Joint Amplitude and Phase Refinement for Monaural Source Separation

Yoshiki Masuyama, *Student Member, IEEE*, Kohei Yatabe, *Member, IEEE*, Kento Nagatomo, and Yasuhiro Oikawa, *Member, IEEE*

***Abstract*—Monaural source separation is often conducted by manipulating the amplitude spectrogram of a mixture (e.g., via time-frequency masking and spectral subtraction). The obtained amplitudes are converted back to the time domain by using the phase of the mixture or by applying phase reconstruction. Although phase reconstruction performs well for the true amplitudes, its performance is degraded when the amplitudes contain error. To deal with this problem, we propose an optimization-based method to refine both amplitudes and phases based on the given amplitudes. It aims to find time-domain signals whose amplitude spectrograms are close to the given ones in terms of the generalized alpha-beta divergences. To solve the optimization problem, the alternating direction method of multipliers (ADMM) is utilized. We confirmed the effectiveness of the proposed method through speech-nonspeech separation in various conditions.**

***Index Terms*—Phase reconstruction, spectrogram consistency, mixture consistency, alpha-beta divergences.**

I. Introduction

MONAURAL source separation (MSS) aims to decompose a single-channel mixture signal into each source signal. While some attempts of time-domain MSS have gained attention recently [1], [2], the majority of MSS is conducted in the time-frequency (T-F) domain [3]–[5]. Ordinary T-F domain methods use the short-time Fourier transform (STFT) and focus on amplitude manipulation, which is often realized by applying non-negative T-F masks [6]–[9]. The estimated amplitude of each source is converted back to the time domain by the inverse STFT (iSTFT) with the phase of the mixture. Due to the use of the mixed phase, the obtained signals contain interference even when the amplitudes are ideally separated.

To tackle this problem, various phase reconstruction methods have been presented [10]–[16]. The Griffin–Lim algorithm (GLA) [11] is a popular method for phase reconstruction from a single amplitude spectrogram. GLA modifies the phase of each separated signal based on the *STFT consistency:* the reconstructed complex STFT coefficient should retain the neighborhood relation caused by the overlapped window of STFT [12]. The multiple input spectrogram inversion (MISI) [13] further considered the *mixture consistency* [17]: a sum of separated signals should coincide with the mixture. MISI achieved remarkable results with the true amplitudes, and various extensions have been presented [18]–[21]. In MSS, however, the estimated amplitudes often contain error, which significantly impairs the performance of MISI. This is because it keeps the given amplitudes and only attempts to reconstruct phases that are appropriate for the amplitudes in terms of STFT and mixture consistencies. To improve the robustness to the error of the given amplitudes, the amplitudes should also be modified jointly with the phases.

In this letter, we propose an optimization-based method that jointly refines amplitudes and phases based on the STFT and mixture consistencies. The optimization problem aims to find the separated time-domain signals whose amplitude spectrograms are close to the given ones while considering the mixture consistency as a regularization. As the measure of dissimilarity of the amplitudes, the generalized alpha-beta divergences [22] are considered, since they have been preferred in MSS [23]. To solve the problem, the alternating direction method of multipliers (ADMM) [24] is utilized. The effectiveness and robustness of the proposed method were confirmed by speech-nonspeech separation using various amplitude estimation methods.

II. Preliminaries

A. Monaural Source Separation in Time-Frequency Domain

Let us consider an observed time-domain signal $\mathbf{y}$ given by

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{s}_k, \tag{1}$$

where $\mathbf{s}_k$ is the kth source signal, and $k = 1, \ldots, K$ is the source index. STFT of a time-domain signal $\mathbf{x}$ is denoted as $\mathbf{X} = \mathcal{G}(\mathbf{x})$. In MSS, many of existing methods focus on the estimation of the amplitude [3]–[5], and the phase of the mixture remains intact. That is, the separated complex STFT coefficient is given by

$$\hat{S}_k[t, f] = A_k[t, f] \frac{Y[t, f]}{|Y[t, f]|}, \tag{2}$$

where $t = 1, \ldots, T$ and $f = 1 \ldots, F$ are respectively the time and frequency indices, $\mathbf{A}_k$ is the estimated amplitude of the kth source, and zero division is replaced by zero. The separated time domain signal is obtained by iSTFT $\mathcal{G}^\dagger$ as $\hat{\mathbf{s}}_k = \mathcal{G}^\dagger(\hat{\mathbf{S}}_k)$.

