Studies on DNN-based phase reconstruction from amplitude spectrograms

深層ニューラルネットを用いたスペクトログラムの位相復元

5119E021　升山 義紀　　指導教員　及川 靖広　教授
MASUYAMA Yoshiki　　Prof. OIKAWA Yasuhiro

概要：音声強調や音声合成などの音響信号処理の多くは，短時間フーリエ変換 (STFT) によって得られる時間周波数表現を介して行われる．STFT の複素係数の振幅に対する処理は古くから研究されてきたが，2010 年頃に位相が処理結果に与える影響が見直されて以降，位相に対する処理も盛んに研究されている．近年，深層ニューラルネット (DNN) による位相復元が複数提案されているが，DNN の学習が困難なことや，膨大なパラメータ数が必要なことが課題である．本研究では，DNN のみで位相を推定するのではなく，信号処理の補助として DNN を利用する枠組みを提案することでこれらの課題を解決する．学習不要な信号処理技術を部分的に用いることで，DNN の学習を安定的に行うことができ，少数のパラメータでも高い性能を実現することを確認した．

キーワード：Griffin–Lim 法，スペクトログラムの無矛盾性，時間周波数解析，音声合成

Keywords: Griffin–Lim algorithm, spectrogram consistency, time-frequency analysis, speech synthesis

1. Introduction

Phase reconstruction of a coefficient of the short-time Fourier transform (STFT) has gained much attention with a wide range of applications. Although the phase of an observed signal is available in some applications, it is unavailable in many applications including text-to-speech. To consider both situations, this thesis considers phase reconstruction that recovers the valid phase only from a given magnitude.

Griffin–Lim algorithm (GLA) has been widely used for phase reconstruction [1]. GLA is based on the spectrogram consistency which requires the complex STFT coefficient to be in the image of STFT. GLA often results in a low-quality signal due to the lack of prior knowledge of the target signal. To leverage the prior knowledge obtained from a training dataset, DNN-based phase reconstruction methods have been studied. They outperformed GLA by using prior knowledge of the target signal, but there still remains the room for improvement in terms of the parameter efficiency and the online extension.

In this thesis, I propose two DNN-based phase reconstruction frameworks. The first one, named *Deep Griffin–Lim Iteration* (DeGLI), incorporates DNN-based denoising into GLA. It takes over the iterative nature of GLA and exploits the two projections used in GLA, which allows one to reduce the number of DNN parameters. The second one is a DNN-based online phase reconstruction framework consisting of two parts: (i) estimation of the phase differences by causal DNNs; and (ii) reconstruction of the phase from the estimated differences. As both parts do not use future information, it woks in real-time.

2. Deep Griffin–Lim Iteration (DeGLI)

In this section, a phase reconstruction framework, named DeGLI, is proposed [2, 3].

2.1 Griffin–Lim algorithm (GLA) [1]

GLA is one of the most popular consistency-based phase reconstruction methods. It is formulated by

$$\mathbf{X}^{[m+1]} = P_{\mathcal{C}}\big(P_{\mathcal{A}}(\mathbf{X}^{[m]})\big), \tag{1}$$

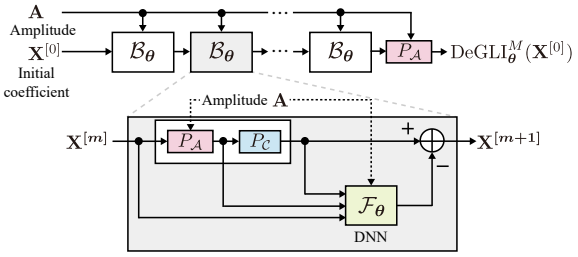

Fig. 1　Overview of DeGLI

where m is the iteration index, and $P_{\mathcal{A}}$ is the projection onto the set of complex STFT coefficients whose magnitude coincides with the given one $\mathbf{A}$:

$$P_{\mathcal{A}}(\mathbf{X})_{\omega,\tau} = A_{\omega,\tau} X_{\omega,\tau}/|X_{\omega,\tau}|. \tag{2}$$

Here, $P_{\mathcal{C}}$ is the projection onto the set of consistent STFT coefficients:

$$P_{\mathcal{C}}(\mathbf{X}) = \mathcal{G}(\mathcal{G}^{\dagger}(\mathbf{X})), \tag{3}$$

where $\mathcal{G}$ denotes STFT, and $\mathcal{G}^{\dagger}$ is its pseudo inverse.

