Interspeech 2024
1-5 September 2024, Kos, Greece

VOICOR: A Residual Iterative Voice Correction Framework for Monaural Speech Enhancement

Rui Cao[1], *Tianrui Wang*[1], *Meng Ge*[2,*], *Andong Li*[3], *Longbiao Wang*[1,*], *Jianwu Dang*[1], *Yungang Jia*[4]

[1]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]National University of Singapore, Singapore [3]Tencent AI Lab, China
[4]Tianjin Branch of National Computer Network Emergency Response Technical Team/Coordination Center of China, Tianjin, China
{caorui_2022,wangtianrui,longbiao_wang}@tju.edu.cn, gemeng@nus.edu.sg

Abstract

Supervised speech enhancement has gained significant progress from recent advancements in neural networks, especially due to their ability to non-linearly fit the diverse representations of target speech, such as waveform or spectrum. However, most of them continue to face issues with degraded speech and residual noise in challenging acoustic scenarios. In this paper, to alleviate the above issues, we propose a residual iterative voice correction framework to further correct the acoustic structure of the speech pre-enhanced by existing enhancement solutions. Concretely, inspired by the chain rule for information, we perform multiple corrections in a residual iterative manner to refine the spectral structure details of pre-enhanced speech. Experimental results on the DNS-Challenge dataset show that our solution consistently improves 0.3+ PESQ score over baselines, with only an additional 1.18 M parameters.

Index Terms: Monaural speech enhancement, articulatory mechanism, deep learning.

1. Introduction

Monaural speech enhancement focuses on extracting the target speech from the noisy mixture, aiming to improve both the quality and intelligibility of speech. Traditional speech enhancement methods usually rely on mathematical formulas derived from assumptions about the statistical characteristics of speech and noise, such as spectral subtraction [1], Wiener filtering [2] and subspace-method [3]. Recent data-driven speech enhancement methods have achieved significant performance gains by employing neural networks to fit the non-linear mapping relationship between the input noisy speech and the target speech, avoiding the traditional statistical assumptions. The existing data-driven direct-fitting methods can be categorized into two streams, namely time-domain and frequency-domain methods. Time-domain methods [4, 5] aim to model the distribution of waveform via speech encoding and decoding operations. In contrast, frequency-domain methods [6–8] work by separating noisy complex spectrum into either magnitude-phase or real-imaginary components to minimize the distance to target speech's complex spectrum.

Recently, researchers have found that these direct-fitting methods continue to face difficulties in addressing speech over-suppression and noise under-suppression [9, 10], which negatively impact hearing perception, especially in challenging acoustic scenarios. Essentially, the direct-fitting methods lack explicit modeling of the acoustic structural patterns of speech. Since the acoustic characteristics inherent in speech lead to corresponding spectral features (e.g., comb-like harmonic structure [11]), they are expected to serve as anchors to help the model distinguish speech from noise. Some studies [11–14] have shown that preserving intrinsic acoustic characteristics of speech (e.g., articulatory and auditory perceptual attributes [15–18]) can ensure that the enhanced speech remains consistent with human hearing perception. As these acoustic characteristics are closely linked to speech production and auditory perception of human speech, refining corresponding speech-specific structural patterns in speech processing can make the enhanced speech sound more natural as compared to direct-fitting solutions. Motivated by this fact, we rethink the speech enhancement task and prompt a research question: Could we systematically correct the unnatural acoustic structure in enhanced speech from existing enhancement solutions to further align with the hearing perception of target speech?

To answer this research question, we propose a residual iterative voice correction framework named VOICOR. Motivated by the multi-stage speech enhancement strategies [19–21], we decouple the estimation of the target spectrum into pre-enhancement and correction components. Building upon any existing enhancement solution as the pre-enhancement component, the correction component is employed to correct the pre-enhanced speech for hearing perception by leveraging the acoustic characteristics of speech. Different from the phase repair via spectrum decoupling of multi-stage strategies, in the correction component, we take the comb-like harmonic structure generated by vocal tract resonance as the key for acoustic structure correction. In addition, inspired by the chain rule for information, we further decouple the correction component into acoustic structure extractor and multiple correctors in a residual iterative manner. By doing so, the proposed VOICOR can recalibrate spectral details by accessing harmonic patterns for further improving the hearing perception of target speech.

