# Maximum a posteriori estimation of spectral gain with harmonic-structure-based phase reconstruction for phase-aware speech enhancement

Yukoh Wakabayashi and Nobutaka Ono
Tokyo Metropolitan University, Tokyo, Japan

***Abstract*—Contrary to a long-held belief, recent work shows that phase spectrum modification plays an important role in speech quality and intelligibility. In this paper, we propose a phase-aware speech enhancement method that consists of our previous phase reconstruction method and a spectral gain estimation method using the phase estimate with a maximum a posteriori criterion. The spectral gain estimation based on a phase estimate improves the estimation accuracy of the target clean speech signal. Experimental results demonstrate that the proposed method improves objective measures for speech perceived quality.**

## I. Introduction

Speech enhancement is one of the most essential techniques in signal processing. In noisy environments, speech corrupted by ambient noise disturbs smooth human-to-human communication and degrades the performance of voice-based applications. Speech enhancement estimates clean speech in noisy environments and resolves these problems. Many speech enhancement methods such as spectral subtraction by Boll [1] and Wiener filtering [2] have been investigated. Spectral gain estimation in the time-frequency domain also has been addressed [3]–[6]. In particular, the minimum mean-square error (MMSE) based method [3] and the maximum a posteriori (MAP) based method [4] are of practical use. Speech enhancement techniques have been generally used to modify the amplitude spectra only, and phase spectra have not been regarded as important features.

Recent study focuses on phase processing after previous research which shows the dependence of speech quality and intelligibility on phase spectrum [7], [8] were presented. Phase reconstruction for speech enhancement includes amplitude-required approaches such as iterative methods [9], [10] and a geometry-based method [11], model-based approaches such as a randomization method [12] and fundamental frequency based methods [13]–[16]. The latter method reported that fundamental frequency based method improves speech quality without requiring a clean amplitude spectrum.

Our previous work [16] revealed that speech enhancement using the harmonic-structure-based phase reconstruction method together with the MMSE amplitude estimator [3] improves an objective perceived speech quality PESQ [17] and a subjective evaluation score more than the baseline method [14] in various noisy environments. This method uses averaged values of the phase distortion (PD) feature [18], which shows harmonic speech phase fluctuations in the time-frequency domain, and the window-phase compensation [14]. Using the PD feature for phase reconstruction estimates harmonic speech phases and the window-phase compensation estimates non-harmonic speech phases.

In this paper, we consider an amplitude estimation method given a speech phase estimate, i.e., phase-aware speech enhancement, to further improve speech quality. Our previous study independently estimates the speech amplitude and phase from noisy observation, while this study estimates the amplitude with using the estimated phase information. Phase-aware speech enhancement techniques were recently advocated in [19], [20]. The authors in [19] derive various MMSE estimators from some cost functions defined by the authors. The MAP estimator [20] is developed from a prior amplitude distribution with phase spectra. Both approaches estimate complex spectra by assuming that phase spectra follows a *von Mises* distribution around an initial phase estimate. In this study, we also focus on the MAP estimator and formulate a spectral gain estimation, not handling the phase estimate as a stochastic value in [20], but as a deterministic one. As a result, this phase-aware MAP (PAMAP) estimator is the same as the estimator of [20].

The remainder of this paper is organized as follows. In Section II, we explain the signal model and a notation. We describe the conventional MAP-based spectral gain estimation in Section III. Section IV introduces the proposed method: brief review of our previous phase reconstruction method for speech enhancement and the PAMAP gain estimator. We evaluate the proposed method through simulation experiments in Section V. Finally, we draw conclusions in Section VI.

## II. Notation

We assume that a speech signal $s(n)$ is corrupted by an additive noise $d(n)$ with time index $n$ as $x(n) = s(n) + d(n)$. The spectral representation is obtained by frame-by-frame processing using windowing and the short-time Fourier transform (STFT) as follows:

$$|X_{k,\tau}|e^{j\phi^X_{k,\tau}} \quad = \quad |S_{k,\tau}|e^{j\phi^S_{k,\tau}} + |D_{k,\tau}|e^{j\phi^D_{k,\tau}}, \qquad (1)$$

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

$\phi^D_{k,\tau}$. The enhanced speech spectrum is reconstructed using the estimated amplitude and phase as $\hat{S}_{k,\tau} = |\hat{S}_{k,\tau}|e^{j\hat{\phi}^S_{k,\tau}}$, where $\hat{A}$ represents the estimation of symbol $A$.