B. Alternating Direction Method of Multipliers (ADMM)

ADMM has been widely utilized for approximately solving the optimization problems of the following form [24]:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \; g(\boldsymbol{\alpha}) + h(\boldsymbol{\beta}) \quad \text{s.t.} \quad \boldsymbol{\alpha} = \mathcal{L}(\boldsymbol{\beta}) \tag{3}$$

Manuscript received August 11, 2020; revised September 27, 2020; accepted October 11, 2020. Date of publication October 15, 2020; date of current version November 11, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomoki Toda. *(Corresponding author: Yoshiki Masuyama.)*

The authors are with Waseda University, Tokyo 169-8555, Japan (e-mail: mas-03151102@akane.waseda.jp; k.yatabe@asagi.waseda.jp; jimijeffericking@akane.waseda.jp; yoikawa@waseda.jp).

Digital Object Identifier 10.1109/LSP.2020.3031464

where $\mathcal{L}$ is a bounded linear operator, and f and g are proper lower-semicontinuous functions. An important aspect of ADMM is that the minimization of each function is conducted separately with an auxiliary variable $\boldsymbol{\gamma}$ as follows:

$$\boldsymbol{\alpha}^{[m+1]} = \mathrm{argmin}_{\boldsymbol{\alpha}}\, g(\boldsymbol{\alpha}) + \frac{\rho}{2}\|\boldsymbol{\alpha} - \mathcal{L}(\boldsymbol{\beta}^{[m]}) + \boldsymbol{\gamma}^{[m]}\|_2^2, \tag{4}$$

$$\boldsymbol{\beta}^{[m+1]} = \mathrm{argmin}_{\boldsymbol{\beta}}\, h(\boldsymbol{\beta}) + \frac{\rho}{2}\|\boldsymbol{\alpha}^{[m+1]} - \mathcal{L}(\boldsymbol{\beta}) + \boldsymbol{\gamma}^{[m]}\|_2^2, \tag{5}$$

$$\boldsymbol{\gamma}^{[m+1]} = \boldsymbol{\gamma}^{[m]} + \boldsymbol{\alpha}^{[m+1]} - \mathcal{L}(\boldsymbol{\beta}^{[m+1]}), \tag{6}$$

where $\rho > 0$ is a hyperparameter, $\|\cdot\|_2$ is the Euclidean norm, and m is the iteration index. Although the global linear convergence is only guaranteed for convex problems with some assumptions [25], its effectiveness for nonconvex problems has been empirically confirmed in many applications [26]–[28].

III. Proposed Method

Instead of directly using the estimated amplitude and phase of the mixture as in Eq. (2), we propose to refine the amplitude and phase via solving an optimization problem with ADMM.

A. Problem Formulation

Our proposed method imposes the following two criteria on separated time-domain signals $(\mathbf{x}_1, \ldots, \mathbf{x}_K)$. The first criterion is that the amplitude spectrograms of the separated signals should be close to the given ones, i.e., $|\mathcal{G}(\mathbf{x}_k)[t,f]| \approx A_k[t,f]$, because $\mathbf{A}_k$ is the key information on each source. The second criterion is the mixture consistency: separated signals should be summed up to the mixture signal as in Eq. (1). We use this criterion as a regularization because [18] showed that the hard constraint of the criterion degrades the separation performance in some cases. Consequently, the proposed method is formulated as the following minimization problem:

$$\min_{(\mathbf{x}_1,\ldots,\mathbf{x}_K)} \sum_{k=1}^{K} D(\mathbf{A}_k \mid \mathcal{G}(\mathbf{x}_k)) + \frac{\lambda}{2\,K}\Big\|\,\mathbf{y} - \sum_{k=1}^{K}\mathbf{x}_k\Big\|_2^2, \tag{7}$$

where $\lambda \geq 0$ is a parameter balancing the two terms.

The first term penalizes the dissimilarity between the given amplitudes and amplitude spectrograms of separated signals:

$$D(\mathbf{A}_k \mid \mathbf{X}_k) = \sum_{t=1}^{T}\sum_{f=1}^{F} d(A_k[t,f] \mid |X_k[t,f]|\,), \tag{8}$$

where $d(a \mid \hat{a})$ is a measure of the dissimilarity between two scalars $a > 0$ and $\hat{a} \geq 0$. That is, the minimization of $D(\mathbf{A}_k \mid \mathbf{X}_k)$ makes $|X_k[t,f]|$ close to $A_k[t,f]$. We use the generalized alpha-beta divergences [22], in particular, the squared Euclidean (EUC) distance, generalized Kullback–Leibler (KL) divergence, dual-Itakura–Saito (dIS) divergence, and dIS divergence for squared variables:

$$d_{\mathrm{EUC}}(a \mid \hat{a}) = \frac{1}{2}(a - \hat{a})^2, \tag{9}$$

$$d_{\mathrm{KL}}(a \mid \hat{a}) = a\log\frac{a}{\hat{a}} - a + \hat{a}, \tag{10}$$

$$d_{\mathrm{dIS}}(a \mid \hat{a}) = \frac{\hat{a}}{a} - \log\frac{\hat{a}}{a} - 1, \tag{11}$$

$$d_{\mathrm{dISs}}(a \mid \hat{a}) = \frac{1}{4}\left(\frac{\hat{a}^2}{a^2} - \log\frac{\hat{a}^2}{a^2} - 1\right). \tag{12}$$

