Iterative Closed-Loop Phase-Aware Single-Channel Speech Enhancement

Pejman Mowlaee, *Member, IEEE*, and Rahim Saeidi, *Member, IEEE*

***Abstract*—Many short-time Fourier transform (STFT) based single-channel speech enhancement algorithms are focused on estimating the clean speech spectral amplitude from the noisy observed signal in order to suppress the additive noise. To this end, they utilize the noisy amplitude information and the corresponding a priori and a posteriori SNRs while they employ the observed noisy phase when reconstructing enhanced speech signal. This paper presents two contributions: i) reconsidering the relation between the phase group delay deviation and phase deviation, and ii) proposing a closed-loop single-channel speech enhancement approach to estimate both amplitude and phase spectra of the speech signal. To this end, we combine a group-delay based phase estimator with a phase-aware amplitude estimator in a closed loop design. Our experimental results on various noise scenarios show considerable improvement in the objective perceived signal quality obtained by the proposed iterative phase-aware approach compared to conventional Wiener filtering which uses the noisy phase in signal reconstruction.**

***Index Terms*—Phase-aware amplitude spectrum estimator, phase spectrum estimation, signal reconstruction, speech enhancements.**

I. Introduction

DESIRED speech signals are often corrupted with some background noise where the recording takes place, resulting in the requirement of a single-channel speech enhancement pre-processor for different speech applications, to name a few: robust automatic speech recognition and speech transmission. The problem has extensively been addressed during the last two decades with some satisfactory performance. While many proposals are dedicated to find a more accurate spectral amplitude estimator, the potential of phase spectrum estimation has often been neglected. More recent studies support the fact that incorporating phase information leads to improved speech enhancement or separation signal quality [1]–[8].

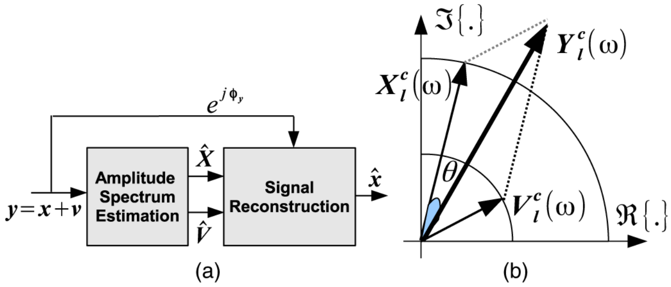

Fig. 1. (a) Block diagram of the conventional single-channel speech enhancement, (b) Vector representation of speech, noise and the resulting noisy complex spectra denoted by $\mathbf{X}_l^c(\omega)$, $\mathbf{V}_l^c(\omega)$, and $\mathbf{Y}_l^c(\omega)$, respectively.

To recover the target speech signal, a linear or non-linear filter is derived based on estimates of speech and noise spectra [9], [10]. This filter is then applied to the magnitude spectrum of the noisy signal. To this end, a noise-suppression rule or an amplitude spectrum estimator (the first block in Fig. 1(a)) is required to estimate the a priori SNR and a posteriori SNR value based on the estimated noise variance. The filter operates in the magnitude domain and emphasizes or attenuates certain frequencies. For signal reconstruction (the second block in Fig. 1(a)), conventional methods directly copy the phase spectrum of the noisy signal.

After introducing phase-aware amplitude estimator in Section II, in Section III we study the group delay deviation constraint for phase estimation. In Section IV an iterative phase-aware speech enhancement method is proposed and experimental results along with conclusions are given in Sections V and VI, respectively.

II. Phase-Aware Versus Conventional Spectral Amplitude Estimator

A. Conventional Spectral Amplitude Estimator

Let $x(n)$ and $v(n)$ be the time domain signals for speech and noise. Then the noisy signal at the lth frame is $y_l(n) = x_l(n) + v_l(n)$ where $n \in [0, N-1]$ and N is the window length. The observed noisy speech STFT is given by: $Y_l^c(\omega) = X_l^c(\omega) + V_l^c(\omega)$ where superscript ^{c} indicates the complex representation for STFT spectra and ω is the frequency. We further define $Y_l(\omega)$, $X_l(\omega)$ and $V_l(\omega)$ as the spectral amplitude for noisy speech, speech and noise at frequency ω and time frame l. The vector representation of the speech and noise spectra in the complex domain is illustrated in Fig. 1(b). Wiener filter has been widely used as a softmask gain function given by:

$$G_l(\omega) = \frac{X_l^2(\omega)}{X_l^2(\omega) + V_l^2(\omega)}, \tag{1}$$

and the time-domain speech signal is given by $\hat{x}_l(n) = \mathrm{DFT}^{-1}\{G_l(\omega)Y_l(\omega)e^{j\phi_{y,l}(\omega)}\}$ with $\phi_{y,l}(\omega)$ as the noisy phase.

