PHASE RECONSTRUCTION IN SINGLE CHANNEL SPEECH ENHANCEMENT BASED ON PHASE GRADIENTS AND ESTIMATED CLEAN-SPEECH AMPLITUDES

Yanjue Song, Nilesh Madhu

IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium
{yanjue.song@ugent.be, nilesh.madhu@ugent.be}

ABSTRACT

Phase *gradients* can help enforce phase *consistency* across time and frequency, further improving the output of speech enhancement approaches. Recently, neural networks were used to estimate the phase gradients from the short-term amplitude spectra of *clean* speech. These were then used to synthesise phase to reconstruct a plausible time-domain signal. However, using purely synthetic phase in speech enhancement yields unnatural-sounding output. Therefore we derive a closed-form phase estimate that combines the synthetic phase with that of the enhanced speech, yielding more natural output. Secondly, we empirically evaluate the benefit of (re-)training the phase gradient estimation networks on the amplitude spectra of the *estimated* clean-speech signal. Lastly we apply our proposed phase enhancement to the output of a *phase-aware* speech enhancement DNN, verifying if an *independent* phase estimator brings additional advantage. Results show that, compared to the baseline, the proposed approach further improves the DNSMOS scores by ≈ 0.1 on average, and significantly in the first quartile on broadband, quasi-stationary noises, where phase enhancement is expected to have maximum benefit. Training phase gradient estimators on estimated speech spectra is additionally beneficial here. Our method even improves the performance of the phase-aware approach, indicating its feasibility as a generic post-processor for speech enhancement.

***Index Terms*—** DNN-based speech enhancement, phase estimation, CRUSE, phase derivatives, phase-aware speech enhancement

1. INTRODUCTION

Phase estimation of the underlying speech signal is a challenging problem in single-channel speech enhancement, because of the lack of discernible structure in phase spectra and the added complication of the 2π periodicity. Classical speech enhancement in the short-time Fourier transform (STFT) domain, therefore, focussed on the estimation of the clean speech amplitude and retained the phase of the noisy input. Typically this was justified by assuming a uniform distribution for phase in $[0, 2\pi]$ [1], or by assuming a (complex) Gaussian distribution for the signal [2], whereby the noisy phase was the MMSE-optimal estimate. Yet, as discussed in [3], the importance of phase estimation in speech enhancement cannot be ignored.

Prior work: Interestingly, while the phase spectra do not show structure, the phase *derivatives* demonstrate clear patterns. This was first systematically studied and exploited in [4,5] to blindly estimate the phase of speech *harmonics* during voiced speech. The core idea here was to enforce the theoretically derived phase gradient across time frames, for all frequency bins containing speech harmonics, and to ensure consistency across *frequency* by factoring in the influence of the window function. However, the approach requires a robust estimate of the fundamental frequency (F_0) in each frame. Unfortunately, F_0 estimation accuracy degrades at low SNRs, rendering the approach less effective at exactly the conditions where phase enhancement would be most beneficial [6]. Secondly, the underlying assumption is that the spectrum is purely harmonic during voiced speech. However, when we have mixed excitation or spectra where higher harmonics are not integer multiples of F_0, phase reconstruction leads to a metallic-sounding output due to over-excitation.

Similar to classical approaches, DNN-based methods operating in the STFT domain either estimate a denoising mask or perform spectral mapping to clean speech amplitudes. To include phase, mask estimation has been extended into the complex domain as, e.g., the complex ideal ratio mask [7] or phase sensitive mask [8]. Spectral mapping has been similarly extended to estimate the complex coefficients of the clean speech, instead of the real-valued amplitude, in the rectangular [9] or polar form [10,11]. However, as shown in [10], the complex extensions make only small changes to the phase, compared to the magnitudes, implying that the DNNs do not sufficiently learn the distribution of the clean speech phase, but likely achieve a local optimum in the MMSE sense, as in the statistical methods.

