Simultaneous Declipping and Beamforming via Alternating Direction Method of Multipliers

Yoshiki Masuyama, Tomoro Tanaka, Kohei Yatabe, Tsubasa Kusano, Yasuhiro Oikawa
Department of Intermedia Art and Science, Waseda University, Tokyo, Japan

***Abstract*—Audio declipping aims to reconstruct the original signal from a clipped observed signal. Clipping has a negative effect on subsequent multichannel processing such as beamforming. This is because the channel-wise nonlinear distortion deteriorates the spatial covariance matrix in the time-frequency domain. Hence, it is important to consider multichannel audio declipping that explicitly takes into account the subsequent audio signal processing. However, most of the existing methods have been developed for the single-channel case. In this paper, we propose a joint optimization method for declipping and beamforming. The multichannel declipped signal and the spatial filter are jointly optimized to sparsify the output of beamforming. Our experimental results show the effectiveness of the proposed method compared to the beamforming with prior declipping.**

***Index Terms*—Multichannel audio declipping, source separation, sparsity, weighted ℓ_1 norm, nonconvex optimization,**

I. Introduction

Clipping is one of the common nonlinear distortions of audio signals, which is caused by an inadequate dynamic range of an acquisition pipeline. Specifically, samples of the signal that exceed the range are replaced by the maximum or minimum threshold. This nonlinear distortion deteriorates the perceptual quality [1] and degrades the performance of subsequent audio signal processing [2]. To address these problems, various audio declipping methods have been presented, including the autoregressive-model-based method [3], nonnegative matrix factorization (NMF)–based method [4], deep neural network (DNN)–based method [5], and sparsity-based methods [6]–[11]. In particular, the sparsity-based methods have been intensively studied in recent years to improve the quality of reconstructed signals [10], [11].

Various electronic devices (e.g., tablets, video cameras, and smart speakers) acquire multichannel audio signals. Some applications using these devices, such as audio monitoring, can receive benefit from declipping. Since declipping allows them to set a higher gain, audio signals emitted from far away can be captured with less quantization error. However, most of the existing methods for audio declipping have been developed for the single-channel case. In the multichannel case, the inter-channel dependencies of the signal should be helpful, and therefore some multichannel methods have been presented [12], [13]. By leveraging the inter-channel dependencies, the multichannel methods outperformed single-channel declipping methods.

Declipping of multichannel audio signals is important for subsequent multichannel audio signal processing. For instance, beamforming and dereverberation significantly rely on the inter-channel dependencies provided as the spatial covariance matrix and the acoustic transfer function [14]–[17]. Such inter-channel dependencies are collapsed by channel-wise nonlinear distortion, and thus declipping is desired. A straightforward approach is the cascading strategy: declipping is firstly applied, and then the multichannel audio signal processing is performed. However, in this strategy, declipping is independent and does not take into account the objective of the subsequent multichannel processing. It should be advantageous to perform declipping and that processing simultaneously.

In this paper, we propose an optimization-based method for simultaneous declipping and beamforming. As an example of the multichannel audio signal processing, this paper considers target speech extraction based on beamforming. The proposed method jointly optimizes the spatial filter and the declipped signals to sparsify the beamformer output. The sparsity of speech signals is considered because it is effective for both declipping [11] and source separation [18]–[20]. We adopt the proximal alternating direction method of multipliers (ADMM) [21], [22] for the joint optimization. In our experiment, the proposed method improved the source-to-distortion ratio (SDR) and source-to-interference ratio (SIR) [23] of the extracted signals more than those of the cascading strategy.

II. Preliminaries

A. Model of Observed Signal

Let a time-domain signal $\mathbf{x}_m = [x_m(1), \ldots, x_m(L)]^\mathsf{T}$ be observed by M microphones, where $m = 1, \ldots, M$ is the index of microphones, L is the length of the signal, and $(\cdot)^\mathsf{T}$ denotes the transpose. This signal is assumed to be given by

$$x_m(l) = (h_m * s)(l) + n_m(l), \tag{1}$$

where $\mathbf{h}_m$ is the room impulse response (RIR) from the target speaker to the mth microphone, $\mathbf{s}$ is the target signal, $\mathbf{n}_m$ is the sum of all interferers observed by the mth microphone, and $*$ denotes the convolution.

Assuming a hard clipping caused by the limitation of an acquisition pipeline, the observed signal is expressed as

$$y_m(l) = \begin{cases} \eta & (x_m(l) \geq \eta) \\ x_m(l) & (-\eta < x_m(l) < \eta) \\ -\eta & (x_m(l) \leq -\eta) \end{cases}, \tag{2}$$

where $\eta > 0$ is the threshold of the hard clipping. In this paper, we aim to estimate the declipped target signal $(h_m * s)(l)$ from the clipped audio mixture $y_m(l)$.