Manuscript received September 12, 2013; revised October 13, 2013; accepted October 15, 2013. The work of P. Mowlaee was supported by the European project DIRHA (FP7-ICT-2011-7-288121). The work of R. Saeidi was supported by the European project BBfor2 (FP7 2007-2013) under Grant 238803 and was performed while with the Radboud University Nijmegen, The Netherlands during 2011–2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chandra Sekhar Seelamantula.

P. Mowlaee is with the Signal Processing and Speech Communication Laboratory, Graz University of Technology, A-8010 Graz, Austria (e-mail: pejman.mowlaee@tugraz.at).

R. Saeidi was with the Centre for Language and Speech Technology, Radboud University Nijmegen, 6525 HT Nijmegen, The Netherlands. He is now with the Speech and Image Processing Unit, University of Eastern Finland, Joensuu, Finland (email:rahim.saeidi@uef.fi).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2013.2286748

B. Phase-Aware Amplitude Estimator

Assuming a complex Gaussian distribution for the spectral coefficient as presented in [2], [6], [7] the phase-aware spectral amplitude estimator is given by:

$$\hat{X}_{\phi_x} = \sqrt{\frac{2}{\beta_1}}\frac{D_{-2}(z)}{D_{-1}(z)}, \text{ where } z = -\frac{2Y\cos(\phi_y - \phi_x)}{\sqrt{2\beta_1}\sigma_v^2}, \quad (2)$$

where $D_{-\nu}(\cdot)$ is the parabolic cylinder function of order ν and for simplicity we dropped l and ω, and we have $\beta_1 = 1/\sigma_v^2 + 1/2\sigma_x^2$ where $E\{V^2\} = \sigma_v^2$ as the noise PSD with complex Gaussian distribution, for the joint distribution for Y and Φ_Y.

III. Phase Estimation for Signal Reconstruction

Early studies in [11] addressed the problem of estimating signals from their modified magnitude spectrum showing that under certain restriction on the window and signal, it is possible to uniquely find the signal by iteratively minimizing a mean square error criterion. More recently, the authors in [4] suggested to impose additional constraint of minimizing the inconsistency in the complex spectrum leading to a *consistent* wiener filter. Recently, in [3], we presented a solution to the phase estimation problem using both geometry and the phase group delay deviation property. In this Section, we relate the phase deviation concept derived in [12] and the phase group delay deviation [13] with Cramer Rao lower bound (CRLB) for phase estimation [14].

A. Relationship Between Phase Group Delay Deviation and Phase Deviation

A complex exponential $A_0 e^{j\omega_0 n}$, windowed by an N-point symmetric window has a constant group delay of $\frac{N-1}{2}$. Assume the noisy signal composed of two exponentials $Y^c(\omega) = X^c(\omega) + V^c(\omega)$ where $X^c(\omega) = A_x e^{j\phi_x} W(\omega - \omega_x)$ and $V^c(\omega) = A_v e^{j\phi_v} W(\omega - \omega_v)$ with A_x and A_v as amplitude of exponentials for speech and noise at frequencies ω_x and ω_v, respectively, while ϕ_x and ϕ_v are their corresponding initial phase values, and $W(\omega)$ is the spectral analysis window. Then from the definition the group delay $\tau(\omega) = -\mathrm{Im}\frac{\partial}{\partial\omega}\log Y^c(\omega)$ and similar to [13] we obtain:

$$\tau(\omega) = -\mathrm{Im}\left[\frac{\partial}{\partial\omega}\log X^c(\omega) + \frac{\partial}{\partial\omega}\log\left(1 + \frac{e^{j\psi(\omega)}}{\mathrm{SNR}(\omega)}\right)\right], \quad (3)$$

with $\psi(\omega) = \phi_v(\omega) - \phi_x(\omega)$ and we define $\mathrm{SNR}(\omega) = \frac{X(\omega)}{V(\omega)}$ as the instantaneous local signal-to-noise ratio at frequency ω. As an extreme case, for $\mathrm{SNR}(\omega) \to \infty$ we get the noisy phase as a reliable estimate of the clean speech phase for very high SNR regions. This observation supports common choice of the noisy phase as minimum mean square error (MMSE) estimate of the clean phase [10].