STFT-based methods are attractive for practical applications, due to their interpretability and tunability. To solve the associated phase enhancement problem, therefore, the use of temporal and spectral phase derivatives (gradients) to estimate the phase of clean speech have come under renewed focus. Specifically interesting - as they are closely related to the denoising problem - are approaches that can estimate the phase given only the *clean* magnitude spectra. Such approaches are based on implicit relations between the spectral and temporal phase gradients and the STFT amplitude spectra. In [12] phase is retrieved in a non-causal and purely signal theoretic framework, by integrating the gradients across time and frequency. In contrast, in [13, 14], DNNs are first trained to predict the mapping from amplitude spectra of clean speech to the phase derivatives along time and frequency. At inference, the phase at any particular STFT bin is obtained by integrating the estimated phase derivatives along time or frequency. The estimated phases from each direction of integration is fused to obtain a single, consistent phase value for that STFT bin, Whereas [13] heuristically weights the estimate from each direction, in proportion to the magnitude along that path, [14] proposes an elegant, analytic solution by posing the fusion as an optimisation problem, to be solved independently at each STFT bin.

Contributions: We first note that the goal in [14] is to *synthesise* the phase of clean speech, given its *clean* STFT amplitudes. In speech enhancement, applying this synthetic phase to the enhanced speech amplitude makes the result sound unnatural compared to the target speech signal in the noisy mixture. This is the inspiration for our work. We derive a closed-form phase estimate, combining the synthetic phase with that of the enhanced speech, leading to more natural output. Secondly, whereas the phase-gradient estimators of [14] are trained on amplitude spectra of clean speech, we empirically evaluate the benefit of training these estimators in matched

conditions, using the *estimated* amplitude spectra. Lastly, we study the added value of our proposed approach in *phase-aware* enhancement approaches. This would indicate the feasibility of the approach as a generic post-processor in STFT-based speech enhancement.

The paper is organised as follows: the signal model and the baseline enhancement networks are discussed in Sec. 2, along with the concept of phase derivatives. Next, we briefly summarise the approach of [14] and derive the analytic solution for the phase estimate in the speech enhancement context (Sec. 3). Following this, we evaluate the proposed approach on the DNS challenge test set and, thereby, also draw conclusions on the questions raised previously. The key take-aways are summarised in the conclusion.

2. STFT DOMAIN SPEECH ENHANCEMENT

Assuming an additive mixing model at the microphone, the observed signal is represented in the STFT domain as:

$$Y(l,m) = H(m)S(l,m) + N(l,m) \tag{1}$$

where the clean speech $S(l,m)$ is degraded by the background noise $N(l,m)$ and possible reverberation introduced by the room transfer function $H(m)$. The l and m are the frame index and the frequency bin index, respectively. Speech enhancement is obtained by estimating a time-frequency (TF) mask (or *gain*) $G(l,m)$ which, applied to $Y(l,m)$, yields the clean speech spectrum estimate:

$$\widehat{S}(l,m) = G(l,m)Y(l,m)\,, \tag{2}$$

from which the time-domain signal $\widehat{s}(k)$ is obtained by inverse Fourier transform and overlap-add. The estimated mask, $G(l,m)$ can be real-valued, or complex-valued in the case of phase-aware extensions. In the former case, the noisy phase is used for the final speech estimate.

2.1. DNN baselines: CRUSE and Complex CRUSE (C-CRUSE)

Convolutional, recurrent encoder-decoder networks with skip connections (commonly called UNets) are widely used for single- and multi-channel speech enhancement in the STFT domain as they offer a good balance between computational efficiency and performance. Specifically, we select CRUSE [15] as the baseline for predicting the real-valued TF mask from the noisy magnitude.

For the *phase-aware* baseline, we extend CRUSE as follows: the input features are formed by concatenating the real and imaginary parts of the noisy spectrogram along the channel dimension. Two output channels are obtained containing, respectively, the real and imaginary part of the complex mask. The final mask is then: $G(l,m) = G_R(l,m) + jG_I(l,m)$. To allow phase to be modelled in the full range $[0, 2\pi]$, the hyperbolic tangent activation function is used in the final layer. The loss function is identical to CRUSE.

2.2. Phase derivatives

In the polar form, the complex spectrogram, $S(l,m)$, of clean speech can be written in terms of the amplitude $A(l,m)$ and phase $\Phi(l,m)$ as $S(l,m) = A(l,m)\exp\big(j\Phi(l,m)\big)$. The phase derivative along frequency and time is then approximated as:

$$\Delta_f\Phi(l,m) = \Phi(l,m) - \Phi(l,m-1) \quad \text{and} \tag{3}$$

$$\Delta_t\Phi(l,m) = \Phi(l,m) - \Phi(l-1,m)\,. \tag{4}$$

As the STFT is computed on windowed, overlapped, time-segments, there is an additional offset term in the phase that is proportional to the frame shift (Q). Because of the 2π periodicity, this offset can distort the structure in the temporal phase difference. To avoid this, [5] proposes to modulate the STFT into the baseband. If M is the frame length, the baseband-modulated phase is given by:

