NON-GRIFFIN–LIM TYPE SIGNAL RECOVERY FROM MAGNITUDE SPECTROGRAM

Ryusei Nakatsu *Daichi Kitahara* *Akira Hirabayashi*

Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan

ABSTRACT

Speech and audio signal processing frequently requires to recover a time-domain signal from the magnitude of a spectrogram. Conventional methods inversely transform the magnitude spectrogram with a phase spectrogram recovered by the Griffin–Lim algorithm or its accelerated versions. The short-time Fourier transform (STFT) perfectly matches this framework, while other useful spectrogram transforms, such as the constant-Q transform (CQT), do not, because their inverses cannot be computed easily. To make the best of such useful spectrogram transforms, we propose an algorithm which recovers the time-domain signal without the inverse spectrogram transforms. We formulate the signal recovery as a nonconvex optimization problem, which is difficult to solve exactly. To approximately solve the problem, we exploit a stochastic convex optimization technique. A well-organized block selection enables us both to avoid local minimums and to achieve fast convergence. Numerical experiments show the effectiveness of the proposed method for both STFT and CQT cases.

***Index Terms*—** Spectrogram, signal recovery, Griffin–Lim algorithm, constant-Q transform, nonconvex stochastic optimization.

1. INTRODUCTION

Spectrogram is a useful expression for analysis of speech and audio signals. Many applications, including *sound source separation* [1]–[3] and *automatic music transcription* [4], [5], are performed in the spectral-domain using spectrograms rather than in the original time-domain. A spectrogram is typically defined as a complex matrix and can be divided into *magnitude* and *phase spectrograms*. Even though both components constitute the original spectrogram, only the magnitude spectrogram is manipulated in some applications [6]–[9]. For example, to extract a specific target sound from a mixed sound, the magnitude spectrogram of the target sound is estimated from that of the mixed sound by applying a well-designed time-frequency mask. Then, a time-domain signal, which is an estimate of the target sound, is recovered through the inverse transform from the estimated magnitude spectrogram with the observed mixed phase spectrogram [10].

However, the magnitude spectrogram of the recovered signal using the mixed phase will be different from the estimated magnitude. To bring the magnitude spectrogram of the recovered signal closer to the estimated one, we have to find a more appropriate phase spectrogram. This problem is often called *phase recovery* or *phase retrieval*. In general, the size of a sound signal is relatively large, and thus the *PhaseLift* approach [11]–[13] is difficult to use since it requires huge memory resources. A well-known method for the phase recovery of the sound signal is the Griffin–Lim algorithm (GLA) [14], which performs *alternating projections*. Since its convergence is slow, accelerated versions are proposed [15], [16]. Particularly, the method in [16] uses the *alternating direction method of multipliers (ADMM)* [17].

This work was supported in part by JSPS Grants-in-Aid Grant Number for Early-Career Scientists JP19K20361 (e-mail: d-kita@fc.ritsumei.ac.jp).

In the above Griffin–Lim (GL) type approach [14]–[16], the inverse of the spectrogram transform has to be computed both for update of the phase spectrogram and for the post-processing to recover a time-domain signal from the estimated complex spectrogram. The short-time Fourier transform (STFT), the most commonly used spectrogram transform, perfectly matches this framework because its inverse is computed easily thanks to the nature of the discrete Fourier transform. However, other useful spectrogram transforms such as the constant-Q transform (CQT) [18], [19] and wavelet transforms [20], which are often used for analysis of overtone components, are not fit for this framework because their inverses cannot be computed easily.

To make the best of such useful spectrogram transforms, we propose an algorithm which can directly recover a time-domain signal from the magnitude spectrogram without computing the inverse of the spectrogram transform. To this end, we first formulate the time-domain signal recovery using the time-domain signal itself while GL type algorithms [14]–[16] used a complex spectrogram as a variable to describe the problem. The formulated problem is nonconvex and difficult to solve exactly. Instead, we approximately solve the problem by applying a stochastic convex optimization technique to avoid local minimums and achieve fast convergence. Specifically, we use the *stochastic dual coordinate ascent ADMM (SDCA-ADMM)* [21], which exploits an inexact augmented Lagrangian and thus can avoid the inverse computation for any spectrogram transform. In the proposed method, the spectrogram is divided into several blocks because of the stochastic optimization. By randomly choosing *multiple blocks* in a well-organized novel manner, the proposed method can recover the time-domain signal with the same convergence performance as in [16]. Numerical experiments show the effectiveness of the proposed non-GL type signal recovery for STFT and CQT spectrograms.

2. GRIFFIN–LIM TYPE PHASE RECOVERY

Let $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{C}$ be the sets of all real, nonnegative real, and complex numbers, respectively. The imaginary unit is denoted by $\imath \in \mathbb{C}$. For any $c \in \mathbb{C}$, $\bar{c}$, $|c| \in \mathbb{R}_+$, and $\angle c \in (-\pi, \pi]$ stand for the complex conjugate, magnitude, and phase of c, respectively. We write vectors with boldface small letters and matrices with capital letters. We denote the transpose, adjoint, and inverse operators by $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\mathrm{H}}$, and $(\cdot)^{-1}$, respectively. The composition of mappings and the Hadamard product are denoted by $\circ$ and $\odot$. The floor function is denoted by $\lfloor \cdot \rfloor$.