Similar to [13], we define the group delay deviation (GDD) at frequency ω denoted by $\Delta\tau(\omega)$ as the deviation in group delay contributed by the superposition of the two harmonics relative to the constant group delay, and we obtain

$$\Delta\tau(\omega) = \frac{\partial}{\partial\omega}\left(\mathrm{Im}\left[\log\left(1 + \frac{e^{j\psi(\omega)}}{\mathrm{SNR}(\omega)}\right)\right]\right). \quad (4)$$

Let $z = 1 + 1/\mathrm{SNR}(\omega)e^{j\psi(\omega)}$. Then from the definition of the logarithm for a complex number $z = re^{j\phi}$ we have $\log z = \ln|z| + j\mathrm{Arg}(z)$, where $\mathrm{Arg}(z) = \phi \in (-\pi, \pi]$ is the principal value of the argument and $r = |1 + \mathrm{SNR}^{-1}(\omega)\cos\psi(\omega) + j\mathrm{SNR}^{-1}(\omega)\sin\psi(\omega)|$ and $\phi = \tan^{-1}(\frac{\sin\psi(\omega)}{\mathrm{SNR}(\omega)+\cos\psi(\omega)})$. Using these in (4) we get:

$$\Delta\tau(\omega) = \frac{\partial}{\partial\omega}\left(\tan^{-1}\left(\frac{\sin\psi(\omega)}{\mathrm{SNR}(\omega) + \cos\psi(\omega)}\right)\right). \quad (5)$$

The argument inside $\tan^{-1}(\cdot)$ function is the same as the phase deviation denoted by $\phi_{\mathrm{dev}}(\omega)$ defined in [12] as the amount of phase change in radian for the speech signal due to noise given by:

$$\tan\phi_{\mathrm{dev}}(\omega) = \frac{\sin\psi(\omega)}{\mathrm{SNR}(\omega) + \cos\psi(\omega)}. \quad (6)$$

Similar to [12], assuming Gaussian distribution for the noise $V(\omega)$ and a voiced speech segment with separated enough harmonics, the maximum phase deviation is given by

$$\phi_{\mathrm{dev,max}}(\omega) = \arcsin\left(\sqrt{\pi/(2\mathrm{SNR}(\omega))}\right). \quad (7)$$

Clearly, for $\mathrm{SNR}(\omega) \to \infty$ we obtain $\phi_{\mathrm{dev}}(\omega) = 0$ which means in the estimation there is no deviation from the clean signal phase. Finally using (6) in (5) we obtain:

$$\Delta\tau(\omega) = \frac{\partial\phi_{\mathrm{dev}}(\omega)}{\partial\omega}, \quad (8)$$

which indicates the relationship between the group delay deviation and the phase deviation in terms of phase difference and local SNR. When the local SNR becomes too low, the phase deviation increases. In [12] it is shown that some roughness in the synthesized speech is to be recognized when phase deviation $\phi_{\mathrm{dev}} > 0.679$ that is roughly the threshold of perception. For moderate to high SNR regions ($\mathrm{SNR} \geq 6$ (dB)), where according to (7) the phase deviation is still small enough, the phase group delay deviation exhibits a minimum, which is in accordance with previous findings in [13] that small group delay deviation is primarily contributed by a single sinusoid. Furthermore, from an estimation theory standpoint, the estimation variance for one sinusoidal phase, given its frequency and independent of our knowledge regarding its amplitude can be derived from the CRLB and is given as [14]:

$$\mathrm{CRLB}(\phi_x|\omega_x) = \frac{2\sigma_v^2}{NA_x^2}, \quad (9)$$

where for one sinusoid we define $\mathrm{SNR} = \frac{A_x^2}{2\sigma_v^2}$ with A_x and ω_x as the amplitude and frequency of the sinusoid observed in noise, σ_v^2 as the noise variance and N is the data length. The estimation error variance is clearly governed by the inverse of the local SNR and is directly related to N. Similar to (5), for high enough SNR ($\mathrm{SNR}(\omega) \to \infty$) the phase estimation variance tends to zero. Furthermore, the larger the data length, the lower the phase estimation error variance, explaining the improved phase estimation performance obtained in [15] for large window length.

