HARMONIC PHASE ESTIMATION IN SINGLE-CHANNEL SPEECH ENHANCEMENT USING VON MISES DISTRIBUTION AND PRIOR SNR

Josef Kulmer and Pejman Mowlaee

Signal Processing and Speech Communication Lab
Graz University of Technology, Graz, Austria
josef.kulmer@tugraz.at pejman.mowlaee@tugraz.at

ABSTRACT

In single-channel speech enhancement the spectral amplitude of the noisy signal is often modified while the noisy spectral phase is directly employed for signal reconstruction. Recently, additional improvement in speech enhancement performance has been reported when the noisy phase is modified. In this work, we propose a Bayesian estimator for phase of harmonics given the noisy speech. The proposed estimator relies on the fundamental frequency and the signal-to-noise ratio at harmonics. Throughout our experiments, we evaluate the performance of the proposed phase enhancement in comparison with the noisy phase, a benchmark and the clean phase as the upper-bound. The proposed method leads to joint improvement in quality and intelligibility at different SNRs and noise types.

***Index Terms*—** Phase estimation, Bayesian estimation, harmonic model, speech enhancement, Von Mises distribution.

1. INTRODUCTION

Many previous studies in single-channel speech enhancement have been concentrated on deriving estimators for spectral amplitude while they suggest to copy the noisy phase in signal reconstruction (see e.g. [1] for a detailed review). The importance of phase for an improved speech enhancement performance has been demonstrated [2–12] . These improvements require a reliable clean phase estimate from the noisy speech. Therefore, finding an improved estimate for the spectral phase at harmonics is of great importance.

In this paper, we propose a novel Bayesian phase estimator assuming a von Mises distribution on phase. The proposed estimator relies on the harmonic model and depends on the fundamental frequency and signal-to-noise ratio at harmonics. We quantify the effectiveness of the proposed phase estimator in single-channel speech enhancement via comparing its performance versus the noisy phase (lower-bound), clean phase (upper-bound) and a benchmark [7].

The paper is arranged as follows: section 2 presents an overview on previous phase and MAP estimators, section 3 presents the proposed phase estimator, section 4 presents results and section 5 concludes on the work.

The work was partially funded by the K-Project ASD in the context of COMET Competence Centers for Excellent Technologies by BMVIT, BMWFJ, Styrian Business Promotion Agency (SFG), Province of Styria Government of Styria and the Technology Agency of the City of Vienna (ZIT). The programme COMET is conducted by Austrian Research Promotion Agency (FFG).

2. RELATION TO PREVIOUS WORKS

2.1. Previous Phase Estimation Methods

The problem of phase estimation dates back to 1980's when several attempts were made to estimate a time-domain signal from a modified STFT amplitude [13, 14]. Griffin and Lim (GL) first proposed an iterative method to reconstruct a time domain signal from a given modified STFT amplitude in a minimum mean square error sense [14]. In [15], a partial phase reconstruction was proposed where the GL solution was confined to a limited set of signal components resulting in an improved performance. In [5], additional constraint on the consistency between the STFT magnitude and phase was suggested leading to consistent Wiener filter. The method required the ad-hoc parameter adjustments. The aforementioned GL-based methods, are limited in performance due to their dependency on accurate estimates of the source amplitude spectra, thus, their performance is limited. Furthermore, many iterations are required to reach a reasonable signal reconstruction quality [12].

More recently, in [8] a phase estimation method was proposed relying on the geometry. The group delay deviation constraint on the spectral phase was employed in order to remove the ambiguity in the phase candidates. In [4], the geometric approach was extended to other time-frequency constraints defined on the spectral phase which was shown useful to derive an iterative closed-loop phase-aware speech enhancement [6]. In [16], a phase enhancement strategy based on randomization of the spectral phase was proposed. Successful reduction of musical noise was demonstrated in the auto focusing noise suppression application. The method required ad-hoc adjustments of the key parameters. The authors in [7] proposed to modify the phase values by incorporating the temporal constraints on the phase at voiced frames and compensating the phase response of the window across frequency. The method requires an accurate fundamental frequency estimate and a quite reliable voice activity detection, and its performance depends on the previous time-frequency phase estimates. Improvement in the perceived quality was reported for low signal-to-noise ratio obtained at the expense of a buzzy speech quality in particular at higher harmonics [3, 17]. A similar idea in [17] was used to estimate the noise in between harmonics and reported improved noise reduction at voiced frames, at the expense of undesirable artifacts and buzzyness in the reconstructed speech [17].