2.1. Griffin–Lim Algorithm

Given a discrete-time signal $x[t]$, let $X := (\hat{x}_{i,j}) \in \mathbb{C}_{\updownarrow}^{L \times J}$ be the spectrogram defined by the short-time Fourier transform (STFT), as

$$\hat{x}_{i,j} := \sum_{\tau=0}^{L-1} w[\tau] x[(j-1)\xi + \tau] e^{-\imath \frac{2\pi(i-1)\tau}{L}},$$

where L is the frame length, J is the number of frames, ξ ($< L$) is the shift of a window function $w[\tau]$, and $\mathbb{C}_{\updownarrow}^{L \times J} := \{(\hat{x}_{i,j}) \in \mathbb{C}^{L \times J} \mid \hat{x}_{1,j} \in \mathbb{R}\ (\forall j)$ and $\overline{\hat{x}}_{i,j} = \hat{x}_{L-i+2,j}\ (\forall j$ and $i = 2, 3, \ldots, \lfloor \frac{L+2}{2} \rfloor)\}$

stands for the *spectrogram space*. We denote STFT by a linear mapping $\mathcal{S} : \mathbb{R}^{(J-1)\xi+L} \ni \boldsymbol{x} \mapsto \mathcal{S}(\boldsymbol{x}) = X \in \mathbb{C}_{\updownarrow}^{L\times J}$. Note that, in this definition, the spectrogram X includes $\hat{x}_{i,j}$ s.t. $i \in [\lfloor \frac{L+2}{2} \rfloor + 1, L]$.

Since $\xi < L$, the dimension JL of the spectrogram space $\mathbb{C}_{\updownarrow}^{L\times J}$ is greater than that of the time signal space given by $(J-1)\xi + L$. Therefore the *range* of STFT $\mathcal{R}_{\mathcal{S}} := \{X \in \mathbb{C}_{\updownarrow}^{L\times J} \mid \exists \boldsymbol{x}\ X = \mathcal{S}(\boldsymbol{x})\}$ is a low-dimensional subspace of $\mathbb{C}_{\updownarrow}^{L\times J}$. By exploiting this property, GLA finds a spectrogram $X \in \mathbb{C}_{\updownarrow}^{L\times J}$ which has an appropriate phase for a given magnitude spectrogram $M \in \mathbb{R}_{+,\updownarrow}^{L\times J}$ ($\subset \mathbb{C}_{\updownarrow}^{L\times J}$) in the intersection of $\mathcal{R}_{\mathcal{S}}$ and $\mathcal{M} := \{X \in \mathbb{C}_{\updownarrow}^{L\times J} \mid |X| := (|\hat{x}_{i,j}|) = M\}$. More specifically, GLA solves the following feasibility problem [14]:

$$\text{find } X \in \mathcal{M} \cap \mathcal{R}_{\mathcal{S}} \tag{1}$$

by alternating projections

$$X^{(l+1)} = \mathcal{P}_{\mathcal{R}_{\mathcal{S}}}(\mathcal{P}_{\mathcal{M}}(X^{(l)})), \tag{2}$$

where $\mathcal{P}_{\mathcal{M}}$ is the projection onto $\mathcal{M}$ given by

$$\mathcal{P}_{\mathcal{M}}(X) = M \odot \Theta_X \tag{3}$$

with the phase matrix $\Theta_X := (e^{\imath \angle \hat{x}_{i,j}}) \in \mathbb{C}_{\updownarrow}^{L\times J}$ of X, and $\mathcal{P}_{\mathcal{R}_{\mathcal{S}}}$ is the projection onto $\mathcal{R}_{\mathcal{S}}$ given by

$$\mathcal{P}_{\mathcal{R}_{\mathcal{S}}}(X) = \mathcal{S} \circ (\mathcal{S}^{\mathrm{H}} \circ \mathcal{S})^{-1} \circ \mathcal{S}^{\mathrm{H}}(X). \tag{4}$$

The computation of $\mathcal{P}_{\mathcal{M}}$ in (3) is very easy since we only have to replace the magnitude of X with M. In the right-hand side of (4), $(\mathcal{S}^{\mathrm{H}} \circ \mathcal{S})^{-1} \circ \mathcal{S}^{\mathrm{H}}$ is equivalent to the inverse STFT (ISTFT). Note that $\mathcal{S}^{\mathrm{H}} \circ \mathcal{S}$ for STFT becomes a diagonal matrix, and its inverse is computed simply by the reciprocal of each diagonal component. The corresponding matrix for CQT does not, however, become diagonal. Hence, GLA is executable for STFT, but not for CQT (see Sect. 2.3).