B. Phase Estimation Using Group Delay Function

In [3] we proposed a solution for the phase estimation problem using both geometry and group delay deviation minimization. In that work it is shown that by using quantized spectral amplitudes for two sources the phase estimation method still performs better than employing the noisy phase. In present work we employ the phase estimation for scenario where the prior assumption on the knowledge of the magnitude

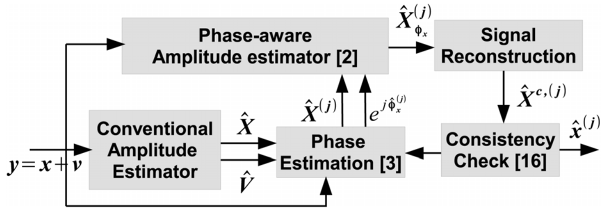

Fig. 2. Block diagram for the proposed closed-loop single-channel speech enhancement algorithm. The numbers written in the blocks refer to the references.

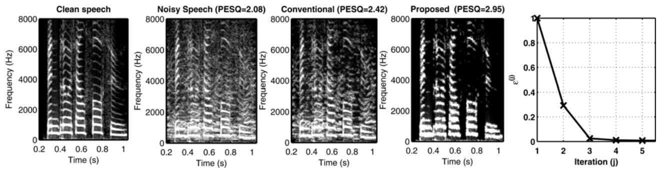

Fig. 3. Spectrogram analysis for female speech saying "*bin blue at L four soon*" selected from the GRID corpus [1] corrupted with babble noise at $\mathrm{SNR} = 5\ (\mathrm{dB})$, comparing (from left to right) spectrograms for the clean speech, noisy speech, Wiener filter with noisy phase, the proposed method after four iterations, (right) convergence behavior for the proposed method showing inconsistency difference across $J = 5$ iterations for input $\mathrm{SNR} = 5\ (\mathrm{dB})$.

spectra of both speech and noise are relaxed and estimated values are utilized.

We emphasize the fact that for signal components of $\mathrm{SNR} \geq 6\ (\mathrm{dB})$, the phase deviation is small, and therefore the use of the noisy signal phase would not result in significant loss in speech quality. On the other hand, for spectral components of SNRs lower than 6 (dB), by employing the aforementioned phase estimation approach, we replace the noisy phase with the estimated phase spectrum. From CRLB analysis, when speech amplitude or noise estimate are correctly estimated, the phase estimation error variance gets relatively small, leading to an improved speech enhancement. To capture high SNR signal components, in this work for the sake of simplicity we employ peak picking. The number of signal components for phase estimation varies based on the number of spectral peaks taken by peak picking. No discrimination is made between voiced/unvoiced frames in phase estimation step.

IV. Proposed Phase-Aware Closed-Loop Solution

We propose a phase-aware speech enhancement solution where the amplitude and phase spectra of speech signal are iteratively estimated. The block diagram of the proposed closed-loop single-channel speech enhancement algorithm is shown in Fig. 2. The procedure is described as follow: For initialization, an enhanced speech signal provided by a conventional method is required. The Wiener filtered speech amplitude estimate is exploited to provide a phase spectrum estimate using the phase group delay property employed at spectral peaks (as presented in Section III-B). The previously obtained estimated phase spectrum is then fedback as the input for the phase-aware amplitude estimator given by (2) in order to improve the spectral amplitude estimate. The estimated phase-aware amplitude together with phase estimate are used to build the complex spectrogram $\hat{\mathbf{X}}^{(j)}_{\hat{\phi}_x}$. The loop is closed with applying Griffin and Lim rule [16]. For next iterations $j \geq 1$, the input of the phase estimation module is provided by applying Griffin and Lim rule [4] on top of the latest complex spectrum.

