Griffin–Lim phase reconstruction using short-time Fourier transform with zero-padded frame analysis

Yukoh Wakabayashi* and Nobutaka Ono*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: {wakayuko, onono}@tmu.ac.jp

***Abstract*—In this paper, we present the short-time Fourier transform (STFT) with zero-padded frame analysis to introduce frequency redundancy into a time–frequency representation, and we investigate its application to phase reconstruction by the Griffin–Lim algorithm. Recent studies on phase reconstruction have suggested that the use of a small STFT frame shift improves the performance of phase reconstruction techniques, which implies that increasing the temporal redundancy of the time–frequency representation makes phase reconstruction easier. Motivated by this, we consider the STFT with zero padding to increase the redundancy of the STFT along the frequency axis. The linearity of this transform and its inverse enables the use of the Griffin–Lim algorithm on the time–frequency domain. We evaluate the performance of phase reconstruction using the STFT with and without zero padding using the spectral distance and perceptual evaluation of speech quality as the criteria. The experimental results show that increasing the frequency redundancy with zero padding improves the phase reconstruction performance similarly to using a small frame shift.**

I. Introduction

The effect of phase information on speech quality has motivated further research on phase-aware signal processing [1]–[3]. It has been reported that phase awareness improves the performance of speech enhancement and source separation. The authors of [4]–[8] consider the harmonic structure of speech and propose a method of estimating the harmonic phase of speech for speech enhancement. Complex-valued spectral gain estimations using the minimum mean square error criterion and the maximum a posteriori criterion are discussed in [9], [10]. In [11], a time–frequency (TF) mask is estimated by using a deep neural network employing the amplitude and phase as input features. Deep clustering [12] has also been used to estimate a real-valued mask for source separation using the phase as a feature. In addition, nonnegative matrix factorization has been extended to handle complex-valued spectra [13]. The phase of speech is reconstructed for source separation and speech synthesis in [14] and [15], respectively.

Most of the above signal processing studies are carried out in the TF domain using the short-time Fourier transform (STFT), which consists of frame segmentation, windowing, and conducting the discrete Fourier transform (DFT). It is well known that the STFT representation has redundancy due to the overlap of frames, that is, a segmented signal sample is included at multiple frames. The Griffin–Lim algorithm (GLA) [16] is a popular phase reconstruction method, and many derivative methods have been proposed [17]–[19]. This algorithm reconstructs the phase by performing iterative calculations with the STFT and inverse STFT (iSTFT) using the redundancy of the STFT. The GLA reconstructs the phase by exploiting this redundancy. In addition, many phase reconstruction methods, including the harmonic-structure-based phase reconstructions [4], [8], employ an STFT frame shift of one-eighth or one-quarter of the frame length. The use of a small frame shift is equivalent to increasing the redundancy of the TF representation along the time axis, which results in improved phase reconstruction performance. These results suggest that increasing the redundancy of the representation is a good approach to estimating the phase.

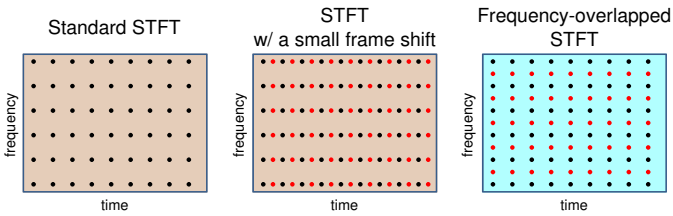


Fig. 1. Conceptual diagrams showing differences among STFT, STFT with a small frame shift, and frequency-overlapped STFT.

In this paper, we consider an STFT that is redundant along the frequency axis, the so-called "frequency-overlapped" STFT, as shown in Fig. 1. It is expected that the increase in the frequency redundancy in the STFT will also make phase reconstruction easier, similarly to the STFT with a small frame shift. We implement the increase in the frequency redundancy by using the well-known zero padding. That is, in this paper, we investigate whether the zero padding improves phase reconstruction by the GLA.