2. Problem Formulation

For supervised learning, the Information Bottleneck principle suggests that the goal is to efficiently capture and represent the relevant information in the input variable about the output-label-variable [22]. Specific to supervised frequency-domain speech enhancement, the goal is to estimate the complex spectrum $\hat{S}$ of the target spectrum S from noisy input X [23]. Based on the above discussions, we rethink the spectrum estimation and express it as an incremental convergence process. That is to say, the estimated spectrum $\hat{S}$ should contain as much effective information of the target spectrum S as possible. Therefore, the desired objective Γ can be formulated as maximizing the mutual information between $\hat{S}$ and S, shown as,

$$\Gamma = I(\hat{S}; S) \tag{1}$$

In terms of the objective in direct-fitting methods, the estimated spectrum may not entirely align with the articulatory

* denotes the corresponding author.

10.21437/Interspeech.2024-53

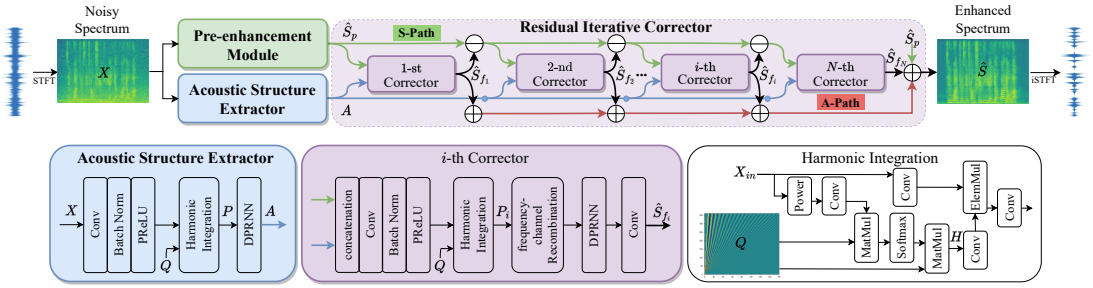


Figure 1: *An overview of our* VOICOR*, which consists of a pre-enhancement module, an acoustic structure extractor, and residual iterative corrector.* VOICOR *corrects the acoustic structure of pre-enhanced speech in a residual iterative manner. The block details are given below the overall framework.*

principle of natural human speech, which typically manifests as the degraded acoustic structure (e.g., incomplete harmonic) and residual noise in the spectrum. To recalibrate the spectrum, we decouple the estimation of target spectrum further into pre-enhancement term $\hat{S}_p$ obtained from existing enhancement solutions and acoustic structure correction term $\hat{S}_f$. The correction term $\hat{S}_f$ can be utilized to repair the pre-enhanced result $\hat{S}_p$, thereby improving the hearing perception of final output $\hat{S}$. In other words, if we assume there is statistical dependence between the distributions of $\hat{S}_p$ and $\hat{S}_f$, the remaining information of the target spectrum S that has not been fully captured by the pre-enhancement term $\hat{S}_p$ can be supplemented by incorporating the correction term $\hat{S}_f$. Thus, Eq (1) can be rewritten as,

$$\Gamma = I(\hat{S}_p, \hat{S}_f; S) \tag{2}$$

Given the chain rule for mutual information, Eq (2) can be further derived via N sub-items for practical implementation. This derivation process is represented as,

$$\begin{aligned}\Gamma &= I(\hat{S}_p; S) + I(\hat{S}_f; S|\hat{S}_p) \\ &= I(\hat{S}_p; S) + \sum_{i=1}^{N} I(\hat{S}_{f_i}; S|\hat{S}_p, \hat{S}_{f_1}, ..., \hat{S}_{f_{i-1}})\end{aligned} \tag{3}$$

From the above equation, it is clear that existing direct-fitting speech enhancement solutions primarily focus on optimizing the first term in Eq (3). Unlike the existing direct-fitting solutions, our solution aims to explore whether it is possible to correct the unnatural acoustic structures in existing solutions using additional information from the second term in Eq (3).