The generalized alpha-beta divergences have been successfully applied to various nonnegative matrix factorization (NMF)–based audio applications [23], [28]–[31], and the use of the KL and dIS divergences often improves their performances [31]. Note that all of the measures given in Eqs. (9)–(12) are convex for the second argument $\hat{a}$. Although the proposed formulation in Eq. (7) is non-convex due to taking absolute value in Eq. (8), we expect that the convexity of d is helpful for solving the optimization problem.

The second term of Eq. (7) enforces the mixture consistency on the separated time-domain signals. Note that the STFT consistency is implicitly considered by handling the separated signals in the time domain.

B. ADMM for Proposed Optimization Problem in Eq. (7)

To solve the optimization problem in Eq. (7) using ADMM, it is reformulated as the following equivalent problem:

$$\min_{\substack{(\mathbf{x}_1,\ldots,\mathbf{x}_K)\\(\mathbf{Z}_1,\ldots,\mathbf{Z}_K)}} \sum_{k=1}^{K} D(\mathbf{A}_k \mid \mathbf{Z}_k) + \frac{\lambda}{2\,K}\Big\|\,\mathbf{y} - \sum_{k=1}^{K}\mathbf{x}_k\Big\|_2^2, \tag{13a}$$

$$\text{s.t.} \quad \mathbf{Z}_k = \mathcal{G}(\mathbf{x}_k) \quad \forall k \in \{1,\ldots,K\}. \tag{13b}$$

This problem is a special case of Eq. (3) obtained by regarding $(\mathbf{Z}_1,\ldots,\mathbf{Z}_K)$ and $(\mathbf{x}_1,\ldots,\mathbf{x}_K)$ as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. Thus, it can be solved by iterating Eqs. (4)–(6). To do so, we must solve the subproblems in Eqs. (4) and (5) as follows.

1) Update of $(\mathbf{Z}_1,\ldots,\mathbf{Z}_K)$ via Eq. (4): The first subproblem is obtained by replacing $g(\boldsymbol{\alpha})$ with $\sum_{k=1}^{K} D(\mathbf{A}_k \mid \mathbf{Z}_k)$:

$$\min_{(\mathbf{Z}_1,\ldots,\mathbf{Z}_K)} \sum_{k=1}^{K} D(\mathbf{A}_k \mid \mathbf{Z}_k) + \frac{\rho}{2}\|\mathbf{Z}_k - \mathbf{V}_k\|_{\mathrm{Fro}}^2 \tag{14}$$

where $\mathbf{V}_k = \mathcal{G}(\mathbf{x}_k^{[m]}) - \mathbf{U}_k^{[m]}$, $\mathbf{U}_k$ is an auxiliary variable corresponding to $\boldsymbol{\gamma}$, and $\|\cdot\|_{\mathrm{Fro}}$ is the Frobenius norm. Since this subproblem is separable for t, f, and k, the following scalar-valued proximity operator [32] gives the update formula:

$$\begin{aligned} Z_k^{[m+1]}[t,f] &= \mathrm{argmin}_Z\; d(A_k[t,f] \mid |Z|\,) + \frac{\rho}{2}|Z - V_k[t,f]|^2, \\ &= \mathrm{prox}_{d(A_k[t,f]||\cdot|)/\rho}(V_k[t,f]), \end{aligned} \tag{15}$$

where the proximity operators for the aforementioned alpha-beta divergences can be calculated as follows[1]

$$\mathrm{prox}_{d_{\mathrm{EUC}}(a||\cdot|)/\rho}(v) = \frac{a + \rho|v|}{1+\rho}\frac{v}{|v|}, \tag{16}$$

$$\mathrm{prox}_{d_{\mathrm{KL}}(a||\cdot|)/\rho}(v) = \frac{\zeta + \sqrt{\zeta^2 + 4a\rho}}{2\rho}\frac{v}{|v|}, \tag{17}$$

$$\mathrm{prox}_{d_{\mathrm{dIS}}(a||\cdot|)/\rho}(v) = \frac{\eta + \sqrt{\eta^2 + 4a^2\rho}}{2a\rho}\frac{v}{|v|}, \tag{18}$$

$$\mathrm{prox}_{d_{\mathrm{dISs}}(a||\cdot|)/\rho}(v) = \frac{|v| + \sqrt{|v|^2 + \xi/\rho}}{\xi}\frac{v}{|v|}, \tag{19}$$