The remainder of this paper is organized as follows. We explain the zero-padded frame analysis and define "frequency-overlapped" STFT in Section II. In Section III, we also define its inverse transform and introduce an application of this STFT to phase reconstruction with the GLA. We evaluate the phase reconstruction accuracy with the zero-padded frame analysis through experiments in Section IV. Finally, we draw conclusions in Section V.

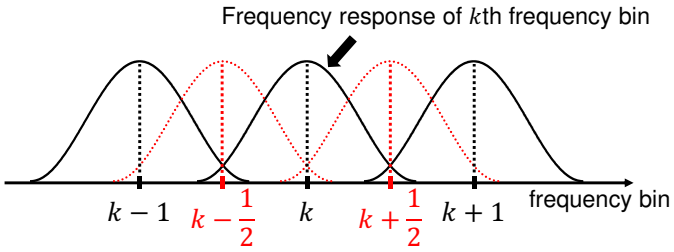


Fig. 2. Conceptual illustration of increase in STFT redundancy in the frequency domain.

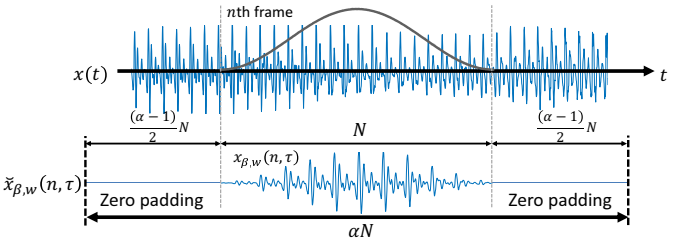


Fig. 3. Illustration of zero-padded frame analysis.

II. Increasing Redundancy of STFT along Frequency Axis

We consider increasing the redundancy of the STFT along the frequency axis in our proposed alternative TF representation. It is first noted that the STFT can be interpreted as a type of filter bank. That is, the kth frequency spectrum is obtained from the filtered signal using a bandpass filter with the central frequency corresponding to the kth bin, as shown in Fig. 2. This bandpass filter is derived from the analysis window function and complex exponential function in the DFT. We can regard the partial overlap between these filters as a redundancy along the frequency axis. Namely, increasing this overlap increases the redundancy. First, we introduce the standard STFT and iSTFT in Section II-A. Next, we define and formulate the aforementioned transform in Section II-B.

A. Standard STFT and its inverse transform

Let t, n, k, and N be the time sample index, frame index, frequency bin index, and frame length, respectively. Then, the STFT of a time-domain signal $x(t)$ with a frame shift of N/β is formulated as

$$X_\beta(n,k) = \sum_{\tau=0}^{N-1} x_{\beta,w}(n,\tau)(F_N)^{k\tau}, \tag{1}$$

$$x_{\beta,w}(n,\tau) = w(\tau)x(\tau + n\frac{N}{\beta}), \tag{2}$$

where $F_N = \exp(-j2\pi/N)$, $w(\tau)$ is the analysis window function, and j is the imaginary unit. The signal $\tilde{x}(t)$ is inversely synthesized using the iSTFT and an overlapping process as

$$\tilde{x}(t) = \sum_n \tilde{w}(t - n\frac{N}{\beta})\tilde{x}_n(t - n\frac{N}{\beta}), \tag{3}$$

$$\tilde{x}_n(\tau) = \frac{1}{N}\sum_{k=0}^{N-1} X_\beta(n,k)(F_N)^{-k\tau}, \tag{4}$$

where $\tilde{w}(\tau)$ is the synthesis window and $\tilde{x}(t)$ is identically equal to $x(t)$ when the following condition is satisfied:

$$\sum_n w(t - n\frac{N}{\beta})\tilde{w}(t - n\frac{N}{\beta}) = 1. \tag{5}$$

B. Formulation of STFT with zero-padded frame analysis

Here, we consider increasing the fineness of the frequency bin index $k = 0, \cdots, N-1$ in (1) as illustated in Figs. 1 and 2. This idea is simply realized as follows by replacing k by a new variable k/α $(0, \cdots, \alpha N - 1)$:

$$X'_\beta(n,k) = \sum_{\tau=0}^{N-1} x_{\beta,w}(n,\tau)(F_N)^{\frac{k}{\alpha}\cdot\tau} \tag{6}$$

