Beyond Clean Phase: Using Silence-Generating Phase for DNN-Based Speech Enhancement

Lars Thieling and Peter Jax
Institute of Communication Systems (IKS)
RWTH Aachen University, Germany
{thieling,jax}@iks.rwth-aachen.de

***Abstract*—Speech enhancement algorithms usually operate in the short-time Fourier transform (STFT) domain and only enhance the magnitude spectrum, while adopting the noisy phase for synthesis. This is because the phase has often been considered unimportant. However, recent findings have proven otherwise, leading to an improved enhancement by considering the phase either implicitly or explicitly. In this paper, we propose a phase-aware extension of our recently published two-stage speech enhancement approach. It comprises, among other improvements, an additional explicit phase estimation stage whose structure is inspired by the fundamental ideas of our work on phase reconstruction. Unlike most phase-aware approaches, we do not estimate the clean phase but propose a novel combined consistent-inconsistent phase (CIP). It corresponds to a silence-generating phase for the noise-dominated time-frequency (TF) parts and thus allows noise reduction without modifying the magnitude spectrum at all. We show that this new CIP can provide a significant performance improvement compared to the clean phase. Experimental results confirm the effectiveness of our proposed extensions, ultimately leading to improved speech quality (PESQ, DNSMOS) and speech distortion (segmental SNR).**

***Index Terms*—Speech enhancement, phase estimation, signal reconstruction, neural network**

I. Introduction

In many applications, such as speech communication or automatic speech recognition, speech signals are often degraded by background noise, resulting in reduced perceived quality and intelligibility. This problem is usually solved by using signal processing or deep learning-based speech enhancement algorithms that aim to extract the target speech from the single-channel noisy mixture. These algorithms typically operate in the short-time Fourier transform (STFT) domain by applying the discrete Fourier transform (DFT) on windowed segments of the input signal. In this STFT domain, they, e.g., apply direct spectral mapping or time-frequency (TF) masking [1]–[3]. Conventionally, the enhancement mainly focused on improving the magnitude spectrum because the phase was generally considered hard to estimate [4] and less important compared to the magnitude [5], [6]. However, since the distorted phase is reused for signal reconstruction, the overall quality is limited.

With the increase in computational power of speech communication devices and the recent advances in deep learning, there has been an increasing number of studies that also consider the phase for speech enhancement (see [7]). For example, some studies apply phase reconstruction after enhancing the magnitude spectrum [8]. More recent deep learning-based approaches estimate the phase implicitly by using a loss in the complex spectrum or waveform domain [1], [2], [9]. Some studies also add an explicit phase loss term or use an explicit subnetwork for phase estimation [4], [10]. In this paper, we propose a novel modular phase-aware deep speech enhancement approach consisting of three stages: A *mask estimation stage* generates a real-valued mask that is used as prior information in the subsequent *magnitude* and *phase estimation stages*. Specifically, we propose a phase-aware extension of our recently published two-stage speech enhancement approach [3] by exploiting key ideas of our deep neural network (DNN)-based phase reconstruction [11].

II. Contributions and Relation to Previous Work

We propose four major modifications to our previous work: 1) preparation of the system for real-time application, 2) an updated loss function for the masking model, and 3) a new phase estimation stage that estimates 4) a novel combined consistent-inconsistent phase (CIP).

For our original speech enhancement approach, no real-time requirements have been imposed, so it can only be used for offline applications. However, for applications such as telecommunications, real-time capability is crucial. Therefore, we apply causal convolutions instead of standard convolutions in our neural networks and limit the lookahead in the magnitude and phase estimation stages to two frames, corresponding to $2 \cdot 5\,\mathrm{ms} = 10\,\mathrm{ms}$ (modification 1).

In our previous work, we used the mean squared error (MSE) as loss for the mask estimation model. However, the MSE is only indirectly related to the achievable performance since it ignores the influence of the mask on the magnitude spectrum. As a remedy, we update the loss by a magnitude spectrum approximation (MSA) that optimizes the achieved magnitude spectrum after application of the mask (modification 2).

So far, our system only enhanced the magnitude while adopting the distorted phase for synthesis, which limits the achievable performance. For this reason, we propose a novel phase estimation stage (modification 3). As mentioned earlier, there exist approaches that consider phase estimation either implicitly or explicitly. However, most of these approaches consider instantaneous phase differences in each TF entry separately. They neglect the continuity of neighboring entries, although, as we will see in this work, the phase differences of overlapping frames play a particularly crucial role (see

Sec. III). Hence, similar to [11], in this work we first estimate the phase differences of adjacent time frames and frequency bins, and then apply a causal phase reconstruction method adapted for speech enhancement which combines the estimated phase differences.