[1]The proximity operators of Eqs. (9)–(12) for a real number are summarized in [31], [33]. Here, we extend them to a complex number. When v is zero, the right-hand side of Eqs. (16)–(19) cannot be defined because of zero division. The proximity operator given in Eq. (15) is set-valued in such a situation, but it must be a scalar to substitute into $Z_k[t,f]$. We simply replace $v/|v|$ by zero, and thus $Z_k[t,f]$ becomes zero. Note that zero division almost never occurs because iSTFT and STFT spread components in the T-F domain.

where $\zeta = |v|\rho - 1$, $\eta = a|v|\rho - 1$, and $\xi = 2 + 1/(a^2\rho)$.

2) Update of $(\mathbf{x}_1, \ldots, \mathbf{x}_K)$ *via Eq. (5):* The second subproblem related to the mixture consistency is obtained by replacing $h(\boldsymbol{\beta})$ with $\frac{\lambda}{2K}\big\|\mathbf{y} - \sum_{k=1}^K \mathbf{x}_k\big\|_2^2$:

$$\min_{(\mathbf{x}_1,\ldots,\mathbf{x}_K)} \frac{\lambda}{2K}\Big\|\mathbf{y} - \sum_{k=1}^K \mathbf{x}_k\Big\|_2^2 + \frac{\rho}{2}\sum_{k=1}^K \|\mathcal{G}(\mathbf{x}_k) - \mathbf{W}_k\|_{\mathrm{Fro}}^2, \quad (20)$$

where $\mathbf{W}_k = \mathbf{Z}_k^{[m+1]} + \mathbf{U}_k^{[m]}$. Assuming that the adjoint of STFT $\mathcal{G}^*$ coincides with iSTFT $\mathcal{G}^\dagger$, its solution is given by

$$\mathbf{x}_k^{[m+1]} = \mathcal{G}^\dagger(\mathbf{W}_k) + \frac{\lambda}{K(\lambda+\rho)}\left(\mathbf{y} - \sum_{k=1}^K \mathcal{G}^\dagger(\mathbf{W}_k)\right). \quad (21)$$

3) Update of $(\mathbf{U}_1, \ldots, \mathbf{U}_K)$ *via Eq. (6):* The update formula for the auxiliary variables can be obtained as

$$\mathbf{U}_k^{[m+1]} = \mathbf{U}_k^{[m]} - \mathcal{G}(\mathbf{x}_k^{[m+1]}) + \mathbf{Z}_k^{[m+1]}. \quad (22)$$

The whole algorithm for solving the reformulated optimization problem in Eq. (13) is summarized in Algorithm 1. While an arbitrary initialization is allowable, one simple initialization is setting $\mathbf{x}_k^{[1]} = \hat{\mathbf{s}}_k$ and $\mathbf{U}_k^{[1]}$ to the zero matrix.

Algorithm 1: ADMM for solving Eq. (13).

Input: $\mathbf{y}$, $(\mathbf{A}_1, \ldots, \mathbf{A}_K)$, $(\mathbf{x}_1^{[1]}, \ldots, \mathbf{x}_K^{[1]})$, $(\mathbf{U}_1^{[1]}, \ldots, \mathbf{U}_K^{[1]})$, $\lambda > 0$, $\rho > 0$
Output: $(\mathbf{x}_1^{[M+1]}, \ldots, \mathbf{x}_K^{[M+1]})$
for $m = 1$ to M **do**
  for $k = 1$ to K, $t = 1$ to T, $f = 1$ to F **do**
    Calculate $Z_k[t,f]^{[m+1]}$ by one of Eqs. (16)–(19).
  end for
  Calculate $(\mathbf{x}_1^{[m+1]}, \ldots, \mathbf{x}_K^{[m+1]})$ by Eq. (21).
  for $k = 1$ to K **do**
    $\mathbf{U}_k^{[m+1]} = \mathbf{U}_k^{[m]} - \mathcal{G}(\mathbf{x}_k^{[m+1]}) + \mathbf{Z}_k^{[m+1]}$
  end for
end for

C. Relation to Existing Methods

In this subsection, relation between existing methods and the proposed method is discussed.

