PHASE RECONSTRUCTION METHOD BASED ON TIME-FREQUENCY DOMAIN HARMONIC STRUCTURE FOR SPEECH ENHANCEMENT

Yukoh Wakabayashi, Takahiro Fukumori, Masato Nakayama, Takanobu Nishiura, and Yoichi Yamashita

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

ABSTRACT

Speech enhancement in noisy environments has been widely investigated by modifying only the amplitude spectrum of the speech signal, while the phase spectrum, which is regarded as an unimportant feature, is ignored. However, it has recently been reported that the phase spectrum plays an important role in the intelligibility and quality of speech. We propose a speech-enhancement method with phase reconstruction, which estimates inartificial phase spectrum by using the time-frequency feature called phase distortion, though a conventional phase reconstruction estimates artificial one. The objective experimental results indicate improvement in speech quality with and efficiency of the proposed method.

Index Terms— Phase reconstruction, Phase distortion, Speech enhancement, Harmonic structure, Fundamental frequency

1. INTRODUCTION

The improvement in engineering techniques has enabled speech communication and speech recognition anywhere with the use of smart phones and tablet PCs. In these cases, speech is corrupted by additive background noise, which makes it difficult to communicate smoothly and maintain performance of automatic speech recognition. Therefore, it is important to estimate clean speech from degraded speech through speech enhancement, thus, many studies on speech enhancement have been conducted [1–5]. The speech-enhancement process is generally done in the time-frequency domain. The amplitude spectrum has been particularly addressed, while the phase spectrum has not been regarded as an important feature [6, 7]. Even spectral subtraction [1] and minimum mean square error - short time spectral amplitude (MMSE-STSA) estimator [2], which are well known speech enhancement techniques, do not reconstruct the phase spectrum. The reasons the phase spectrum has been ignored are as follows.

- The spectral amplitude is known to be more important in speech communication than the phase spectrum [6, 7].
- Ephraim and Malah have reported that the noisy phase (unprocessed) is the MMSE optimal estimate [2].

This work was supported by JSPS KAKENHI Grant Number 16J07584, 26280065, 15K16030, and 16K16094.

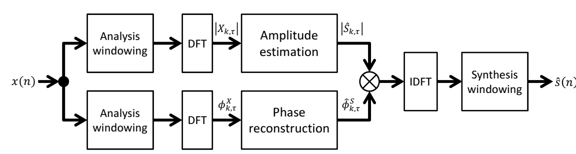


Fig. 1. Block diagram of speech enhancement with phase reconstruction.

- The phase wrapping issue makes it difficult to analyze the characteristics of the phase spectrum.

However, Paliwal et al. [8] have recently argued that the clean phase spectrum improves speech intelligibility and the perceptual evaluation of speech quality (PESQ) [9]. Moreover, many studies on the importance of the phase spectrum in speech intelligibility and quality have been conducted [10–13]. These investigations enabled the widespread study of phase reconstruction in speech enhancement. Figure 1 shows a block diagram of speech enhancement with phase reconstruction.

Griffin et al. [14] and Le Roux et al. [15] proposed iterative phase-reconstruction methods based on the discrete Fourier transform (DFT) and inverse DFT (IDFT). Sugiyama et al. [16] proposed noise reduction with phase randomization. Gerkmann et al. [17–21] and Mowlaee et al. [22–24] proposed fundamental frequency-based methods. This fundamental frequency-based approach involves a simple algorithm and has received much attention. In this paper, we focus on this approach and propose a phase-reconstruction method. The short-time Fourier transform phase improvement (STFTPI) method [17], which is used as a baseline method, reconstructs artificial phase spectrum, which causes buzzy speech quality. We tackle this problem and improve speech quality by estimating inartificial phase spectrum.

The remainder of this paper is organized as follows. In Section 2, we review related work regarding phase reconstruction for speech enhancement. In Section 3, we introduce a signal model and define symbols. We describe the proposed phase-reconstruction method in Section 4 and discuss the evaluation of the method in Section 5. Finally, we draw conclusions in Section 6.