B. Single-channel Audio Declipping Based on Sparsity

A sparsity-based audio declipping problem has been typically formulated as the following optimization problem [8][1]:

$$\min_{\mathbf{z}_m} \ \mathcal{S}(\mathcal{G}(\mathbf{z}_m)) \quad \text{s.t.} \ \mathbf{z}_m \in \mathcal{X}_m, \tag{3}$$

where $\mathcal{G}(\cdot)$ denotes the short-time Fourier transform (STFT), $\mathcal{S}(\cdot)$ is a sparsity-promoting function such as the ℓ_1 norm, and $\mathbf{z}_m$ is a declipped signal. In the constraint of (3), $\mathcal{X}_m$ is the set of time-domain signals that satisfy the clipping consistency [10] defined as follows:

$$\mathcal{X}_m = \left\{ \mathbf{z}_m \in \mathbb{R}^L \;\middle|\; \begin{array}{ll} z_m(l) \geq \eta & (y_m(l) = \eta) \\ z_m(l) = y_m(l) & (-\eta < y_m(l) < \eta) \\ z_m(l) \leq -\eta & (y_m(l) = -\eta) \end{array} \right\}. \tag{4}$$

The original signal can be reconstructed to some extent by sparsifying the signal under the clipping consistency constraint because the clipping collapses the sparsity of the original signal. We refer the reader to survey papers [10], [11] for more details of audio declipping for the single-channel case.

C. Distortionless Beamforming

Target speech extraction has been performed by beamforming, which can be efficiently implemented in the time-frequency (T-F) domain. Let STFT coefficients of the unclipped mixture $\mathbf{x}_m$ be represented as

$$\boldsymbol{\chi}(t,f) = [\mathcal{G}(\mathbf{x}_1)(t,f), \ldots, \mathcal{G}(\mathbf{x}_M)(t,f)]^{\mathsf{T}}, \tag{5}$$

where $\mathcal{G}(\mathbf{x}_m)(t,f)$ is the (t,f)th entry of the STFT coefficients of $\mathbf{x}_m$, and $t = 1, \ldots, T$ and $f = 1, \ldots, F$ are the time and frequency indices, respectively. Beamforming extracts the target signal by a spatial filter $\boldsymbol{\omega}(f) = [\omega_1(f), \ldots, \omega_M(f)]^{\mathsf{T}}$:

$$\phi(t,f) = \boldsymbol{\omega}(f)^{\mathsf{H}} \boldsymbol{\chi}(t,f), \tag{6}$$

where $(\cdot)^{\mathsf{H}}$ denotes the Hermitian transpose. One of the most popular beamformers is the minimum power distortionless response (MPDR) beamformer [24], which is obtained by solving the following optimization problem:

$$\min_{\boldsymbol{\omega}(f)} \ \sum_{t=1}^{T} |\boldsymbol{\omega}(f)^{\mathsf{H}} \boldsymbol{\chi}(t,f)|^2 \quad \text{s.t.} \ \boldsymbol{\omega}(f) \in \mathcal{A}_f, \tag{7}$$

where $\mathcal{A}_f$ is the set of spatial filters that satisfy the distortionless constraint,

$$\mathcal{A}_f = \{\boldsymbol{\omega}(f) \in \mathbb{C}^M \mid \boldsymbol{\omega}(f)^{\mathsf{H}} \boldsymbol{\alpha}(f) = 1\}, \tag{8}$$

and $\boldsymbol{\alpha}(f) = [1, \alpha_2(f), \ldots, \alpha_M(f)]^{\mathsf{T}}$ is the relative transfer function (RTF) of the target signal estimated in advance.

Although the MPDR beamformer has been widely used, its performance may not be superb due to the residual interferers. To tackle this problem, the sparse distortionless response (SPDR) beamformer [25] has been presented, where the ℓ_1 norm of the extracted signal (i.e., $\sum_{t=1}^{T} |\boldsymbol{\omega}(f)^{\mathsf{H}} \boldsymbol{\chi}(t,f)|$) is minimized for each frequency under the distortionless constraint. As a result of exploiting the sparsity of the target signal, SPDR outperforms MPDR.

[1]Although the audio declipping problem in (3) was proposed for the single-channel case [8], we add the subscript m for consistency of this paper. This sparsity-based method is thus conducted independently to each channel.

III. Simultaneous Declipping and Beamforming

In this section, we propose an optimization-based method for simultaneous declipping and beamforming. Under the constraints considered in the single-channel audio declipping and the distortionless beamforming, the proposed method sparsifies the output of beamforming via the proximal ADMM.