1) Relation to GLA and MISI: GLA [11] and MISI [13] reconstruct phase based on the STFT consistency, where MISI further considers the mixture consistency. They are realized by the following iterative algorithm:

$$Z_k^{[m+1]}[t,f] = A_k[t,f]\frac{\mathcal{G}(\mathbf{x}_k^{[m]})[t,f]}{|\mathcal{G}(\mathbf{x}_k^{[m]})[t,f]|}, \quad (23)$$

$$\mathbf{x}_k^{[m+1]} = \mathcal{G}^\dagger(\mathbf{Z}_k^{[m+1]}) + \frac{\kappa}{K}\left(\mathbf{y} - \sum_{k=1}^K \mathcal{G}^\dagger(\mathbf{Z}_k^{[m+1]})\right). \quad (24)$$

where GLA and MISI correspond to $\kappa = 0$ and $\kappa = 1$, respectively. The difference between the algorithm given in Eqs. (23)–(24) and the proposed one is as follows: (1) the proposed algorithm introduces an auxiliary variable $\mathbf{U}_k$ based on ADMM; and (2) the proposed algorithm modifies the amplitude of $\mathbf{Z}_k$ from $\mathbf{A}_k$ as in Eq. (15) while GLA and MISI keeps $\mathbf{A}_k$ as in Eq. (23). The second difference allows refining the amplitude jointly with the phase. To refine the amplitude, the proposed method additionally takes $O(TF)$ computation, due to Eqs. (16)–(19) applied to each T-F bin. Note that the computational costs of GLA, MISI, and the proposed method are dominated by STFT and iSTFT which take $O(TF\log F)$. Hence, the additional calculation of the proposed method does not affect the order of the overall computational cost.

In Eq. (24), MISI exactly enforces the mixture consistency by setting $\kappa = 1$, i.e., $\mathbf{y} = \sum_{k=1}^K \mathbf{x}_k$. On the other hand, GLA does not enforce the consistency, i.e., $\mathbf{x}_k = \mathcal{G}^\dagger(\mathbf{Z}_k)$. The proposed algorithm induces the mixture consistency softly. Specifically, κ corresponds to $\lambda/(\lambda+\rho)$ in Eq. (21), where $\mathbf{Z}_k^{[m+1]}$ in Eq. (24) is replaced by $\mathbf{W}_k$ in Eq. (21). Here, $\lambda/(\lambda+\rho)$ takes 0 and 1 when $\lambda = 0$ and $\lambda \to \infty$, respectively. Note that GLA can be derived from the proposed formulation in Eq. (7) with d_{EUC} and $\lambda = 0$ by applying a majorization-minimization algorithm [34]. MISI can also be obtained from Eq. (7) with d_{EUC} at the limit of λ tending to ∞ [18], [19]. Thus, the proposed method can be viewed as an extension of GLA and MISI in three aspects: extending d_{EUC} to generalized alpha-beta divergences, considering the mixture consistency softly, and applying ADMM.

2) Relation to CWF: While the aforementioned phase reconstruction methods aim to keep the given amplitudes, some studies attempt to refine both amplitudes and phases [35]–[37]. In the literature, the consistent Wiener filter (CWF) achieved promising performance by combining the STFT consistency with the usual Wiener filtering. When $K = 2$, CWF[2] can be formulated by a function similar to d_{EUC} but considers the variables in terms of complex numbers [37]:

$$\min_{\mathbf{x}_k} \sum_{t=1}^T \sum_{f=1}^F \nu[t,f]|\mu_k[t,f] - \mathcal{G}(\mathbf{x}_k)[t,f]|^2, \quad (25)$$

where $\boldsymbol{\mu}_k$ is the result of the usual Wiener filtering, and $\nu[t,f] = (1/A_1^2[t,f]) + (1/A_2^2[t,f])$ is a weight. Since the phase of $\boldsymbol{\mu}_k$ is that of the mixture, the source signals $(\mathbf{s}_1, \ldots, \mathbf{s}_K)$ are not the optimal solution of Eq. (25) even when the true amplitudes are known. In contrast, the proposed formulation given in Eq. (7) only measures the similarity of the amplitudes. Consequently, the source signals are the global optimal solution whenever $A_k[t,f] = |S_k[t,f]|$.

IV. Experimental Evaluation

A. Experimental Conditions

The effectiveness of the proposed method was investigated through speech-nonspeech separation as in [18], [37]. We used the subset of the TIMIT dataset [38] provided in [39], which contains 200 gender-balanced utterances. We added 4 kinds of noises from CHiME-3 [40] to the utterances. The sampling rate was 16 kHz. STFT was implemented with the canonical tight window of the Hann window [41] whose length was 32 ms, and the shift size was 8 ms. The proposed method was compared with GLA [11], MISI [13], modified MISI (MMISI) [18], and CWF [37]. In our early experiment, the performance of MISI and CWF saturated before 50 iterations, while that of the other methods gradually varied after 50 iterations. To balance them, all methods were iterated 100 times. Their parameters were manually tuned to optimum. For the proposed method, $\rho = 10$ and $\lambda = 1000$.