In addition, we define the symbols for the harmonic component ($h = 0, \cdots, H-1$) as $k_h = \underset{k}{\operatorname{argmin}} |k - \kappa_h|$, $\kappa_h = f_{h,\tau}K/F_s$, where $k_h$ is a frequency bin index with respect to the $h$-th harmonic frequency at frame $\tau$, $\kappa_h$ is a non-integral value in the frequency bin scale. We omit the time frame index $\tau$ from these definitions to simplify the notation. $H$ and $F_s$ are the harmonic number and sampling frequency, respectively. the $h$-th harmonic frequency and fundamental frequency at $\tau$ are $f_{h,\tau} = (h+1)f_{0,\tau}$ and $f_{0,\tau}$, respectively.

## III. Conventional MAP-based gain estimation

Spectral gain estimation with the MAP criterion was introduced by Lotter and Vary [4]. They assume that real and imaginary parts of the speech complex spectrum are independent and Gaussian distributed. They also assume that phase spectra are uniformly distributed and amplitude spectra $A$ have the following prior distribution:

$$p(A) = \frac{\mu^{\nu+1}}{\Gamma(\nu+1)} \frac{A^\nu}{\sigma_s^{\nu+1}} \exp\left\{-\frac{\mu A}{\sigma_s}\right\}, \tag{2}$$

where $\sigma_s^2$ is the standard deviation of speech, $\Gamma(\cdot)$ is the gamma function, $\nu$ and $\mu$ are shape parameters. Under these assumptions, the maximization of a posterior probabilities $p(A|X)$ and $p(\phi^S|X)$ gives the amplitude and phase estimators. Then, the phase estimate is the unprocessed phase, and the amplitude estimate is represented by the following spectral gain:

$$G_{k,\tau} = u + \sqrt{u^2 + \frac{\nu}{2\gamma_{k,\tau}}}, \tag{3}$$

$$u = \frac{1}{2} - \frac{\mu}{4\sqrt{\xi_{k,\tau}\gamma_{k,\tau}}}, \tag{4}$$

where $\xi_{k,\tau}$ and $\gamma_{k,\tau}$ respectively indicate the a priori and a posteriori SNRs.

## IV. Phase-aware speech enhancement

In this section, we describe the proposed speech enhancement technique. This technique, as shown in Fig. 1, consists of five blocks: noise power estimator, calculation of a posteriori SNR, a priori SNR estimator, phase estimator, and spectral gain estimator. A spectral gain is generally estimated from the noise estimator and the a priori and a posteriori SNR estimators. In the proposed scheme, the spectral gain is estimated from these three estimates and an additional phase estimate. The phase estimate is calculated from the fundamental frequency $f_{o,\tau}$ referring to the a priori SNR, as described in [16]. In the two following sub-sections, we briefly review the phase reconstruction method and explain the phase-aware spectral gain estimation.

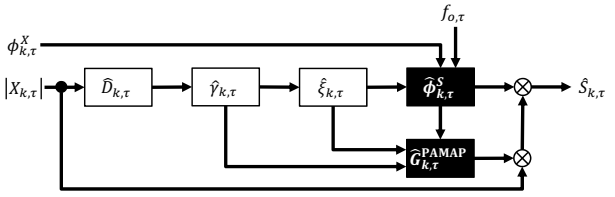



Fig. 1. Block diagram of phase-aware speech enhancement.