A. Proposed Formulation

When an M-channel clipped audio mixture $(\mathbf{y}_1, \ldots, \mathbf{y}_M)$ is observed, we can estimate the declipped target signal by a cascading strategy. First, the channel-wise audio decelipping in (3) is conducted to estimate a declipped mixture $(\mathbf{x}_1, \ldots, \mathbf{x}_M)$. Then, the distortionless beamforming in (7) is applied to STFT of the declipped mixture. However, in this strategy, each of the unclipped audio mixture $\mathbf{x}_m$ is assumed to be sparse, which is not as appropriate as in the single-source case. In addition, the declipping does not take into account the subsequent beamforming.

To address these problems, we propose simultaneous declipping and beamforming so that the spatially filtered signal is sparsified (instead of the mixture). Taking over the clipping consistency constraint in (3) and the distortionless constraint in (7), the proposed optimization problem is formulated as

$$\min_{\substack{(\boldsymbol{\omega}(1),\ldots,\boldsymbol{\omega}(F),\\ \mathbf{z}_1,\ldots,\mathbf{z}_M)}} \ \sum_{t=1}^{T} \sum_{f=1}^{F} \lambda(f) \, |\boldsymbol{\omega}(f)^{\mathsf{H}} \boldsymbol{\gamma}(t,f)|, \tag{9a}$$

$$\text{s.t.} \quad \boldsymbol{\gamma}(t,f) = [\mathcal{G}(\mathbf{z}_1)(t,f), \ldots, \mathcal{G}(\mathbf{z}_M)(t,f)]^{\mathsf{T}}, \quad \boldsymbol{\omega}(f) \in \mathcal{A}_f, \quad \mathbf{z}_m \in \mathcal{X}_m, \tag{9b}$$

where $\lambda(f) > 0$ is a weight calculated in advance. This optimization problem estimates a declipped signal such that the output of beamforming is sparse. When the observed signal is not clipped (i.e., $\mathcal{X}_m = \{\mathbf{x}_m\}$), the proposed method coincides with SPDR[2], and thus it can be interpreted as an extension of SPDR for clipped signals.

The cost function promotes sparsity of the extracted target signal in the T-F domain by penalizing its weighted ℓ_1 norm in (9a). We define the weight $\lambda(f)$ as follows:

$$\lambda(f) = \frac{\varepsilon}{\varepsilon + \kappa(f)/\mathrm{Max}(\kappa(1), \ldots, \kappa(F))}, \tag{10}$$

$$\kappa(f) = \mathrm{Mean}(|\hat{\phi}(1,f)|, \ldots, |\hat{\phi}(T,f)|), \tag{11}$$

where $\varepsilon > 0$ is a parameter for adjusting the weight, and $\mathrm{Max}(\cdot)$ and $\mathrm{Mean}(\cdot)$ return the maximum and mean of the inputted tuple, respectively. Here, $\hat{\phi}(t,f)$ is the target signal estimated by an existing method. This weight becomes smaller as the magnitude of the estimated target $|\hat{\phi}(t,f)|$ becomes larger so that components in that frequency are not penalized so much. We stress that use of the weight in (10)–(11) is unique to the proposed method that jointly performs declipping and beamforming.

Our proposed method differs from the existing multichannel declipping methods [12], [13] in the following two respects.

[2]The optimization problem of SPDR is independent for each frequency. Hence, the frequency-wise weight can be omitted in the optimization.

First, the proposed method uses RTF to handle the inter-channel dependency of the target signal, while the existing methods do not use such information. Second, they are based on models of a multichannel mixture, e.g., structured sparsity across channels. In contrast, the proposed method assumes the sparsity only on the target signal and does not require the sparsity of interferers. Recently, a joint declipping and separation method based on nonnegative tensor factorization has been proposed for a single-channel case [26]. The proposed method handles a multichannel case and is based on beamforming.

B. ADMM for Solving Optimization Problem in (9)

In the proposed method, the nonconvex optimization problem in (9) must be solved. To minimize the weighted ℓ_1 norm of the output signal under the constraints, we use the proximal ADMM [21], [22]. It can handle multiple cost functions and constraints separately, similar to the typical ADMM [27] whose effectiveness has been confirmed in various applications of nonconvex optimization [8], [28]–[31].

