its online implementation makes it possible to use an alternative initialization scheme, exploiting phase relationships over time. In particular, we propose here to use a phase model that arise from modeling audio signals as mixtures of sinusoids [32], [33]. It can be shown [25] that the phase $\boldsymbol{\varphi}_j$ of a source represented as a mixture of slowly-varying sinusoids follows the relationship:

$$\boldsymbol{\varphi}_{j,t}(f) = \boldsymbol{\varphi}_{j,t-1}(f) + 2\pi l \boldsymbol{\nu}_{j,t}(f), \tag{26}$$

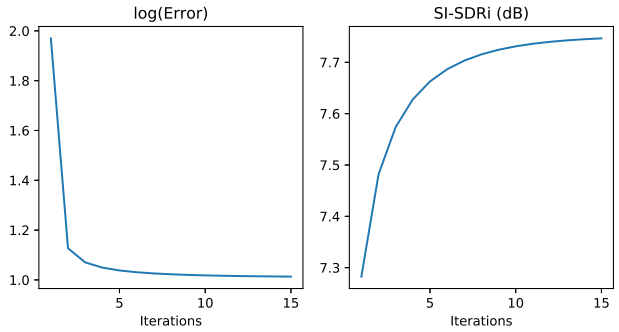

Fig. 2. Error (3) and SI-SDRi over iterations in the Estim. setting.

where $\boldsymbol{\nu}_{j,t}(f)$ is the normalized frequency in channel f and time frame t. We propose to use this model as a phase initialization scheme for oMISI (i.e., at line 4 in Algorithm 1). The frequencies ν are estimated from the magnitude spectra using quadratic interpolation around each magnitude peak [25], [34].

V. Experimental evaluation

A. Dataset and protocol

For evaluation, we consider a single-channel speech separation task. We use the Danish hearing in noise test dataset [35]. We consider three speaker pairs denoted MF, MM and FF, where M and F stand for male and female respectively, in order to cover all gender combinations. All audio files were recorded with a sampling rate of 44.1 kHz, and down-sampled at 16 kHz in our experiments. The STFT is computed using a 16 ms long Hann window, 50 % overlap, and a zero-padding factor of 2. The synthesis window is defined as in [28], so that the STFT is Hermitian, as discussed in Section III-C.

Two scenarios are considered. The Oracle scenario uses the ground truth magnitude spectra of the sources. In the Estim. scenario, they are estimated using the DNN described in [20] (where the interested reader can find details on the DNN architecture and training). This DNN predicts a soft mask that is applied to the mixture to yield magnitude estimates. We compare oMISI to its offline counterpart and to the amplitude mask (AM) used as a baseline [20].

The separation quality is measured with the scale-invariant signal-to-distortion ratio improvement (SI-SDRi) [36]. We provide some audio excerpts[1] for a subjective evaluation, and the code for reproducing the Oracle scenario experiments [2].

B. Results

We present in Fig. 2 the objective loss (3) and SI-SDRi over iterations for the MISI algorithm using the MF pair in the Estim. setting (similar results are obtained using the other speaker combinations). We observe a non-increasing cost function over iterations, which empirically confirms the convergence of MISI. The SI-SDRi appears to saturate at 15 iterations, thus we present hereafter the results obtained with this value. Since with oMISI, each frame is processed $(K+1)$ times more as in MISI, we reduce the number of iterations to $15/(K+1)$ for a fair comparison. Finally, note that the SI-SDRi can further increase in the Oracle setting.

TABLE I
Average SI-SDRi in dB (higher is better). Bold fonts correspond to the best performance among online techniques.

		MF		MM		FF	
	Latency	Estim.	Oracle	Estim.	Oracle	Estim.	Oracle
AM	16 ms	7.5	8.8	5.7	7.3	5.1	7.5
MISI	offline	7.9	23.8	6.2	22.3	5.4	22.9
oMISI							
mix	16 ms (K=0)	7.7	16.4	6.1	15.8	5.4	16.9
mix	24 ms (K=1)	7.9	20.2	6.2	19.4	5.4	19.6
mix	32 ms (K=2)	**7.9**	**21.4**	**6.2**	**20.4**	**5.4**	20.6
sin	24 ms (K=1)	7.8	15.2	6.2	14.6	5.4	**20.7**

The separation results are reported in Table I. We first remark that MISI improves the performance over the baseline by approximately 0.4 dB in the Estim. scenario. In the Oracle scenario, the improvement is more significant ($\approx$ 15 dB), which highlights the room for improvement for phase recovery. Besides, in the Estim. scenario, we observe that oMISI performs as well as MISI with $K = 1$. The performance of oMISI drops slightly for $K = 0$, which was expected as no future frame is taken into account: nonetheless, it is still improved compared to AM, and the drop in comparison to the offline method is quite small. Finally, the performance does not further improve for $K = 2$ in the Estim. scenario. These results demonstrate the potential of oMISI for real-time applications.