### A. Harmonic-structure-based phase reconstruction

We reported a phase reconstruction method based on speech harmonic structure for speech enhancement [16]. With this method, harmonic phase spectra are estimated by using the time-frequency fluctuations through the temporally-averaged PD feature. Non-harmonic phase spectra are estimated with the window-phase compensation with integer $\delta \in \left[-\kappa_0/2,\ \kappa_0/2\right]$ in voiced interval. The estimations are formulated as

$$\hat{\phi}^S_{k_h,\tau} = \begin{cases} \phi^X_{k_h,\tau}, & \hat{\xi}_{k_h,\tau} > \xi_{\text{thre}}, \\ \tilde{\Phi}_{h,\tau} + \phi^X_{k_{h-1},\tau} + \phi^X_{k_0,\tau}, & \text{otherwise}, \end{cases} \tag{5}$$

$$\tilde{\Phi}_{h,\tau} = \angle \exp\left(j \sum_{t=\tau-\Delta}^{\tau+\Delta} \Phi_{k_h,t}\right), \tag{6}$$

$$\Phi_{k_h,\tau} = \phi^X_{k_h,\tau} - \phi^X_{k_{h-1},\tau} - \phi^X_{k_0,\tau}, \tag{7}$$

$$\hat{\phi}^S_{k_h+\delta,\tau} = \begin{cases} \phi^X_{k_h+\delta,\tau}, & \hat{\xi}_{k_h+\delta,\tau} > \xi_{\text{thre}}, \\ \hat{\phi}^S_{k_h,\tau} - \phi^W_{k_h-\kappa_h} + \phi^W_{k_h-\kappa_h+\delta}, & \text{otherwise}, \end{cases} \tag{8}$$

where the $\xi_{\text{thre}}$, $\Delta$, $j$, and $\angle z$ terms are the threshold value of the a priori SNR, the parameter for temporal averaging, an imaginary unit $\sqrt{-1}$, and the phase of $z \in \mathbb{C}$, respectively. The PD feature is defined as (7), and the harmonic and non-harmonic phase estimations are respectively shown in (5)–(7) and (8). The phase property $\phi^W_\omega$ of the window function in continuous-valued frequency $\omega$ is calculated via discrete-time Fourier transform (see (18) of [16]). The a priori SNR estimate $\hat{\xi}_{k_h,\tau}$ is used as a reliability criterion for the phase spectrum at $\tau$ and $k_h$.

### B. PAMAP-based spectral gain estimation

We consider the maximization of a posterior probability given an observed speech $X$ and a phase estimate $\hat{\phi}^S$. Then, using Bayes rule, we obtain

$$\begin{aligned} |\hat{S}|^{\mathrm{PAMAP}} &= \underset{A}{\operatorname{argmax}}\, \mathcal{J} = \underset{A}{\operatorname{argmax}}\, p(A|X, \hat{\phi}^S) \\ &= \underset{A}{\operatorname{argmax}}\, p(X|A, \hat{\phi}^S) p(A). \end{aligned} \tag{9}$$

By substituting the prior (2) into (9) under the Gaussian assumption of the noise, we have

$$\mathcal{J} \propto A^\nu \exp\left\{-\frac{\mu A}{\sigma_s^2} - \frac{|X - Ae^{j\hat{\phi}^S}|^2}{\sigma_d^2}\right\}, \tag{10}$$

where $\sigma_d^2$ is the standard deviation of noise. By differentiating the logarithm of (10) and setting it to zero, we obtain the

speech amplitude spectrum using the gain estimator as

$$|\hat{S}_{k,\tau}^{\mathrm{PAMAP}}| = G_{k,\tau}^{\mathrm{PAMAP}}|X_{k,\tau}|, \quad (11)$$

$$G_{k,\tau}^{\mathrm{PAMAP}} = u + \sqrt{u^2 + \frac{\nu}{2\gamma_{k,\tau}}}, \quad (12)$$

$$u = \frac{1}{2}\cos(\phi_{k,\tau}^{X} - \hat{\phi}_{k,\tau}^{S}) - \frac{\mu}{4\sqrt{\xi_{k,\tau}\gamma_{k,\tau}}}. \quad (13)$$

Despite a different derivation, the resulting estimator is identical to the estimator in [20].

