Show & Tell: Phase-Aware Single-channel Speech Enhancement

Pejman Mowlaee[†]*, Mario Kaoru Watanabe*[†] *and Rahim Saeidi*[‡]

[†] Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria
[‡] Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
pejman.mowlaee@tugraz.at mario.watanabe@student.tugraz.at rahim.saeidi@let.ru.nl

Abstract

Many short-time Fourier transform (STFT) based single-channel speech enhancement algorithms are focused on estimating clean speech spectral amplitude from the noisy observed signal in order to suppress the additive noise. To this end, the state-of-the-art speech enhancement algorithms, employ noisy amplitude information and correspondingly a priori and a posteriori SNRs while they use no information about the speech phase spectrum. In this show and tell we demonstrate our recent progress on developing novel ideas towards incorporating phase information in solving single-channel speech enhancement problem.

Index Terms: Single-channel speech enhancement, phase estimation for signal reconstruction, amplitude estimation.

1. Introduction

A single-channel speech enhancement algorithm is of great importance in many practical applications to name a few: to design a robust automatic speech recognition system, speech transmission in mobile telephony, and in enhancing the perceived signal quality of the desired signals recorded in adverse noise scenarios.

Earlier studies emphasize on the higher importance of deriving an accurate estimator for clean speech amplitude spectrum rather than the phase spectrum [1]. More recent studies, however, report certain improvement in speech enhancement or separation signal quality [2–8], speech intelligibility [9–12] and automatic speech recognition accuracy [13, 14] by utilizing enhanced phase spectrum. Assuming circular symmetric joint probability density function (pdf) for amplitude and phase spectra, phase is uniformly distributed and independent of the amplitude spectrum. Under such assumption, the MMSE optimal estimate of the clean speech phase is equal to the noisy phase [15–17], obtained when the temporal dynamics are ignored. Taking the temporal dynamics into account, the speech phase and amplitude spectra are not generally independent. For example, in [18] the deviation in phase group delay (GD) was shown to follow the spectral amplitude behavior. Furthermore the phase spectrum of the original signal can be estimated by iteratively performing short-time Fourier transform (STFT) and inverse STFT, given the spectral amplitudes only [19].

Figure 1 shows the block diagram of a typical single-channel speech enhancement algorithm composed of two stages: i) a spectral amplitude estimator like Wiener filter [20] followed by ii) a signal reconstruction stage, where the noisy phase itself is often employed for signal reconstruction. Some recent studies support the evidence that further improvement in speech enhancement performance is achievable by incorporating speech phase spectrum at spectral amplitude estimation stage [3, 8, 21]. Furthermore, for the signal reconstruction stage (see stage 2 in Figure 1), recent studies for single-channel speech enhancement and single-channel source separation applications reported potential improvement achievable by replacing the noisy phase with an estimated one [2, 4, 5, 22–28].

This work was partially funded by the European project DIRHA (FP7-ICT-2011-7-288121) and by the ASD (Acoustic Sensing & Design) with joint support from Speech Processing Solutions GmbH Vienna, BMVIT, BMWFJ, Styrian Business Promotion Agency (SFG), and the Government of Styria

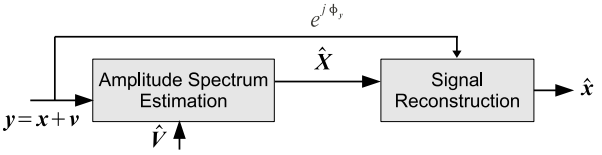

Figure 1: Block diagram of the conventional single-channel speech enhancement, composed of two stages: amplitude estimation (first block), and signal reconstruction (second block).

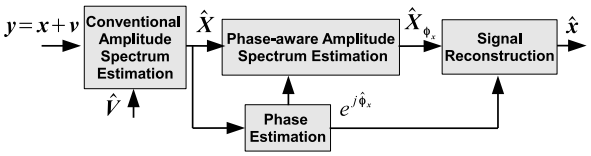

Figure 2: Block diagram of the proposed phase-aware approach.