$$= \sum_{\tau=0}^{\alpha N-1} \breve{x}_{\beta,w}(n,\tau)(F_{\alpha N})^{(\tau-\Delta)k}, \tag{7}$$

where $\Delta = (\alpha-1)N/2$ and $\breve{x}_{\beta,w}(n,\tau)$ is defined as

$$\breve{x}_{\beta,w}(n,\tau) = \begin{cases} x_{\beta,w}(n,\tau-\Delta), & \tau = \Delta,\cdots,\Delta+N-1, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

These equations show that increasing the fineness of k is equivalent to the STFT of the αN-length augmented signal padded with Δ-length zeros on both sides, $\breve{x}_{\beta,w}(n,\tau)$, as illustrated in Fig. 3. Therefore, we define the "frequency-overlapped" STFT of $x(t)$ using the parameters α and β, which respectively control the redundancy of frequency and time, as the follows:

$$X_{\alpha,\beta}(n,k) = \mathbf{STFT}_{\alpha,\beta}\left[x(t)\right] \tag{9}$$

$$= \sum_{\tau=0}^{\alpha N-1} \breve{x}_{\beta,w}(n,\tau)(F_{\alpha N})^{\tau k}, \tag{10}$$

where the origin of the phase is set at the foremost point in the αN-length augmented frame. Note that $\mathbf{STFT}_{1,\beta}$ is equivalent to the standard STFT (using a nonzero-padded frame) with a frame shift of N/β.

III. Phase Reconstruction using STFT with Zero-Padded Frame Analysis

We consider phase reconstruction using GLA with the aforementioned transform. Nakamura and Kameoka [20] proposed a GLA with the wavelet transform, which suggests that any linear TF representation can be used for phase reconstruction with the GLA if the inverse transform is a linear operation. Therefore, we also define the inverse of the STFT with zero-padded frame analysis. We can simply formulate the inverse

Algorithm 1 Griffin–Lim phase reconstruction with $\mathbf{STFT}_{\alpha,\beta}$

Input: $|\mathbf{X}^{[0]}| \in \mathbb{R}^{\alpha N \times T}$, where T is total frame number
Output: $\hat{x}(t)$
initial random phase $\mathbf{\Phi} \in \mathbb{R}^{\alpha N \times T}$
$\mathbf{X}^{[1]} = |\mathbf{X}^{[0]}| \odot \mathbf{\Phi}$
$x^{[1]}(t) = \mathbf{iSTFT}_{\alpha,\beta}\left[\mathbf{X}^{[1]}\right]$
for $i = 1 : I-1$ **do**
$\mathbf{Y}^{[i]} = \mathbf{STFT}_{\alpha,\beta}\left[x^{[i]}(t)\right]$
$\mathbf{X}^{[i+1]} = |\mathbf{X}^{[0]}| \odot \exp\left\{j\angle\mathbf{Y}^{[i]}\right\}$
$x^{[i+1]}(t) = \mathbf{iSTFT}_{\alpha,\beta}\left[\mathbf{X}^{[i+1]}\right]$
end for
return $\hat{x}(t) = x^{[I]}(t)$

transform of $X_{\alpha,\beta}(n,k)$ as follows:

$$\tilde{x}(t) = \mathbf{iSTFT}_{\alpha,\beta}\left[X_{\alpha,\beta}\right] \tag{11}$$

$$= \sum_n \tilde{w}(t - n\frac{N}{\beta})\tilde{x}_n(t - n\frac{N}{\beta}), \tag{12}$$

$$\tilde{x}_n(\tau) = \breve{x}_n(\tau + \Delta), \quad \tau = 0, \cdots, N-1, \tag{13}$$

$$\breve{x}_n(\tau) = \frac{1}{\alpha N}\sum_{k=0}^{\alpha N-1} X_{\alpha,\beta}(n,k)(F_{\alpha N})^{-\tau k}, \tag{14}$$

where (13) is the term regarding the consistency of the STFT with zero-padded frame analysis: from the definitions of (8) and (10), $\breve{x}(\tau)$ in (14) obtained from $X_{\alpha,\beta}(n,k)$, which we call a consistent transform, requires the following numerical restriction:

$$\breve{x}_n(\tau) = 0, \quad 0 \le \tau \le \Delta - 1 \cup \Delta + N \le \tau \le \alpha N - 1. \tag{15}$$