$$\Psi(l,m) = \Phi(l,m) + \psi_0(l,m)\,, \tag{5}$$

with $\psi_0(l,m) \equiv -2\pi lm\frac{Q}{M}$. The *baseband* phase difference is then:

$$\Delta_t\Psi(l,m) = \Psi(l-1,m) - \Psi(l,m)\,. \tag{6}$$

Note that, given the baseband phase difference $\Delta_t\Psi(l,m)$, it is easy to compute $\Delta_t\Phi(l,m)$. Finally, if either $\Phi(l,m-1)$ or $\Phi(l-1,m)$ are available and the corresponding phase differences from (3) resp. (4) can be estimated, $\Phi(l,m)$ can be computed.

3. PHASE ESTIMATION

3.1. Estimating $\Delta_f\Phi(l,m)$ and $\Delta_t\Psi(l,m)$

As shown in [12], phase derivatives are connected to the log magnitude spectra. While it is possible to estimate these derivatives analytically, it would require limiting assumptions on the evolution of phase across time and frequency and exhibit the same drawbacks as [5]. In contrast, data-driven methods [13, 14], where DNNs are utilised to learn this relationship, offer more robust estimates.

Two separate UNets, each with one fully connected bottleneck layer and 1×1 add-skip connection, are used to predict the two phase differences, respectively. Training targets are the phase differences of clean speech, $\Delta_f\Phi(l,m)$ and $\Delta_t\Psi(l,m)$. Due to the periodic nature of phase, cosine loss functions, defined below, are adopted.

$$\mathcal{L}_f = \sum_{l,m}\Big(1-\cos\Big(\widehat{\Delta_f\Phi}(l,m) - \Delta_f\Phi(l,m)\Big)\Big) \tag{7a}$$

$$\mathcal{L}_t = \sum_{l,m}\Big(1-\cos\Big(\widehat{\Delta_t\Psi}(l,m) - \Delta_t\Psi(l,m)\Big)\Big) \tag{7b}$$

3.2. Phase retrieval from clean speech amplitudes

As independent networks are employed to predict the phase differences across time and frequency, we obtain two estimates of $\Phi(l,m)$ - one for each integration path. A final, consistent phase estimate is obtained by fusing the individual results. This is formulated very elegantly in [14] as an optimisation problem, allowing $\Phi(l,m)$ to be computed recursively and in an analytical manner, once the phase differences themselves are estimated. We briefly summarise this, using the same notation as [14], before deriving our extension.

Define $V(l,m) \equiv \frac{S(l,m)}{S(l-1,m)}$, which, clearly, is linked to the temporal phase difference. Inserting $\widehat{\Delta_t\Phi}(l,m)$, and the *known* clean speech amplitudes $A(l,m)$, we obtain an estimate of $V(l,m)$ as:

$$\widehat{V}(l,m) = \frac{A(l,m)}{A(l,m-1)}\exp\big(j\,\widehat{\Delta_t\Phi}(l,m)\big). \tag{8}$$

Denote by $\boldsymbol{z}_l = \big[z(l,0), z(l,1), \ldots, z(l,M')\big]^T$, the complex spectrum *estimate* of clean speech for the $M' = M/2$ positive frequencies of frame l. The reason for using $z(l,m)$ instead of $S(l,m)$ will become clear presently. Given the clean speech spectral estimate $\widehat{\boldsymbol{S}}_{l-1} = [\widehat{S}(l-1,0), \widehat{S}(l-1,1), \ldots, \widehat{S}(l-1,M')]^T$ of the prior frame, $\boldsymbol{z}_l$ can be obtained by minimising:

$$\mathcal{J}_t(\boldsymbol{z}_l, \widehat{\boldsymbol{S}}_{l-1}, \widehat{\boldsymbol{V}}_l) = \big\|\boldsymbol{z_l} - \widehat{\boldsymbol{V}}_l \odot \widehat{\boldsymbol{S}}_{l-1}\big\|^2_{\boldsymbol{\Lambda}_l}\,. \tag{9}$$

$\widehat{\boldsymbol{V}}_l = [\widehat{V}(l,0), \widehat{V}(l,1), \ldots, \widehat{V}(l,M')]^T$, and $\odot$ is the Hadamard product. The notation $\big\|\boldsymbol{e}\big\|^2_{\boldsymbol{\Lambda}}$ is the weighted inner product: $\boldsymbol{e}^H\boldsymbol{\Lambda}\boldsymbol{e}$.