2.2 Concept of DeGLI

Overview of DeGLI is illustrated in Fig. 1. It consists of the same multiple sub-block $\mathcal{B}_{\boldsymbol{\theta}}$:

$$\mathbf{X}^{[m+1]} = \mathcal{B}_{\boldsymbol{\theta}}(\mathbf{X}^{[m]}) \tag{4}$$

$$= \mathbf{Z}^{[m]} - \mathcal{F}_{\boldsymbol{\theta}}(\mathbf{X}^{[m]}, \mathbf{Y}^{[m]}, \mathbf{Z}^{[m]}), \tag{5}$$

where $\mathbf{Y}^{[m]} = P_{\mathcal{A}}(\mathbf{X}^{[m]})$, $\mathbf{Z}^{[m]} = P_{\mathcal{C}}(\mathbf{Y}^{[m]})$, $\mathcal{F}_{\boldsymbol{\theta}}$ is a DNN for denoising, and $\boldsymbol{\theta}$ is the parameters of the DNN. The entire DeGLI is defined as

$$\mathrm{DeGLI}^{M}_{\boldsymbol{\theta}}(\mathbf{X}^{[0]}) = P_{\mathcal{A}}(\mathcal{B}_{\boldsymbol{\theta}}(\cdots \mathcal{B}_{\boldsymbol{\theta}}(\mathcal{B}_{\boldsymbol{\theta}}(\mathbf{X}^{[0]})))), \tag{6}$$

where $\mathbf{X}^{[0]}$ is an initial coefficient. Thanks to leveraging the projections and using a single DNN iteratively, one can reduce the number of DNN parameters comparing to existing DNN-based phase reconstruction.

2.3 Experimental evaluation of DeGLI

DeGLI was compared with GLA, WaveNet and WaveGlow in a subjective test with 16 participants. The training used 12500 utterances in the LJ speech dataset, and the evaluation used other 40 utterances.

By inserting the DNN, DeGLI outperformed GLA as shown in Fig. 2. The score of DeGLI was compa-

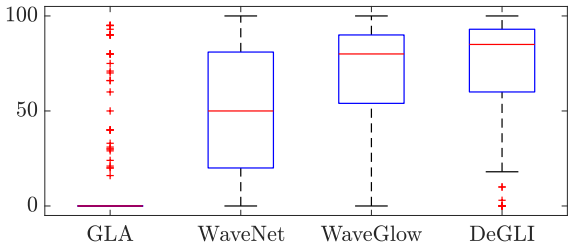

Fig. 2 Subjective scores of reconstructed signals

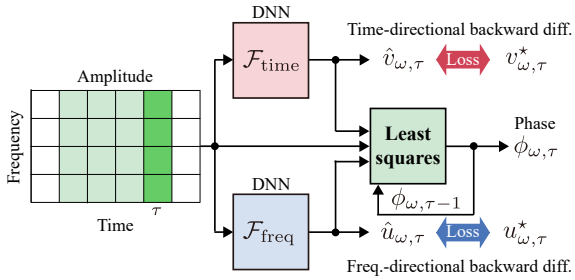

Fig. 3 Overview of two-stage online phase reconstruction

rable to WaveGlow even though the number of DNN parameters was reduced to less than 0.5%.

3. Online phase reconstruction via DNN

Online phase reconstruction is crucial for a wide range of applications including the incremental speech synthesis. To this end, a DNN-based online phase reconstruction framework is presented in this section.

3.1 Concept of proposed two-stage online phase reconstruction

The proposed DNN-based online phase reconstruction consists of two parts as shown in Fig. 3. First, the phase differences with respect to time and frequency are estimated by causal DNNs. Second, the phase is reconstructed from the estimated phase differences in a frame-by-frame manner. As both parts do not depend on future information, the proposed framework works in real-time without any look-ahead frames.

Let $\mathcal{W}(\cdot)$ be the wrapping operator, and $v_{\omega,\tau} = \mathcal{W}(\phi_{\omega,\tau} - \phi_{\omega,\tau-1})$ and $u_{\omega,\tau} = \mathcal{W}(\phi_{\omega,\tau} - \phi_{\omega-1,\tau})$ are the backward differences of phase with respect to time and frequency, respectively. These phase differences at the τth frame are estimated by two DNNs, $\mathcal{F}_{\text{time}}$ and $\mathcal{F}_{\text{freq}}$, as follows:

$$\hat{\mathbf{v}}_{\tau} = \mathcal{F}_{\text{time}}(\boldsymbol{\Psi}_{\tau}, \boldsymbol{\theta}_{\text{time}}), \tag{7}$$

$$\hat{\mathbf{u}}_{\tau} = \mathcal{F}_{\text{freq}}(\boldsymbol{\Psi}_{\tau}, \boldsymbol{\theta}_{\text{freq}}), \tag{8}$$

where $\boldsymbol{\Psi}_{\tau}$ is the input feature calculated from the current and previous magnitudes. A periodic loss function should be used to train the DNNs because the true phase differences have the ambiguity of $2n\pi$ [4].

