SIGNAL RECONSTRUCTION FROM MEL-SPECTROGRAM BASED ON BI-LEVEL CONSISTENCY OF FULL-BAND MAGNITUDE AND PHASE

Yoshiki Masuyama,[1] *Natsuki Ueno,*[1] *Nobutaka Ono,*[1]

[1]Tokyo Metropolitan University, Japan

ABSTRACT

We propose an optimization-based method for reconstructing a time-domain signal from a low-dimensional spectral representation such as a mel-spectrogram. Phase reconstruction has been studied to reconstruct a time-domain signal from the full-band short-time Fourier transform (STFT) magnitude. The Griffin–Lim algorithm (GLA) has been widely used because it relies only on the redundancy of STFT and is applicable to various audio signals. In this paper, we jointly reconstruct the full-band magnitude and phase by considering the bi-level relationships among the time-domain signal, its STFT coefficients, and its mel-spectrogram. The proposed method is formulated as a rigorous optimization problem and estimates the full-band magnitude based on the criterion used in GLA. Our experiments demonstrate the effectiveness of the proposed method on speech, music, and environmental signals.

Index Terms— Phase reconstruction, waveform synthesis, mel-spectrogram, bi-level consistency, proximal splitting methods.

1. INTRODUCTION

Phase reconstruction of short-time Fourier transform (STFT) coefficients has been studied for decades [1–6]. When the STFT magnitude is synthesized or processed in the time–frequency domain, the corresponding phase is required to convert it to the time domain by using the inverse STFT. While various phase reconstruction methods have been developed [5, 6], the Griffin–Lim algorithm (GLA) [1] has been widely used because it works without any assumptions on the target signal. GLA leverages the redundancy of STFT and reconstructs the phase based on the consistency between complex STFT coefficients and time-domain signals. Recently, variants of GLA have been developed from the point of view of optimization algorithms [2–4] and through integration with deep neural networks [7, 8]. These methods have shown promising performance when the given magnitude does not contain errors.

Recently, text-to-speech [9–11] and voice conversion [12, 13] pipelines predict a low-dimensional spectral representation, such as a mel-spectrogram, and reconstruct a time-domain signal from the phaseless representation. The auditory-motivated representation efficiently preserves the essential information and is easier to predict than the full-band magnitude. The second stage of the pipelines requires mel-spectrogram inversion that reconstructs a time-domain signal from the given mel-spectrogram. While neural vocoders directly reconstruct a time-domain signal [14–16], recent studies reconstruct complex STFT coefficients and incorporate the inverse STFT [17, 18] to reduce computational complexity. These studies are relevant to phase reconstruction, and phase reconstruction is still used for synthesizing more general audio signals [19, 20].

When we use phase reconstruction in mel-spectrogram inversion, we first reconstruct the full-band magnitude and then estimate
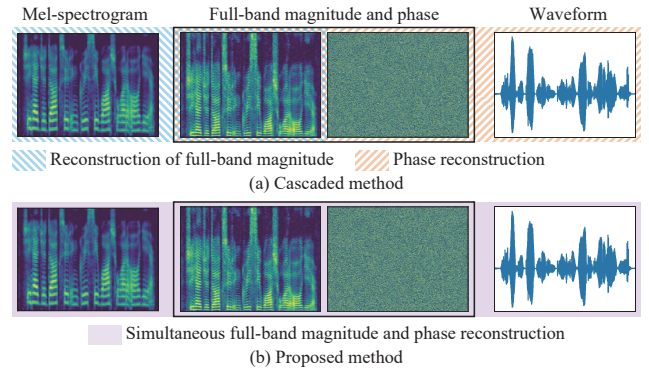

Figure 1: Illustrations of the (a) cascaded method and (b) simultaneous optimization method for reconstructing the audio signal from a given mel-spectrogram.

the phase from the reconstructed full-band magnitude as illustrated in Fig. 1 (a). We call such a two-stage method the cascaded method. While the non-negative least squares method has been widely used to reconstruct the full-band magnitude, it reconstructs the magnitude frame-by-frame and cannot obtain the original full-band magnitude in common settings. The reconstruction error deteriorates the performance of the cascaded method because the phase is estimated to be consistent with the reconstructed full-band magnitude.