2. Concept From the Technical Point-of-view

2.1. Phase-independent Spectral Amplitude Estimation

Let $x_l(n)$ and $v_l(n)$ as time domain signals for speech and noise at the lth frame. 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 ^{c} super index indicates the complex representation for STFT spectra where $Y_l^c(\omega) = Y_l(\omega)e^{j\phi_{y,l}(\omega)}$ with $Y_l(\omega)$ and $\phi_{y,l}(\omega)$ as amplitude and phase spectra of noisy signal, respectively, ω as the frequency. To reconstruct the enhanced signal we need to associate the magnitude and phase spectra and take the inverse STFT. To this end, Wiener filter has been widely used as the softmask gain function and is given by [20]:

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

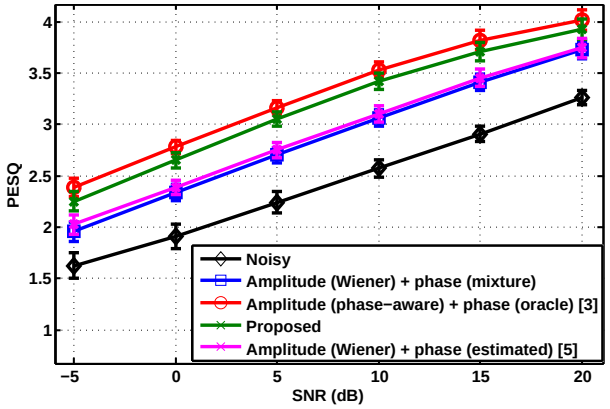

Figure 3: The PESQ results obtained for speech corrupted with babble noise for SNR $\in [-5, 20]$ with 5 decibels step, averaged over ten speakers. The performance of phase-aware amplitude estimator and estimated phase is compared to mixture phase and phase-independent Wiener filter.

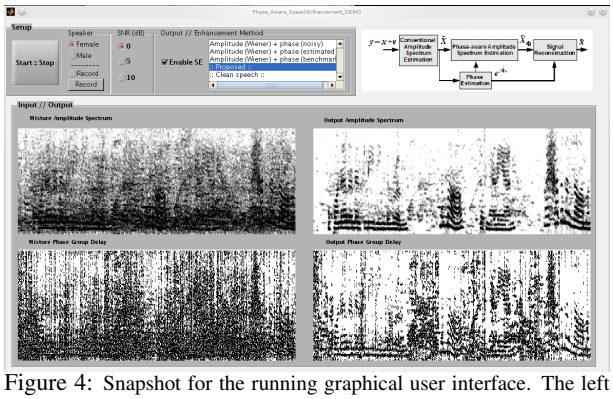

Figure 4: Snapshot for the running graphical user interface. The left panel shows the female speech corrupted by babble noise at SNR $= 0$ (dB). The right panel shows the enhanced signal by the proposed method.

The time-domain signal estimate for the ith signal at the lth frame is given by $\hat{x}_l(n) = \text{DFT}^{-1}\{G_l(\omega)Y_l(\omega)e^{j\phi_{y,l}(\omega)}\}$.

2.2. Phase-Aware Amplitude Estimator

Assuming a complex Gaussian distribution for the spectral coefficient as presented in [3]($\mu = 0.5, \beta = 1$ [8]) the phase-aware spectral amplitude estimator is given by:

$$\hat{X}_{\phi_{x,l}}(\omega) = \sqrt{\frac{2}{\beta_{1,l}}}\frac{D_{-2}(z)}{D_{-1}(z)},\ z = -\frac{2Y_l(\omega)}{\sqrt{2\beta_{1,l}\sigma^2_{v,l}}}\cos(\phi_{\text{dev}}(\omega)), \tag{2}$$

with $\phi_{\text{dev}}(\omega) = \phi_{y,l}(\omega) - \phi_{x,l}(\omega)$ as the phase deviation of clean speech phase due to noise addition with $D_{-\nu}(\cdot)$ is the parabolic cylinder function of order ν and $\beta_{1,l} = 1/\sigma^2_{v,l} + 1/2\sigma^2_{x,l}$ with $E\{V^2_{v,l}(\omega)\} = \sigma^2_{v,l}$ as the noise PSD with a complex Gaussian distribution for the joint distribution of $Y_l(\omega)$ and $\phi_{y,l}(\omega)$.

2.3. Phase Estimation for Signal Reconstruction

In [5], we presented a solution to resolve the phase estimation problem using both geometry and group delay minimization. The phase group delay deviation constraint helped to select the correct phase candidate which was unknown due to the ambiguity in the sign of the phase difference between the two spectra.

Following [5], for given the speech and noise magnitude spectra, the steps required for group delay based phase estimation at spectral peaks are: 1) use the geometry constraint to find the two most probable phase values for each time frame and frequency bin (as given by (16) in [5] shown in Fig. 3), and 2) use the phase group delay deviation constraint in order to estimate phase spectrum out of the two candidates (given by Equation (17) in [5]). The reconstructed signal is

$$\hat{x}_l(n) = \text{DFT}^{-1}\{\hat{G}_{\hat{\phi}_{x,l}}(\omega)Y_l(\omega)e^{j\hat{\phi}_{x,l}(\omega)}\}, \tag{3}$$

where $\hat{G}_{\hat{\phi}_{x,l}}(\omega)$ is given by employing phase-aware amplitude estimate $\hat{X}_{\hat{\phi}_{x,l}}(\omega)$ together with $\hat{V}_l(\omega)$ in (1). The prior assumption on the knowledge of the speech and noise magnitude spectra required for phase estimation were relaxed to their quantized versions in [5], still leading to improved perceived signal quality compared to noisy phase or other benchmarks. Here, to have speech phase spectrum estimate, we drive our phase estimation approach using the conventional Wiener filter as the initial amplitude spectrum estimate while as our noise estimate the first noise-only frames are used.