The second part of the proposed framework reconstructs the phase from the phase differences estimated by the DNNs. Since it is not easy to directly handle the phase differences with the ambiguity, they are converted to the ratios of complex STFT coefficients:

$$\hat{\nu}_{\omega,\tau} = A_{\omega,\tau}/A_{\omega,\tau-1}\mathrm{e}^{i\hat{v}_{\omega,\tau}}, \tag{9}$$

$$\hat{\upsilon}_{\omega,\tau} = A_{\omega,\tau}/A_{\omega,\tau-1}\mathrm{e}^{i\hat{u}_{\omega,\tau}}. \tag{10}$$

The obtained ratios are insensitive to the ambiguity of the estimated phase differences.

Using the ratios of complex STFT coefficients, an online phase reconstruction is formulated as a weighted lest squares problem in the complex domain:

$$\tilde{\mathbf{x}}_{\tau} = \arg\min_{\tilde{\mathbf{x}}_{\tau}} \sum_{\omega=0}^{K-1} \lambda_{\omega,\tau} |\tilde{x}_{\omega,\tau} - \hat{\nu}_{\omega,\tau} x_{\omega,\tau-1}|^2 + \sum_{\omega=1}^{K-1} \gamma_{\omega,\tau} |\tilde{x}_{\omega,\tau} - \hat{\upsilon}_{\omega,\tau} \tilde{x}_{\omega-1,\tau}|^2, \tag{11}$$

$$x_{\omega,\tau} = A_{\omega,\tau} \tilde{x}_{\omega,\tau} / |\tilde{x}_{\omega,\tau}|, \tag{12}$$

where $\lambda_{\omega,\tau} > 0$ and $\gamma_{\omega,\tau} > 0$ are the reliability of the estimated phase differences with respect to time and frequency, respectively. The final estimate of phase $\phi_{\omega,\tau}$ is obtained by taking complex argument of $x_{\omega,\tau}$.

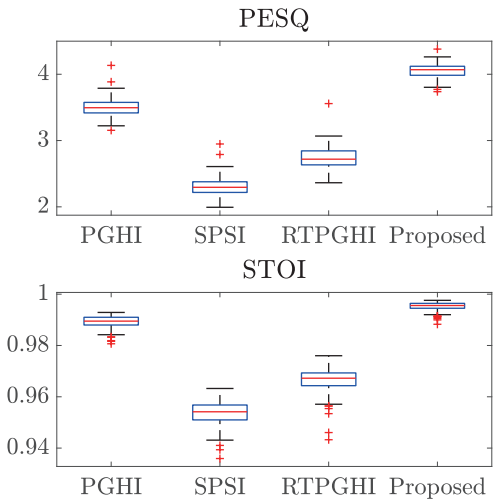

Fig. 4 PESQ and STOI of reconstructed signals

3.2 Experimental evaluation of DNN-based online phase reconstruction

The proposed framework was compared with two online phase reconstruction methods: the single pass spectrogram inversion (SPSI) and the real-time phase gradient heap integration (RTPGHI). The offline version of RTPGHI (PGHI) was also evaluated.

Objective measures of the reconstructed signals are illustrated in Fig. 4. The proposed framework outperformed not only online but also offline methods.

4. Conclusion

In this thesis, two DNN-based phase reconstruction frameworks are presented. Both of them obtain benefits from the strong modeling capability of DNN and the efficiency of signal processing techniques. Thanks to this, the proposed frameworks reconstruct high-quality speech signals compared with either of one.

References

[1] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," IEEE TASSP, vol.32, no.2, pp.236–243, Apr. 1984.

[2] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Deep Griffin–Lim iteration," Proc. IEEE ICASSP, pp.61–65, May 2019.

[3] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa and N. Harada, "Deep Griffin–Lim Iteration: Trainable Iterative Phase Reconstruction Using Neural Network," IEEE JSTSP, vol.25, no.1, pp.37–50, Jan. 2021.

[4] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Phase reconstruction based on recurrent phase unwrapping with deep neural networks," Proc. IEEE ICASSP, pp.826–830, May 2020.