To address this problem, we explore optimization-based methods that jointly reconstruct the full-band magnitude and phase as depicted in Fig. 1 (b). We integrate the optimization problems for full-band magnitude reconstruction and phase reconstruction into a single problem. We further modify its cost function, which makes the method less sensitive to a hyperparameter. To solve the problems, we adopt the inertial proximal alternating linearized minimization (iPALM) [21] that can handle non-convex and non-smooth functions. In our experiments with speech signals, the proposed methods outperformed the cascaded method in terms of PESQ [22] and the extended STOI (ESTOI) [23]. We also demonstrated the advantage of the proposed method on music signals and foley sounds[1].

2. PRELIMINARIES

2.1. Phase Reconstruction with Original-scale Magnitude

Let the STFT coefficients of an audio signal $\boldsymbol{x}$ be $\boldsymbol{X} = \mathcal{G}(\boldsymbol{x}) \in \mathbb{C}^{F \times T}$, where F and T are the number of frequency bins and time frames, respectively. As STFT is a redundant transform in common settings, the image of STFT $\mathcal{C}$ is a linear subspace of $\mathbb{C}^{F \times T}$. If complex STFT coefficients $\boldsymbol{X}$ are not in $\mathcal{C}$, the magnitude of the

---

[1]yoshikimas.github.io/signal-reconstruction-from-mel-spectrogram

reconstructed signal differs from the original one as follows:

$$|\mathcal{G}(\mathcal{G}^{\dagger}(\boldsymbol{X}))| \neq |\boldsymbol{X}|, \tag{1}$$

where $(\cdot)^{\dagger}$ denotes the Moore-Penrose pseudo-inverse, and $|\cdot|$ denotes entry-wise absolute value. Note that $\mathcal{G}^{\dagger}(\cdot)$ is the popular inverse STFT. The discrepancy in (1) is adverse in many audio applications since the magnitude of $\boldsymbol{X}$ is processed to the desired one.

To tackle this problem, phase reconstruction, or spectrogram inversion, is formulated as the following problem [3]:

$$\text{Find } \boldsymbol{X} \text{ s.t. } \boldsymbol{X} \in \mathcal{C} \cap \mathcal{A}, \tag{2}$$

where $\mathcal{A}$ is a set of STFT coefficients whose magnitude is equal to the given one $\boldsymbol{A} \in \mathbb{R}_+^{F \times T}$:

$$\mathcal{A} = \{\boldsymbol{X} \in \mathbb{C}^{F \times T} \mid |\boldsymbol{X}| = \boldsymbol{A}\} . \tag{3}$$

When $\boldsymbol{A}$ is artificially generated, $\mathcal{C} \cap \mathcal{A}$ can be empty. In such cases, there is no solution for (2), and we can consider the following optimization problem instead:

$$\min_{\boldsymbol{X} \in \mathcal{C}} \mathcal{L}(\boldsymbol{X}, \boldsymbol{A}) = \||\boldsymbol{X}| - \boldsymbol{A}\|^2, \tag{4}$$

where $\|\cdot\|$ denotes the Frobenius norm. The magnitude of the reconstructed signal $\mathcal{G}^{\dagger}(\boldsymbol{X})$ keeps the given one $\boldsymbol{A}$ as much as possible in terms of the squared Frobenius norm.

2.2. Griffin–Lim Algorithm (GLA)

GLA reconstructs the phase based on the redundancy of STFT and is implemented as alternating projections onto the sets in (2) [1]:

$$\boldsymbol{X}^{[k+1]} = P_{\mathcal{C}}\left(P_{\mathcal{A}}\left(\boldsymbol{X}^{[k]}\right)\right), \tag{5}$$

where k is the iteration index, $P_{\mathcal{S}}(\cdot)$ is the projection onto a set $\mathcal{S}$:

$$P_{\mathcal{S}}(\boldsymbol{Y}) = \underset{\boldsymbol{X}}{\arg\min} \iota_{\mathcal{S}}(\boldsymbol{X}) + \|\boldsymbol{X} - \boldsymbol{Y}\|^2, \tag{6}$$

and the indicator function $\iota_{\mathcal{S}}(\cdot)$ is defined as

$$\iota_{\mathcal{S}}(\boldsymbol{X}) = \begin{cases} 0 & (\boldsymbol{X} \in \mathcal{S}) \\ \infty & (\boldsymbol{X} \notin \mathcal{S}) \end{cases}. \tag{7}$$