2.2. Previous MAP Estimators

In the following, we list the previous MAP estimators for speech enhancement; In [18,19] joint-MAP (JMAP) short-time spectral amplitude (STSA) and phase estimator conditioned on the observed com-

plex STFT coefficient was derived and it was shown that the noisy phase is the MAP phase estimate (similar conclusion was drawn in the JMAP demonstrated in [20]). Finally in [21], under a uniform phase prior distribution, the MAP estimator in log-spectral domain was derived and the MAP phase estimate was shown to be the noisy phase. In all these methods, the simplifying assumptions are twofold: i) the independence assumption in the joint amplitude and phase distribution, and ii) the assumption of a uniform phase prior distribution. In the following, we derive the MAP estimator of harmonic phase assuming a von Mises distribution phase prior.

3. PROPOSED HARMONIC PHASE ESTIMATION

As our signal model, we assume that the observed noisy speech is modeled as sum of harmonics corrupted in additive noise. The noisy signal is segmented into frames denoted by $y(n,l)$ where $n \in [0, N-1]$ and l denotes the frame index with frame length N:

$$y(n,l) = \sum_{h=1}^{H_l} A(h,l)\cos(h\omega_0(l)n + \theta(h,l)) + \nu(n,l), \quad (1)$$

where $y(n,l)$, $\nu(n,l)$ and $\omega_0(l)$ denote the observed noisy speech, the additive noise and the normalized fundamental frequency at frame l, respectively, with H_l as the number of harmonics. Each harmonic $h \in [1, H_l]$ is characterized by the harmonic triple of amplitude $A(h,l)$, frequency $h\omega_0(l)$ and phase $\theta(h,l)$. In this work, we aim at enhancing the phase without modification of the amplitude. In the following we derive the *maximum a posteriori* estimator for harmonic phase denoted by $\hat{\theta}(h,l)$ for the case of one sinusoid in noise and in Section 3.2 extend it to sum of harmonics as in speech.

3.1. MAP Phase Estimator using Von Mises Phase Prior

Consider one harmonic of the clean speech signal as:

$$\bar{y}(n) = A\cos(h\omega_0 n + \theta) + \nu(n), \quad (2)$$

characterized by the sinusoidal triple parameters, i.e. $\{A, h\omega_0, \theta\}$ with defining the observation vector as $\bar{\mathbf{y}} = \{\bar{y}(n)\}_{n=0}^{N-1}$. The MAP solution for $\hat{\theta}$ is obtained by solving the following:

$$\theta_{MAP} = \arg\max_{\theta} \frac{p(\bar{\mathbf{y}}|\theta)p(\theta)}{p(\bar{\mathbf{y}})} = \arg\max_{\theta} p(\bar{\mathbf{y}}|\theta)p(\theta). \quad (3)$$

Given the observed noisy signal and under white Gaussian noise assumption for $\nu(n)$, for the sinusoidal phase values $p(\bar{\mathbf{y}}|\theta)$ is:

$$p(\bar{\mathbf{y}}|\theta) = c_0 \exp\left\{-\frac{1}{\sigma^2}\sum_{n=0}^{N-1}\Big(\bar{y}(n) - A\cos(h\omega_0 n + \theta)\Big)^2\right\}, \quad (4)$$

with σ^2 as the noise variance and $c_0 = (2\pi\sigma^2)^{-\frac{N}{2}}$. The novelty of this work lies in incorporating a more generalized prior distribution known as von Mises than the typically used uniform phase prior assumption for phase. The von Mises distribution (also used in speech analysis/synthesis [22]) is the maximum entropy distribution for a given circular mean (μ_c) and concentration (κ) [23, Section 3.5.4]. Here, in order to take into account the uncertainty in the prior phase information in our proposed MAP phase estimator, we incorporate the von Mises distribution given by:

$$\theta \sim \mathcal{VM}(\mu_c, \kappa) \quad ; \quad p(\theta) = \frac{\exp\left(\kappa\cos(\theta - \mu_c)\right)}{2\pi I_0(\kappa)}, \quad (5)$$

where $I_\nu(\cdot)$ is the modified Bessel function of the first kind of order ν. Plugging (5) and (4) in (3) and discarding the constants, we get:

$$L(\theta) = -\frac{1}{\sigma^2}\sum_{n=0}^{N-1}\big(\bar{y}(n) - A\cos(h\omega_0 n + \theta)\big)^2 + \kappa\cos(\theta - \mu_c). \quad (6)$$