We assessed those methods in four conditions of the amplitude estimation: spectral subtraction (SS) [3]–based ratio masking, supervised NMF (SNMF) [42]–based ratio masking, ideal

[2]CWF enforces the STFT consistency either as a constraint or a regularization. Eq. (25) corresponds to the case of the constraint.

TABLE I
COMPARISON OF SI-SDR, PESQ, AND STOI AVERAGED OVER 200 MIXTURES, WHERE "OBSERVED" USED THE PHASE OF THE MIXTURE WITH THE ESTIMATED AMPLITUDE. SHADED CELLS WITH AND WITHOUT BOLD FONTS INDICATE THE HIGHEST AND SECOND HIGHEST SCORES, RESPECTIVELY

	SI-SDR												
	SNR = 0 dB				SNR = 5 dB				SNR = 10 dB				Average
T-F Mask	SS	SNMF	IRM	tIAM	SS	SNMF	IRM	tIAM	SS	SNMF	IRM	tIAM	
Observed	4.88	9.07	11.89	11.90	9.80	12.65	15.06	15.37	14.36	15.93	18.35	18.99	13.19
GLA	4.68	8.33	12.32	14.06	9.56	11.66	15.12	17.57	14.05	14.52	18.10	21.16	13.43
MISI	4.82	9.01	13.12	17.89	9.75	12.53	16.12	20.88	14.30	15.64	19.24	24.15	14.79
MMISI	4.88	9.07	11.91	15.03	9.80	12.66	15.08	18.30	14.36	15.93	18.38	21.72	13.93
CWF	**6.30**	**9.73**	13.01	13.45	**10.75**	**13.13**	16.16	16.50	**14.97**	**16.54**	19.46	19.68	14.14
Prop. (EUC)	4.75	8.96	13.01	17.89	9.69	12.52	16.10	20.93	14.27	15.71	19.30	24.22	14.78
Prop. (KL)	5.17	9.15	13.63	**18.72**	9.97	12.67	**16.74**	**22.04**	14.41	15.88	**19.97**	**25.56**	**15.33**
Prop. (dIS)	4.25	9.32	**13.74**	17.21	9.49	12.74	**16.74**	20.09	14.08	16.05	19.89	23.06	14.72
Prop. (dISs)	5.11	9.33	13.67	16.89	9.93	12.76	16.68	19.84	14.28	16.09	19.83	22.85	14.77
	Wide-band PESQ												
	SNR = 0 dB				SNR = 5 dB				SNR = 10 dB				Average
T-F Mask	SS	SNMF	IRM	tIAM	SS	SNMF	IRM	tIAM	SS	SNMF	IRM	tIAM	
Observed	1.16	1.54	3.12	3.11	1.37	1.91	3.45	3.44	1.73	2.35	3.74	3.74	2.55
GLA	1.15	1.53	3.91	4.17	1.36	1.90	4.05	4.25	1.72	2.33	4.17	4.34	2.91
MISI	1.15	1.53	3.69	4.18	1.36	1.91	3.91	4.27	1.72	2.34	4.08	4.35	2.87
MMISI	1.16	1.54	3.13	3.61	1.37	1.91	3.46	3.86	1.73	2.35	3.75	4.06	2.66
CWF	**1.25**	**1.69**	3.25	3.33	**1.50**	**2.03**	3.55	3.62	1.87	2.41	3.80	3.84	2.68
Prop. (EUC)	1.15	1.52	3.52	4.11	1.36	1.89	3.78	4.22	1.72	2.32	4.00	4.32	2.83
Prop. (KL)	1.13	1.56	**3.96**	**4.36**	1.32	1.95	**4.09**	**4.39**	1.68	2.38	**4.21**	**4.43**	**2.95**
Prop. (dIS)	1.18	1.63	3.86	4.06	1.45	2.01	3.97	4.11	1.89	**2.43**	4.09	4.19	2.91
Prop. (dISs)	1.20	1.63	3.83	4.05	1.48	2.01	3.96	4.10	**1.91**	**2.43**	4.09	4.18	2.91
	STOI												
	SNR = 0 dB				SNR = 5 dB				SNR = 10 dB				Average
T-F Mask	SS	SNMF	IRM	tIAM	SS	SNMF	IRM	tIAM	SS	SNMF	IRM	tIAM	
Observed	**0.700**	0.818	0.927	0.942	**0.805**	0.884	0.945	0.957	**0.886**	0.927	0.962	0.971	0.894
GLA	0.698	0.817	**0.951**	0.978	0.803	0.884	**0.960**	0.981	0.885	0.927	0.970	0.986	0.903
MISI	0.698	0.818	0.945	0.980	0.803	0.885	0.957	0.983	0.885	0.928	0.969	0.988	0.903
MMISI	**0.700**	0.818	0.928	0.958	**0.805**	0.884	0.945	0.969	**0.886**	0.927	0.962	0.979	0.897
CWF	0.684	0.813	0.925	0.928	0.798	0.881	0.948	0.950	0.884	0.926	0.967	0.968	0.889
Prop. (EUC)	0.697	0.817	0.942	0.977	0.803	0.884	0.955	0.982	0.885	0.927	0.969	0.988	0.902
Prop. (KL)	0.694	**0.822**	0.949	**0.983**	0.803	**0.887**	**0.960**	**0.986**	**0.886**	**0.929**	**0.972**	**0.991**	**0.905**
Prop. (dIS)	0.667	0.819	0.944	0.971	0.779	0.883	0.957	0.976	0.869	0.926	0.971	0.983	0.895
Prop. (dISs)	0.666	0.818	0.943	0.968	0.781	0.883	0.956	0.974	0.873	0.927	0.970	0.982	0.895