The projection onto the set $\mathcal{C}$ is given by

$$P_{\mathcal{C}}(\boldsymbol{X}) = \mathcal{G}(\mathcal{G}^{\dagger}(\boldsymbol{X})). \tag{8}$$

Meanwhile, the projection onto the set $\mathcal{A}$ is defined as

$$P_{\mathcal{A}}(\boldsymbol{X}) = \boldsymbol{A} \odot \boldsymbol{X} \oslash |\boldsymbol{X}|, \tag{9}$$

where $\odot$ and $\oslash$ are entry-wise multiplication and division, respectively, and we set $0/|0| = 0$ as in [3, 4].

We can derive the iterative procedure of GLA in (5) by applying the projected gradient method to (4) [24, 25][2]:

$$\boldsymbol{X}^{[k+1]} = P_{\mathcal{C}}(\boldsymbol{X}^{[k]} - \mu \nabla_{\boldsymbol{X}} \mathcal{L}(\boldsymbol{X}^{[k]}, \boldsymbol{A})), \tag{10}$$

where $\mu \in \mathbb{R}_+$ is the step size, and the gradient for $\mathcal{L}(\boldsymbol{X}, \boldsymbol{A})$ is given by

$$\nabla_{\boldsymbol{X}} \mathcal{L}(\boldsymbol{X}, \boldsymbol{A}) = \boldsymbol{X} - P_{\mathcal{A}}(\boldsymbol{X}). \tag{11}$$

By substituting (11) to (10) and setting μ to 1, the projected gradient method coincides with the alternating projections in (5). This interpretation of GLA is more relevant to the proposed method.

---

[2]The cost function in (4) has non-smooth points, where we set the gradient to zero and obtain $\nabla_{\boldsymbol{X}} \mathcal{L}(\boldsymbol{X}, \boldsymbol{A})$ in (11). With abuse of terminology, we still refer to it as the gradient for simplicity as in [24].

2.3. Mel-Spectrogram Inversion

In recent text-to-speech pipelines [9–11], a mel-spectrogram has been used as an intermediate representation. Mel-spectrogram inversion aims to reconstruct a time-domain signal from a mel-spectrogram $\boldsymbol{M} \in \mathbb{R}_+^{B \times T}$ given by $\boldsymbol{E}\boldsymbol{A}$, where B is the number of mel bins, and $\boldsymbol{E} \in \mathbb{R}^{B \times F}$ converts the frequency scale.

Since phase reconstruction has been successfully addressed, a simple method for mel-spectrogram inversion is to reconstruct the full-band magnitude in advance and then apply a phase reconstruction method. The reconstruction of the full-band magnitude can be formulated as follows:

$$\min_{\boldsymbol{Y}} \frac{1}{2}\|\boldsymbol{E}\boldsymbol{Y} - \boldsymbol{M}\|^2, \tag{12a}$$

$$\text{s.t. } \boldsymbol{Y} \in \mathcal{N}, \tag{12b}$$

where $\mathcal{N} = \mathbb{R}_+^{F \times T}$. The cascaded method for mel-spectrogram inversion with (12) has been used in popular audio analysis packages including `Librosa` [26]. The optimization problem in (12) did not consider the relationship between adjacent time frames in STFT, and its solution is not unique in common settings. As a result, the reconstructed full-band magnitude $\boldsymbol{Y}$ cannot be consistent with a time-domain signal, which deteriorates the performance of the subsequent phase reconstruction.

3. PROPOSED MEL-SPECTROGRAM INVERSION

In this section, we present two mel-spectrogram inversion methods that estimate full-band magnitude and phase jointly.

3.1. Joint Full-band Magnitude and Phase Reconstruction

As discussed in Section 2.3, the cascaded mel-spectrogram inversion method reconstructs the full-band magnitude frame-by-frame and results in sub-optimal performance due to the inconsistency in the reconstructed full-band magnitude. Meanwhile, we can reduce the inconsistency by leveraging the redundancy of STFT such as the signal overlaps in adjacent time frames. We thus simultaneously consider the bi-level relationships: the relationship between a mel-spectrogram and a full-band magnitude; and the relationship between full-band STFT coefficients and a time-domain signal as shown in Fig. 1 (b). The former and latter relationships are induced by $\boldsymbol{M} = \boldsymbol{E}\boldsymbol{A}$ and STFT, respectively.