Most speech enhancement approaches try to directly estimate the clean spectrum, along with the clean phase. However, in TF regions without speech, estimating the clean phase is not meaningful. In particular, if the associated enhanced magnitude spectrum in this region is not accurately estimated, i.e., is non-zero, this may leave residual noise. We therefore propose to estimate a novel combined consistent-inconsistent phase (CIP) instead (modification 4). While it corresponds to the clean phase in speech-dominated regions, it uses a silence-generating phase for the noise-dominated parts (see Sec. III).

III. Silence-Generating Phase

A common task in speech processing is to reconstruct a signal solely from a given magnitude spectrogram. The Griffin-Lim algorithm [12] is frequently used for this purpose. It aims to find the closest consistent STFT spectrogram using iteratively alternating STFT and inverse STFT (ISTFT) operations, where the term "consistent spectrogram" refers to a spectrogram that can be obtained as the STFT spectrogram of a time-domain signal. Le Roux et al. [13], [14] have modified this algorithm to produce maximally inconsistent spectrograms that can be resynthesized through ISTFT as silence, assuming that the same window is used for analysis and synthesis. As already mentioned by Le Roux [14], the problem of finding a maximally inconsistent spectrogram becomes trivial for the case of a rectangular window with an overlap of, e.g., 50% or 75%: it suffices to add π to the phase of every other frame. As will be shown below, this trivial solution can also be applied for window functions $\mathrm{w}(k)$ with perfect reconstruction, i.e., for those satisfying the Princen-Bradley criterion [15]

$$\mathrm{w}^2\big(k+\tfrac{L}{2}\big)+\mathrm{w}^2(k)=1 \text{ for } k=0,1,\ldots,\tfrac{L}{2}-1 \quad (1)$$

when using a frame length L and frame shift R such that

$$L/R \in \left\{4\kappa \,|\, \kappa \in \mathbb{N}^+\right\}. \quad (2)$$

To show this, we first consider the perfect cancellation criterion for the weighted overlap-add synthesis method [12]

$$x'(k)=\frac{\sum_{\lambda=-\infty}^{\infty}\mathrm{w}(k-\lambda R)\,\tilde{y}_{\mathrm{w}}(\lambda,k)}{\sum_{\lambda=-\infty}^{\infty}\mathrm{w}^2(k-\lambda R)} \stackrel{!}{=} 0, \quad (3)$$

where λ is the frame index, $\tilde{y}_{\mathrm{w}}(\lambda,k) = \mathrm{IDFT}(Y_{\mathrm{w}}(\lambda,\mu)\cdot \mathrm{e}^{\mathrm{j}\pi\lambda}) = \mathrm{w}(k-\lambda R)\,y(k)\,(-1)^{\lambda}$ with $Y_{\mathrm{w}}(\lambda,\mu)$ being the STFT of a given time signal $y(k)$, and $\mathrm{w}(k)$ denotes the analysis/synthesis window. Thus, we obtain perfect cancellation independent of $y(k)$ if

$$\sum_{\lambda=\lambda_0}^{\lambda_0+\frac{L}{R}-1}\mathrm{w}^2(k-\lambda R)\,(-1)^{\lambda}\stackrel{!}{=}0 \quad (4)$$

holds for the overlap region of each successive L/R frames, i.e., $k=k_0,k_0+1,\ldots,k_0+R-1$ with $k_0=\lambda_0 R+L-R$.

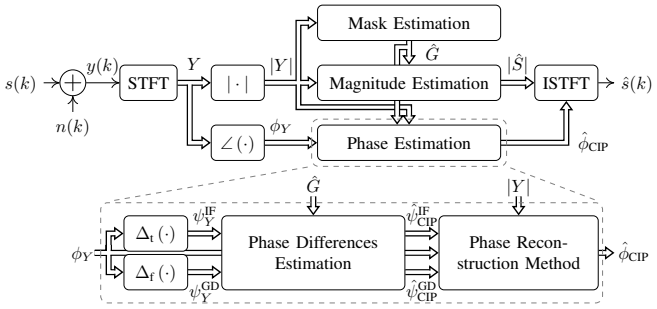

Fig. 1. Block diagram of the proposed speech enhancement system. The frame index λ and frequency bin index μ of all quantities in the TF domain are omitted for simplicity.