The MAP solution is given by taking the derivative of $L(\theta)$ with respect to θ and setting it equal to zero. Similar to [24, Section 7.5], here we assume that $\sum_{n=0}^{N-1}\sin(h\omega_0 n + \theta)\cos(h\omega_0 n + \theta) \approx 0$ for $h\omega_0$ not close to 0 or π and for N being large enough, we obtain

$$\frac{dL(\theta)}{d\theta} = -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} 2A\bar{y}(n)\sin(h\omega_0 n + \theta) - \kappa\sin(\theta - \mu_c). \quad (7)$$

Using $\sin(a \pm b) = \sin(a)\cos(b) \pm \cos(a)\sin(b)$ we get

$$\begin{aligned}\frac{dL(\theta)}{d\theta} = \cos(\theta)&\left(\frac{2A}{\sigma^2}\sum_{n=0}^{N-1}\big(\bar{y}(n)\sin(h\omega_0 n)\big) - \kappa\sin(\mu_c)\right)\\ &+ \sin(\theta)\left(\frac{2A}{\sigma^2}\sum_{n=0}^{N-1}\big(\bar{y}(n)\cos(h\omega_0 n)\big) + \kappa\cos(\mu_c)\right).\end{aligned} \quad (8)$$

The MAP solution denoted by θ_{MAP} is then given by solving for θ resulting in $\frac{dL(\theta)}{d\theta} = 0$ leading to the following equation:

$$\frac{\sin(\theta_{MAP})}{\cos(\theta_{MAP})} = \frac{-\frac{2A}{\sigma^2}\sum_{n=0}^{N-1}\big(\bar{y}(n)\sin(h\omega_0 n)\big) + \kappa\sin(\mu_c)}{\frac{2A}{\sigma^2}\sum_{n=0}^{N-1}\big(\bar{y}(n)\cos(h\omega_0 n)\big) + \kappa\cos(\mu_c)}, \quad (9)$$

which finally results in the following MAP solution for phase estimate:

$$\boxed{\theta_{MAP} = \tan^{-1}\frac{-\frac{2A}{\sigma^2}\sum_{n=0}^{N-1}\bar{y}(n)\sin(h\omega_0 n) + \kappa\sin(\mu_c)}{\frac{2A}{\sigma^2}\sum_{n=0}^{N-1}\bar{y}(n)\cos(h\omega_0 n) + \kappa\cos(\mu_c)}} \quad (10)$$

The MAP phase estimate is a function of the following parameters: parameters of von Mises prior (μ_c and κ), data length (N) and the local signal-to-noise ratio ($\frac{A}{\sigma^2}$). As an extreme scenario of large SNRs ($A \gg \sigma^2$), the MAP estimator asymptotically degenerates to the ML estimate given by the noisy DFT phase sampled at the harmonic frequency [24, p. 168]. At such high SNR scenario, the noisy phase is more weighted rather than the mean value. This is possible by incorporating a low κ in von Mises prior considered in the proposed estimator. On the other hand, the additional dependency on the harmonic SNR (given by $\frac{A}{\sigma^2}$) serves as a reliability check mechanism about the estimated phase implying that at low SNRs, the proposed MAP estimator relies only on the mean value μ_c. The von Mises distribution phase prior enables to sweep a flexible framework to take into account the uncertainty in phase estimation captured by the mean and variance. A low κ leads to a uniform phase prior while a large concentration $\kappa \to \infty$ resembles a delta Dirac denoting a high certainty in the estimated phase (at high SNRs).

3.2. Extension to Speech Signal

Here we extend the derived MAP phase estimate for one harmonic in noise given in (10) to sum of harmonics as observed in speech signal. The estimator is applied to each frame l and at each harmonic h individually. Therefore, using (10) we obtain the estimator of harmonic phase as in (11).