To apply the proximal ADMM, we reformulate the optimization problem in (9) to the following form:

$$\min_{\substack{(\boldsymbol{\omega}(1),\ldots,\boldsymbol{\omega}(F),\\ \boldsymbol{\Gamma}_1,\ldots,\boldsymbol{\Gamma}_M,\boldsymbol{\Psi})}} \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{\Psi}) + \sum_{f=1}^{F} \iota_{\mathcal{A}_f}(\boldsymbol{\omega}(f)) + \sum_{m=1}^{M} \iota_{\mathcal{C}_m}(\boldsymbol{\Gamma}_m), \quad (12a)$$

$$\text{s.t.} \quad \psi(t,f) = \boldsymbol{\omega}(f)^{\mathsf{H}}\boldsymbol{\gamma}(t,f), \quad (12b)$$

where $\psi(t,f)$ and $\gamma_m(t,f)$ are the (t,f)th entries of $\boldsymbol{\Psi} \in \mathbb{C}^{T\times F}$ and $\boldsymbol{\Gamma}_m \in \mathbb{C}^{T\times F}$, respectively, the first term

$$\mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{\Psi}) = \sum_{t=1}^{T}\sum_{f=1}^{F} \lambda(f)\,|\psi(t,f)| \quad (13)$$

is the weighted ℓ_1 norm in (12a), and

$$\iota_{\mathcal{Q}}(x) = \begin{cases} 0 & (x \in \mathcal{Q}) \\ \infty & (x \notin \mathcal{Q}) \end{cases} \quad (14)$$

is the indicator function with respect to a set $\mathcal{Q}$. The constraints in (9b) are recast into the indicator functions in (12a), where $\mathcal{C}_m$ is the set of STFT coefficients that are obtained from a signal in the set of clipping consistency $\mathcal{X}_m$ [see (4)]:

$$\mathcal{C}_m = \{\boldsymbol{\Gamma}_m \in \mathbb{C}^{T\times F} \mid \exists\ \mathbf{z}_m \in \mathcal{X}_m,\ \boldsymbol{\Gamma}_m = \mathcal{G}(\mathbf{z}_m)\}. \quad (15)$$

With $\rho > 0$, the augmented Lagrangian of (12) is given by

$$\begin{aligned}\mathcal{L}_\rho(\widetilde{\boldsymbol{\omega}}, \widetilde{\boldsymbol{\Gamma}}, \boldsymbol{\Psi}, \boldsymbol{\Theta}) &= \mathcal{S}_{\boldsymbol{\lambda}}(\boldsymbol{\Psi}) + \sum_{f=1}^{F} \iota_{\mathcal{A}_f}(\boldsymbol{\omega}(f)) + \sum_{m=1}^{M} \iota_{\mathcal{C}_m}(\boldsymbol{\Gamma}_m) \\ &\quad + \sum_{t=1}^{T}\sum_{f=1}^{F} \bar{\theta}(t,f)\left(\boldsymbol{\omega}(f)^{\mathsf{H}}\boldsymbol{\gamma}(t,f) - \psi(t,f)\right) \\ &\quad + \frac{\rho}{2}\sum_{t=1}^{T}\sum_{f=1}^{F} |\boldsymbol{\omega}(f)^{\mathsf{H}}\boldsymbol{\gamma}(t,f) - \psi(t,f)|^2, \quad (16)\end{aligned}$$

where $\boldsymbol{\Theta} \in \mathbb{C}^{T\times F}$ is a dual variable, $\widetilde{\boldsymbol{\omega}} = (\boldsymbol{\omega}(1),\ldots,\boldsymbol{\omega}(F))$, $\widetilde{\boldsymbol{\Gamma}} = (\boldsymbol{\Gamma}_1,\ldots,\boldsymbol{\Gamma}_M)$, and $\bar{(\cdot)}$ denotes the complex conjugate.

By using the augmented Lagrangian given in (16), the proximal ADMM for (12) can be written as follows[3]:

$$\begin{aligned}\widetilde{\boldsymbol{\Gamma}}^{[k+1]} &\leftarrow \operatorname*{argmin}_{\widetilde{\boldsymbol{\Gamma}}}\ \mathcal{L}_\rho(\widetilde{\boldsymbol{\omega}}^{[k]}, \widetilde{\boldsymbol{\Gamma}}, \boldsymbol{\Psi}^{[k]}, \boldsymbol{\Theta}^{[k]}) \\ &\qquad + \frac{\beta}{2}\sum_{m=1}^{M}\left\|\boldsymbol{\Gamma}_m - \boldsymbol{\Gamma}_m^{[k]}\right\|_{\mathrm{Fro}}^2, \quad (17)\end{aligned}$$

$$\begin{aligned}\widetilde{\boldsymbol{\omega}}^{[k+1]} &\leftarrow \operatorname*{argmin}_{\widetilde{\boldsymbol{\omega}}}\ \mathcal{L}_\rho(\widetilde{\boldsymbol{\omega}}, \widetilde{\boldsymbol{\Gamma}}^{[k+1]}, \boldsymbol{\Psi}^{[k]}, \boldsymbol{\Theta}^{[k]}) \\ &\qquad + \frac{\beta}{2}\sum_{f=1}^{F}\left\|\boldsymbol{\omega}(f) - \boldsymbol{\omega}(f)^{[k]}\right\|_2^2, \quad (18)\end{aligned}$$