Assuming that (1) and (2) are satisfied, we can rewrite the left part of (4) as

$$\sum_{\lambda=\lambda_0}^{\lambda_0+\frac{L}{2R}-1}\mathrm{w}^2(k-\lambda R)\left((-1)^{\lambda}-(-1)^{\lambda+\frac{L}{2R}}\right)+(-1)^{\lambda+\frac{L}{2R}}, \quad (5)$$

which is zero since $(-1)^{\lambda+\frac{L}{2R}}=(-1)^{\lambda}$.

Thus, if L and R satisfy (2) and the window function satisfies (1), we can define a silence-generating phase for a given consistent spectrogram $X(\lambda,\mu)=\mathrm{STFT}(x(k))$ as

$$\tilde{\phi}_X(\lambda,\mu) \coloneqq \phi_X(\lambda,\mu)+\pi\lambda, \quad (6)$$

where $\phi_X(\lambda,\mu)=\angle X(\lambda,\mu)$ is the phase spectrum of X.

IV. Proposed System Architecture

Our proposed speech enhancement system is depicted in Fig. 1. It aims to extract the target speech $s(k)$ from the noisy mixture $y(k)=s(k)+n(k)$ using three estimation stages. Similar to our previous work [3], a spectral mask $\hat{G}$ is estimated initially, which is then used to estimate the denoised magnitude spectrum $|\hat{S}|$ in another stage. Moreover, we propose a novel phase estimation stage that receives the noisy phase $\phi_Y=\angle Y$ in addition to $|Y|$ and $\hat{G}$. Inspired by our work on phase reconstruction [11], this stage consists of a network estimating phase differences followed by a phase reconstruction method.

Each stage contains an individually trained neural network whose general architecture is shown in Table I. It is similar to that in our previous work [3] but uses causal convolutions. The CConv2D layers consist of a 2D causal convolution followed by a batch normalization and an individual activation function, which depends on the considered network. For a given CConv2D(F, 3×3, 2) layer, F specifies the number of filters, 3×3 is the kernel size, and 2 corresponds to the stride along the feature axis. Each ResBlock(F, 3×3) layer consists of a skip connection and two stacked CConv2D(F, 3×3, 1) layers, the second of which always uses a linear activation.

A. Mask Estimation Stage

Similar to our previous work, the network used to estimate the mask from the noisy magnitude has an encoder-decoder architecture according to Table I with $C_{\mathrm{in}}=1$ input channel, $C_{\mathrm{out}}=1$ output channel, $F_{\mathrm{init}}=F/4$, $F=64$, and $B=8$. This results in a height of $H_{\mathrm{res}}=H/4$ in the residual blocks. Except for the linear unit in the second CConv2D of each

TABLE I
COMMON NETWORK ARCHITECTURE WHERE THE DOWN- AND UPSAMPLING LAYERS (HIGHLIGHTED IN GRAY) ARE ONLY USED IN THE MASK ESTIMATION. CCONV2D CAN BE "STANDARD" OR "GATED".

	Layer	Activation size
	Input	$H \times$ None $\times C_{\text{in}}$
	CConv2D(F_{init}, 9×9, 1)	$H \times$ None $\times F_{\text{init}}$
	CConv2D($F/2$, 3×3, 2)	$H \times$ None $\times F/4$
	CConv2D(F, 3×3, 2)	$H/2 \times$ None $\times F/2$
$B\times$	ResBlock(F, 3×3)	$H_{\text{res}} \times$ None $\times F$
	$\vdots$	$\vdots$
	ResBlock(F, 3×3)	$H_{\text{res}} \times$ None $\times F$
	Upsampling $(2, 1)$	$H/2 \times$ None $\times F/2$
	CConv2D($F/2$, 3×3, 1)	$H/2 \times$ None $\times F/2$
	Upsampling $(2, 1)$	$H \times$ None $\times F/2$
	CConv2D($F/4$, 3×3, 1)	$H \times$ None $\times F/4$
	CConv2D(C_{out}, 9×9, 1)	$H \times$ None $\times C_{\text{out}}$

ResBlock layer and the sigmoid function in the output layer, the exponential linear unit (ELU) is adopted as activation function for all CConv2D layers.