Let $w(n)$ as the window function and $Y(k,l) = \mathcal{F}\{w(n)y(n,l)\}$

$$\hat{\theta}(h,l) = \tan^{-1} \frac{-\frac{2A(h,l)}{\sigma^2(h,l)} \sum_{n=0}^{N-1} y(n,l) \sin(h\omega_0(l)n) + \kappa(h,l) \sin(\mu_c(h,l))}{\frac{2A(h,l)}{\sigma^2(h,l)} \sum_{n=0}^{N-1} y(n,l) \cos(h\omega_0(l)n) + \kappa(h,l) \cos(\mu_c(h,l))}. \quad (11)$$

as the DFT of the noisy input. We define $|Y(k,l)|$ and $\phi = \angle Y(k,l)$ as the noisy spectral amplitude and phase, respectively with $k \in [0, K-1]$ as the frequency bin and K as the DFT length. In order to synthesize the final phase-enhanced time-domain signal, we need to transform the MAP harmonic phase estimates given in (11) to the STFT domain. This is done by the modification of the frequency bins lying within the main-lobe width of the analysis window. The enhanced STFT phase $\hat{\phi}(k,l)$ is then given by:

$$\hat{\phi}(\lfloor h\omega_0(l)K \rceil + i, l) = \hat{\theta}(h,l), \forall i \in [-N_p(l)/2, N_p(l)/2], \quad (12)$$

where $N_p(l) = \min(N_w, \omega_0(l)K/(2\pi))$ denotes the minimum value of either the main-lobe width of the analysis window N_w or the frequencies close to the neighboring harmonic. The phase values between the harmonics not lying within the analysis window are not modified in order to preserve information at plosives or fricatives frames. The protection of these frames is important in order to preserve a high speech intelligibility performance (see the results in Section 4). Finally, the phase-enhanced time-domain signal is obtained by applying the inverse DFT on

$$\hat{Y}(k,l) = |Y(k,l)| e^{j\hat{\phi}(k,l)}, \quad (13)$$

followed by the overlap-add procedure.

3.3. Von Mises Distribution Parameter Estimation

In this section the estimation of the parameters of the phase distribution is presented. Under the assumption that a reliable fundamental frequency estimate is given, as shown in [25], the spectral phase $\psi(h,l)$ of each harmonic is estimated by a linear interpolation of the spectral phase values $\phi(k,l)$ along the frequency as depicted in Figure 1. Using the phase decomposition principle in [22], the spectral phase $\psi(h,l)$ is decomposed as follows:

$$\psi(h,l) = hS \sum_{l'=0}^{l} \omega_0(l') + \Psi(h,l), \quad (14)$$

where the first term is linear phase characterized by the fundamental frequency $\omega_0(l)$ and the frame shift S. The second term captures the phase contribution by the vocal tract and the glottis source. Similar to [26], we fit a von Mises distribution on $\Psi(h,l)$ to characterize the statistical behavior as the mean and variance given in the following:

$$z(h,l) = \frac{1}{R} \sum_{l'=l-R/2}^{l+R/2} e^{j\Psi(h,l')}, \quad (15)$$

$$\mu_c(h,l) = \angle z(h,l), \quad (16)$$

$$\sigma_c^2(h,l) = 1 - |z(h,l)|, \quad (17)$$

where R denotes the number of frames within 20 ms to capture the short-time stationarity of speech. The mean value after addition of the linear phase denoted by $\mu_c(h,l)$ at each frame is obtained by:

$$\mu_c(h,l) = hS \sum_{l'=0}^{l} \omega_0(l') + \mu(h,l). \quad (18)$$

The concentration $\kappa(h,l)$ is calculated by inversion of the following relation [27] as proposed in [28]:

$$\sigma_c^2(h,l) = 1 - \frac{I_1(\kappa(h,l))}{I_0(\kappa(h,l))}. \quad (19)$$

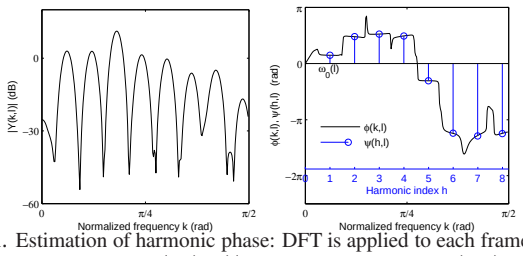

Fig. 1. Estimation of harmonic phase: DFT is applied to each frame $y(n,l)$ (left) spectral amplitude $|Y(k,l)|$ (right) spectral phase $\phi(k,l)$ along frequencies k, and the harmonic phase $\psi(h,l)$ is estimated by linear interpolation of $\phi(k,l)$ dependent on the normalized fundamental frequency $\omega_0(l)$.