2.2. Griffin–Lim Like Phase Recovery via ADMM

GLA assumed that $\mathcal{M} \cap \mathcal{R}_{\mathcal{S}} \neq \varnothing$ on the problem of (1). In general, the estimated magnitude spectrogram M includes some errors, and thus $\mathcal{M} \cap \mathcal{R}_{\mathcal{S}} \neq \varnothing$ is not guaranteed. To obtain an accurate phase spectrogram even if $\mathcal{M} \cap \mathcal{R}_{\mathcal{S}} = \varnothing$, Masuyama *et al.* proposed an algorithm which finds the spectrogram closest to $\mathcal{R}_{\mathcal{S}}$ among spectrograms belonging to $\mathcal{M}$ by the following optimization problem [16]:

$$\underset{X\in\mathcal{M}}{\text{minimize}}\ \frac{1}{2}\|X - \mathcal{P}_{\mathcal{R}_{\mathcal{S}}}(X)\|_{\mathrm{F}}^2, \tag{5}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. To solve the problem of (5) faster than GLA, they adopted ADMM [17], which is briefly summarized as follows. For two proper lower-semicontinuous convex functions $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ and $g : \mathbb{R}^k \to \mathbb{R} \cup \{\infty\}$, let us

$$\underset{\boldsymbol{x}\in\mathbb{R}^m, \boldsymbol{z}\in\mathbb{R}^k}{\text{minimize}}\ f(\boldsymbol{x}) + g(\boldsymbol{z}) \quad \text{subject to } A\boldsymbol{x} + B\boldsymbol{z} = \boldsymbol{c}, \tag{6}$$

where $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{n\times k}$, and $\boldsymbol{c} \in \mathbb{R}^n$. For the problem of (6), an augmented Lagrangian function is constructed as

$$\mathcal{L}_\rho(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{u}) := f(\boldsymbol{x}) + g(\boldsymbol{z}) - \boldsymbol{u}^{\mathrm{T}}(A\boldsymbol{x} + B\boldsymbol{z} - \boldsymbol{c}) + \frac{\rho}{2}\|A\boldsymbol{x} + B\boldsymbol{z} - \boldsymbol{c}\|_2^2,$$

where $\boldsymbol{u} \in \mathbb{R}^n$ is the dual variable and $\rho > 0$. By minimizing $\mathcal{L}_\rho$ in terms of $\boldsymbol{x}$ and $\boldsymbol{z}$ alternately, and by increasing $\mathcal{L}_\rho$ in terms of $\boldsymbol{u}$, the following update formulas converging to the solution are given:

$$\begin{cases} \boldsymbol{x}^{(l+1)} = \underset{\boldsymbol{x}\in\mathbb{R}^m}{\text{argmin}}\ f(\boldsymbol{x}) + \frac{\rho}{2}\left\|A\boldsymbol{x} + B\boldsymbol{z}^{(l)} - \boldsymbol{c} - \frac{1}{\rho}\boldsymbol{u}^{(l)}\right\|_2^2, \\ \boldsymbol{z}^{(l+1)} = \underset{\boldsymbol{z}\in\mathbb{R}^k}{\text{argmin}}\ g(\boldsymbol{z}) + \frac{\rho}{2}\left\|A\boldsymbol{x}^{(l+1)} + B\boldsymbol{z} - \boldsymbol{c} - \frac{1}{\rho}\boldsymbol{u}^{(l)}\right\|_2^2, \\ \boldsymbol{u}^{(l+1)} = \boldsymbol{u}^{(l)} - \rho(A\boldsymbol{x}^{(l+1)} + B\boldsymbol{z}^{(l+1)} - \boldsymbol{c}). \end{cases} \tag{7}$$

When using ADMM, it is important to check whether or not argmin in the first and second lines of (7) can be computed easily.

Although the problem of (5) is nonconvex due to $\mathcal{M}$, Masuyama *et al.* applied (7) and proposed the following GL like phase recovery:

$$\begin{cases} X^{(l+1)} = \mathcal{P}_{\mathcal{M}}(Z^{(l)} + \frac{1}{\rho}U^{(l)}), \\ Z^{(l+1)} = \frac{1}{\rho+1}\left[\rho X^{(l+1)} - U^{(l)} + \mathcal{P}_{\mathcal{R}_{\mathcal{S}}}(X^{(l+1)} - \frac{1}{\rho}U^{(l)})\right], \\ U^{(l+1)} = U^{(l)} - \rho(X^{(l+1)} - Z^{(l+1)}), \end{cases} \tag{8}$$

by defining $\boldsymbol{x}$ in (6) as $X \in \mathbb{C}_{\updownarrow}^{L\times J}$, $\boldsymbol{z}$ as $Z \in \mathbb{C}_{\updownarrow}^{L\times J}$, f as the indicator function of $\mathcal{M}$, g as the cost function in (5), A as the identity matrix, B as the negative identity matrix, and $\boldsymbol{c}$ as the zero matrix. This algorithm achieved better convergence performance than GLA, but $\mathcal{P}_{\mathcal{R}_{\mathcal{S}}}$ appears again in the second line. Thus this is not executable for CQT.