Instead of using the MSE on the masks, we propose to use the magnitude spectrum approximation (MSA), i.e.,

$$\mathcal{L}_{\text{MSA}} = \frac{1}{KQ} \sum_{\lambda=0}^{K-1} \sum_{\mu=0}^{Q-1} |\hat{G}(\lambda,\mu) \cdot |Y(\lambda,\mu)| - |S(\lambda,\mu)||^2 \quad (7)$$

as loss, where K is the number of frames and $Q = N/2 + 1$ with N being the DFT size. This way, the mask is optimized with respect to its effect on $|\hat{S}'| \mathrel{\widehat{=}} \hat{G} \cdot |Y|$, which establishes a more direct relation to its noise reduction performance.

B. Magnitude Estimation Stage

As proposed in [3], $\hat{G}$ is not used directly for noise reduction, but as prior information in the magnitude estimation model that delivers the enhanced magnitude spectrum $|\hat{S}|$. That is, $\hat{G}$ provides a rough estimate of where speech and noise dominate, and the magnitude estimation model can process these regions differently and focus on local structures in the spectrum. To this end, the model uses the network architecture described in Table I with $F_{\text{init}} = F = 64$, $B = 2$, and omitting the down- and upsampling layers, i.e., $H_{\text{res}} = H$. Furthermore, all standard causal convolutions are replaced by gated causal convolutions. The activation functions are chosen as in the mask estimation model, only the sigmoid function in the output layer is replaced by a linear one. While we still have $C_{\text{out}} = 1$ output channel, the mask is used as an additional input channel to the network, i.e., $C_{\text{in}} = 2$. The MSE between $|\hat{S}|$ and $|S|$ is used as loss function and two additional input frames are adopted as lookahead.

C. Phase Estimation Stage

In our previous work [3], only the magnitude has been enhanced and the noisy phase has been used for synthesis, which inevitably leads to a suboptimal solution. In this paper, we propose a novel phase estimation stage using the general system architecture from our work on phase reconstruction [11]. That is, we first estimate phase differences and then apply a suitable phase reconstruction method for combination. However, instead of estimating the clean phase ϕ_S, we propose a new combined consistent-inconsistent phase (CIP)

$$\phi_{\text{CIP}}(\lambda,\mu) \coloneqq \angle\Big(G(\lambda,\mu)\, \mathrm{e}^{\mathrm{j}\phi_S(\lambda,\mu)} + (1-G(\lambda,\mu))\, \mathrm{e}^{\mathrm{j}\tilde{\phi}_Y(\lambda,\mu)}\Big), \quad (8)$$

where $G \mathrel{\widehat{=}} |S|/|Y|$ is the ideal spectral magnitude mask and $\tilde{\phi}_Y$ is the silence-generating phase for the noisy spectrum Y according to (6). Hence, ϕ_{CIP} corresponds to the clean phase in speech-dominated parts and the silence-generating phase in noise-dominated parts. Thus, noise reduction can be achieved by solely replacing the noisy phase by our proposed ϕ_{CIP}, which, as shown in Sec. V-A, leads to similar or even better results than combining the clean magnitude $|S|$ with the noisy phase spectrum ϕ_Y.

While only the magnitude spectrum of a time signal was available in our work on phase reconstruction [11], here in speech enhancement we can exploit the noisy phase ϕ_Y as another source of information. Especially for high SNRs, this can be very helpful because in this case the noisy phase corresponds approximately to the clean one. It is therefore reasonable to consider the noisy phase for enhancement. In this work, we propose to use it in two ways. First, it is used in order to extract the input features for the phase differences estimation network. Second, it is used as a kind of supporting estimate in the phase reconstruction method.

1) Phase Differences Estimation: The phase differences along time and frequency, namely the instantaneous frequency (IF) and group delay (GD), are estimated by a neural network based upon the IF and GD of the noisy spectrum Y, i.e., ψ_Y^{IF} and ψ_Y^{GD}. More precisely, as proposed in [11], the shift corrected and wrapped IFs and GDs are always employed as input and output features.[1] Besides ψ_Y^{IF} and ψ_Y^{GD}, the estimated mask $\hat{G}$ is used as a third input to our phase differences estimation network, i.e., $C_{\text{in}} = 3$, and the $C_{\text{out}} = 2$ desired phase differences $\psi_{\text{CIP}}^{\text{IF}}$ and $\psi_{\text{CIP}}^{\text{GD}}$ are output. Except for the number of input and output channels, the remaining neural network hyperparameters and the used lookahead are identical to those in the magnitude estimation from Sec. IV-B. In TF regions where the corresponding magnitude spectrum used for synthesis is close to zero, the phase is of little relevance. To account for this during training, we add a weighting to the regularized cosine loss function [11], i.e.,