2. RELATED WORK

There are two approaches related to phase reconstruction for speech enhancement. One involves amplitude-required methods. The iterative phase-spectrum-estimation method by Griffin and Lim based on the DFT and IDFT is a well known phase-reconstruction method. The consistent Wiener filtering proposed by Le Roux et al. [15] is also an iterative method that estimates the consistently complex spectrum with iteration of the DFT and IDFT. However, these iterative methods require a clean spectral amplitude, and their estimation accuracy significantly depends on the amplitude spectrum. Moreover, these methods have high computational complexity, which makes real-time processing difficult. Mowlaee and Saedi [25] reported phase reconstruction based on the relationship of geometry, but this reconstruction also requires a clean amplitude spectrum, which is a difficult condition to satisfy, just like with iterative methods.

The other approach has recently received much attention in phase reconstruction. This involves model-based methods, particularly fundamental frequency-based methods. Compared to amplitude-required methods, fundamental frequency-based methods are advantageous in that the clean amplitude spectrum is not required and the computational complexity is low. Gerkmann et al. [17] and Mowlaee et al. [22] proposed fundamental frequency-based methods. While phase decomposition and spectral smoothing with relative phase shift were used for phase reconstruction [22], the harmonic signal model was used and phase reconstruction along time and frequency was defined [17]. With the STFTPI [17], it is assumed that the fundamental frequency changes slowly over time in voiced duration. The STFTPI defines two simple algorithms along time and frequency as follows:

$$\hat{\phi}^S_{\tilde{\kappa}[h,\tau],\tau} = \hat{\phi}^S_{\tilde{\kappa}[h,\tau],\tau-1} + \Omega^k_{h,\tau} L, \tag{1}$$

$$\hat{\phi}^S_{\tilde{\kappa}[h,\tau]+\delta,\tau} = \hat{\phi}^S_{\tilde{\kappa}[h,\tau],\tau} - \phi^W_{\tilde{\kappa}[h,\tau]-\kappa[h,\tau]} + \phi^W_{\tilde{\kappa}[h,\tau]-\kappa[h,\tau]+\delta}, \tag{2}$$

where L is a frame shift, ϕ^W_ω is the phase property of a window function in frequency ω with discrete-time Fourier transform, $\Omega^k_{h,\tau}$ is a normalized angular frequency, $\kappa[h,\tau]$ is a non-integral value in the frequency bin scale with respect to the h-th harmonic frequency at time frame τ, and $\tilde{\kappa}[h,\tau]$ is a frequency bin with respect to the h-th harmonic frequency at time frame τ. These symbols are defined in detail in Sec. 3. Equation (1) shows the relationship between phase spectra at time frames τ and $\tau-1$, while Eq. (2) represents phase estimation adjacent to the harmonic component based on window-phase compensation with integer $\delta \in \left[-\kappa[0,\tau]/2,\ \kappa[0,\tau]/2\right]$.

Equations (1) and (2) improve the PESQ, while they reconstruct an artificial phase, which causes buzzyness at higher harmonics. Moreover, the estimation accuracy depends on the initial value in Eq. (1). The STFTPI method uses the noisy phase at the onset of a voiced duration as the initial estimate. However, the signal-to-noise ratio (SNR) in the harmonic component, especially higher harmonics, may be low and unreliable. Furthermore, the phase spectra in each harmonic component are independently estimated. These processes may also cause buzzyness.

3. SIGNAL MODEL AND NOTATIONS

We assume that a speech signal $s(n)$ is corrupted by an additive noise $d(n)$ with time sample index n as $x(n) = s(n) + d(n)$. The noisy speech spectrum $X_{k,\tau}$ is given by segmentation, windowing, and DFT of the noisy speech $x(n)$. The $X_{k,\tau}$ is represented as

$$|X_{k,\tau}|e^{j\phi^X_{k,\tau}} = |S_{k,\tau}|e^{j\phi^S_{k,\tau}} + |D_{k,\tau}|e^{j\phi^D_{k,\tau}}, \tag{3}$$

where $k = 0, \cdots, K-1$ and τ are the frequency bin and time frame, respectively. Here, $|X_{k,\tau}|$, $|S_{k,\tau}|$, and $|D_{k,\tau}|$ are the amplitude spectrum of the noisy speech signal, clean speech signal, and additive noise signal, respectively. The $\phi^X_{k,\tau}$, $\phi^S_{k,\tau}$, and $\phi^D_{k,\tau}$ are the phase spectra of these signals.