ratio masking (IRM) [6], and truncated ideal amplitude masking (tIAM) [43]. In the first three conditions, amplitudes were estimated by the ratio masking: $A_k[t,f] = M_k[t,f]|Y[t,f]|$, where $M_k[t,f] = \Lambda_k[t,f]/(\sum_{k=1}^{2}\Lambda_k[t,f])$ is the ratio mask. Here, $\Lambda_1[t,f]$ and $\Lambda_2[t,f]$ are nonnegative scalars related to the amplitudes of speech and noise, respectively, as follows. In the SS-based ratio masking, $\mathbf{\Lambda}_k$ was calculated in a blind setting, i.e., the amplitudes of speech and noise at each T-F bin were unknown[3]. The SNMF-based ratio masking was conducted in a semi-blind setting, where the true amplitudes were not utilized directly [4]. IRM and tIAM were calculated from the true amplitudes.

B. Experimental Results

The separated signals were evaluated by SI-SDR [44], wide-based PESQ [45], and STOI [46]. These scores averaged over 200 mixtures are summarized in Table I. As mentioned in the introduction, MISI performed well when the amplitude estimation was accurate, i.e., IRM and tIAM. MISI was overtaken by MMISI, a robust variant of MISI, in the blind/semi-blind settings, SS and SNMF. CWF achieved the highest SI-SDR in SS and SNMF, but resulted in the lowest SI-SDR in the case of tIAM. These results indicate that the optimal method depends on the accuracy of amplitudes.

In contrast to those methods, the proposed method worked well for both accurate and inaccurate amplitudes. In particular, the proposed method with KL divergence [Prop. (KL)] outperformed the existing methods in terms of the average scores over the four conditions. For 100 iterations, MISI and Prop. (KL) took 2.31 s and 3.54 s, respectively .[5] With similar computational time, SI-SDR of MISI and Prop. (KL) resulted in 14.78 dB (with 50 iterations) and 15.20 dB (with 30 iterations), respectively. That is, the proposed method performed well even with less number of iterations. Note that the use of divergences other than KL performed better in some situations. Hence, the proposed method has a potential of further improvement by using a more appropriate divergence.

V. CONCLUSION

We proposed an amplitude and phase refinement method for MSS in the T-F domain. The proposed method is formulated as an optimization problem, and an efficient algorithm was developed based on ADMM for solving it. The proposed method using the KL divergence as a dissimilarity measure of amplitudes performed well in speech-nonspeech separation.

[3]In the SS-based ratio masking, the amplitude of the noise is replaced by its time average: $\Lambda_2[t,f] = (1/T)\sum_{t=1}^{T}|S_2[t,f]|$. Then, the amplitude of the speech was calculated as follows [3] $\Lambda_1[t,f] = \max(Y[t,f] - \Lambda_2[t,f], 0)$.

[4]In the SNMF-based ratio masking, first, KL-NMF was applied to the amplitude spectrogram of each source $|S_k[t,f]|$ for 300 iterations, where the rank was set to 30. Then, SNMF [42] was applied to the amplitude of the mixture for obtaining $\mathbf{\Lambda}_k$, where the basis of NMF were fixed.

[5]We measured run time using a signal whose length was the average in the dataset (3.14 s) and AMD Ryzen 7 1800X with MATLAB 2020a.