That is, if $X_{\alpha,\beta}(n,k)$ is not consistent, this restriction is not satisfied; therefore, inconsistent elements, namely, nonzero elements appear in the segments. Accordingly, we prune these elements in (13) before the overlap-add process. In addition, note that this definition is equivalent to the use of a zero-padded synthesis window. Namely, if any analysis window is selected, the perfect reconstruction condition in (5) is always satisfied owing to the zeros on both sides of this synthesis window.

We apply the STFT with zero-padded frame analysis to phase reconstruction with the GLA. This STFT increases the redundancy of the TF representation, whereas the GLA estimates the consistent phase such that it decreases the redundancy. As a result, it is expected that the zero padding will provide more information for reconstructing the consistent phase with the GLA. Concretely, the number of nonzero elements on both sides of $\breve{x}_n(\tau)$, which originate from the phase replacement in the GLA, is reduced during the iteration because they become zero as originally defined. This algorithm is given in Algorithm 1, where the operation $\odot$ is the element-wise product.

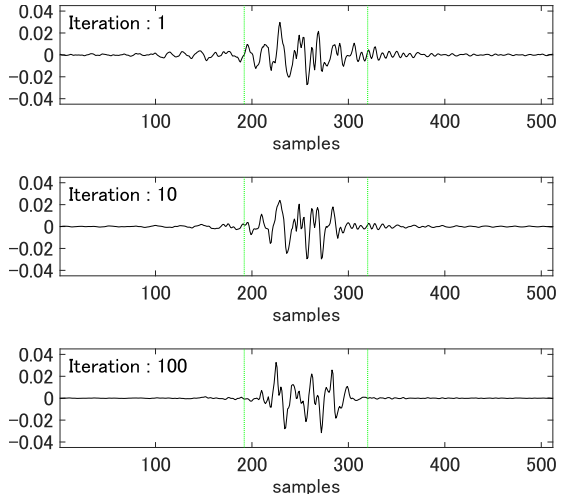
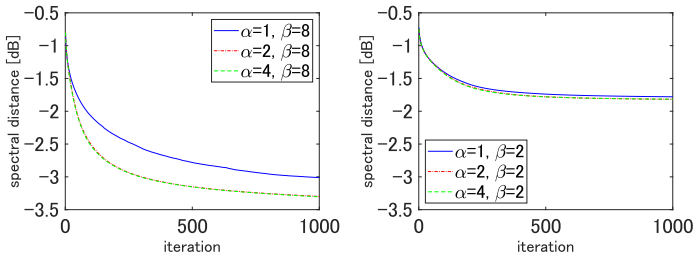


Fig. 4. Segmented waveforms in a frame, $\breve{x}_n(\tau)$ in (14), with the iSTFT ($\alpha = 4$), after 1, 10, and 100 iterations during the GLA, where the green vertical lines show both bounds of the frame.



(a) $N = 512$ with a $1/8$ shift (b) $N = 512$ with a $1/2$ shift

Fig. 5. Spectral distance as a function of the number of iterations of the GLA, where N is the frame length.

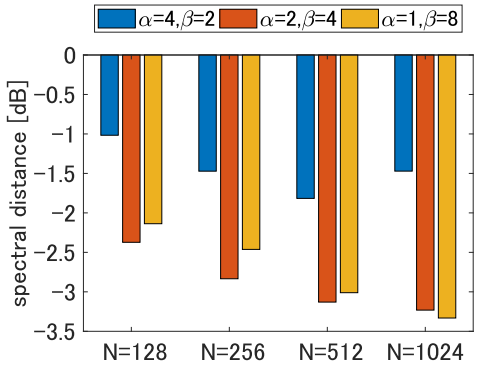


Fig. 6. Spectral distance after 1,000 iterations for an equal number of measurements, the product of the parameters α and β is 8 in this case, and N indicates the frame length.