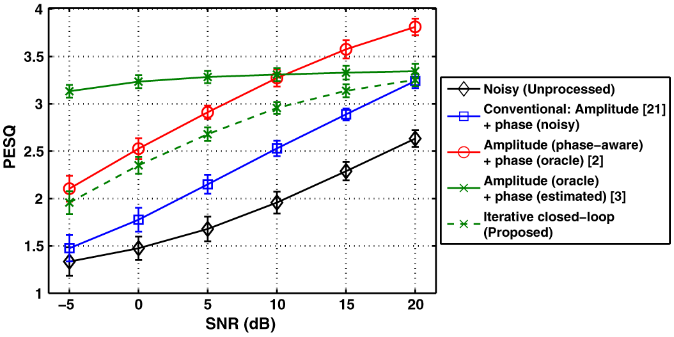

Fig. 4. PESQ results versus input signal-to-noise ratio measured in decibels.

As the convergence criterion we select the inconsistency constraint as in [4] defined in complex domain as: $\mathrm{F}(\mathbf{X}^{c,(j)}) = \mathrm{STFT} \circ \mathrm{iSTFT}(\mathbf{X}^{c,(j)}) - \mathbf{X}^{c,(j)}$, where $\mathbf{X}^{c,(j)} = \mathbf{X}^{(j)} e^{j\hat{\phi}_x^{(j)}}$ is the complex spectrogram of the speech signal obtained at the jth iteration. To investigate the convergence of the proposed method at each iteration, we measure the difference of inconsistency of the complex spectrogram obtained at each jth iteration defined as $D^{(j)} = \|\mathrm{F}(\mathbf{X}^{c,(j)})\|_2^2 - \|\mathrm{F}(\mathbf{X}^{c,(j-1)})\|_2^2$. As our stopping criterion, we check the number of iterations required to achieve the relative error defined as $\varepsilon^{(j)} = \frac{|D^{(j)}|}{\|\mathrm{F}(\mathbf{X}^{c,(j)})\|_2^2} < \varepsilon$ with $\varepsilon = 0.05$. The final enhanced speech signal is given by:

$$\hat{\mathbf{x}}^{(J)} = \mathrm{DFT}^{-1}\left\{\hat{\mathbf{X}}^{(J)}_{\hat{\phi}_x} e^{j\hat{\phi}_x^{(J)}}\right\}. \tag{10}$$

V. Experimental Results

A. Database and Experiment Setup

The speech signals are extracted from GRID [17] and TIMIT database [18] while noise is taken from NOISEX-92 [19]. As our frame setup, we chose a Hamming window length of 32 ms with a frameshift of 4 ms at 8 kHz sampling frequency. The noise is estimated online using a voice activity detector where we use the improved minima controlled recursive averaging (IMCRA) proposed in [20]. For the conventional method, speech is estimated using decision-directed approach in [21]. To initialize the noise tracker we use the first noise-only frames.

B. Spectrogram Analysis

Fig. 3 shows spectrograms obtained by employing the proposed approach on a noisy female speech signal corrupted with babble noise at $\mathrm{SNR} = 5\ (\mathrm{dB})$. The last panel in Fig. 3 shows the convergence behavior of the proposed iterative approach for $J = 5$ total number of iterations in terms of the relative error $\varepsilon^{(j)}$. The results are shown for an input $\mathrm{SNR} = 5\ (\mathrm{dB})$, averaged over segments from ten speakers in GRID corpus [17]. In the following experiments, we set the maximum number of iterations to $J = 4$. The results of the proposed method after fourth iterations are shown and compared with those from the clean speech reference, noisy speech (unprocessed), and conventional method using noisy phase. The phase-aware solution recovers considerably more harmonic structure of the speech signal compared to the phase-independent approach[1]. This results in considerable improvement in terms of the perceived speech quality justified by the PESQ score reported at the title of Fig. 3.

[1] Some audio wave files are available at the following link: http://www.spsc.tugraz.at/SPL2013phase