In detail, to estimate the full-band magnitude and phase jointly, we integrate the optimization problems in (4) and (12) as follows:

$$\min_{\boldsymbol{X}, \boldsymbol{Y}} \frac{1}{2}\mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}) + \frac{\lambda}{2}\|\boldsymbol{E}\boldsymbol{Y} - \boldsymbol{M}\|^2, \tag{13a}$$

$$\text{s.t. } \boldsymbol{X} \in \mathcal{C}, \ \boldsymbol{Y} \in \mathcal{N}, \tag{13b}$$

where $\lambda \in \mathbb{R}_+$ is a hyperparamter. The reconstruction of the full-band magnitude considers its bi-level relationships with the mel-spectrogram and with the time-domain signal.

As another formulation, we can replace the first term in (13a) to a constraint: $\boldsymbol{Y} = |\boldsymbol{X}|$. The obtained optimization problem can be reformulated as the following unconstrained optimization problem:

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{E}|\mathcal{G}(\boldsymbol{x})| - \boldsymbol{M}\|^2. \tag{14}$$

While we can apply the gradient descent method to (14) as in GLA, we will demonstrate that applying the iPALM algorithm for (13) improves the quality of the reconstructed signal faster in Section 4.1.

3.2. Squared Distance to Set as Mel-spectrogram Fidelity

In (13a), the two cost functions are defined on different domains: the full-band magnitude and the mel-spectrogram, which makes it difficult to tune λ intuitively. We thus replace the second term in (13a) by the squared distance to the following set $\mathcal{M}$:

$$\frac{1}{2}d_{\mathcal{M}}^2(\boldsymbol{Y}) = \inf_{\boldsymbol{Z} \in \mathcal{M}} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Z}\|^2, \tag{15}$$

$$\mathcal{M} = \{\boldsymbol{Z} \in \mathbb{R}^{F \times T} \mid \boldsymbol{E}\boldsymbol{Z} = \boldsymbol{M}\}, \tag{16}$$

This cost function is computed on the full-band magnitude as the first term of (13a). The gradient for (15) is given by [27]

$$\nabla \frac{1}{2}d_{\mathcal{M}}^2(\boldsymbol{Y}) = \boldsymbol{Y} - P_{\mathcal{M}}(\boldsymbol{Y}), \tag{17}$$

$$P_{\mathcal{M}}(\boldsymbol{Y}) = \boldsymbol{Y} - \boldsymbol{E}^\dagger(\boldsymbol{E}\boldsymbol{Y} - \boldsymbol{M}). \tag{18}$$

By using the cost function in (15), we formulate another optimization problem for mel-specrogram inversion:

$$\min_{\boldsymbol{X},\boldsymbol{Y}} \frac{1}{2}\mathcal{L}(\boldsymbol{X},\boldsymbol{Y}) + \frac{\lambda}{2}d_{\mathcal{M}}^2(\boldsymbol{Y}), \tag{19a}$$

$$\text{s.t. } \boldsymbol{X} \in \mathcal{C}, \ \boldsymbol{Y} \in \mathcal{N}. \tag{19b}$$

We will demonstrate that (19) is relatively insensitive to the value of λ than (13) in Section 4.1.

3.3. iPALM Algorithm for (19)

To solve the optimization problem in (19), in this paper, we adopt the iPALM algorithm [21]. Let us consider an optimization problem with three cost functions:

$$\min_{\boldsymbol{X},\boldsymbol{Y}} \mathcal{F}_1(\boldsymbol{X}) + \mathcal{F}_2(\boldsymbol{Y}) + \mathcal{H}(\boldsymbol{X},\boldsymbol{Y}), \tag{20}$$

where the cost functions can be non-convex. We assume the following proximity operator for $\mathcal{F}_1(\cdot)$ and $\mathcal{F}_2(\cdot)$ is calculated efficiently:

$$\operatorname{prox}_{\mathcal{F}}(\boldsymbol{X}) = \operatorname*{argmin}_{\boldsymbol{Z}} \mathcal{F}(\boldsymbol{Z}) + \frac{1}{2}\|\boldsymbol{Z} - \boldsymbol{X}\|^2. \tag{21}$$