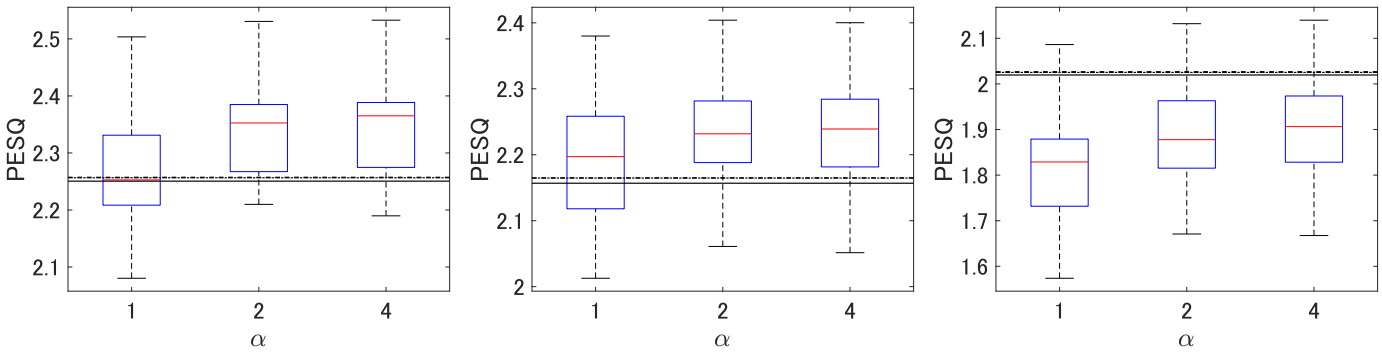
IV. Evaluation

A. Setup

We conducted two experiments to evaluate the phase reconstruction performance of the STFT with zero-padded frame analysis compared with that of the standard STFT ($\alpha = 1$) in different cases: the given amplitude spectrum in the GLA is an oracle and corrupted by noise. In the latter case, we suppose an application to speech enhancement, i.e., the estimated amplitude is obtained from a noisy amplitude corrupted by a white noise of 0 dB from the NOISEX-92 database [21].

TABLE I
AVERAGE SPECTRAL DISTANCE [DB] AFTER 1,000 ITERATIONS IN THE ORACLE AMPLITUDE CASE, WHERE N IS THE FRAME LENGTH.

N	128			256			512			1024		
α	1	2	4	1	2	4	1	2	4	1	2	4
$\beta = 2$	−0.97	−1.02	−1.02	−1.37	−1.47	−1.47	−1.78	−1.82	−1.82	−1.44	−1.48	−1.47
$\beta = 4$	−2.16	−2.37	−2.32	−2.71	−2.83	−2.83	−3.08	−3.13	−3.13	−3.19	−3.23	−3.22
$\beta = 8$	−2.14	−2.77	−2.76	−2.46	−2.92	−2.96	−3.01	−3.30	−3.30	−3.33	−3.46	−3.46



(a) $N = 256$ with $1/8$ shift ($\beta = 8$) (b) $N = 256$ with $1/4$ shift ($\beta = 4$) (c) $N = 256$ with $1/2$ shift ($\beta = 2$)

Fig. 7. Box plots of PESQ score in speech enhancement application, where N is the frame length, and the black horizontal lines show the PESQ score of the combination between the estimated amplitude and noisy phase with $\alpha = 1$ (solid), $\alpha = 2$ (chain), and $\alpha = 4$ (dashed).