2.3. Constant-Q Transform

The constant-Q transform (CQT) [18], [19] is useful for analysis of overtone components because it analyzes frequency components in the following geometric progression with the common ratio $2^{1/r}$:

$$f_i := 2^{\frac{i-1}{r}} f_{\min} > 0 \quad (i = 1, 2, \ldots, I_{\max}),$$

where $f_{\max} := f_{I_{\max}} < f_{\mathrm{s}}/2$, f_{s} is the sampling frequency, and r is often set to 12 or 24. The name 'CQT' is from the fact a quality factor

$$Q_i := \left[\!\!\left[q\frac{f_i}{\delta f_i} \right]\!\!\right] := \left[\!\!\left[q\frac{f_i}{f_{i+1} - f_i} \right]\!\!\right] = \left[\!\!\left[q\frac{1}{2^{1/r} - 1} \right]\!\!\right] = Q$$

is *constant* independently of i, where $q \in (0, 1]$ and $[\![\cdot]\!]$ is the rounding function. For example, if $r = 24$ and $q = 0.82$, then $Q = 28$.

Components of a CQT spectrogram $X := (\hat{x}_{i,j})$ are defined by

$$\hat{x}_{i,j} := \frac{1}{\sqrt{L_i}} \sum_{\tau=0}^{L_i-1} w_i[\tau] x[(j-1)\xi_i + \tau] e^{-\imath \frac{2\pi Q \tau}{L_i}},$$

for $i = 1, 2, \ldots, I_{\max}$, where $L_i := [\![Qf_{\mathrm{s}}/f_i]\!]$ is the frame length and ξ_i ($< L_i$) is the shift of a window function $w_i[\tau]$ for the ith frequency. We also define $\hat{x}_{2I_{\max}-i+1,j} := \overline{\hat{x}}_{i,j}$ for $i = 1, 2, \ldots, I_{\max}$ as components corresponding to $-f_i$. Note that a CQT spectrogram X is a matrix only if $\forall i\ \xi_i = \xi < L_{I_{\max}}$ holds, and in this case the spectrogram space is defined as $\mathbb{C}_{\updownarrow}^{2I_{\max}\times J} := \{(\hat{x}_{i,j}) \in \mathbb{C}^{2I_{\max}\times J} \mid \overline{\hat{x}}_{i,j} = \hat{x}_{2I_{\max}-i+1,j}\ (\forall j \text{ and } i = 1, 2, \ldots, I_{\max})\}$, where J is the number of frames. If we denote CQT by a linear mapping $\mathcal{S}$, $\mathcal{S}^{\mathrm{H}} \circ \mathcal{S}$ is not a diagonal matrix and ICQT $(\mathcal{S}^{\mathrm{H}} \circ \mathcal{S})^{-1} \circ \mathcal{S}^{\mathrm{H}}$ is not available.

3. NON-GRIFFIN–LIM TYPE SIGNAL RECOVERY WITHOUT THE INVERSE SPECTROGRAM TRANSFORM

To sum up the discussion in Sect. 2, the GL type *phase recovery* algorithms [14]–[16] require the computation of $(\mathcal{S}^{\mathrm{H}} \circ \mathcal{S})^{-1}$, which is easy for STFT because $\mathcal{S}^{\mathrm{H}} \circ \mathcal{S}$ is a simple diagonal matrix. However, when $\mathcal{S}$ stands for CQT, it is difficult to compute $(\mathcal{S}^{\mathrm{H}} \circ \mathcal{S})^{-1}$, and to apply the update formulas of (2) and (8). In addition, even if the solution $\widehat{X}$ to the optimization problem such as (1) or (5) can be obtained, the inverse transform $(\mathcal{S}^{\mathrm{H}} \circ \mathcal{S})^{-1} \circ \mathcal{S}^{\mathrm{H}}$ has to be computed as the post-processing to recover a time-domain signal $\boldsymbol{x}$ from $\widehat{X}$.

To eliminate the inverse transform $(\mathcal{S}^{\mathrm{H}}\circ\mathcal{S})^{-1}\circ\mathcal{S}^{\mathrm{H}}$ as the post-processing, we reformulate the optimization problem of (5) as an essentially equivalent problem using the time-domain signal itself:

$$\underset{\boldsymbol{x}\in\mathbb{R}^{(J-1)\xi+L}}{\text{minimize}}\ \frac{1}{2}\big\|M-|\mathcal{S}(\boldsymbol{x})|\big\|_{\mathrm{F}}^2. \tag{9}$$

We propose a non-GL type *signal recovery* algorithm which reconstructs $\boldsymbol{x}$ by solving (9) without computing $(\mathcal{S}^{\mathrm{H}}\circ\mathcal{S})^{-1}$. To this end, we use a stochastic convex optimization technique SDCA-ADMM.