$$\phi(t,f)^{[k+1]} \leftarrow \boldsymbol{\omega}(f)^{[k+1]\mathsf{H}}\boldsymbol{\gamma}(t,f)^{[k+1]} \quad \forall\,(t,f), \quad (19)$$

$$\boldsymbol{\Psi}^{[k+1]} \leftarrow \mathcal{T}_{\boldsymbol{\lambda}}\Big(\boldsymbol{\Phi}^{[k+1]} + \frac{1}{\rho}\boldsymbol{\Theta}^{[k]}\Big), \quad (20)$$

$$\boldsymbol{\Theta}^{[k+1]} \leftarrow \boldsymbol{\Theta}^{[k]} + \rho(\boldsymbol{\Phi}^{[k+1]} - \boldsymbol{\Psi}^{[k+1]}), \quad (21)$$

where $\beta > 0$, $\|\cdot\|_{\mathrm{Fro}}$ is the Frobenius norm, $\|\cdot\|_2$ is the Euclidean norm, and $k = 1,\ldots,K$ is the iteration index. The weighted soft-thresholding $\mathcal{T}_{\boldsymbol{\lambda}}(\cdot)$ in (20) is given by

$$\mathcal{T}_{\boldsymbol{\lambda}}(\boldsymbol{\Xi})(t,f) = \mathrm{Max}\Big(1 - \frac{\lambda(f)}{|\xi(t,f)|}, 0\Big)\,\xi(t,f). \quad (22)$$

We apply the projected gradient method (PGM) to (17) for updating $\widetilde{\boldsymbol{\Gamma}}$. Its detail is given in Appendix A. The analytic solution of the subproblem in (18) can be obtained by solving the Karush–Kuhn–Tucker (KKT) system. Detail of the updating formula of $\widetilde{\boldsymbol{\omega}}$ is given in Appendix B.

IV. Experimental Evaluation

The effectiveness of the proposed method was evaluated in target speech extraction from clipped multichannel audio mixtures. The proposed method (`Prop`) was compared with direct beamforming (`MPDR`, `SPDR`), cascaded declipping and beamforming (`D-MPDR`, `D-SPDR`), and beamforming applied to the oracle unclipped mixtures (`U-MPDR`, `U-SPDR`).

A. Experimental Settings

As dry source signals of both targets and interferers, utterances from the Voice Conversion Challenge (VCC) 2018 dataset [32] were used[4]. They were resampled at 16 kHz. To simulate convolutive mixtures in a rectangular room, the `pyroomacoustics` toolbox [34] was used. The room size was 5.0 m × 3.5 m × 2.5 m, and the reverberation time was uniformly distributed in $[0.2, 0.4]$ s. A circular microphone array with 3 channels, whose radius was 10 cm, was located at the center of the room. The target speaker and one interference talker were randomly located 1.0 m from the array center.

Each multichannel audio mixture was scaled such that the maximum magnitude of the whole mixture became 1.0 in the time domain. Then, the mixtures were clipped at $\eta \in \{0.1, 0.2, 0.3, 0.5\}$. For each clipping level, 10 mixtures

[3]For simplicity, some options of the proximal ADMM are omitted from (17)–(21). See [21], [22] for details and the condition for convergence.

[4]The VCC 2018 dataset has been used for evaluation of not only voice conversion but also multichannel sound source separation [33].