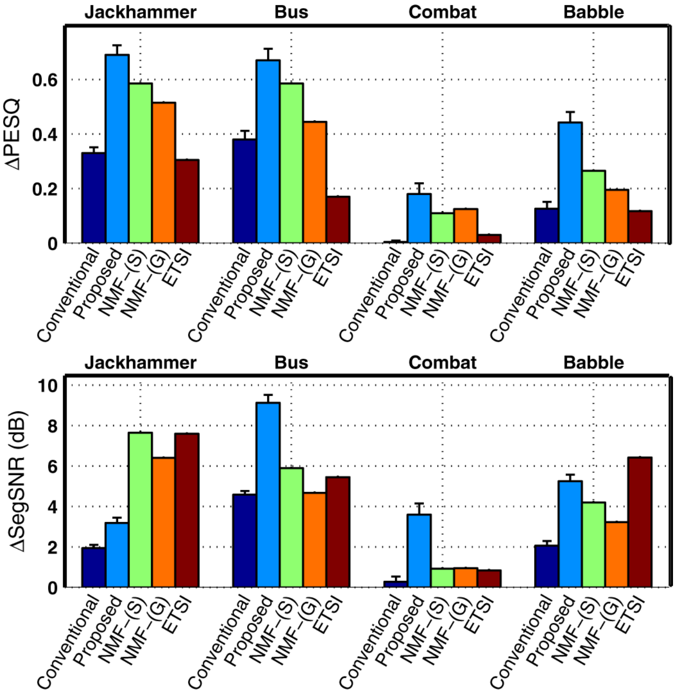

Fig. 5. Speech enhancement performance for four noise types: jackhammer, bus, combat and babble averaged over 20 sentences of TIMIT database mixed at 0 (dB) input SNR.

C. PESQ Evaluation

Fig. 4 shows the PESQ results averaged over fifty utterances (10 speakers with 5 sentences per speaker) selected from the GRID corpus. The utterances are contaminated by babble noise at different SNR levels. The PESQ results obtained by the proposed iterative phase-aware method in Eq. (10) are shown in dashed green line. We also report the performance obtained using conventional method described in Section V-A where noisy phase is used for signal reconstruction as given by Eq. (1) (solid blue line). For further analysis of the results, we also include two upper-bounds obtained from our previous studies: i) phase-aware amplitude estimation given the clean phase spectrum [2] calculated in Eq. (2), and ii) speech signal synthesized using estimated phase given the oracle magnitude spectrum [3]. The proposed method significantly improves the perceptual quality for the low SNR region compared to the phase-independent method. Feedback of the estimated phase spectrum for amplitude estimation and signal reconstruction has the potential of providing considerable improvement in the overall speech enhancement performance.

Using the segments contaminated by factory2 noise at $\mathrm{SNR} = 0$ (dB), the $\Delta\mathrm{PESQ}$ results (compared to noisy signal) of 0.52 ± 0.06 and 0.31 ± 0.07 for the proposed (Fig. 3) and conventional approach (Fig. 1) are achieved, respectively. The corresponding results for white noise are 1.00 ± 0.08 and 0.50 ± 0.07. The resulting average real time factor (RTF), computed as, $\mathrm{RTF} = (\text{processing time [s]})/(\text{length of audio file [s]})$ that was achieved on a standard PC [2] running Windows 7 Enterprise (64 bit) and Matlab 7.9.0 is 1.6 for conventional Wiener filter and 9.2 for the proposed algorithm.

D. Results for More Noise Types and TIMIT Speech Corpus

To investigate the robustness of the proposed method against different noise types and a different speech corpus, we conduct another experiment: We extract 20 utterances (2 utterance per 10 speakers) from the TIMIT database [18] as for speech signals and we employed jackhammer, bus, babble and combat noise. The results are averaged over 20 utterances and reported in Fig. 5c for input SNR of 0 (dB). As our evaluation criteria we report the relative improvement in PESQ [22] and segmental SNR [23]. As for benchmark methods, we compare with two non-negative matrix factorization (NMF) methods in [24]: NMF-(S): speaker-specific speech models and NMF-(G): using non-speaker-specific speech model trained on a mixed gender group. We further include results obtained by ETSI front end [25]. Averaging over all four noise types, it is observed that the proposed method achieves 0.50 ± 0.04 relative improvement in PESQ compared to the maximum 0.38 reported in [24] using NMF methods and 0.21 ± 0.02 obtained by the conventional method. Similarly, the proposed method achieves segmental SNR improvement of 5.29 ± 0.38 (dB) compared to 2.22 ± 0.21 (dB) obtained by conventional method and the maximum 4.67 (dB) obtained by NMF methods [24].

ΔPESQ results show similar trends between methods for all four noise scenarios while segmental SNR results departs from these trends for jackhammer and babble noise. This is due to the fact that SNR-based measures are very sensitive to differences in spectral gain normalization, time delay between the signals to evaluate, and time frame setup; resulting in artifacts [26] and outliers [27]. As indicated in [28], segmental SNR measure, among several other objective measures, shows the lowest correlation with listening tests. As our proposed method modifies the amplitude and phase in a phase-aware way, hence segmental SNR metric shows high sensitivity. Therefore, judgement based on segmental SNR results and discrepency with PESQ metric should be made with caution.