We define the symbols for the harmonic component ($h = 0, \cdots, H-1$) as follows:

$$\tilde{\kappa}[h,\tau] = \underset{k}{\operatorname{argmin}} \left| k - \kappa[h,\tau] \right|, \tag{4}$$

$$\kappa[h,\tau] = \frac{K}{2\pi}\Omega^k_{h,\tau}, \tag{5}$$

$$\Omega^k_{h,\tau} = \underset{\Omega_{h,\tau}}{\operatorname{argmin}} \left| \Omega_k - \Omega_{h,\tau} \right|, \tag{6}$$

where H and K are the harmonic number and DFT point, respectively. The $\Omega_k = 2\pi k/K$ and $\Omega_{h,\tau} = 2\pi f_{h,\tau}/F_s$ are normalized angular frequencies. The $f_{h,\tau} = (h+1)f_{0,\tau}$ and $f_{0,\tau}$ are the h-th harmonic frequency and fundamental frequency at τ, respectively, F_s is the sampling frequency, and $\hat{\cdot}$ represents the estimation of symbol $\cdot$.

4. PROPOSED PHASE-RECONSTRUCTION METHOD

We propose a phase-reconstruction method based on the phase behavior in the time-frequency domain. Figure 2 shows the conceptual diagram of the proposed method. The phase spectra in the harmonic components are estimated from the relationship of harmonics and temporal behavior. First, we assume that the phase spectrum of the speech is modeled as a summation of a minimum-phase term, linear-phase term, and source-shape term as Degottex and Erro did in their study [26]. Then, the phase distortion (PD) feature in the h-th harmonic component $\Pi_{\tilde{\kappa}[h,\tau],\tau}$ is defined as

$$\Pi_{\tilde{\kappa}[h,\tau],\tau} = \phi^S_{\tilde{\kappa}[h,\tau],\tau} - \phi^S_{\tilde{\kappa}[h-1,\tau],\tau} - \phi^S_{\tilde{\kappa}[0,\tau],\tau}. \tag{7}$$

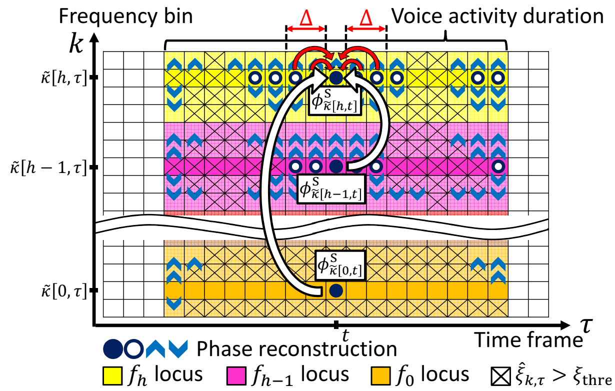


Fig. 2. Conceptual diagram of proposed phase-reconstruction method.

The PD feature represents the relationship between the phase spectra in the h-th and $(h-1)$-th harmonic components through the phase spectrum in the 0-th harmonic component, i.e., fundamental frequency. Degottex and Erro discussed an important temporal behavior of the PD feature [26], which is the PD feature's temporal constancy $\Pi_{\tilde{\kappa}[h,\tau],\tau} = \Pi_h$ in voiced duration. From this constancy and Eq. (7), the phase spectrum in the h-th harmonic component is derived as

$$\phi^S_{\tilde{\kappa}[h,\tau],\tau} = \Pi_h + \phi^S_{\tilde{\kappa}[h-1,\tau],\tau} + \phi^S_{\tilde{\kappa}[0,\tau],\tau}. \tag{8}$$