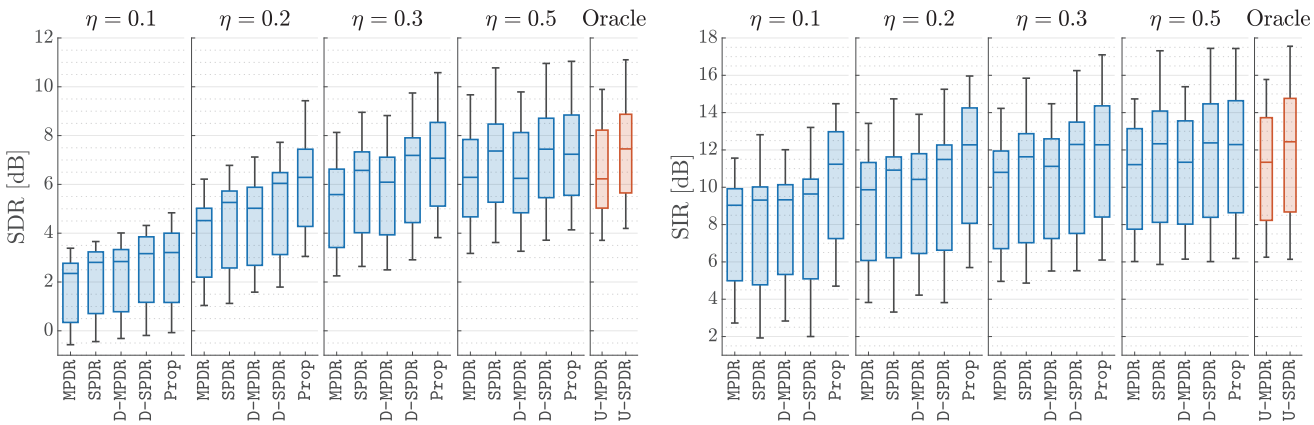

Fig. 1. Comparison of SDR and SIR of the extracted target signals with differnet clipping levels η. Blue boxes represent the performance of the only beamforming (MPDR, SPDR), the cascading strategies (D-MPDR, D-SPDR), and the proposed joint optimization (Prop). Red boxes are for MPDR and SPDR to the oracle unclipped mixtures (U-MPDR, U-SPDR). The central lines indicate the median, and the boxes correspond to the first and third quartiles.

were generated. Target speech extraction was performed for both of the two utterances in the mixture independently. Thus, we evaluated 20 extracted signals for each clipping level.

For the cascading strategy, the sparsity-promoting function for declipping was the ℓ_1 norm. We used eigenvalue decomposition to estimate RTF of the target signal from another unclipped signal uttered from the same position as the target speaker. For the proposed method, the weight $\boldsymbol{\lambda}$ in (10) and the initial values were calculated by cascading the channel-wise declipping and MPDR, i.e., D-MPDR. Other parameters were set to $\rho = 1$, $\beta = 0.001$, and $\varepsilon = 0.001$. The main proximal ADMM and PGM for updating $\widetilde{\boldsymbol{\Gamma}}$ were iterated at 1000 and 30 times, respectively. STFT was implemented with the 64 ms Hann window with a 16 ms shift. The performance of target speech extraction was measured by SDR and SIR [23].

B. Experimental Results

The experimental results are illustrated in Fig. 1. As shown in the red boxes, SPDR outperformed MPDR in the case of unclipped mixtures, which indicates the effectiveness of exploiting the sparsity of the target signal. When $\eta = 0.5$ (i.e., clipping level was moderate), both the cascading strategy (D-SPDR) and the joint optimization (Prop) achieved the performance similar to that for the oracle case (U-SPDR). Prop outperformed the other methods when the clipping was more severe, i.e., $\eta \in \{0.1, 0.2, 0.3\}$. This indicates the effectiveness of the joint optimization. Especially, as the result of sparsifying the extracted signal, the median of SIR was improved more than 1.5 dB over D-SPDR when $\eta = 0.1$.

V. Conclusion

In this paper, we presented a joint optimization method for declipping and beamforming. In the proposed method, the multichannel declipped signal and the spatial filter are optimized to sparsify the extracted signal via the proximal ADMM. Our experimental results show the effectiveness of the joint optimization compared to the straightforward cascading strategy. Our future work includes the joint optimization method for declipping and blind source separation.

Appendix

A. Projected Gradient Method for Updating $\widetilde{\boldsymbol{\Gamma}}$ in (17)

In the proximal ADMM, $\widetilde{\boldsymbol{\Gamma}}$ is updated in (17) by minimizing the following convex function $\mathcal{M}^{[k]}(\cdot)$:

$$\mathcal{M}^{[k]}(\widetilde{\boldsymbol{\Gamma}}) = \sum_{m=1}^{M} \iota_{\mathcal{C}_m}(\boldsymbol{\Gamma}_m) + \frac{\beta}{2} \sum_{m=1}^{M} \left\| \boldsymbol{\Gamma}_m - \boldsymbol{\Gamma}_m^{[k]} \right\|_{\mathrm{Fro}}^2 + \frac{\rho}{2} \sum_{t=1}^{T} \sum_{f=1}^{F} |\boldsymbol{\omega}(f)^{[k]\mathsf{H}} \boldsymbol{\gamma}(t,f) - \nu(t,f)^{[k]}|^2, \quad (23)$$

where $\nu(t,f)^{[k]} = \psi(t,f)^{[k]} - \theta(t,f)^{[k]}/\rho$. To minimize $\mathcal{M}^{[k]}(\widetilde{\boldsymbol{\Gamma}})$, PGM iteratively updates $(\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_M)$ as follows:

$$\boldsymbol{\upsilon}(t,f) \leftarrow \boldsymbol{\gamma}(t,f) - \mu\Big(\boldsymbol{\Pi}(f)\boldsymbol{\gamma}(t,f) - \boldsymbol{\vartheta}(t,f)\Big) \ \forall (t,f), \quad (24)$$