VI. Conclusion

In this letter we showed that incorporating the knowledge of the estimated speech spectral phase in a closed-loop manner leads to improved perceived signal quality compared to previous phase-independent solutions. To this end, we proposed a closed-loop single-channel speech enhancement algorithm where we combined phase estimation with phase-aware amplitude estimator. The performance of the proposed method was compared to the Wiener filter with noisy phase.

The current results justify the effectiveness of the proposed closed-loop phase-aware approach as an interesting alternative that can push the limits of previous phase-independent solutions employed for long. The proposed method performs close to two bounds on the speech enhancement performance achieved by taking into account prior information about speech amplitude or phase spectra; in terms of PESQ, at low SNRs, the proposed approach asymptotically reaches the performance exhibited by the phase-aware amplitude estimator given by the oracle phase values while at high SNRs, the proposed method performs close to that obtained by exploiting the estimated phase given the oracle amplitude spectrum prior.

[2] equipped with a Core i7 CPU clocked at 2.94 GHz, and with 4 GB of RAM

References

[1] D. Gunawan and D. Sen, “Iterative phase estimation for the synthesis of separated sources from single-channel mixtures,” *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 421–424, May 2010.

[2] P. Mowlaee and R. Saeidi, “On phase importance in parameter estimation in single-channel speech enhancement,” in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7462–7466.

[3] P. Mowlaee, R. Saiedi, and R. Martin, “Phase estimation for signal reconstruction in single-channel speech separation,” in *Proc. Int. Conf. Spoken Language Processing*, 2012.

[4] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama, “Consistent wiener filtering: Generalized time-frequency masking respecting spectrogram consistency,” in *Proc. LVA ICA 2010*, 2010, pp. 89–96.

[5] M. K. Watanabe and P. Mowlaee, “Iterative sinusoidal-based partial phase reconstruction in single-channel source separation,” in *Proc. 14th Int. Conf. Spoken Language Processing*, 2013, pp. 832–836.

[6] P. Mowlaee, M. Watanabe, and R. Saeidi, “Show & tell: Phase-aware single-channel speech enhancement,” in *14th Annu. Conf. Int. Speech Communication Association*, 2013, pp. 1872–1874.

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

[8] N. Sturmel and L. Daudet, “Iterative phase reconstruction of wiener filtered signals,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2012, pp. 101–104.

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

[10] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, “Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, 2007.

[11] M. Hayes, J. Lim, and A. Oppenheim, “Signal reconstruction from phase or magnitude,” *IEEE Trans. Acoust., Speech, Signal Processi.*, vol. ASSP-28, no. 6, pp. 672–680, Dec. 1980.

[12] P. Vary, “Noise suppression by spectral magnitude estimation mechanism and theoretical limits,” *Signal Process.*, vol. 8, no. 4, pp. 387–400, 1985.

[13] A. P. Stark and K. K. Paliwal, “Group-delay-deviation based spectral analysis of speech,” in *INTERSPEECH*, 2009, pp. 1083–1086.

[14] D. Rife and R. R. Boorstyn, “Single tone parameter estimation from discrete-time observations,” *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 591–598, 1974.

[15] K. K. Paliwal and L. Alsteris, “Usefulness of phase spectrum in human speech perception,” in *Proc. Eurospeech*, 2003, pp. 2117–2120.

[16] D. Griffin and J. Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 236–243, Feb. 1984.

[17] M. Cooke, J. R. Hershey, and S. J. Rennie, “Monaural speech separation and recognition challenge,” *Comput. Speech Lang.*, vol. 24, no. 1, pp. 1–15, 2010.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren*, DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*, 1993.

[19] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[20] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

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

[22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” *Speech Commun.*, vol. 2, pp. 749–752, Aug. 2001.

[23] P. Loizou*, Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC, 2007.

[24] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, “Speech denoising using nonnegative matrix factorization with priors,” in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4029–4032.

[25] Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms Tech. Rep. ETSI ES 202 050 V1.1.3, 2003.

[26] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[27] J. H. L. Hansen and B. L. Pellom, “An effective quality evaluation protocol for speech enhancement algorithms,” in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 1998, pp. 2819–2822.

[28] Y. Hu and P. C. Loizou, “Evaluations of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.