We selected 10 utterances from the TIMIT database [22] as the target speech samples with a sampling rate of 16 kHz. We used the Hann window for spectrum analysis and the window that satisfies a perfect reconstruction condition [16] for synthesis. We ran the GLA for 1000 iterations with frame lengths of 8, 16, 32, and 64 ms and frame shifts of one-eighth, one-quarter, and one-half of the frame length, that is, $\beta = 8, 4, 2$, and compared the standard STFT ($\alpha = 1$) and the STFT with zero padding ($\alpha = 2, 4$). We performed the experiment ten times for every target speech with a different initial random phase, which is uniformly distributed for an unbiased evaluation. In the estimated amplitude case, after we estimated the amplitude spectrogram with an MMSE-STSA estimator [23], we conducted the GLA using the spectrogram. When the spectral gain was estimated, we used the oracle noise power. In the oracle amplitude case, we evaluated the quality of phase reconstruction in terms of the spectral distance (SD) formulated as

$$\mathrm{SD} = 10\log_{10}\left\{\left|\left||\mathbf{X}_{1,4}| - |\hat{\mathbf{X}}_{1,4}|\right|\right|_{\mathrm{F}}^{2} \Big/ ||\mathbf{X}_{1,4}||_{\mathrm{F}}^{2}\right\}, \quad (16)$$

where $\mathbf{X}_{1,4}$ and $\hat{\mathbf{X}}_{1,4}$ are the original and the estimated spectrograms using the standard STFT with a $1/4$ shift, that is, $\mathbf{X}_{1,4} = \mathbf{STFT}_{1,4}[s(t)]$ and $\hat{\mathbf{X}}_{1,4} = \mathbf{STFT}_{1,4}[\hat{x}(t)]$, where $s(t)$ is a target speech, and $||\cdot||_{\mathrm{F}}^2$ is the Frobenius norm. In the estimated amplitude case, we used the perceptual evaluation of speech quality (PESQ) [24] to determine speech quality.

B. Experimental results and discussion

First, we confirm that the GLA achieves the constancy of the STFT with zero-padded frame analysis along the frequency axis. Fig. 4 shows $\breve{x}_n(\tau)$ in (14) after 1, 10, and 100 iterations during the GLA with the STFT ($\alpha = 4$). As shown in the figure, both sides of the segmented waveforms converge to zero with increasing number of iterations, as expected.

Fig. 5 and Table I show the average SD as a function of the number of iterations and after 1000 iterations, respectively. These results show that the smaller the frame length and frame shift, the higher the phase reconstruction performance with zero-padded frame analysis. On the other hand, none of cases with a shift of half of the frame length reduced the SD. This is because the small amount of redundancy in the shift of half the frame length makes it difficult to estimate the phase using the GLA. Large frame lengths provide sufficient information to reconstruct consistent phases; therefore, the improvement in the SD for the standard STFT ($\alpha = 1$) with a large frame length becomes close to that of the STFT ($\alpha = 2, 4$). Note that the number of iterations required for convergence using the zero-padded STFT is smaller than that using the standard STFT in Fig. 5; however, the calculation cost of an iteration becomes expensive owing to the increase in the DFT length αN.

The bar graph in Fig. 6 illustrates the SD after 1000 iterations for an equal number of measurements, i.e., the product of the parameters α and β is the same, $\alpha\beta = 8$ in this case. This condition equalizes the number of TF points that these two parameters affect. Fig. 6 shows the following two findings: the half-shift case ($\beta = 2$) has a lower performance than the other cases, whereas the performance with the combination of the STFT and zero padding is slightly higher than that with the standard STFT, except for $N = 1024$. Overall, the use of the STFT with zero padding achieves the same or slightly better performance than the standard STFT for an equal number of measurements.

Fig. 7 illustrates box plots of PESQ scores of 100 samples (10 sources and 10 trials) after 1,000 iterations in the case of using estimated amplitudes with frame length $N = 256$ (16 ms). The three black lines (solid, chain, and dashed) indicate the PESQ score when using the estimated amplitude and noisy phase with $\alpha = 1, 2, 4$, respectively. Note that the chain and dashed lines ($\alpha = 2, 4$) overlap in this figure. As shown, using zero padding in the GLA improves PESQ, except for the half shift. This result is similar to the case of the oracle amplitude, which shows that the GLA with zero-padded frame analysis improves the phase reconstruction performance. In addition, when other frame lengths, e.g., $N = 128, 512, 1024$, were used in the analysis, the PESQs using the noisy phase and phase reconstruction using the GLA with $\alpha = 2, 4$ were the same, except for the half shift, whereas the PESQ with $\alpha = 1$ was lower than the PESQ with the noisy phase, just like Fig. 7(c). The combination of increasing frequency and temporal redundancies provides a better phase reconstruction performance than the standard STFT with a small frame shift. However, a shift smaller than one-half, that is, the redundancy of the STFT along the time axis, is required to accurately reconstruct the phase even when zero-padded frame analysis is used.