## V. Experimental evaluation

### A. Setup

We conducted objective evaluation experiments to evaluate the proposed phase-aware speech enhancement method in comparison with the conventional spectral gain estimators [3], [4]. We randomly selected 20 utterances consisting of 10 male and 10 female speakers from the ATR 216 phoneme balanced words [21] 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 speech was generated by mixing the clean speech samples and additive noise signals at SNRs ranging from 0 to 15 dB. White and babble noises were selected from NOISEX-92 [22]. We used the Hann window for amplitude spectrum analysis and the Blackman window for phase analysis. It was reported that this approach improves PESQ compared with using only one window [7]. The Blackman window has a large dynamic range and is suitable for phase spectrum analysis [23]. The noisy phase (unprocessed) and clean phase are used as a baseline in order to compare them with conventional speech enhancement and to confirm the upper bound performance, respectively.

The a priori SNR $\hat{\xi}_{k,\tau}$ was estimated using the decision-directed method [3] with a smoothing factor of $0.98$, and the noise-power estimate was calculated using the weighted noise estimator [24]. The parameters $\xi_{\mathrm{thre}}$ and $\Delta$ in (5)–(8) were both set to 1. The pitch estimation filter with amplitude compression (PEFAC) [25] was used to estimate the fundamental frequency $f_{0,\tau}$ as in [13], [14]. We used a voiced probability estimated by the PEFAC to estimate voiced intervals with threshold probability of 0.5. The number of harmonic components $H$ is defined as the largest integer less than $4000/f_{0,\tau}$, and the phase spectra are reconstructed up to 4 kHz in voiced intervals. To evaluate the performance of speech quality, we used the PESQ and the segmental SNR in objective evaluations. The higher these metrics, the higher speech quality and intelligibility are. We also used the UnRMSE [26] to evaluate phase estimation accuracy. The lower the UnRMSE, the higher the phase estimation accuracy.

### B. Experimental results

Fig. 2 illustrates the PESQ, segmental SNR (SegSNR), and UnRMSE improvements in the white and babble noise environments. The figure shows the improvement relative to the noisy speech signal (the unprocessed amplitude and phase spectra). Here, (+n), (+e), and (+c) indicate that the noisy, estimated, and clean phases are used as the phase value for signal reconstruction, respectively. Figs. 2(a) and (d) show that MAP+n and MAP+e, and our previous method (MMSE+e) improve PESQ most in the white and babble noise environments, respectively. PAMAP+e improves PESQ the second most in both noise environments; PAMAP+e improves PESQ compared with MMSE+e in the white noise environment, while MAP+e slightly improves it more than PAMAP+e. The PESQ result in the babble noise environment has the opposite trend, namely, PAMAP+e improves this metric over MAP+e and degrades it compared with MMSE+e. In contrast, as shown in Figs. 2(b) and (e), the SegSNR is slightly improved by PAMAP+e more than other cases, except for PAMAP+c, in both environments. These results show that PAMAP can reconstruct the speech signal itself, while this method does not always improve its perceived quality in respect of PESQ. These results are caused by the phase estimation accuracy, which is shown by the following observations: the results of PESQ and SegSNR indicate that PAMAP+c outperforms other methods in any other environments, and the UnRMSE difference between the clean and estimated phase cases is very large, as shown in Figs. 2(c) and (f). These results imply that PAMAP improves perceived quality measures compared with our previous work in many cases, but the phase estimation accuracy does not improve much over MAP. It is necessary to modify the phase estimation to improve estimation accuracy.

## VI. Conclusion

We proposed a phase-aware speech enhancement method based on the combination of the PAMAP gain estimator and the PD-based method to further improve speech quality over our previous method. The experimental results indicate that the proposed method improves segmental SNR and PESQ more than our previous work, namely, the MMSE gain estimator and the phase reconstruction method. However, the proposed method does not improve PESQ compared with the MAP gain estimator. Our future work will involve improving the phase reconstruction method and considering an integrative complex gain estimation approach between amplitude and phase estimations to further improve speech quality.

## Acknowledgment

This work was supported by the SECOM Science and Technology Foundation and JSPS KAKENHI Grant Number 18H06482.