Both parameters $\mu_c(h,l)$ and $\kappa(h,l)$ are estimated for voiced and unvoiced regions of speech. In case of unvoiced regions the $\mu_c(h,l)$ has no explanatory power as $\kappa(h,l) \to 0$ indicates a large variance in phase showing a uniform distribution. Thus, at unvoiced regions, the phase estimate in (11) relies more on noisy phase than $\mu_c(h,l)$.

4. RESULTS

4.1. Experiment Setup

As experiment setup, we randomly chose 50 utterances spoken by 20 speakers (10 male and 10 female) from GRID corpus [29] corrupted by white and babble noise from NOISEX-92 [30] at SNRs between 0 to 15 decibels. As the evaluation criteria, similar to [7, 26], we report the perceived speech quality and speech intelligibility quantified by the perceptual evaluation of speech quality (PESQ) [31] and the short-time objective intelligibility measure (STOI) [32], respectively. The prior SNR at harmonics is provided by interpolation of the SNR obtained by the amplitude estimator MMSE-STSA [33] and the minimum statistics as the noise estimator [34]. The fundamental frequency is estimated using PEFAC [35]. As analysis window, we found that a Blackman window is advantageous as it exhibits a large side-lobe rejection ratio, leading to the best performance. The frame length is set to 24 ms. We chose a frame shift of 2 ms ($R = 10$) as it allows for a more accurate von Mises parameter estimation.

4.1.1. Proof-of-Concept Experiment

The STFT phase exhibits no structure [26], therefore alternative phase representations are required for qualitative evaluation of the estimated phase versus the clean phase as reference. To this end, we look at the following representations: 1) spectrogram demonstrating the influence of replacing noisy phase with the modified phase spectrum, 2) group delay demonstrating the correctness of the estimated phase across frequency showing a harmonic structure [36], and 3) phase variance to assess the quality of the synthesized speech [22].

Figure 2 shows the graphical justification for the proof of concept experiment carried out on a male utterance corrupted with white noise at SNR = 5 (dB). From the spectrogram (top row) it is observed that the spectral amplitude is enhanced after reconstructing the phase-enhanced signal via overlap-and-add. This is justified by the recovery of the harmonic structure observed in the phase-

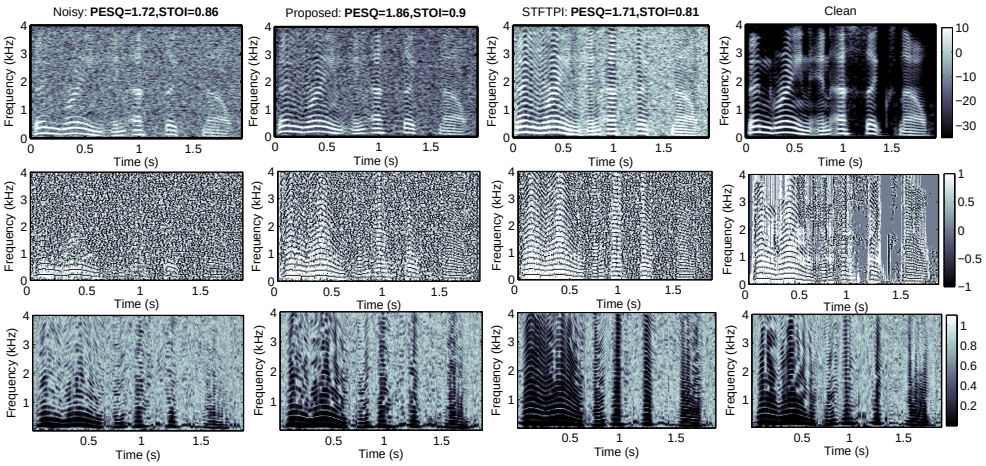

Fig. 2. Spectrogram (top), group delay (middle), and circular variance (bottom) shown for a male utterance corrupted with white noise at SNR $= 5$ (dB). From left to right: noisy (unprocessed), MAP phase-enhanced, STFTPI [7] and clean (oracle). The noise reduction is achieved by phase-only enhancement.