3.1. Stochastic Dual Coordinate Ascent ADMM

SDCA-ADMM exploits the characteristic that the dual variable $\boldsymbol{u}$ in (7) converges to the optimal solution of the *dual problem* [21]. The convex optimization problem to be considered is to

$$\underset{\boldsymbol{x}\in\mathbb{R}^m}{\text{minimize}}\ \sum_{i=1}^{n} f_i(\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x})+g(B^{\mathrm{T}}\boldsymbol{x}), \tag{10}$$

where $A:=(\boldsymbol{a}_1,\boldsymbol{a}_2,\dots,\boldsymbol{a}_n)\in\mathbb{R}^{m\times n}$, $B\in\mathbb{R}^{m\times h}$, and $f:\mathbb{R}^n\ni\boldsymbol{u}\mapsto\sum_{i=1}^n f_i(u_i)\in\mathbb{R}\cup\{\infty\}$ and $g:\mathbb{R}^h\to\mathbb{R}\cup\{\infty\}$ are supposed to be a data fidelity and a regularization term, respectively. The dual problem for the primal problem of (10) is to

$$\underset{\boldsymbol{u}\in\mathbb{R}^n,\boldsymbol{v}\in\mathbb{R}^h}{\text{minimize}}\ \sum_{i=1}^{n} f_i^*(u_i)+g^*(\boldsymbol{v})\quad\text{subject to } A\boldsymbol{u}+B\boldsymbol{v}=\boldsymbol{0}, \tag{11}$$

where f_i^* and g^* are the *conjugate function* of f_i and g, respectively. By solving (11) with the ADMM update formulas as in (7), the *dual* variable $\boldsymbol{x}$ converges to the solution to the problem of (10), because the relationship between $\boldsymbol{x}$ and $\boldsymbol{u}$ in (7) is reversed.

In many cases, however, it is difficult to directly adopt the above idea because argmin in the first and second lines of (7) cannot be computed easily. To overcome the difficulty, SDCA-ADMM uses an *inexact* augmented Lagrangian function, which is defined by adding two proximal terms as

$$\begin{aligned}\widehat{\mathcal{L}}_\rho^{(l)}(\boldsymbol{u},\boldsymbol{v},\boldsymbol{x}):=f^*(\boldsymbol{u})+g^*(\boldsymbol{v})-\boldsymbol{x}^{\mathrm{T}}(A\boldsymbol{u}+B\boldsymbol{v})\\ +\frac{\rho}{2}\|A\boldsymbol{u}+B\boldsymbol{v}\|_2^2+\frac{1}{2}\|\boldsymbol{u}-\boldsymbol{u}^{(l)}\|_{G_A}^2+\frac{1}{2}\|\boldsymbol{v}-\boldsymbol{v}^{(l)}\|_{G_B}^2,\end{aligned}$$

where $G_A:=\rho(\eta_A I_n-A^{\mathrm{T}}A)$ and $G_B:=\rho(\eta_B I_h-B^{\mathrm{T}}B)$ are positive semidefinite matrices to avoid the computation of $(A^{\mathrm{T}}A)^{-1}$ and $(B^{\mathrm{T}}B)^{-1}$. SDCA-ADMM minimizes $\widehat{\mathcal{L}}_\rho^{(l)}$ in terms of $\boldsymbol{u}$ and $\boldsymbol{v}$ alternately and increases it in terms of $\boldsymbol{x}$. Note that the first term in (11) is separated in terms of u_i so that the stochastic optimization can be easily introduced as follows. SDCA-ADMM splits the index set $\mathcal{I}:=\{1,2,\dots,n\}$ for f_i^* into b subsets $\{\mathcal{I}_d\}_{d=1}^b$ such that $\mathcal{I}_d\neq\varnothing$, $\bigcup_{d=1}^b\mathcal{I}_d=\mathcal{I}$ and $\mathcal{I}_d\cap\mathcal{I}_{d'}=\varnothing$ $(d\neq d')$. Then, SDCA-ADMM updates the variables by

$$\left\lfloor\begin{aligned}&\text{Choose one index } d\in\{1,2,\dots,b\} \text{ uniformly at random,}\\ &\boldsymbol{q}^{(l+1)}=\boldsymbol{v}^{(l)}+\tfrac{1}{\rho\eta_B}B^{\mathrm{T}}\big[\boldsymbol{x}^{(l)}-\rho(A\boldsymbol{u}^{(l)}+B\boldsymbol{v}^{(l)})\big],\\ &\boldsymbol{v}^{(l+1)}=\boldsymbol{q}^{(l+1)}-\tfrac{1}{\rho\eta_B}\operatorname{prox}_{\rho\eta_B g}(\rho\eta_B\boldsymbol{q}^{(l+1)}),\\ &\boldsymbol{p}_{\mathcal{I}_d}^{(l+1)}=\boldsymbol{u}_{\mathcal{I}_d}^{(l)}+\tfrac{1}{\rho\eta_{A_{\mathcal{I}_d}}}A_{\mathcal{I}_d}^{\mathrm{T}}\big[\boldsymbol{x}^{(l)}-\rho(A\boldsymbol{u}^{(l)}+B\boldsymbol{v}^{(l+1)})\big],\\ &\boldsymbol{u}_{\mathcal{I}_d}^{(l+1)}=\boldsymbol{p}_{\mathcal{I}_d}^{(l+1)}-\tfrac{1}{\rho\eta_{A_{\mathcal{I}_d}}}\operatorname{prox}_{\rho\eta_{A_{\mathcal{I}_d}}f_{\mathcal{I}_d}}(\rho\eta_{A_{\mathcal{I}_d}}\boldsymbol{p}_{\mathcal{I}_d}^{(l+1)}),\\ &\boldsymbol{x}^{(l+1)}=\boldsymbol{x}^{(l)}-\gamma\rho\big[A\boldsymbol{u}^{(l+1)}+B\boldsymbol{v}^{(l+1)}\\ &\qquad\qquad-(1-\phi)(A\boldsymbol{u}^{(l)}+B\boldsymbol{v}^{(l)})\big],\end{aligned}\right. \tag{12}$$