To obtain $\boldsymbol{z}_l$ from the phase gradient across *frequency*, we define $U(l,m) \equiv \frac{S(l,m)}{S(l,m-1)}$. Given $\widehat{\Delta_f\Phi}(l,m)$, we estimate $U(l,m)$ as:

$$\widehat{U}(l,m) = \frac{A(l,m)}{A(l,m-1)}\exp\big(j\,\widehat{\Delta_f\Phi}(l,m)\big). \tag{10}$$

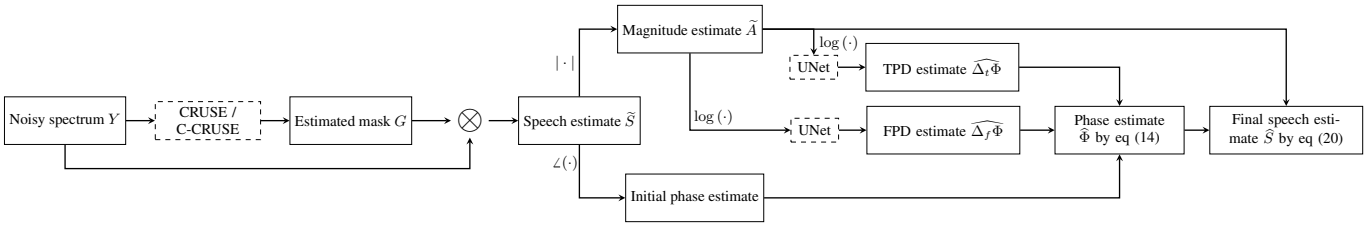

Fig. 1: Block diagram of the proposed speech enhancement system with phase reconstruction. Dashed boxes represent neural networks whereas solid boxes indicate data contained. The frame index l and the frequency bin index m have been dropped for conciseness.

Defining the sparse $(M'-1)\times M'$ matrix $\boldsymbol{D}_l$ as:

$$D_l(m,m') = \begin{cases} -\widehat{U}(l,m+1) & m' = m \\ 1 & m' = m+1 \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

it is clear that $\boldsymbol{z}_l$ can be estimated by minimising:

$$\mathcal{J}_f(\boldsymbol{z}_l, \widehat{\boldsymbol{U}}_l) = \left\|\boldsymbol{D}_l\boldsymbol{z}_l\right\|^2_{\boldsymbol{\Gamma}_l}. \tag{12}$$

Minimising the combined cost functions in (9) and (12) :

$$\mathcal{J}(\boldsymbol{z}_l) = \left\|\boldsymbol{z}_l - \widehat{\boldsymbol{V}}_l \odot \widehat{\boldsymbol{S}}_{l-1}\right\|^2_{\boldsymbol{\Lambda}_l} + \left\|\boldsymbol{D}_l\boldsymbol{z}_l\right\|^2_{\boldsymbol{\Gamma}_l}. \tag{13}$$

yields $\widehat{\boldsymbol{z}}_l = \left(\boldsymbol{\Lambda}_l + \boldsymbol{D}_l^H\boldsymbol{\Gamma}_l\boldsymbol{D}_l\right)^{-1}\boldsymbol{\Lambda}_l\left(\widehat{\boldsymbol{V}}_l \odot \widehat{\boldsymbol{S}}_{l-1}\right)$, which is the jointly optimal estimate. From this, the phase estimate is obtained as:

$$\widehat{\Phi}(l,m) = \angle\widehat{z}(l,m). \tag{14}$$

where $\angle z$ calculates the phase of a complex coefficient z. This is combined with the amplitude $A(l,m)$ to yield $\widehat{S}(l,m)$.

The weights $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$ indicate the reliability of the two predicted phase differences. Since the network prediction accuracy is related to the spectral magnitude, it is proposed in [14] to define diagonal weighting matrices as:

$$\boldsymbol{\Lambda}_l(i,i) = (A(l-1,i)A(l,i))^p \quad \text{and} \tag{15}$$

$$\boldsymbol{\Gamma}_l(i,i) = \gamma\cdot(A(l,i)A(l,i+1))^p, \tag{16}$$

where p is the magnitude compression factor, and γ is the extra factor to balance two different estimates.