The iPALM algorithm solves the optimization problem in (20) by iterating the following procedure:

$$\underline{\boldsymbol{X}}^{[k]} = \boldsymbol{X}^{[k]} + \alpha(\boldsymbol{X}^{[k]} - \boldsymbol{X}^{[k-1]}), \tag{22a}$$

$$\boldsymbol{X}^{[k+1]} = \operatorname{prox}_{\mathcal{F}_1/\tau_1^{[k]}}\left(\underline{\boldsymbol{X}}^{[k]} - \frac{1}{\tau_1^{[k]}}\nabla_{\boldsymbol{X}}\mathcal{H}(\underline{\boldsymbol{X}}^{[k]}, \boldsymbol{Y}^{[k]})\right), \tag{22b}$$

$$\underline{\boldsymbol{Y}}^{[k]} = \boldsymbol{Y}^{[k]} + \beta(\boldsymbol{Y}^{[k]} - \boldsymbol{Y}^{[k-1]}), \tag{22c}$$

$$\boldsymbol{Y}^{[k+1]} = \operatorname{prox}_{\mathcal{F}_2/\tau_2^{[k]}}\left(\underline{\boldsymbol{Y}}^{[k]} - \frac{1}{\tau_2^{[k]}}\nabla_{\boldsymbol{Y}}\mathcal{H}(\boldsymbol{X}^{[k+1]}, \underline{\boldsymbol{Y}}^{[k]})\right), \tag{22d}$$

where $\alpha \in \mathbb{R}_+$ and $\beta \in \mathbb{R}_+$ are the inertial parameters. When both inertial parameters are zero, the iterative procedure coincides with the original PALM algorithm [28].

To use the iPALM algorithm for (19), we replace its constraints with indicator functions and reformulate it into the form of (20):

$$\min_{\boldsymbol{X},\boldsymbol{Y}} \iota_{\mathcal{C}}(\boldsymbol{X}) + \iota_{\mathcal{N}}(\boldsymbol{Y}) + \mathcal{I}(\boldsymbol{X},\boldsymbol{Y}), \tag{23}$$

Algorithm 1 iPALM for mel-pectrogram inversion

Input: $\boldsymbol{X}^{[-1]}, \boldsymbol{X}^{[0]}, \boldsymbol{Y}^{[0]}, \lambda, \alpha$
Output: $\boldsymbol{X}^{[K]}$

    for $k = 0, \ldots, K-1$ **do**
        $\underline{\boldsymbol{X}}^{[k]} = \boldsymbol{X}^{[k]} + \alpha(\boldsymbol{X}^{[k]} - \boldsymbol{X}^{[k-1]})$
        $\boldsymbol{X}^{[k+1]} = P_{\mathcal{C}}(P_{\mathcal{Y}^{[k]}}(\underline{\boldsymbol{X}}^{[k]}))$
        $\boldsymbol{Z}^{[k]} = 1/(1+\lambda)|\boldsymbol{X}^{[k+1]}| + \lambda/(1+\lambda)P_{\mathcal{M}}(\boldsymbol{Y}^{[k]})$
        $\boldsymbol{Y}^{[k+1]} = P_{\mathcal{N}}(\boldsymbol{Z}^{[k]})$
    end for

---

where $\mathcal{I}(\boldsymbol{X},\boldsymbol{Y})$ is given by

$$\mathcal{I}(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{2}\mathcal{L}(\boldsymbol{X},\boldsymbol{Y}) + \frac{\lambda}{2}d_{\mathcal{M}}^2(\boldsymbol{Y}). \tag{24}$$