3. Proposed Full System

Innovative Aspects: Here in this show and tell, we aim at demonstrating the impact of phase information in the *full enhancement system*. Figure 2 shows the block diagram of our proposed system. Given the initial Wiener filtered speech spectral amplitude estimate, we estimate the speech phase spectrum using group delay constraints as described in [5]. The estimated phase values are then passed to a phase-aware MMSE amplitude estimator in [3], where a more accurate speech spectral amplitude is provided. The so-obtained estimated amplitude and phase spectra are used to reconstruct the final enhanced speech signal. We show that incorporating the estimated phase spectrum and employing it in phase-aware amplitude estimator leads to considerable improvement in the perceived signal quality.

Results: Figure 3 shows the PESQ results averaged over segments of ten speakers selected from GRID corpus corrupted with babble noise, where the input signal-to-noise ratio is swept within the range of -5 to 20 decibels with 5 decibels step. The proposed method significantly improves the perceptual quality for low SNR region compared to phase-independent amplitude or phase estimation methods.

Presentation Format: The demonstrator runs on a Laptop. Headphones are provided for listeners. The sample video demonstration for the MATLAB Graphical User Interface (GUI) is available at `http://www2.spsc.tugraz.at/people/pmowlaee/Video_GUI.wmv` to represent the contents of the show and tell session. In short, we start with playing a noise corrupted speech signal, followed by a Wiener filter enhanced signal. Then we switch to the proposed phase-aware approach. Figure 4 shows a snapshot from the graphical user interface while running for a female speech signal corrupted with babble noise at SNR $= 5$ decibels[1]. Top panel shows the spectrogram for noisy and enhanced signals on the left, while the bottom panel shows the phase group delay representation for the two signals, respectively. The phase-aware solution successfully recovers the harmonic structure of the speech signal compared to phase-independent approach.

[1]Audio wave files and spectrograms are downloadable at: `http://www.spsc.tugraz.at/showandtellphase`.

4. References

[1] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 30, no. 4, pp. 679–681, 1982.

[2] D. Gunawan and D. Sen, "Iterative phase estimation for the synthesis of separated sources from single-channel mixtures," *Signal Processing Letters, IEEE*, vol. 17, no. 5, pp. 421 –424, May 2010.

[3] P. Mowlaee and R. Saeidi, "On phase importance in parameter estimation in single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2013, pp. 7462 – 7466.

[4] 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.

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

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

[7] 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.

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

[9] L.D. Alsteris and K.K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, pp. 727–736, 2006.

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

[11] L.D. Alsteris and K.K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Elsevier Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

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

[13] R. Schluter, R. S. Uter, and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 133–136.

[14] P. Aarabi, *Phase-Based Speech Processing*, World Scientific Publishing, 2006.

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

[16] R. C. Hendriks and R. Martin, "MAP estimators for speech enhancement under normal and rayleigh inverse Gaussian distributions," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 918–927, 2007.

[17] 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 Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

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

[19] D. Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 2, pp. 236 – 243, Apr 1984.

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

[21] P. Mowlaee and R. Martin, "On phase importance in parameter estimation for single-channel source separation," in *The International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[22] N. Sturmel, L. Daudet, and L. Girin, "Phase-based informed source separation of music," in *Proc. 13th International Conference on Digital Audio Effects (DAFx-12)*, Sept. 2011, pp. 375–386.

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

[24] N. Sturmel, L. Daudet, and L. Girin, "Phase-based informed source separation of music," in *Proc. 13th International Conference on Digital Audio Effects (DAFx-12)*, Sept. 2012.

[25] J. Le Roux, Kameoka. H., N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. 13th International Conference on Digital Audio Effects (DAFx-10)*, Sept. 2010, pp. 397–403.

[26] 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 *9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2010, pp. 89–96.

[27] M. K. Watanabe and P. Mowlaee, "Iterative sinusoidal-based partial phase reconstruction in single-channel source separation," in *Proceedings of the 14th International Conference on Spoken Language Processing*, 2013.

[28] M. Krawczyk and T. Gerkmann, "STFT phase improvement for single channel speech enhancement," in *International Workshop on Acoustic Signal Enhancement; Proceedings of IWAENC*, 2012, pp. 1–4.