3. Proposed VOICOR

3.1. Overview

Motivated by the redefinition of speech enhancement in Sec. 2, we borrow the idea from chain rule and propose a residual iterative voice correction framework called VOICOR, as illustrated in Fig. 1. Our VOICOR views the voice correction process as a multi-variate fitting problem, where the non-linear correction function can be parameterized by the residual iterative corrector in Fig. 1. In practice, based on decoupling the desired objective into pre-enhancement term and iterative corrections in Eq (3), we thus parameterize the final output $\hat{S}$ and correspondingly decouple the forward computation $\mathcal{F}_u(X)$ as,

$$\hat{S} = \mathcal{F}_u(X) = \hat{S}_p + \sum_{i=1}^{N} \boldsymbol{R_i}(A, \hat{S}_p - \sum_{j=0}^{i-1} \hat{S}_{f_j}) \tag{4}$$

where $\hat{S}_p$, A, and $\hat{S}_{f_i}$ represent the features respectively produced by pre-enhancement module (PEM), acoustic structure extractor (ASE), and i-th corrector $\boldsymbol{R_i}$ of residual iterative corrector (RIC), with the special condition that $\hat{S}_{f_0} = 0$, as shown in Fig. 1. We implement the chain rule in Eq (3) as dual paths in a residual iterative manner. Dual paths consist of the A-Path and S-Path, marked in red and green in Eq (4), respectively.

Fig. 1 shows the architecture of the proposed VOICOR. The input is the noisy complex spectrum computed by Short-Time Fourier Transform (STFT), denoted as $X \in \mathbb{R}^{T \times 2 \cdot F}$, where T and $2 \cdot F$ denote the number of time frames and frequency bins (real and imaginary parts). Any spectrum estimation approach can serve as the PEM, and its result $\hat{S}_p \in \mathbb{R}^{T \times 2 \cdot F}$ is considered as a preliminary enhanced output. The modulated feature $A \in \mathbb{R}^{T \times C \times F}$ with the highlighted acoustic structure information extracted from X is further utilized by the RIC to perform structural corrections on $\hat{S}_p$, through the residual iterative manner. ASE and RIC work together as a correction component, acting on the PEM in a post-processing style.

3.2. Acoustic Structure Extractor

The acoustic structure extractor is designed to extract acoustic structural features based on the articulatory mechanism and conducts dimensionality expansion on the channel dimension during the processing. Concretely, we stack Conv block, Harmonic Integration [24], and DPRNN [25] to form ASE. Given the noisy spectrum X, the Conv block first roughly filters the input, which consists of a causal 2D convolution (Conv) followed by Batch Normalization and PReLU. Then, Harmonic Integration actively learns to calculate the location of harmonics in the spectrum, according to a comb-pitch conversation matrix $Q \in \mathbb{R}^{L \times F}$ with L pitch candidates, as shown in Fig. 1. The process can be formulated as,

$$H = \text{softmax}(\text{Conv}(X_{\text{in}}^2) \cdot Q^T) \cdot Q, \tag{5}$$

where $H \in \mathbb{R}^{T \times C \times F}$ can be seen as a reorganized representation of Q, which contains the distribution of harmonics. Then, we can utilize H to highlight the harmonic structure, shown as,

$$P = \text{Conv}(\text{Conv}(X_{\text{in}}) \odot \text{Conv}(H)), \tag{6}$$

where $\odot$ denotes the element-wise multiplication, and P represents the feature with the highlighted structure information. Finally, DPRNN [25] (i.e., DPRNN$(\cdot)$) is used to further generate the acoustic structure feature A with articulatory attribute,

$$A = \text{DPRNN}(P) \tag{7}$$

3.3. Residual Iterative Corrector

The residual iterative corrector is designed to correct the pre-enhancement term $\hat{S}_p$ from PEM using the acoustic structural feature A from ASE. As shown in Fig. 1, the residual iterative corrector consists of multiple interconnected sub-correctors that

collectively establish a sequential chain configuration. Specifically, we take the i-th corrector $\boldsymbol{R_i}$ as an example, given the extracted acoustic structure feature A and the residual spectrum that needs to be corrected after previous $(i-1)$ corrections $\hat{S}_p - \sum_{j=0}^{i-1} \hat{S}_{f_j}$, the current i-th corrector is capable of identifying the harmonic locations through comb-pitch conversation matrix Q, allowing the acoustic structure A to fill in the gaps within the residual spectrum $\hat{S}_p - \sum_{j=0}^{i-1} \hat{S}_{f_j}$ at those specific harmonic locations. The process can be denoted as,

$$P_i = \text{HI}\{\text{BN}(\text{Conv}(\text{Concat}(A, \hat{S}_p - \sum_{j=0}^{i-1} \hat{S}_{f_j}))), Q\}, \quad (8)$$

where $\text{HI}(\cdot)$, $\text{BN}(\cdot)$, and $\text{Concat}(\cdot)$ represent the Harmonic Integration, Batch Normalization, and concatenation, respectively. To strengthen the information interaction between different dimensions (i.e., time, frequency, and channel dimensions), we further employ frequency-channel Recombination (i.e., $\text{FCR}(\cdot)$) and DPRNN followed by a Conv to adjust the spectrum again. Thus, the i-th correction can be defined as,

$$\hat{S}_{f_i} = \boldsymbol{R_i}(A, \hat{S}_p - \sum_{j=0}^{i-1} \hat{S}_{f_j}) = \text{Conv}(\text{DPRNN}(\text{FCR}(P_i))) \quad (9)$$

More details of this processing are shown in Fig. 1.