$$\boldsymbol{\Gamma}_m \leftarrow \mathcal{P}_{\mathcal{C}_m}(\boldsymbol{\Upsilon}_m) \ \forall m, \quad (25)$$

where the (t,f)th entry of $\boldsymbol{\Upsilon}_m$ is $\upsilon_m(t,f)$, $\mu > 0$ is a step size, $\boldsymbol{\Pi}(f) = \rho\boldsymbol{\omega}(f)^{[k]}\boldsymbol{\omega}(f)^{[k]\mathsf{H}} + \beta\mathbf{I}$ ($\mathbf{I}$ is the identity matrix), and $\boldsymbol{\vartheta}(t,f) = \rho\boldsymbol{\omega}(f)^{[k]}\nu(t,f)^{[k]} + \beta\boldsymbol{\gamma}(t,f)^{[k]}$. Here, $\mathcal{P}_{\mathcal{C}_m}(\cdot)$ is the projection onto the set $\mathcal{C}_m$. Assuming that STFT is a parseval tight frame [8], $\mathcal{P}_{\mathcal{C}_m}(\cdot)$ is given by

$$\mathcal{P}_{\mathcal{C}_m}(\boldsymbol{\Gamma}_m) = \mathcal{G}(\mathcal{P}_{\mathcal{X}_m}(\mathcal{G}^{\dagger}(\boldsymbol{\Gamma}_m))), \quad (26)$$

where $\mathcal{G}^{\dagger}(\cdot)$ is the inverse STFT, $\mathcal{P}_{\mathcal{X}_m}(\cdot)$ is given by

$$\mathcal{P}_{\mathcal{X}_m}(\mathbf{z}_m)(l) = \begin{cases} \mathrm{Max}(z_m(l), \eta) & (y_m(l) = \eta) \\ y_m(l) & (-\eta < y_m(l) < \eta) \\ \mathrm{Min}(z_m(l), -\eta) & (y_m(l) = -\eta) \end{cases}, \quad (27)$$

and $\mathrm{Min}(\cdot)$ returns the minimum of its inputs. PGM in (24)–(25) is used in each iteration of the proximal ADMM. Note that the number of iterations for PGM can be small because the subproblem is allowed to be solved approximately.

B. KKT System for Updating $\widetilde{\boldsymbol{\omega}}$ in (18)

According to (18), the spatial filter $\widetilde{\boldsymbol{\omega}}$ is updated by minimizing the following convex function $\mathcal{N}^{[k]}(\cdot)$:

$$\mathcal{N}^{[k]}(\widetilde{\boldsymbol{\omega}}) = \sum_{f=1}^{F} \iota_{\mathcal{A}_f}(\boldsymbol{\omega}(f)) + \frac{\beta}{2}\sum_{f=1}^{F}\left\|\boldsymbol{\omega}(f) - \boldsymbol{\omega}(f)^{[k]}\right\|_2^2 + \frac{\rho}{2}\sum_{t=1}^{T}\sum_{f=1}^{F}|\boldsymbol{\omega}(f)^{\mathsf{H}}\boldsymbol{\gamma}(t,f)^{[k+1]} - \nu(t,f)^{[k]}|^2, \quad (28)$$

Considering the KKT conditions for the global minimum, we obtain the following linear system for each frequency:

$$\begin{pmatrix} \boldsymbol{\Sigma}(f) & \boldsymbol{\alpha}(f) \\ \boldsymbol{\alpha}(f)^{\mathsf{H}} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\omega}(f)^{\star} \\ \varrho(f)^{\star} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\zeta}(f) \\ 1 \end{pmatrix}, \quad (29)$$

where $\varrho(f)^{\star} \in \mathbb{R}$ is KKT multiplier, $\boldsymbol{\alpha}(f)$ is RTF in (8), and

$$\boldsymbol{\Sigma}(f) = \rho\sum_{t=1}^{T}\boldsymbol{\gamma}(t,f)^{[k+1]}\boldsymbol{\gamma}(t,f)^{[k+1]\mathsf{H}} + \beta\mathbf{I}, \quad (30)$$

$$\boldsymbol{\zeta}(f) = \rho\sum_{t=1}^{T}\boldsymbol{\gamma}(t,f)^{[k+1]}\bar{\nu}(t,f)^{[k]} + \beta\boldsymbol{\omega}(f)^{[k]}. \quad (31)$$

Since this KKT system is nonsingular, the solution to the linear system in (29) can be analytically calculated. Then, $\widetilde{\boldsymbol{\omega}}^{[k+1]}$ is obtained as $(\boldsymbol{\omega}(1)^{\star}, \ldots, \boldsymbol{\omega}(F)^{\star})$.