3.3. Phase reconstruction for speech enhancement

When applying the above approach for estimating the phase for speech enhancement, two points should be considered: first, the magnitude spectra used for obtaining the phase gradients and estimating the phase are obtained from the preceding speech enhancement system. Thus, they are imperfect and possibly contain artefacts due to residual noise and speech distortion. Hence, it might be advantageous to train the DNNs for phase gradient estimation on these estimated speech amplitudes. Second, using the purely synthetic estimate from (14) leads to an output that sounds unnatural compared to the speech in the noisy mixture. Therefore, to obtain natural-sounding audio, the initial phase available from the speech enhancement stage should be incorporated in the phase estimator. We propose to do this by including an additional term in the cost function in (13), which penalises large deviations from the estimated speech spectrum obtained after the speech enhancement stage. Denote by $\widetilde{\boldsymbol{S}}_l = [\widetilde{S}(l,0), \widetilde{S}(l,1), \ldots, \widetilde{S}(l,M')]^T$ the enhanced speech at frame l, from the baseline speech enhancement. The *spectral deviation* cost to be added to (13) can be expressed as:

$$\mathcal{J}_s(\boldsymbol{z}_l, \widetilde{\boldsymbol{S}}_l) = \left\|\boldsymbol{z}_l - \widetilde{\boldsymbol{S}}_l\right\|^2_{\boldsymbol{\Omega}_l}. \tag{17}$$

Since only the current frame is relevant to this distance, we propose to construct $\boldsymbol{\Omega}_l$ as the diagonal matrix:

$$\boldsymbol{\Omega}_l(i,i) = \omega(\widetilde{A}(l,i))^{2p}, \tag{18}$$

consistent with the definition of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Gamma}$. Further ω is a hyperparameter to adjust the contribution of this cost component. This leads to the following estimate of $\boldsymbol{z}_l$ in the context of speech enhancement:

$$\widehat{\boldsymbol{z}}_l = \left(\boldsymbol{\Lambda}_l + \boldsymbol{D}_l^H\boldsymbol{\Gamma}_l\boldsymbol{D}_l + \boldsymbol{\Omega}_l\right)^{-1}\left(\boldsymbol{\Lambda}_l\left(\widehat{\boldsymbol{V}}_l \odot \widehat{\boldsymbol{S}}_{l-1}\right) + \boldsymbol{\Omega}_l\widetilde{\boldsymbol{S}}_l\right) \tag{19}$$

Computing the enhanced phase as in (14), we obtain the final clean speech estimate as:

$$\begin{aligned}\widehat{S}(l,m) &= \left|\widetilde{S}(l,m)\right|\exp\left(j\,\widehat{\Phi}(l,m)\right) \\ &= \left|G(l,m)Y(l,m)\right|\exp\left(j\,\widehat{\Phi}(l,m)\right)\end{aligned} \tag{20}$$

The system is summarised by the block diagram in Fig. 1.

4. EXPERIMENTAL EVALUATION AND DISCUSSION

For CRUSE and C-CRUSE, we adopt the four layer encoder-decoder structure with grouped GRUs (true to [15]), to predict the mask for the initial noise reduction and dereverberation. The networks were trained on the DNS challenge 2021 wideband dataset [16]. We synthesised 140 hours of training data from English speech, with 50% of them in reverberant conditions (T_{60} in the range 0.3 s – 1.3 s). The SNR of the training set was varied between -5 dB and 20 dB. Audio was sampled at 16 kHz. The STFT employs 75% overlapped frames and a square-root Hann window of length $M = 512$ for analysis & synthesis. The enhanced speech signals were evaluated by segmental SNR [17], STOI [18], and DNSMOS P.835 [19].

The UNets for phase gradient estimation comprise three convolutional layers in the encoder and decoder, respectively. Kernels of dimension 2×3 (time, frequency), and strides of 1×2 were used at all layers. The number of channels of each layer were: $16, 32, 32$, which resulted in a 992 unit fully connected layer at the bottleneck. All convolutional layers were followed by the leaky ReLU function with an $\alpha = 0.003$ negative slope. Two sets of UNets were trained: 1) the first set, as originally proposed for phase retrieval, learn the relationship between the *clean* magnitude spectra and the phase derivatives. They are agnostic of the speech enhancement (SE) stage; 2) the second set is adapted to the speech enhancement context, and learn to predict the phase derivatives of the clean speech from the *estimated* magnitude spectra. Both approaches have unique potential advantages: enhancement-agnostic networks need no retraining when switching other components in the pipeline, while networks