By substituting each term of (23) into (22), we yield Algorithm 1, where we set β to 0 and $\tau_1^{[k]}$ to $1/2$ for simplicity[3]. Meanwhile, we fix $\tau_2^{[k]}$ to $1 + \lambda$ based on the partial Lipschitz constant of the gradient of $\mathcal{I}(\cdot,\cdot)$. The proximity operator of each indicator function is the projection by definition, and the projection onto the set $\mathcal{N}$ is computed by taking the maximum with zero entry-wise. According to (22b), the update of $\boldsymbol{X}^{[k]}$ corresponds to the projected gradient method for minimizing $\mathcal{I}(\boldsymbol{X}, \boldsymbol{Y}^{[k]})$ on $\mathcal{C}$. By extending the interpretation of GLA as the projected gradient method in (10), we obtain the update as $P_{\mathcal{C}}(P_{\mathcal{Y}^{[k]}}(\underline{\boldsymbol{X}}^{[k]}))$, where $\mathcal{Y}^{[k]}$ is the set of the STFT coefficients whose magnitude is equal to $\boldsymbol{Y}^{[k]}$ at all time–frequency bins. More precisely, the update of $\boldsymbol{X}$ with the inertial acceleration corresponds to the fast GLA (FGLA) [2].

The iPALM algorithm for (13) requires changing the gradient in (22d) from that for (19). As a result, the update of $\boldsymbol{Z}$ in Algorithm 1 is replaced as follows:

$$\boldsymbol{Z}^{[k]} = \boldsymbol{Y}^{[k]} - \frac{1}{1+\lambda}\left(\boldsymbol{Y}^{[k]} - |\boldsymbol{X}^{[k+1]}| + \lambda\boldsymbol{G}\boldsymbol{Y}^{[k]} - \lambda\boldsymbol{v}\right), \tag{25}$$

where $\boldsymbol{G} = \boldsymbol{E}^{\mathsf{T}}\boldsymbol{E}$, $\boldsymbol{v} = \boldsymbol{E}^{\mathsf{T}}\boldsymbol{M}$, and $(\cdot)^{\mathsf{T}}$ denotes the transpose.

4. EXPERIMENTS

4.1. Evaluation on Speech Signals

In this section, we demonstrate the effectiveness of the proposed methods on speech signals. We used 200 utterances of the TIMIT dataset provided in [29], where the half was for tuning λ and the rest was for evaluation. The sampling rate was 16 kHz. STFT was performed using the Hann window of 1024 with a 256-sample shift, and the number of mel bins was 80, which is popular in recent text-to-speech pipelines [9–11]. We first investigated the effect of λ on the iPALM algorithms for (13) and (19), where the algorithms are abbreviated as Prop-mel and Prop-full, respectively. The number of iterations was 500, and $\alpha = 0.9$. The reconstructed signals were evaluated by PESQ [22], ESTOI [23], and the spectral convergence on mel-spectrogram:

$$\text{SCM} = 20\log_{10}\left(\frac{\|\boldsymbol{E}|\mathcal{G}(\widehat{\boldsymbol{x}})| - \boldsymbol{M}\|}{\|\boldsymbol{M}\|}\right), \tag{26}$$

---

[3]To guarantee the convergence to a critical point, $\tau_1^{[k]}$ should be tuned in each iteration [21]. However, we experimentally confirmed that Algorithm 1 stably works even with $\tau_1^{[k]} = 1/2$ regardless of the iteration index k.
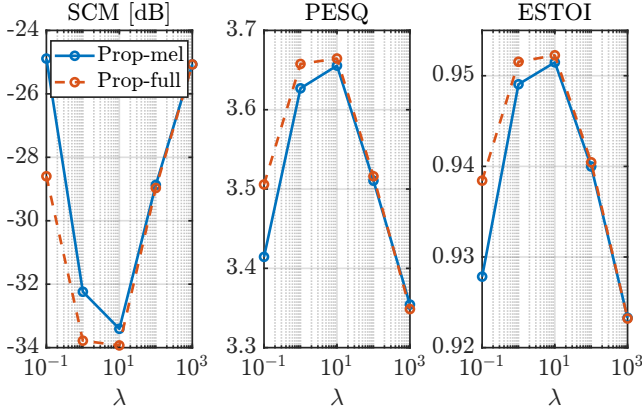
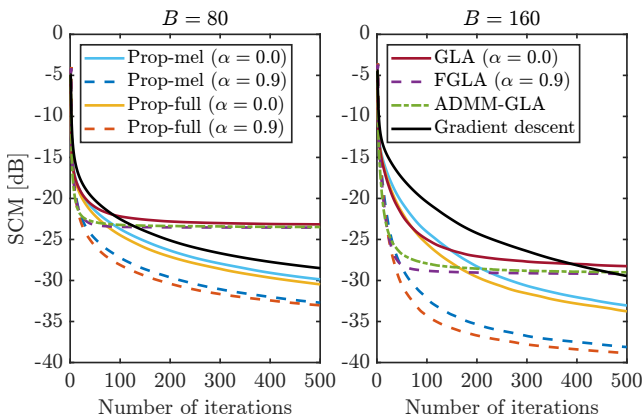Figure 2: Average SCM/PESQ/ESTOI of the signals reconstructed by the proposed methods with different λ.