V. Conclusion

In this paper, we discussed a TF transform considering the redundancy of the STFT and introduced zero padding to the STFT and iSTFT. We additionally proposed a GLA based on these transforms. The use of zero padding provided a better phase reconstruction performance than the standard STFT with a small frame shift in terms of SD in the oracle amplitude case. In speech enhancement applications, the use of zero padding accurately reconstructed phase spectra in terms of PESQ. Future work involves the evaluation of speech quality by a listening test, the use of the zero-padded frame analysis in other applications such as speech synthesis and source separation, and the evaluation of the performance of the applications.

VI. Acknowledgements

This work was supported by the SECOM Science and Technology Foundation and JSPS KAKENHI Grant Number 19H21546.

References

[1] T. Gerkmann, M. Krawczyk, and J. Le Roux, "Phase processing for single-channel speech enhancement," *IEEE Signal Process. Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[2] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," in *ELSEVIER, Speech Commu.*, vol. 81, 2016, pp. 1–29.

[3] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*. Chichester, UK: John Wiley & Sons, 2017.

[4] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 22, no. 12, pp. 1931–1940, 2014.

[5] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 23, no. 8, pp. 1283–1294, 2015.

[6] ——, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information," *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 23, no. 9, pp. 1521–1532, 2015.

[7] Y. Wakabayashi, T. Fukumori, M. Nakayama, T. Nishiura, and Y. Yamashita, "Phase reconstruction method based on time-frequency domain harmonic structure for speech enhancement," in *Proc IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 5560–5565.

[8] ——, "Single-channel speech enhancement with phase reconstruction based on phase distortion averaging," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 26, no. 9, pp. 1559–1569, Sept 2018.

[9] M. Krawczyk and T. Gerkmann, "On MMSE-based estimation of amplitude and complex speech spectral cofficients under phase-uncertainly," *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 24, no. 12, pp. 2251–2262, 2016.

[10] P. Mowlaee, J. Stahl, and J. Kulmer, "Iterative joint MAP single-channel speech enhancement given non-uniform phase prior," *ELSEVIER, Speech Commu.*, vol. 86, no. C, pp. 85–96, Feb. 2017. [Online]. Available: https://doi.org/10.1016/j.specom.2016.11.008

[11] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *Proc IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016, pp. 5220–5224.

[12] G. Wichern and J. L. Roux, "Phase reconstruction with learned time-frequency representations for single-channel speech separation," in *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, 2018, pp. 396–400.

[13] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE/ACM Trans. on Audio, Speech and Lang. Process.*, vol. 21, no. 5, pp. 971–982, May 2013.

[14] P. Magron, R. Badeau, and B. David, "Model-based STFT phase recovery for audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1095–1105, June 2018.

[15] 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)*, 2018, pp. 286–290.

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

[17] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. ISCA Workshop Statistical Perceptual Audition (SAPA)*, 2008, pp. 23–28.

[18] N. Perraudin, P. Balazs, and P. L. Sondergaard, "A fast Griffin-Lim algorithm," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2013, pp. 1–4.

[19] K. Yatabe, Y. Masuyama, and Y. Oikawa, "Rectified linear unit can assist Griffin-Lim phase recovery," in *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, 2018, pp. 555–559.

[20] T. Nakamura and H. Kameoka, "Fast signal reconstruction from magnitude spectrogram of continuous wavelet transform based on spectrogram consistency," in *DAFx*, 2014.

[21] A. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," DRA Speech Res. Unit, Tech. Rep., 1992.

[22] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download," Philadelphia: Linguistic Data Consortium, 1993.

[23] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[24] ITU-T, "Perceptual evaluation of speech quality (PESQ)," *ITU-T Rec. P. 862*, 2001.