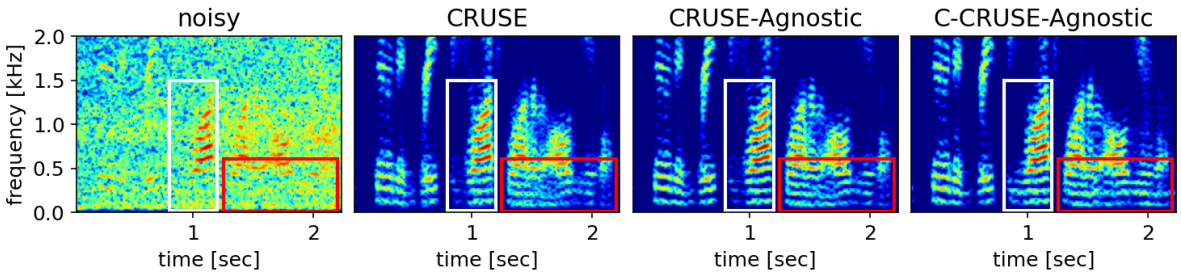

Fig. 2: Comparison of the denoised signals by CRUSE and with the proposed phase reconstruction. Noisy signal: Street noise, -2 dB. Note the clearer harmonic structure after phase reconstruction by the proposed method in the highlighted area. C-CRUSE in combination with phase reconstruction even manages to pick up very weak harmonic structure (white box) and gives a more continuous harmonic spectrum (red box)

trained specific to a certain speech enhancement system might provide better performance, due to matched conditions. We denote the networks as 'SE-Agnostic' and 'SE-Matched' in the sequel.

Optimal values for the compression factor p and the weights γ and ω, were obtained by grid search. Based on the DNSMOS scores on a subset of development data of the DNS2020 challenge [20], with wideband, quasi-stationary noise, the optimal parameters were $[p = 0.3, \gamma = 10, \omega = 5]$ for SE-Agnostic, and $[p = 0.5, \gamma = 10, \omega = 5]$ for SE-Matched.

4.1. Results & Discussion

The optimised system is evaluated on the DNS2021 synthetic test set [16]. Averaged metrics are given in Tab. 1. Compared to the noisy input, a significant improvement is provided by all approaches, and on all metrics. As an *upper bound* on achievable performance, we also present results where the *oracle, clean* phase is used with the estimated speech amplitudes. Comparing CRUSE and its complex extension, we see a marginal benefit of estimating the phase in the enhancement stage - in line with previous works.

More importantly, the proposed phase enhancement improves all quality metrics, compared to *both* corresponding baselines (CRUSE & C-CRUSE). Only STOI is constant for all approaches. This is expected: phase enhancement should not affect the speech *envelope*, which is important for intelligibility. This *sanity check* ensures we do not improve quality at the cost of intelligibility. We also see that the phase-enhanced outputs are *comparable* to using oracle phase – a pleasing result.

Table 1: Averaged instrumental metrics on test set. Best results in bold. Phase enhancement consistently improves all metrics compared to baselines and is comparable to using *oracle* phase.

Method	segSNR [dB]	STOI	DNSMOS	
			OVRL	SIG
Noisy	6.87	0.87	2.53	3.33
CRUSE	13.74	0.93	3.10	3.36
CRUSE-Agnostic	14.30	0.93	3.17	3.43
CRUSE-Matched	14.19	0.93	3.17	3.44
C-CRUSE	13.92	0.93	3.14	3.40
C-CRUSE-Agnostic	**14.45**	0.93	**3.20**	**3.45**
CRUSE-OraclePhase	14.51	0.94	3.17	3.43
C-CRUSE-OraclePhase	14.77	0.94	3.20	3.45

We expect the proposed phase reconstruction to offer maximum benefit with stationary, broadband noise conditions, as such noise typically results in vocoding artefacts between the harmonics after speech enhancement – which phase reconstruction can ameliorate. In such cases, we expect differences between the various configurations to be more evident. Thus, we split the test set into two subsets: **a)** mixtures with stationary or short-term stationary noise, such as car, traffic, babble; **b)** mixtures with sparse, transient noise, such as footsteps, typing, etc. The distribution of the DNSMOS scores are shown in Fig. 3 for both subsets. We now see that on *subset a*, SE-Matched has a bigger margin over the SE-Agnostic. When the noise is less stationary and sparse (*subset b*), using SE-Agnostic is better. We reason that in such cases there are fewer contiguous regions where the speech and noise overlap. Then, SE-Agnostic networks, being trained on clean speech, yield more accurate phase estimates. The averaged results in Tab. 1 may indicate only a small achievable improvement by using the proposed phase reconstruction on CRUSE/C-CRUSE. However, Fig. 3 shows that phase reconstruction manages to boost the signal quality in poor SNR conditions – reflected by the decreased spread and higher minimum in the scores!