where $\gamma>0$, $\eta_{A_{\mathcal{I}_d}}\geq\|A_{\mathcal{I}_d}\|_{\mathrm{op}}^2$, $\eta_B\geq\|B\|_{\mathrm{op}}^2$, $\|\cdot\|_{\mathrm{op}}$ is the operator norm, $\phi=1/b$ is the probability that each u_i is chosen, and prox denotes the *proximity operator*. When the proximity operators of f_i and g are computed easily, the update formulas of (12) can be executed.

3.2. Signal Recovery via SDCA-ADMM

We solve the nonconvex optimization problem of (9) via the SDCA-ADMM update formulas. Define $f(\boldsymbol{u})=\sum_i f_i(u_i)$ in (10) as

$$\sum_{i,j} f_{i,j}(\hat{u}_{i,j}):=\sum_{i,j}\frac{1}{2}(m_{i,j}-|\hat{u}_{i,j}|)^2=\frac{1}{2}\big\|M-|U|\big\|_{\mathrm{F}}^2,$$

g as the zero function, A as $\mathcal{S}^{\mathrm{H}}$, and B as the zero matrix. Then, the dual problem of (9) is given by

$$\underset{U\in\mathbb{C}_{\updownarrow}^{L\times J}}{\text{minimize}}\ \sum_{i,j} f_{i,j}^*(\hat{u}_{i,j})\quad\text{subject to } \mathcal{S}^{\mathrm{H}}(U)=\boldsymbol{0}. \tag{13}$$

To solve the problem of (13) using SDCA-ADMM, we have to carefully choose the index set $\mathcal{I}_d\subset\mathcal{I}:=\{(i,j)\}_{j=1,2,\dots,J}^{i=1,2,\dots,L}$. In the conventional method, only one index set out of the b index sets was randomly chosen as in (12). This strategy not only prevents acceleration but also may get stuck in a local minimum due to the nonconvexity.

To overcome this difficulty, it should be noted that the convergence of (12) is guaranteed if the parameter ϕ agrees with the probability that each component u_i is chosen from $\boldsymbol{u}$. Therefore, instead of updating one of the b non-overlapping index sets, we propose to update *multiple index sets* with a probability $\phi\in(0,1]$, where $\phi=1$ means the non-stochastic optimization. Based on these discussions, we propose the following update formulas for solving (13):

$$\left\lfloor\begin{aligned}&\text{Choose multiple indices } \mathcal{D}\in 2^{\{1,2,\dots,b\}} \text{ with a probability } \phi,\\ &P_{\mathcal{I}_\mathcal{D}}^{(l+1)}=U_{\mathcal{I}_\mathcal{D}}^{(l)}+\tfrac{1}{\rho\eta}\mathcal{S}_{\mathcal{I}_\mathcal{D}}\big(\boldsymbol{x}^{(l)}-\rho\mathcal{S}^{\mathrm{H}}(U^{(l)})\big),\\ &U_{\mathcal{I}_\mathcal{D}}^{(l+1)}=\tfrac{1}{\rho\eta+1}\big(\rho\eta\,P_{\mathcal{I}_\mathcal{D}}^{(l+1)}-M_{\mathcal{I}_\mathcal{D}}\odot\Theta_{P_{\mathcal{I}_\mathcal{D}}^{(l+1)}}\big),\\ &\boldsymbol{x}^{(l+1)}=\boldsymbol{x}^{(l)}-\gamma\rho\mathcal{S}^{\mathrm{H}}\big(U^{(l+1)}-(1-\phi)U^{(l)}\big),\end{aligned}\right. \tag{14}$$

where $\rho>0$, $\gamma>0$, and $\eta\geq\|\mathcal{S}\|_{\mathrm{op}}^2$. Note that there is no computation of the inverse spectrogram transform $(\mathcal{S}^{\mathrm{H}}\circ\mathcal{S})^{-1}\circ\mathcal{S}^{\mathrm{H}}$ in the formulas of (14). Hence, the proposed algorithm is *non-GL type* and can be applied not only to STFT but also to CQT. As shown in Sect. 4.2, we propose to divide the index set along the frequency axis, as $\{(1,j)\}_{j=1}^J$ and $\{(i,j),(L-i+2,j)\}_{j=1}^J$ $(i=2,3\dots,\lfloor\frac{L+2}{2}\rfloor)$.