Figure 3: SCM with respect to the number of iteration.

where $\widehat{\boldsymbol{x}}$ is the reconstructed signal. The objective measures with different λ are shown in Fig. 2. Both methods performed best with $\lambda = 10$, and thus we used this value in the following experiments. Prop-full is relatively insensitive to the value of λ, which will make the hyperparameter tuning easier in other conditions.

Next, we compared the proposed methods with the cascaded method, where `mel_to_stft` in `Librosa` solved the optimization problem in (12) by the L-BFGS-B algorithm [30]. Phase reconstruction was performed by GLA [1], FGLA [2], and ADMM-GLA [3]. We also evaluated the gradient descent method for (14). Fig. 3 shows SCM for various methods per iteration. Regardless of the phase reconstruction methods, the performance of the cascaded method was limited due to the error in the intermediate reconstruction of the full-band magnitude, especially when $B = 80$. The performance improvement of the gradient descent method was slower than the iPALM algorithms with $\alpha = 0.0$, while the computational complexity is the same. The inertial acceleration further improved the performance. Table 1 shows the PESQ and ESTOI after 500 iterations, and Prop-full with $\alpha = 0.9$ performed best.

4.2. Evaluation on Music and Environmental Signals

To demonstrate the effectiveness of the joint optimization, we compared Prop-full with the cascaded method using FGLA. As music signals, 12 song snippets from the MASS dataset[4] were used. The

---
[4] www.upf.edu/web/mtg/mass

Table 1: PESQ and ESTOI with different number of mel bins.

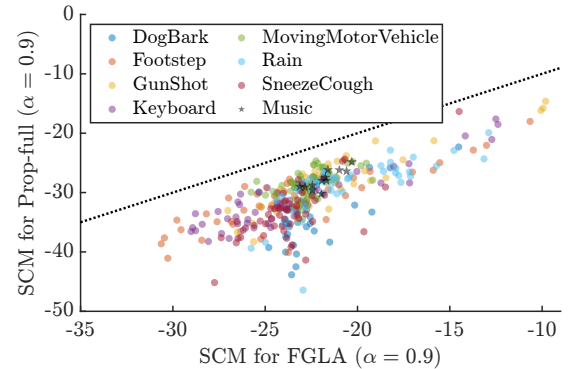|	$B = 80$		$B = 160$	
	PESQ	ESTOI	PESQ	ESTOI
GLA ($\alpha = 0.0$)	3.31	0.91	3.82	0.96
FGLA ($\alpha = 0.9$)	3.30	0.92	3.89	0.97
Gradient descent	3.57	0.94	3.86	0.97
Prop-mel ($\alpha = 0.0$)	3.60	0.94	3.92	0.97
Prop-mel ($\alpha = 0.9$)	3.65	**0.95**	4.05	**0.98**
Prop-full ($\alpha = 0.0$)	3.61	**0.95**	3.94	0.97
Prop-full ($\alpha = 0.9$)	**3.68**	**0.95**	**4.06**	**0.98**

Figure 4: SCM for Prop-full versus SCM for the cascaded method using FGLA. The dotted line indicates equal performance.

sampling rate was 44100 Hz, and mel-spectrograms with 96 bins were computed with the Hann window of 2048 samples and the 256-sample shift [20]. We also investigated the performance on foley sounds from the development sets of DCASE2023 Task 7[5]. We used 50 signals for each of the classes of foley sounds. While the sampling rate was 22050 Hz, the STFT conditions were the same as in Section 4.1, which coincides with the challenge baseline.