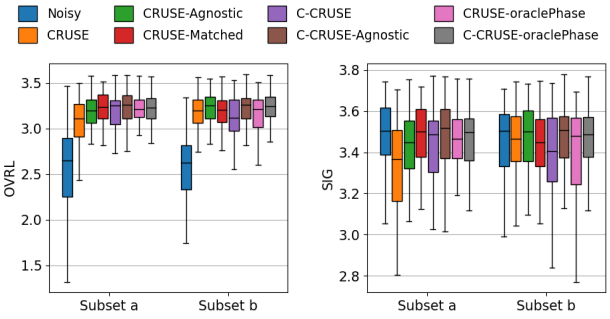

Fig. 3: DNSMOS score distribution, separately on broadband/quasi-stationary and transient/sparse noise subsets from DNS 2021 test set.

5. CONCLUSIONS

We proposed a phase reconstruction method for speech enhancement, based on phase gradients. Using independent DNNs to predict spectral and temporal phase derivatives from the estimated amplitude spectra (from a preceding speech enhancement stage), we obtain two estimates of the phase. A closed-form, analytic solution was derived to fuse these estimates in an MMSE-optimal manner. We further introduced an additional cost term that incorporated phase information present in signal after the speech enhancement stage – which led to a more natural-sounding output. Experimental results validate the quality improvement brought by the proposed phase enhancement - with the performance of the proposed method being comparable to using oracle phase. The proposed phase estimator is also beneficial when used with phase-aware speech enhancement, indicating its feasibility as a generic post-processor in STFT-based speech enhancement frameworks. Lastly, training the phase derivative estimator DNNs specific to the preceding speech enhancement stage is beneficial when noise is (short-term) stationary and broadband. For sparse, transient noises, training the DNNs on clean-speech spectra gives more accurate results. Audio samples can be found at `https://aspireugent.github.io/diff-based-phase-reconstruction-SE/`.

6. REFERENCES

[1] P. Vary and R. Martin, *Digital speech transmission: Enhancement, coding and error concealment*, John Wiley & Sons, 2006.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

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

[4] T. Gerkmann, M. Krawczyk, and R. Rehr, "Phase estimation in speech enhancement—unimportant, important, or impossible?," in *IEEE 27th Convention of Electrical and Electronics Engineers in Israel*. IEEE, 2012, pp. 1–5.

[5] 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, 2014.

[6] P. Vary, "Noise suppression by spectral magnitude estimation -mechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387–400, 1985.

[7] 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, 2015.

[8] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[9] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[10] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 9458–9465.

[11] A. Li, C. Zheng, G. Yu, J. Cai, and X. Li, "Filtering and refining: A collaborative-style framework for single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2156–2172, 2022.

[12] Z. Průša, P. Balazs, P. Søndergaard, and L. Peter, "A noniterative method for reconstruction of phase from STFT magnitude," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1154–1164, 2017.

[13] L. Thieling, D. Wilhelm, and P. Jax, "Recurrent phase reconstruction using estimated phase derivatives from deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7088–7092.

[14] K. Nagatomo Y. Masuyama, K. Yatabe and Y. Oikawa, "Online phase reconstruction via DNN-based phase differences estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 163–176, 2022.

[15] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 656–660.

[16] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 deep noise suppression challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6623–6627.

[17] K. Eneman, A. Leijon, S. Doclo, A. Spriet, M. Moonen, and J. Wouters, "Auditory-profile-based physical evaluation of multi-microphone noise reduction techniques in hearing instruments," in *Advances in Digital Speech Transmission*, pp. 431 – 458. John Wiley & Sons Ltd., New York, 2008.

[18] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.

[19] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2022.

[20] C. K. Reddy, V. Gopal, R. Cutler, R. Cheng E. Beyrami, H. Dubey, S. Matusevych, A. Aazami R. Aichner, S. Braun, et al., "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.