enhanced signal similar to that existing in the clean phase. This harmonic structure was lost in the noisy speech. The group delay plot (middle row) justifies the recovery of the harmonic structure in phase across frequency axis and is enhanced compared to the noisy phase. The visualized improvement reported here is obtained due to the proposed phase-only enhancement without an explicit employment of an amplitude enhancement scheme. Finally, the plots in the bottom row show the circular variance plots confirming that the proposed method successfully reduces the phase variance without over-estimation at higher frequencies as in [7].

Both PESQ and STOI outcomes are shown at the top of each panel in Figure 2. Comparing results versus STFTPI [7] and clean phase reveals that some artificial harmonics are introduced by the STFTPI at higher harmonics leading to some artifacts perceived as buzzyness in the reconstructed signal. Similar observation was reported in [3, 17]. The buzzy speech quality reduces the speech intelligibility as justified by the degraded STOI, lower than noisy signal. The proposed phase estimator, in contrast, balances a trade-off between replacing the noisy phase with an estimated one and on how much to rely on the noisy phase sampled at harmonics. The trade-off leads to a joint improvement in the perceived quality and intelligibility across SNRs as shown in Figure 3.

4.1.2. Perceived Quality and Speech Intelligibility

Figure 3 shows the delta improvement compared to the noisy input in terms of the perceived quality and speech intelligibility. The results are averaged over utterances grouped to white and babble noise scenarios. The results for unprocessed (noisy) and clean phase (oracle) are shown for comparison purposes demonstrating the lower and the upper bounds for the phase estimation performance, respectively. As benchmark, we report the performance of the STFTPI [7] where PEFAC [35] is used for fundamental frequency estimation.

As the input SNR is increased, the performance of the proposed method asymptotically approaches to that provided by clean phase known as the upper-bound phase estimation performance. For white noise scenario, the performance of the MAP estimator is not sensitive to the oracle knowledge of F0 and harmonic SNR. For babble noise scenario, however, the additional improvement obtained by the MAP phase estimator due to the oracle prior knowledge motivates to use a more accurate, F0 and SNR estimator. In overall, the proposed MAP phase estimator consistently improves both perceived quality as well as speech intelligibility across all SNRs and noise types. This is an interesting observation since the conventional amplitude-only speech enhancement methods were previously reported to reduce the speech intelligibility of the noisy speech [37].

The STFTPI method [7] improves PESQ at the expense of a considerable reduction in the speech intelligibility compared to the noisy signal. This can be explained by the introduced artificial harmonics perceived as buzzyness mainly occurring at higher frequencies. This is visually observed by comparing the last two plots shown in Figure 2. Similar observation was reported in [3, 17]. The artificial harmonics produced by STFTPI are not available in the clean signal hence the speech intelligibility by STFTPI degrades. In contrast to [7, 17], the proposed MAP phase estimator leads to a joint improvement in the perceived quality and speech intelligibility.

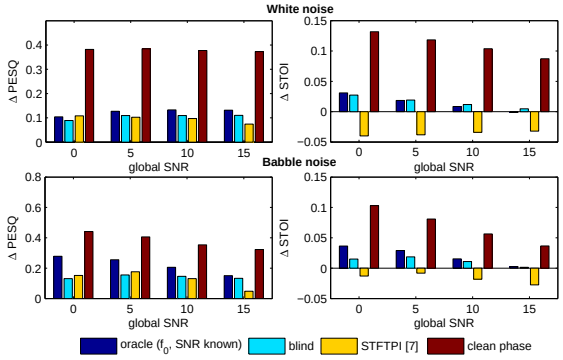

Fig. 3. PESQ and STOI Improvement for (top) white (bottom) babble noise.

5. CONCLUSION

In this paper we derived a Maximum A Posteriori harmonic phase estimator in single-channel speech enhancement. Our experiments showed that when noisy phase is replaced by the proposed MAP phase estimate, instrumental measures show joint improvement in the perceived speech quality and speech intelligibility consistently for various signal-to-noise ratios and noise types. Unlike the benchmark method, the proposed method is less sensitive to fundamental frequency and relies on no voicing state decision which arguably is erroneous at low signal-to-noise ratios and non-stationary noise.