Using one future frame then appears as a good compromise between latency and performance, thus we test the initialization with the sinusoidal phase model with $K = 1$. However, this scheme does not improve the performance of oMISI over using the mixture's phase overall. We suggest that the speakers in the MM and MF pairs are sufficiently orthogonal (i.e., less overlapping) in the TF domain, thus the mixture's phase is a good quality initial estimate. Nonetheless, the Estim. results highlight that the FF pair is the most challenging, and the corresponding Oracle results indicate that this initialization scheme has some potential, provided accurate enough magnitude estimates.

VI. Conclusion

In this paper, we provided the first proof of convergence of the MISI algorithm. We adapted it to operate online without any performance loss, which is an important step towards real-time audio source separation. Future work will focus on deriving alternative spectrogram inversion algorithms based on this theoretical framework, e.g., by replacing the magnitude mismatch distance by a β-divergence, which is more adapted to audio. This algorithm will also be incorporated into an end-to-end framework for time-domain source separation along with learned and more advanced [37] phase models.

VII. Acknowledgements

We thank Roland Badeau for his insight on optimization, and Gaurav Naithani for providing the magnitude estimates.

[1] https://magronp.github.io/demos/spl20_omisi.html
[2] https://github.com/magronp/omisi

REFERENCES

[1] P. Comon and C. Jutten, *Handbook of blind source separation: independent component analysis and applications*. Academic press, 2010.

[2] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. Interspeech 2018*, September 2018.

[3] E. Cano, D. FitzGerald, A. Liutkus, M. D. Plumbley, and F. Stter, "Musical source separation: An introduction," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 31–40, Jan 2019.

[4] D. Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[5] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, June 2000.

[6] 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.

[7] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018.

[8] ——, "Conv-TasNet: Surpassing ideal timefrequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, August 2019.

[9] P. Magron, R. Badeau, and B. David, "Phase recovery in NMF for audio source separation: an insightful benchmark," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015.

[10] 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.

[11] 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.

[12] Z. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018.

[13] P. Magron, K. Drossos, S. I. Mimilakis, and T. Virtanen, "Reducing interference with phase recovery in DNN-based monaural singing voice separation," in *Proc. Interspeech*, September 2018.

[14] Z.-Q. Wang, J. Le Roux, D. Wang, and J. R. Hershey, "End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction," in *Proc. of Interspeech*, September 2018.

[15] G. Wichern and J. Le Roux, "Phase reconstruction with learned time-frequency representations for single-channel speech separation," in *Proc. of IWAENC*, September 2018.

[16] Z. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[17] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[18] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 370–382, May 2019.

[19] 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.

[20] G. Naithani, T. Barker, G. Parascandolo, L. Bramsløw, N. H. Pontoppidan, and T. Virtanen, "Low latency sound source separation using convolutional recurrent neural networks," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2017.

[21] R. Aihara, T. Hanazawa, Y. Okato, G. Wichern, and J. L. Roux, "Teacher-student deep clustering for low-delay single channel speech separation," in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[22] S. Wang, G. Naithani, and T. Virtanen, "Low-latency deep clustering for speech separation," in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[23] 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.

[24] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, September 2011.

[25] 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, June 2018.

[26] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 217–220, March 2013.

[27] B. Yang, "A study of inverse short-time Fourier transform," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2008.

[28] 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.

[29] G. T. Beauregard, X. Zhu, and L. L. Wyse, "An efficient algorithm for real-time spectrogram inversion," in *Proc. International Conference on Digital Audio Effects (DAFx)*, 2005.

[30] X. Zhu, G. T. Beauregard, and L. Wyse, "Real-time iterative spectrum inversion with look-ahead," in *Proc. IEEE International Conference on Multimedia and Expo*, July 2006.

[31] 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.

[32] 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.

[33] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, December 2014.

[34] M. Abe and J. O. Smith, "Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks," in *Audio Engineering Society Convention 117*, May 2004.

[35] J. B. Nielsen and T. Dau, "The Danish hearing in noise test," *International journal of audiology,*, vol. 50, no. 3, pp. 202–8, March 2011.

[36] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR half-baked or well done?" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[37] Z. Průša and P. L. Søndergaard, "Real-time spectrogram inversion using phase gradient heap integration," in *Proc. International Conference on Digital Audio Effects (DAFx)*, September 2016.