Equation (8) represents not only the relationship between harmonic components but the temporal behaviors of the phase spectra. In addition, we assume that the local SNR in the 0-th and 1-th harmonic components is high. In other words, the noisy phase spectra are close to the phase spectra of the clean speech in these components. This assumption and Eq. (8) derive the following phase reconstruction:

$$\hat{\phi}^S_{\tilde{\kappa}[h,\tau],\tau} = \begin{cases} \phi^X_{\tilde{\kappa}[h,\tau],\tau}, & \hat{\xi}_{\tilde{\kappa}[h,\tau],\tau} > \xi_{\mathrm{thre}}, \\ \tilde{\Pi}_{\tilde{\kappa}[h,\tau],\tau} + \phi^X_{\tilde{\kappa}[h-1,\tau],\tau} + \phi^X_{\tilde{\kappa}[0,\tau],\tau}, & \text{otherwise}, \end{cases} \tag{9}$$

$$\tilde{\Pi}_{\tilde{\kappa}[h,\tau],\tau} = \angle \exp\left(j \sum_{t=\tau-\Delta}^{\tau+\Delta} \Pi_{\tilde{\kappa}[h,\tau],t} \right), \tag{10}$$

where $\hat{\xi}_{\tilde{\kappa}[h,\tau],\tau}$ is the estimate of the priori SNR at τ and frequency bin $\tilde{\kappa}[h,\tau]$. The ξ_{thre}, Δ, and j are the threshold value of the priori SNR, parameter for temporal averaging, and an imaginary unit, respectively. These equations show that the phase spectrum at τ and $\tilde{\kappa}[h,\tau]$ is estimated from the phase spectra at adjacent frames and at lower harmonic components. In addition, the phase spectrum between harmonics is estimated using Eq. (2) as follows:

$$\hat{\phi}^S_{\tilde{\kappa}[h,\tau]+\delta,\tau} = \begin{cases} \phi^X_{\tilde{\kappa}[h,\tau],\tau}, & \hat{\xi}_{\tilde{\kappa}[h,\tau],\tau} > \xi_{\mathrm{thre}}, \\ \hat{\phi}^S_{\tilde{\kappa}[h,\tau],\tau} - \phi^W_{\rho} + \phi^W_{\rho+\delta}, & \text{otherwise}, \end{cases} \tag{11}$$

$$\rho = \tilde{\kappa}[h,\tau] - \kappa[h,\tau]. \tag{12}$$

Finally, the enhanced speech spectrum is reconstructed by the estimated amplitude and phase as $\hat{S}_{k,\tau} = |\hat{S}_{k,\tau}| e^{j\hat{\phi}^S_{k,\tau}}$ $(k = 0, \cdots, K-1)$, then the IDFT, synthesis windowing, and overlap-add are conducted to transform enhanced speech in the time domain.

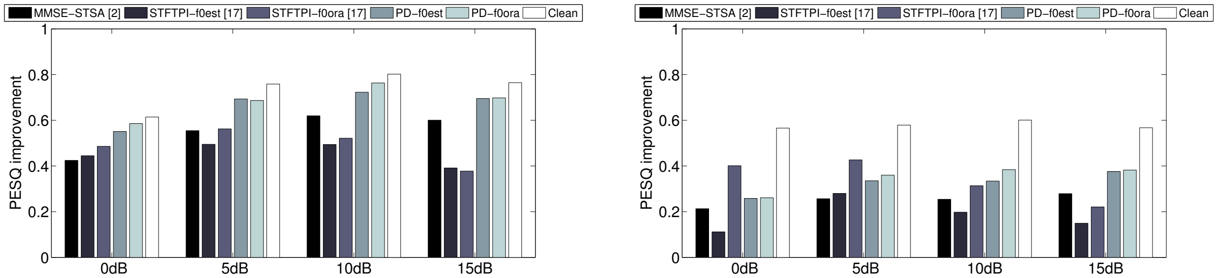
5. EXPERIMENTAL EVALUATIONS

5.1. Experimental setup

We conducted experiments to evaluate the proposed phase-reconstruction method compared with a baseline method (STFTPI). We randomly selected 20 utterances consisting 10 male and 10 female speakers from ATR phoneme-balance 216 words [27] as the clean speech samples. The sampling rate during the experiments was 16 kHz, and the frame length, frame shift, and DFT points were 32 ms, 4 ms, and 512 points, respectively. Noisy speeches were generated by mixing the clean speech samples and additive noise signals at SNRs ranging from 0 to 15 dB. White and babble noise were selected from NOISEX-92 [28] as stationary and non-stationary noise signals, respectively. We used the Hamming window function for amplitude spectrum analysis and the Blackman window function for phase analysis and spectral synthesis, as shown in Fig. 1. The Blackman window has a large dynamic range and is suitable for phase-spectrum analysis.