4. NUMERICAL EXPERIMENTS

To confirm the convergence performance of the proposed algorithm and the validity of the stochastic optimization, we conducted numerical experiments of time-domain signal recovery from a STFT magnitude spectrogram in Sects. 4.1 & 4.2 and from a CQT magnitude spectrogram in Sect. 4.3. Each recovered time-domain signal $\boldsymbol{x}$ was evaluated by the normalized error $100\,\frac{\|M-|\mathcal{S}(\boldsymbol{x})|\|_{\mathrm{F}}}{\|M\|_{\mathrm{F}}}$ [%] of the magnitude spectrogram because it corresponds to the cost function in (9) and phase errors have less effect on human ears. An open source piano sound of 7,000 ms and the Hanning window were used with sampling frequency 16 kHz, frame length 64 ms, and frame shift 32 ms. In all methods (2), (8) and (14), the number of iterations was 10,000.

4.1. Performance Evaluation of the Proposed Method in STFT

We compared the performance of the proposed method and the existing methods [14], [16] under the ideal situation of $\mathcal{M}\cap\mathcal{R}_{\mathcal{S}}\neq\varnothing$.

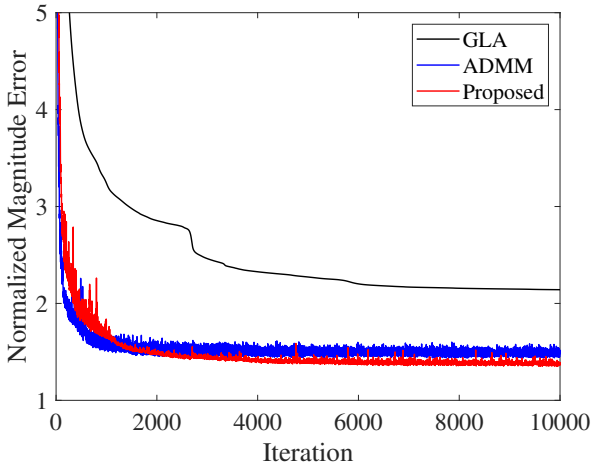

Fig. 1. Normalized error curves for 3 methods (STFT case).

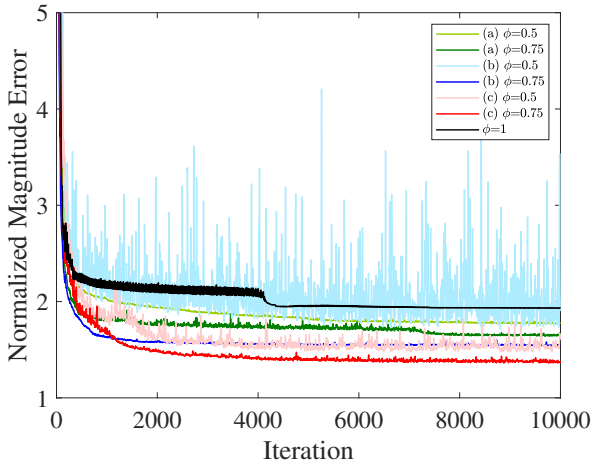

Fig. 2. Normalized error curves for 3 index set divisions (STFT case).

The piano sound was recovered by each method and the normalized error was computed. The parameter in the ADMM method [16] was $\rho = 0.08$ and those in the proposed method were $\gamma = 1$, $\rho = 0.01$, $\eta = 1.001L$, and $\phi = 0.75$. The index set $\mathcal{I} = \{(i,j)\}_{j=1,2,\ldots,J}^{i=1,2,\ldots,L}$ was divided along the frequency axis as mentioned in Sect. 3.2. The parameter values were empirically adjusted for the best results. Figure 1 shows the normalized error curve in terms of the iteration for each method. We can see that the ADMM method and the proposed method indicate faster convergence speed than GLA. We emphasize that the proposed method achieves mostly the same convergence performance as the ADMM method *without ISTFT* $(\mathcal{S}^{\mathrm{H}} \circ \mathcal{S})^{-1} \circ \mathcal{S}^{\mathrm{H}}$.

4.2. Validity Evaluation of the Stochastic Optimization in STFT

We prepared three types of divisions of the index set $\mathcal{I}$: (a) *primal division* $\{(1,j)\}$ $(j = 1,2,\ldots,J)$ and $\{(i,j),(L-i+2,j)\}$ $(i = 2,3\ldots,\lfloor\frac{L+2}{2}\rfloor; j = 1,2,\ldots,J)$ $(b = J\lfloor\frac{L+2}{2}\rfloor)$, (b) *frame division* $\{(i,j)\}_{i=1}^{L}$ $(j = 1,2,\ldots,J)$ $(b = J)$, and (c) *frequency division* $\{(1,j)\}_{j=1}^{J}$ and $\{(i,j),(L-i+2,j)\}_{j=1}^{J}$ $(i = 2,3\ldots,\lfloor\frac{L+2}{2}\rfloor)$ $(b = \lfloor\frac{L+2}{2}\rfloor)$. We also prepared the selection probability ϕ for each division as $\phi = 0.5$ and $\phi = 0.75$. In addition, $\phi = 1$ was also used for comparison to the non-stochastic optimization scenario. Figure 2 shows the normalized error curves in terms of the iteration.[1] All the curves except for that of the frame division with $\phi = 0.75$ are lower than that for $\phi = 1$. Therefore, we can confirm the effectiveness of the stochastic optimization. The best performance was given by the frequency division with $\phi = 0.75$.