References

[1] C. Tan, B. R. J. Moore, and N. Zacharov, "The effect of nonlinear distortion on the perceived quality of music and speech signals," *J. Audio Eng. Soc.*, vol. 51, no. 11, pp. 1012–1031, Nov. 2003.

[2] M. J. Harvilla and R. M. Stern, "Least squares signal declipping for robust speech recognition," in *Interspeech*, Sept. 2014, pp. 2073–2077.

[3] A. J. E. M. Janssen, R. N. J. Veldhuis, and L Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 2, pp. 317–330, Apr. 1986.

[4] Ç. Bilen, A. Ozerov, and P. Pérez, "Audio declipping via nonnegative matrix factorization," in *IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2015, pp. 1–5.

[5] W. Mack and E. A. P. Habets, "Declipping speech using deep filtering," in *IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2019, pp. 200–204.

[6] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "A constrained matching pursuit approach to audio declipping," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2011, pp. 329–332.

[7] S. Kitic, L. Jacques, N. Madhu, M. P. Hopwood, A. Spriet, and C. De Vleeschouwer, "Consistent iterative hard thresholding for signal declipping," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 5939–5943.

[8] S. Kitić, N. Bertin, and R. Gribonval, "Sparsity and cosparsity for audio declipping: A flexible non-convex approach," in *Int. Conf. Latent Var. Anal. Signal Sep. (LVA/ICA)*, Aug. 2015, pp. 243–250.

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

[10] P. Záviška, P. Rajmic, A. Ozerov, and L. Rencker, "A survey and an extensive evaluation of popular audio declipping methods," *IEEE J. Sel. Top. Signal Process.*, vol. 15, no. 1, pp. 5–24, Jan. 2021.

[11] C. Gaultier, S. Kitić, R. Gribonval, and N. Bertin, "Sparsity-based audio declipping methods: Selected overview, new algorithms, and large-scale evaluation," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 29, pp. 1174–1187, Feb. 2021.

[12] A. Ozerov, Ç. Bilen, and P. Pérez, "Multichannel audio declipping," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2016, pp. 659–663.

[13] C. Gaultier, N. Bertin, and R. Gribonval, "Cascade: Channel-aware structured cosparse audio declipper," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 571–575.

[14] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 18, no. 7, pp. 1830–1840, Sept. 2010.

[15] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 21, no. 7, pp. 1369–1380, July 2013.

[16] O. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 32, no. 2, pp. 240–251, Nov. 2014.

[17] T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, and R. Haeb-Umbach, "Jointly optimal denoising, dereverberation, and source separation," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 28, pp. 2267–2282, July 2020.

[18] Y. Li, S. Amari, A. Cichocki, D. W. C. Ho, and S. Xie, "Underdetermined blind source separation based on sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 423–437, Feb. 2006.

[19] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, Aug. 2007.

[20] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 18, no. 7, pp. 1818–1829, Sept. 2010.

[21] R. Shefi and M. Teboulle, "Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization," *SIAM J. Opt.*, vol. 24, no. 1, pp. 269–297, Feb. 2014.

[22] R. I. Boţ and D. K. Nguyen, "The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates," *Math. Oper. Res.*, vol. 45, no. 2, pp. 682–712, Mar. 2020.

[23] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 14, no. 6, pp. 1462–1469, July 2006.

[24] H. L. Van Trees, *Optimum array processing*, John Wiley & Sons, 2002.

[25] S. Emura, S. Araki, T. Nakatani, and N. Harada, "Distortionless beamforming optimized with ℓ_1-norm minimization," *IEEE Signal Process. Lett.*, vol. 25, no. 7, pp. 936–940, July 2018.

[26] Ç. Bilen, A. Ozerov, and P. Pérez, "Solving time-domain audio inverse problems using nonnegative tensor factorization," *IEEE Trans. Signal Process.*, vol. 66, no. 21, pp. 5604–5617, Nov. 2018.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Now Publishers Inc., Jan. 2010.

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

[29] 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.

[30] P. H. Vial, P. Magron, T. Oberlin, and C. Févotte, "Phase retrieval with bregman divergences and application to audio signal recovery," *IEEE J. Sel. Top. Signal Process.*, vol. 15, no. 1, pp. 51–64, Jan. 2021.

[31] T. Kusano, Y. Masuyama, K. Yatabe, and Y. Oikawa, "Designing nearly tight window for improving time-frequency masking," in *Int. Congr. Acoust. (ICA)*, Sept. 2019, pp. 2885–2892.

[32] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Odyssey*, June 2018, pp. 195–202.

[33] L. Li, H. Kameoka, S. Inoue, and S. Makino, "FastMVAE: A fast optimization algorithm for the multichannel variational autoencoder method," *IEEE Access*, vol. 8, pp. 228740–228753, Dec. 2020.

[34] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 351–355.