We used the MMSE-STSA estimator [2] for amplitude enhancement and the STFTPI [17] and proposed method for phase reconstruction. Moreover, the noisy (unprocessed) and clean phases were applied to compare with conventional speech enhancement and confirm the upper-bound performance, respectively. The priori SNR was estimated using the directed-decision method [2] with the MMSE-STSA gain, and the noise power estimate was calculated using the average power spectrum of the noisy speech in the first five unvoiced intervals. The parameters ξ_{thre} in Eqs. (9) and (11) and Δ in Eq. (10) were 5 and 1, respectively. These values were experimentally determined.

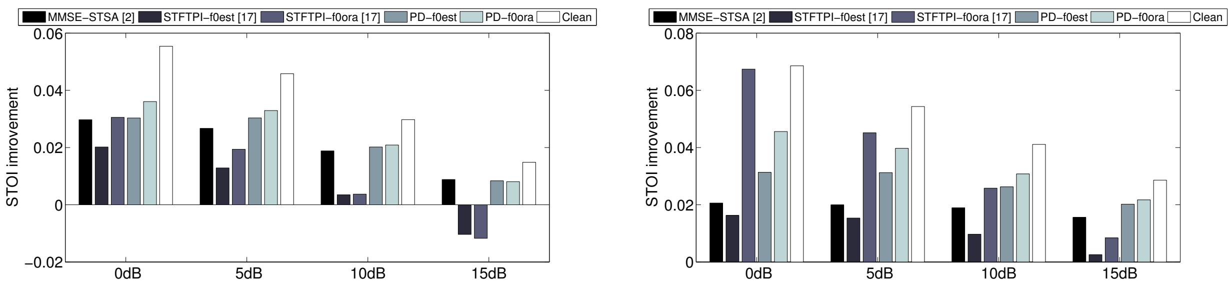
To estimate fundamental frequency f_0 and voice activity duration, we used the pitch estimation filter with amplitude compression (PEFAC) [29] and a simple method based on signal power thresholding, respectively. The number of harmonic components H is defined as $\lfloor 4000/f_{0,\tau} \rfloor$, as defined by Gerkmann et al. [17] and phase reconstructions were conducted up to 4 kHz in voiced duration. Here, $\lfloor \cdot \rfloor$ represents the flooring operation. To evaluate the performance of speech quality and intelligibility, we used the PESQ and short



(a) white noise

(b) babble noise

Fig. 3. PESQ improvement with various methods at various input SNRs.



(a) white noise

(b) babble noise

Fig. 4. STOI improvement with various methods at various input SNRs.

time objective intelligibility measure (STOI) [30]. The higher these indexes, the higher speech quality and intelligibility are.

5.2. Results

Figure 3 shows the PESQ improvement with various methods at various input SNRs. The PESQ improvement is defined as the improvement relative to the noisy speech signal (the unprocessed amplitude and phase spectrum). As shown in Fig. 3 (a), the proposed method (PD-f0ora, PD-f0est) outperformed the baseline method (STFTPI-f0ora, STFTPI-f0est) for white noise. These results suggest that the temporally averaged PD feature from Eqs. (9) and (10) was effective compared with the phase replacement from Eq. (1). The constraint of phase reconstruction by using the priori SNR was also effective. Figure 3 (a) also shows that the two methods depended on the estimation accuracy of f_0, and the phase reconstructions with the estimated f_0 degraded speech quality. However, the proposed method with the estimated f_0 (PD-f0est) was effective compared to the STFTPI with the oracle f_0 (STFTPI-f0ora). In contrast, the proposed method was not effective at low SNRs (0 and 5 dB), as shown in Fig. 3 (b). This is because the assumption described in Sec. 4 was not met in the babble-noise environment, and this non-stationarity degraded the estimation accuracy of the noise power and priori SNR.