4.3. Validity Evaluation of the Stochastic Optimization in CQT

We also conducted numerical experiments of time-domain signal recovery from a CQT magnitude spectrogram by the proposed method. We set the parameters of CQT to $r = 24$, $q = 0.82$, $f_{\min} = 60$ Hz, $f_{\max} = 5{,}922$ Hz (i.e., $I_{\max} = 160$), and $\xi_i = \xi = L_{I_{\max}}/2$. Since we cannot use the frame division if $\xi_i \neq \xi_{i'}$ $(i \neq i')$, we used only the primal division and the frequency division. Figure 3 shows the normalized magnitude error curves in terms of the iteration for the proposed method. The lowest normalized error converged to 0.63 %, which means that the proposed method enables time-domain signal recovery from the CQT magnitude spectrogram. In each type of division of the index set $\mathcal{I}$, there was a trade-off between the convergence speed and the normalized error according to the probability ϕ. When $\phi = 0.5$, the normalized error was lower than when $\phi = 0.75$. On the other hand, when $\phi = 0.75$, the convergence was faster than when $\phi = 0.5$. Since the proposed method makes the stochastic optimization more flexible, it would be best to use relatively large ϕ in the early part and then to use relatively small ϕ in the latter part.

[1] Strictly speaking, the ten point moving average filter is used to make the curves more visible.

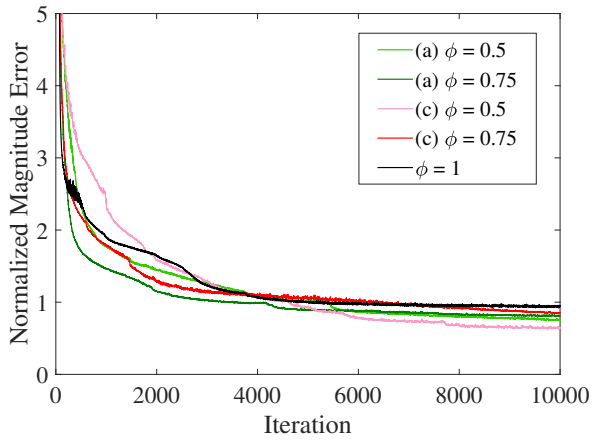

Fig. 3. Normalized error curves for 2 index set divisions (CQT case).

5. CONCLUSION

In this paper, we proposed a time-domain signal recovery algorithm which does not require the computation of the inverse of the spectrogram transform. We first reformulated the signal recovery problem as a nonconvex optimization problem using the time-domain signal itself. To solve this problem, we used a stochastic optimization technique SDCA-ADMM for fast convergence without the inverse transform. Numerical experiments showed that the proposed method with a well-organized division of a cost function along the frequency axis achieves the same excellent convergence performance as the GL type algorithms for a STFT magnitude spectrogram. We also showed that the proposed method is effective for a CQT magnitude spectrogram.

6. REFERENCES

[1] S. Makino, S. Araki, R. Mukai, and H. Sawada, "Audio source separation based on independent component analysis," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)*, Vancouver, Canada, 2004, vol. 5, pp. 668–671.

[2] T. Kim, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[3] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[4] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[5] M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon, "Computer-aided melody note transcription using the Tony software: accuracy and efficiency," in *Proceedings of International Conference on Technologies for Music Notation and Representation (TENOR)*, Paris, France, 2015, pp. 28–30.

[6] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1830–1839, 2008.

[7] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[8] M. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 243–248, 1976.

[9] V. Verfaille, U. Zolzer, and D. Arfib, "Adaptive digital audio effects (A-DAFx): A new class of sound transformations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1817–1831, 2006.

[10] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proceedings of International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 745–751,

[11] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2012.

[12] K. Jaganathan, Y. C. Eldar, and B. Hassibi, "STFT phase retrieval: Uniqueness guarantees and recovery algorithms," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 770–781, 2016.

[13] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval from coded diffraction patterns," *Applied and Computational Harmonic Analysis*, vol. 39, no. 2, pp. 277–299, 2015.

[14] D. W. 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, 1984.

[15] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast Griffin–Lim algorithm," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2013, 4 pages.

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

[17] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.

[18] J. C. Brown, "Calculation of a constant-Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[19] A. Nagathil and R. Martin, "Evaluation of spectral transforms for music signal analysis," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2013, 4 pages.

[20] I. Daubechies, *Ten Lectures on Wavelets*, ser. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia, PA, USA: SIAM, 1992, vol. 61.

[21] T. Suzuki, "Stochastic dual coordinate ascent with alternating direction multiplier method," in *Proceedings of International Conference on Machine Learning (ICML)*, Beijing, China, 2014, 9 pages.