6. REFERENCES

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement*, Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2013.

[2] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Phase importance in speech processing applications," *in Proceedings of the International Conference on Spoken Language Processing*, pp. 1623–1627, 2014.

[3] T. Gerkmann, "MMSE-optimal enhancement of complex speech coefficients with uncertain prior knowledge of the clean speech phase," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4511–4515.

[4] P. Mowlaee and R. Saeidi, "Time-frequency constraint for phase estimation in single-channel speech enhancement," *The International Workshop on Acoustic Signal Enhancement*, pp. 338–342, 2014.

[5] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE signal processing letters*, vol. 20, no. 3, pp. 217 – 220, 2013.

[6] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 12, pp. 1235–1239, Dec. 2013.

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

[8] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proceedings of the International Conference on Spoken Language Processing*, 2012.

[9] K. K. Paliwal, K. K. Wojcicki, and B. J. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[10] C. Chacon and P. Mowlaee, "Least squares phase estimation of mixed signals," *in Proceedings of the 15th International Conference on Spoken Language Processing*, pp. 2705–2709, 2014.

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

[12] P. Mowlaee and M. Watanabe, "Iterative sinusoidal-based partial phase reconstruction in single-channel source separation," in *Proc. Interspeech*, 2013, pp. 832–836.

[13] M. H. Hayes, J. S. Lim, and A. V. Oppenheim, "Signal reconstruction from phase or magnitude," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 6, pp. 672–680, Dec 1980.

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

[15] N. Sturmel and L. Daudet, "Informed source separation using iterative reconstruction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 1, pp. 178–185, 2013.

[16] A. Sugiyama and R. Miyahara, "Phase randomization - a new paradigm for single-channel signal enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7487–7491.

[17] S. P. Patil and J. N. Gowdy, "Exploiting the baseband phase structure of the voiced speech for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 6133–6137.

[18] T. Lotter and P. Vary, "Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model," *EURASIP J. Appl. Signal Process.*, pp. 1110–1126, Jan. 2005.

[19] P. J. Wolfe and S. J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," in *EURASIP JASP on Digital Audio for Multimedia Communications*, 2003, pp. 1043–1051.

[20] B. J. Borgstrom and A. Alwan, "A unified framework for designing optimal stsa estimators assuming maximum likelihood phase equivalence of speech and noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2579–2590, Nov 2011.

[21] J. Hao, H. Attias, S. Nagarajan, T. W. Lee, and T. J. Sejnowski, "Speech enhancement, gain, and noise spectrum adaptation using approximate bayesian estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 24–37, Jan 2009.

[22] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP, Journal on Audio, Speech, and Music Processing - Special Issue: Models of Speech - In Search of Better Representations*, 2014.

[23] K. V. Mardia and P. E. Jupp, *Directional Statistics*, Wiley edition, 1999.

[24] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Prentice Hall, 1993.

[25] R. McAulay and T. Quatieri, "Speech analysis/Synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[26] J. Kulmer, P. Mowlaee, and M. Watanabe, "A probabilistic approach for phase estimation in single-channel speech enhancement using von mises phase priors," in *IEEE Workshop on Machine Learning for Signal Processing*, Sept. 2014.

[27] N. I. Fisher, *Statistical Analysis of Circular Data*, Cambridge University Press, Oct. 1995.

[28] P. Berens, "Circstat: A matlab toolbox for circular statistics," *Journal of Statistical Software*, vol. 31, no. 10, pp. 1–21, 9 2009.

[29] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Acoustical Society of America Journal*, vol. 120, pp. 2421, 2006.

[30] 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," *Technical Report, DRA Speech Research Unit*, 1992.

[31] The ITU Radiocommunication Assembly, "ITU-T P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Tech. Rep., ITU, 2000.

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

[33] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 443–445, 1985.

[34] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul 2001.

[35] S. Gonzalez and M. Brookes, "PEFAC - a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, Feb 2014.

[36] A.P. Stark and K.K. Paliwal, "Group-Delay-Deviation Based Spectral Analysis of Speech," in *Proceedings of the International Conference on Spoken Language Processing*, 2009.

[37] P. C. Loizou and K. Gibak, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, Jan 2011.