$$\mathcal{L}_{\text{wreg}}^{\gamma} = \frac{1}{KQ} \sum_{\lambda=0}^{K-1} \sum_{\mu=0}^{Q-1} (|S(\lambda,\mu)| + \gamma) \cdot d_{\text{reg}}\Big(\Delta\hat{\psi}(\lambda,\mu)\Big), \quad (9)$$

where $d_{\text{reg}}(\cdot)$ denotes the element-wise distance function of the regularized cosine loss $\mathcal{L}_{\text{reg}}$ proposed in [11] and $\gamma \in \mathbb{R}$ is a parameter to adjust the estimation behavior of the networks towards consistent or inconsistent phase. For our experiments, we chose $\gamma = 0.01$ as it leads to a good compromise between consistent and inconsistent phase estimates.

2) Phase Reconstruction Method: In order to obtain a consistent phase spectrum from the estimated phase differences, an appropriate phase reconstruction method is required. For example, one could apply our proposed magnitude-weighted average (MWA) algorithm [11] using $|Y|$ for weighting. That is, the phase at a given TF entry is the weighted average of three estimates calculated from the previously estimated

[1] Note that in case of IF, the shift correction is also often referred to as baseband phase difference transformation as introduced in [16].

phase elements in its causal vicinity, i.e., it depends only on entries from the current and past frames. However, due to the recursive structure, error propagation may occur. Fortunately, we can counteract this by exploiting the noisy phase, e.g., by using it as a fourth estimate for our averaging procedure, i.e., $\varphi_4(\lambda,\mu) = \phi_Y(\lambda,\mu)$ and $\alpha_4(\lambda,\mu) = |Y(\lambda,\mu)|$ according to [11]. This adapted procedure is referred to as supported magnitude-weighted average (SMWA) in the following.

V. Experimental Results and Analysis

The noisy speech data sets were created using clean speech from the VCTK database [17]. Specifically, the speakers were carefully divided into non-overlapping training, validation, and test subsets. Considering these speaker subsets, the training/validation set was created by mixing 1000/100 randomly chosen audio files with the 38 different noise recordings from the DEMAND [18] and QUT [19] databases using SNRs of 0, 5, and 10 dB. For mixing, the SNR is only calculated in sections where speech is present. To allow for an unbiased evaluation, different non-overlapping excerpts from the noise recordings were used for both sets. The test set was created similarly, using 100 randomly selected clean speech files, four unseen noise recordings (car, crossroads, office, and pub noise) from the ETSI database [20], and SNRs of 0, 5, and 10 dB. In total, this resulted in 112.4 h of data in the training, 10.6 h in the validation, and 1.2 h in the test set.

All audio files were resampled to $f_s = 16\,\text{kHz}$. For the STFT and the ISTFT, an $L = 320$ samples square-root Hann window with 75% overlap, i.e., $R = 80$ samples window shift, and $N = 320$ frequency bins was used such that all conditions for the silence-generating phase (see Sec. III) were satisfied. For the encoder-decoder architecture comprising the down- and upsampling layers shown in Table I, divisibility of the input dimension H by 4 is required. Therefore, the feature axes of all inputs to the neural networks were padded by reflection to $H = 164$. During training, the number of input frames was fixed to 320 to get mini-batches of constant dimension.

All models were trained at least 10 epochs on the shuffled training data, using an Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a learning rate of 0.001, and a batch size of 8. The models were evaluated on the validation set in the middle and at the end of each epoch so that the best iteration could be selected based on the minimum validation loss after training.

We employed segmental SNR (SNR_{seg} in dB) [21], wideband PESQ [22], STOI [23], and DNSMOS [24] as our quality measures. While the first three measures are intrusive, i.e., the clean signal is available for comparison, DNSMOS is non-intrusive, using the noisy signal for evaluation only.