REFERENCES

[1] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. Int. Soc. Music Inf. Retrieval*, Sep. 2018, pp. 334–340.

[2] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[5] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[6] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 7092–7096.

[7] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, pp. 246–250.

[8] M. Kolbæk, D. Yu, Z. H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[9] Z. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2018, pp. 686–690.

[10] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[11] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[12] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. Int. Conf. Digit. Audio Effects (DAFx)*, Sep. 2010, pp. 397–403.

[13] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 421–424, May 2010.

[14] N. Sturmel and L. Daudet, "Iterative phase reconstruction of Wiener filtered signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 101–104.

[15] P. Magron, R. Badeau, and B. David, "Model-based STFT phase recovery for audio source separation," *IEEE/ACM Trans. Audio, Speech Lang. Proc.*, vol. 26, no. 6, pp. 1095–1105, Jun. 2018.

[16] Z. Ni and M. I. Mandel, "Mask-dependent phase estimation for monaural speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2020, pp. 7269–7273.

[17] S. Wisdom *et al.*, "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2019, pp. 900–904.

[18] D. Wang, H. Kameoka, and K. Shinoda, "A modified algorithm for multiple input spectrogram inversion," in *Proc. Interspeech*, Sep. 2019, pp. 4569–4573.

[19] P. Magron and T. Virtanen, "Online spectrogram inversion for low-latency audio source separation," *IEEE Signal Process. Lett.*, vol. 27, pp. 306–310, Jan. 2020.

[20] Z.-Q. Wang, J. Le Roux, D. L. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, Sep. 2018, pp. 2708–2712.

[21] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 2, pp. 370–382, May 2019.

[22] A. Cichocki, S. Cruces, and S. Amari, "Generalized alpha-beta divergences, and their application to robust nonnegative matrix factorization," *Entropy*, vol. 13, no. 1, pp. 134–170, Jan. 2011.

[23] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with NMF: Divergences, constraints, and algorithms," *Audio Source Separation*. Berlin, Germany: Springer, Mar. 2018, pp. 1–24.

[24] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, vol. 3, Delft, The Netherlands: Now Publishers, Jan. 2010.

[25] W. Deng and W. Yin, "On the global and linear convergence of the generalized alternating direction method of multipliers," *J. Sci. Comput.*, vol. 66, no. 3, pp. 889–916, Mar. 2016.

[26] P. Záviška, P. Rajmic, O. Mokrý, and Z. Průša, "A proper version of synthesis-based sparse audio declipper," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2019, pp. 591–595.

[27] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Griffin–Lim like phase recovery via alternating direction method of multipliers," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 184–188, Jan. 2019.

[28] D. L. Sun and C. Févotte, "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence," *IEEE Int. Conf. Acoust., Speech, Signal Process.*, Jul. 2014, pp. 6201–6205.

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

[30] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 3, pp. 780–791, Mar. 2007.

[31] H. Kagami and M. Yukawa, "Supervised nonnegative matrix factorization with Dual-Itakura–Saito and Kullback–Leibler divergences for music transcription," in *Proc. Eur. Signal Process. Conf.*, Dec. 2016, pp. 1138–1142.

[32] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Opt.*, vol. 1, no. 3, pp. 127–239, Jan. 2014.

[33] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs, "A variational formulation for frame-based inverse problems," *Inverse Problem*, vol. 23, no. 4, pp. 1495–1518, Jun. 2007.

[34] K. Oyamada, H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, and H. Ando, "Generative adversarial network-based approach to signal reconstruction from magnitude spectrograms," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2514–2518.

[35] N. Sturmel and L. Daudet, "Informed source separation using iterative reconstruction," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 1, pp. 178–185, Jan. 2013.

[36] 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. Int. Conf. Latent Variable Anal. Signal Separation*, Sep. 2010, pp. 89–96.

[37] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, Mar. 2013.

[38] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," *Natl. Inst. Stand. Technol.*, Gaithersburg, MD, USA, 1993.

[39] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*. Hoboken, NJ, USA: Wiley, 2016.

[40] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task, and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, Dec. 2015, pp. 504–511.

[41] A. J. E. M. Janssen and T. Strohmer, "Characterization and computation of canonical tight windows for Gabor frames," *J. Fourier Anal. Appl.*, vol. 8, no. 1, pp. 1–28, Jan. 2002.

[42] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation*, Sep. 2007, pp. 414–421.

[43] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2015, pp. 708–712.

[44] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR—Half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2019, pp. 626–630.

[45] "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," ITU-T Rec. P.862.2, Geneva, Switzerland, 2005.

[46] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process,*, vol. 19, no. 7, pp. 2155–2136, Sep. 2011.