The STOI improvement with various methods at various input SNRs is shown in Fig. 4. From Fig. 4 (a), the results for white noise indicate that the proposed method outperformed the STFTPI at high SNRs (10 and 15 dB). Figure 4 (b) shows that the STFTPI achieved high intelligibility, while it significantly depended on the f_0 value. By contrast, the proposed method did not largely depend on the f_0 value, and the STOI with this method was stable regardless of the SNR. The dependence was also caused by the difference between the phase replacement from Eq. (1) and phase estimation from Eq. (9).

6. CONCLUSIONS

We proposed a phase-reconstruction method based on the priori SNR and PD feature. The PD feature represents the relationship between the phase spectra in harmonics and temporal behaviors. We used the priori SNR as a reliability index of the phase spectrum. The experimental results indicate that the PESQ with the proposed method outperformed that with the STFTPI for white noise at both high and low SNRs and for babble noise at high SNRs. While the STFTPI achieved high STOI at low SNRs, the proposed method improved STOI at high SNRs. Future work will involve improving the proposed method, introducing a noise estimator, and conducting experiments for various types of noise and objective experiments.

7. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[2] 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.

[3] P.J. Wolfe and S.J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *EURASIP Journal on Applied Signal Process.*, vol. 2003, no. 10, pp. 1043–1051, 2003.

[4] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[5] R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, and K. Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 20, no. 7, pp. 2080–2094, 2012.

[6] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. of IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[7] D.L. Wang and J.S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. on Acoust., Speech and Signal Process.*, vol. 30, no. 4, pp. 679–681, 1982.

[8] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *ELSEVIER, Speech Commu.*, vol. 53, no. 4, pp. 465–494, 2011.

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

[10] M. Kazama, S. Gotoh, M. Tohyama, and T. Houtgast, "On the significance of phase in the short term Fourier spectrum for speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 127, no. 3, pp. 1432–1439, 2010.

[11] T. Gerkmann, M. Krawczyk, and J. Le Roux, "Phase processing for single-channel speech enhancement," *IEEE Signal Process. Magazine*, pp. 55–66, 2015.

[12] M.S.E. Langarani, H. Veisi, and H. Sameti, "The effect of phase information in speech enhancement and speech recognition," in *The 11th. Int. Conf. on Info. Sci., Signal Process. and their applications*, 2012, pp. 1446–1447.

[13] T. Gerkmann, M. Krawczyk, and R. Rehr, "Phase estimation in speech enhancement unimportant,important,or impossible?," in *Proc. IEEE Conv. Elect. Electron. Eng. Israel*, Eilat, Israel, 2012.

[14] D.W. Griffin and J.S. 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.

[15] 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.

[16] A. Sugiyama and R. Miyahara, "Phase randomization - a new paradigm for single-channel signal enhancement," in *Proc IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 7487–7491.

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

[18] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 129–132, 2013.

[19] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4199–4208, 2014.

[20] M. Krawczyk, R. Rehr, and T. Gerkmann, "Phase-sensitive real-time capable speech enhancement under voiced-unvoiced uncertainty," in *Proc. EURASIP Eur. Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, 2013.

[21] M. Krawczyk and T. Gerkmann, "STFT phase improvement for single channel speech enhancement," in *Proc. Int. Workshop Acoust. Echo, Noise Control (IWAENC)*, Aachen, Germany, 2012.

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

[23] J. Kulmer, P. Mowlaee, and M. Watanabe, "A probabilistic approach for phase estimation in single-channel speech enhancement using von Mises phase priors," in *Proc IEEE Workshop Mach. Learn. Signal Process.*, 2014, pp. 1–6.

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

[25] P. Mowlaee and R. Saeidi, "Time-frequency constraint for phase estimation in single-channel speech enhancement," in *Proc. Int. Workshop Acoust. Signal Enhance. (IWAENC)*, 2014, pp. 338–342.

[26] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP J. Audio, Speech, Music Process.*, vol. 2014:38, no. 1, Oct. 2014.

[27] K. Takeda, Y. Sagisaka, and S. Katagiri, "Acoustic-phonetic labels in a Japanese speech database," *European Conference on Speech Technology*, vol. 2, pp. 13–16, 1987.

[28] 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," Tech. Rep., DRA Speech Res. Unit, 1992.

[29] S. Gonzalez and M. Brookes, "PEFAC - a pitch estimation algorithm robust to high levels of noise," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 22, no. 2, pp. 518–530, 2014.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.