A. Theoretically Achievable Performance

In order to demonstrate the efficiency of the proposed CIP, we first examine the theoretically achievable scores, i.e., without considering any kind of estimation. The scores calculated on the test set using different combinations of magnitude and phase spectra are reported in Table II. If the magnitude spectrum would be perfectly estimated, i.e., $|\hat{S}| = |S|$, the clean phase ϕ_S naturally achieves the best results. Using ϕ_Y or our oracle ϕ_{CIP} according to (8), the scores degraded only slightly. Assuming that the magnitude in noise reduction is usually not perfectly estimated, these differences are likely negligible.

TABLE II
Theoretically achievable test scores averaged over all SNR and noise conditions when using different combinations of magnitude and phase spectra for ISTFT.

Magn.	Phase	PESQ	STOI	SNR_{seg}	DNSMOS		
					OVRL	SIG	BAK
$\|S\|$	ϕ_S	**4.64**	**1.00**	–	**3.18**	**3.49**	**4.00**
	ϕ_{CIP}	4.27	0.89	14.32	3.08	3.38	3.99
	ϕ_Y	3.97	0.92	8.38	3.07	3.39	3.96
$\|Y\|$	ϕ_S	2.06	0.76	−4.68	2.35	3.20	2.53
	ϕ_{CIP}	**4.16**	**0.93**	**11.94**	**3.02**	**3.35**	**3.92**
	ϕ_Y	1.85	0.67	−7.93	2.08	2.81	2.25

When the noisy magnitude spectrum $|Y|$ is used for reconstruction, significantly worse results are obtained for ϕ_S and ϕ_Y. The results for ϕ_{CIP}, however, remain very similar to the clean magnitude case. Due to its silence-generating behavior in the noise-dominated parts, our ϕ_{CIP} clearly outperforms the clean ϕ_S. Furthermore, it can be seen that using $|Y|$ with ϕ_{CIP} even better results compared to $|S|$ with ϕ_Y can be achieved for the intrusive measures. Contrary to the common assumption that enhancing the magnitude is more important than the phase, this result demonstrates the strong potential of phase processing in speech enhancement.

B. Overall System Evaluation

The actual achieved test scores after training the neural networks are shown in Fig. 2. In order to evaluate the influence of the different stages, we compared our proposed extended speech enhancement approach (ETSSE), including the phase estimation stage, with the approach (TSSE) that does not enhance the phase but uses the noisy phase ϕ_Y for reconstruction similar to [3]. For ETSSE, we used two different phase reconstruction methods, namely MWA from [11] and SMWA from Sec. IV-C. While $\text{TSSE}_{\text{Magn}}$ also uses $|\hat{S}|$ from the magnitude estimation stage, $\text{TSSE}_{\text{Mask}}$ uses $\hat{G}$ from the mask estimation stage for magnitude enhancement, i.e., $|\hat{S}'| \mathrel{\widehat{=}} \hat{G} \cdot |Y|$. In order to have a state-of-the-art reference algorithm, we reimplemented the implicitly phase-aware fully convolutional recurrent network (FCRN) based on [2] and trained it using a gradient norm clipping of 0.1.

On average, the best results are obtained by $\text{ETSSE}_{\text{SMWA}}$, i.e., including the phase estimation stage using the proposed SMWA. Both $\text{ETSSE}_{\text{MWA}}$ and $\text{ETSSE}_{\text{SMWA}}$ improve the obtained $\text{DNSMOS}_{\text{bak}}$ scores compared to $\text{TSSE}_{\text{Magn}}$, which is likely due to the estimation of CIP that corresponds to the silence-generating phase for noise-dominated parts. The spectral components in these TF regions, erroneously left in the magnitude estimation, are consequently attenuated by the phase enhancement. For SNR_{seg}, a significant degradation can be observed using MWA for reconstruction, i.e., $\text{ETSSE}_{\text{MWA}}$, which can be explained by the phase error propagation in the recursive structure of MWA. Using the noisy phase as prior information in the proposed SMWA, i.e., $\text{ETSSE}_{\text{SMWA}}$, counteracts this problem and leads to improved SNR_{seg} values