SCMs for Prop-full and the cascaded method using FGLA are shown in Fig. 4. While the performance of both methods was diverse, Prop-full achieved better performance regardless of the sound class. This result confirms the effectiveness of the proposed method not only for speech but also for general audio signals.

5. CONCLUSION

In this paper, we propose to jointly estimate the full-band magnitude and phase based on the bi-level consistency. We explore two optimization problems and derive the iPALM algorithms for them. Our experimental results on speech, music, and environmental signals show the advantage of the proposed method for general audio signals. We will improve the optimization algorithms to reduce the number of iterations for obtaining a signal with sufficient quality.

6. ACKNOWLEDGMENT

---
[5] dcase.community/challenge2023/task-foley-sound-synthesis

7. REFERENCES

[1] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, pp. 236–243, Apr. 1984.

[2] N. Perraudin, P. Balazs, and P. L. Søndergaard, "A fast Griffin–Lim algorithm," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2013, pp. 1–4.

[3] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Griffin–Lim like phase recovery via alternating direction method of multipliers," *IEEE Signal Process. Lett.*, vol. 26, pp. 184–188, Jan. 2019.

[4] T. Peer, S. Welker, and T. Gerkmann, "Beyond Griffin–Lim: Improved iterative phase retrieval for speech," in *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, Sept. 2022, pp. 1–5.

[5] Z. Průša, P. Balazs, and P. L. Søndergaard, "A noniterative method for reconstruction of phase from STFT magnitude," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 1154–1164, May 2017.

[6] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, and H. Saruwatari, "Phase reconstruction from amplitude spectrograms based on von–Mises-distribution deep neural network," in *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, Sept. 2018, pp. 286–290.

[7] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Deep Griffin–Lim iteration," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2019, pp. 61–65.

[8] ——, "Deep Griffin–Lim iteration: Trainable iterative phase reconstruction using neural network," *IEEE J. Sel. Top. Signal Process.*, vol. 15, pp. 37–50, Jan. 2021.

[9] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2018, pp. 4779–4783.

[10] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2018.

[11] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2021.

[12] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC3: Examining and improving CycleGAN-VCs for mel-spectrogram conversion," in *Proc. Interspeech*, Oct. 2020, pp. 2017–2021.

[13] T. Hayashi, W. C. Huang, K. Kobayashi, and T. Toda, "Non-autoregressive sequence-to-sequence voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, June 2021, pp. 7068–7072.

[14] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2019, pp. 3617–3621.

[15] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, Dec. 2019.

[16] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2020.

[17] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "ISTFT-NET: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2022, pp. 6207–6211.

[18] J. J. Webber, C. Valentini-Botinhao, E. Williams, G. E. Henter, and S. King, "Autovocoder: Fast waveform generation from a learned speech representation using differentiable digital signal processing," *arXiv:2211.06989*, 2022.

[19] Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, "Onoma-to-wave: Environmental sound synthesis from onomatopoeic words," *APSIPA Trans. Signal, Inf. Process.*, vol. 11, May 2022.

[20] B. D. Giorgi, M. Levy, and R. Sharp, "Mel spectrogram inversion with stable pitch," in *Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, Dec. 2022, pp. 233–239.

[21] T. Pock and S. Sabach, "Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems," *SIAM J. Imag. Sci.*, vol. 9, pp. 1756–1787, 2016.

[22] *P.862.2: Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, ITU-T Std. P.862.2, 2007.

[23] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 2009–2022, Aug. 2016.

[24] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, "A nonconvex approach for phase retrieval: Reshaped wirtinger flow and incremental algorithms," *J. Mach. Learn. Represent.*, vol. 18, pp. 1–35, Nov. 2017.

[25] P. Chen, A. Fannjiang, and G. R. Liu, "Phase retrieval with one or two diffraction patterns by alternating projection with null initialization," *J. Fourier Anal. Appl.*, vol. 24, pp. 719–758, June 2018.

[26] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in python," in *Proc. Python Science Conf.*, July 2015, pp. 18–24.

[27] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, pp. 127–239, Jan. 2014.

[28] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Progr.*, vol. 146, pp. 459–494, July 2013.

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

[30] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM J. Sci. Comput.*, vol. 16, no. 5, pp. 1190–1208, Sept. 1995.