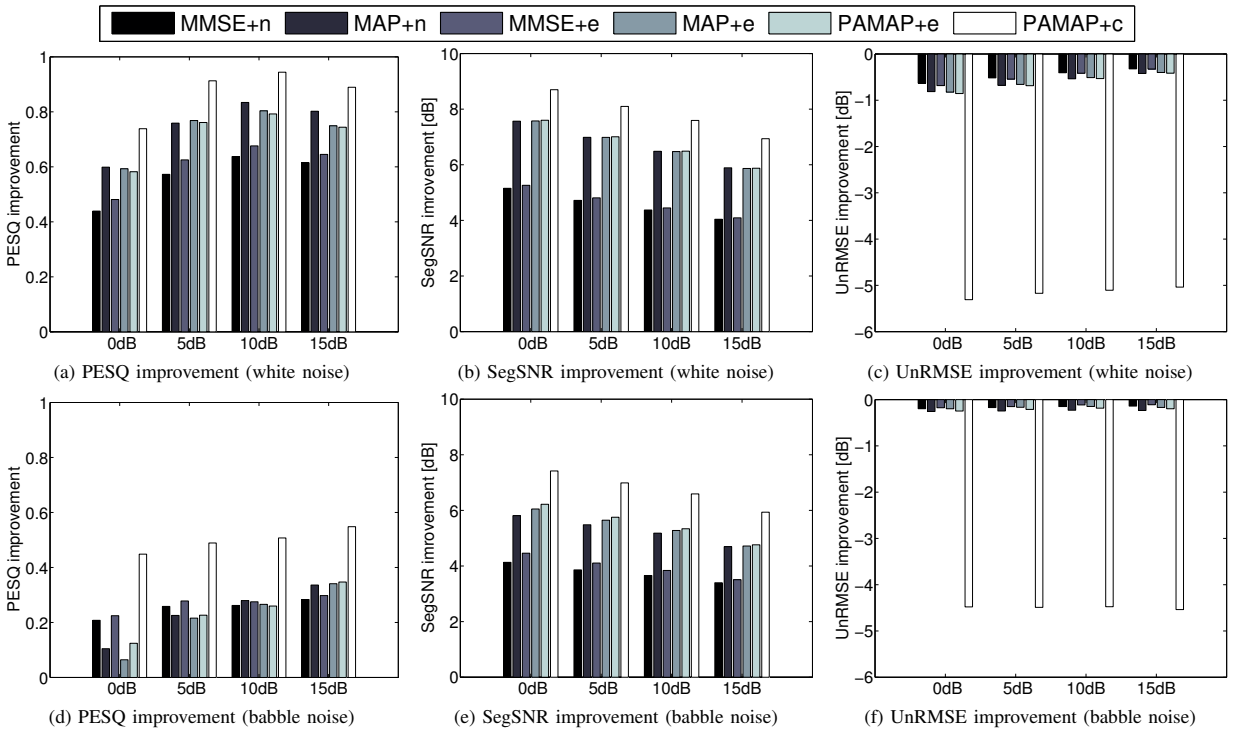
## References

[1] S. 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] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, Inc., 2007.

[3] 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.

[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.




(a) PESQ improvement (white noise) (b) SegSNR improvement (white noise) (c) UnRMSE improvement (white noise)

(d) PESQ improvement (babble noise) (e) SegSNR improvement (babble noise) (f) UnRMSE improvement (babble noise)

Fig. 2. PESQ, SegSNR, and UnRMSE improvements relative to noisy speech, with the MMSE, MAP, and phase-aware MAP (PAMAP) estimators for amplitude modification and with (+e) the proposed estimator for phase reconstruction, (+n) an unprocessed phase, and (+c) an oracle phase at various input SNRs.


[5] L. Lightburn, E. D. Sena, A. Moore, P. A. Naylor, and M. Brookes, "Improving the perceptual quality of ideal binary masked speech," in *Proc IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 661–665.

[6] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[7] 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.

[8] 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.

[9] 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.

[10] J. L. 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.*, 2017, pp. 5560–5565.

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

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

[18] 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, pp. 1–16, Oct. 2014.

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

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

[21] 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.

[22] 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.

[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] M. Kato, A. Sugiyama, and M. Serizawa, "Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 89, no. 2, pp. 43–53, 2006.

[25] 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.

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