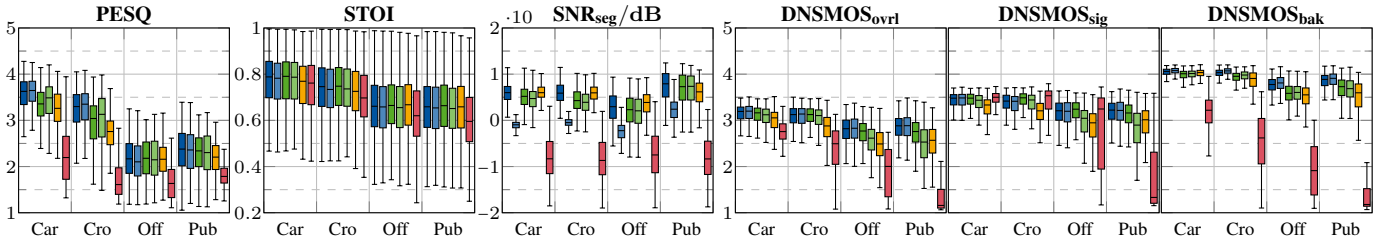

Fig. 2. Achieved results for the different noise signals (car, crossroads, office, pub) from the test set: $\text{ETSSE}_{\text{SMWA}}$ (■; proposed), $\text{ETSSE}_{\text{MWA}}$ (■; proposed), $\text{TSSE}_{\text{Magn}}$ (■; similar to [3]), $\text{TSSE}_{\text{Mask}}$ (■; similar to [3]), FCRN (■; based on [2]), Noisy (■). Outliers are not shown for visual clarity.

compared to $\text{TSSE}_{\text{Magn}}$. Beyond that, $\text{ETSSE}_{\text{SMWA}}$ can also improve the PESQ and $\text{DNSMOS}_{\text{ovrl}}$ scores compared to $\text{TSSE}_{\text{Magn}}$, while achieving similar STOI values. Except for PESQ and $\text{DNSMOS}_{\text{bak}}$, $\text{TSSE}_{\text{Magn}}$ achieves better results than $\text{TSSE}_{\text{Mask}}$. This is similar to the findings in our previous work, but the difference between both is slightly smaller this time, likely due to the new MSA loss function in the masking model, which considers the influence of the estimated mask on the magnitude. Except for SNR_{seg}, all TSSE and ETSSE approaches achieve better results than FCRN on average. Exemplary audio samples are available online[2].

VI. Conclusions

In this paper, we presented a phase-aware extension of our recently published two-stage speech enhancement approach. Particularly, we proposed an additional phase estimation stage whose basic structure originates from our work on phase reconstruction. That is, we first estimate phase differences and then use a phase reconstruction method for combination. For both steps, we have provided suitable modifications for use in speech enhancement. In particular, we proposed to estimate a novel combined consistent-inconsistent phase (CIP), which enables noise reduction by solely modifying the phase. We showed that CIP can lead to results even beyond those obtained with the clean phase when the noisy magnitude spectrum is used for reconstruction. Furthermore, we demonstrated that our proposed phase-aware extension further improved the overall achieved results.

References

[1] Z.-Q. Wang, G. Wichern, and J. Le Roux, "On the compensation between magnitude and phase in speech separation," *IEEE Signal Processing Letters*, vol. 28, pp. 2018–2022, 2021.

[2] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Fully convolutional recurrent networks for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6674–6678.

[3] L. Thieling and P. Jax, "Two-stage speech enhancement using gated convolutions," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[4] D. Kim, H. Han, H.-K. Shin, S.-W. Chung, and H.-G. Kang, "Phase continuity: Learning derivatives of phase spectrum for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6942–6946.

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

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

[7] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

[8] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[10] T. Peer and T. Gerkmann, "Phase-aware deep speech enhancement: It's all about the frame length," *JASA Express Letters*, vol. 2, no. 10, p. 104802, Oct. 2022.

[11] 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)*, 2021, pp. 7088–7092.

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

[13] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. Interspeech*, Jan. 2008, pp. 23–28.

[14] J. Le Roux, "Phase-controlled sound transfer based on maximally-inconsistent spectrograms," in *Proc. of the Acoustical Society of Japan Spring Meeting*, no. 1-Q-51, Mar. 2011.

[15] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1153–1161, 1986.

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

[17] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017.

[18] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, p. 3591, 2013.

[19] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Interspeech*, 2010, pp. 3110–3113.

[20] "Speech processing, transmission and quality aspects (STQ); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database," ETSI EG 202 396-1, 2008.

[21] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, 2002.

[22] ITU, "Rec. p.862.2: Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," 2018.

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

[24] C. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[2] http://iks.rwth-aachen.de/qr/eusipco2023-etsse