3.4. Dual Paths in Residual Iterative Manner

As an instantiation of the chain rule in Eq (3), we propose a residual iterative manner to realize collaborative interaction between each module in VoiCor. To maximally simulate the interaction of mutual information in Eq (3), we introduce S-Path and A-Path as two forward computation paths to implement the residual iterative manner. As mentioned before, the S-Path (green line in Fig. 1) is a residual connection [26] established between pre-enhancement term $\hat{S}_p$ and j-th corrector output $\hat{S}_{f_j}$, which performs subtraction operation, formulated as $\hat{S}_p - \sum_{j=0}^{i-1} \hat{S}_{f_j}$. The A-Path (red line in Fig. 1) performs addition operation between $\hat{S}_p$ and i-th corrector output $\hat{S}_{f_i}$, formulated as $\hat{S}_p + \sum_{i=1}^{N} \hat{S}_{f_i}$. By filtering out the corrected spectral feature $\hat{S}_{f_j}$ from $\hat{S}_p$, the S-Path enables each corrector to progressively focus on the remaining spectral details that PEM and previous correctors fail to capture. Through the A-Path, each correction term $\boldsymbol{R_i}(A, \hat{S}_p - \sum_{j=0}^{i-1} \hat{S}_{f_j})$ is skip-connected with the $\hat{S}_p$ to form the final output $\hat{S}$ of VoiCor, ensuring direct optimization of each module during the end-to-end training. In this way, multiple correctors are stacked collectively to achieve systematic recalibration of the acoustic structure.

3.5. End-to-End Training

The entire framework is trained in an end-to-end manner. To this end, the integration of the articulatory attribute into the framework, along with the incorporation of auditory perceptual attribute in the loss function, can be viewed as mutually reinforcing elements. We introduce a combination of time-domain sample-level SI-SNR [27] and human psychoacoustic perception-based PMSQE [28] as the overall loss function,

$$\mathcal{L} = \mathcal{L}_{SI-SNR} + \mathcal{L}_{PMSQE} \quad (10)$$

4. Experimental Setup

We conduct experiments on the data taken from the 2020 DNS-Challenge [29], which has 500 hours of clean speech from 2150 speakers and over 180 hours of noise from 150 classes. We generate 100 hours of noisy speech for the ablation and generalizability experiments and 300 hours for the final comparison, respectively. The signal-to-noise ratio (SNR) ranges from -5 dB to 20 dB. The dataset is partitioned into a training set and a validation set in a 4:1 ratio. For testing audio, the SNR is between -5 dB and 30 dB, comprising a total of 10 hours of noisy speech.

The 32 ms Hanning window with 25% overlap and 512-point STFT are used. The channel numbers of ASE and each corrector are {6,12} and {6,2}, respectively. The kernel size of convolution in Conv block and Harmonic Integration are (5,2) and (1,3), respectively. The stride of all convolution is (1,1). The head numbers of Harmonic Integration and frequency-channel Recombination are 4, 8, and 1, respectively. Model training is conducted using PyTorch with the Adam optimizer. The initial learning rate is set to 0.001, and a 0.75 learning rate decay will be applied if the validation loss does not decrease for 3 consecutive epochs.

To provide a comprehensive assessment, we employ the number of the parameters (Para.), perceptual evaluation of speech quality (PESQ) [30] for speech quality, scale-invariant signal-to-distortion ratio (SI-SDR) [31] for speech distortion, and short-time objective intelligibility measure (STOI) [32] for intelligibility. In the last three metrics, higher values indicate better performance.

5. Results and Analysis

5.1. Ablation Study

We utilize DPCRN [7] as the PEM (i.e., VoiCor [DPCRN]) for the ablation study. As shown in Table 1, with the number of correctors N increasing from 1 to 4, objective metrics show an initial rise followed by a slight decrease in trend. Compared with the enhanced results utilizing only DPCRN as PEM without correction (i.e., N=0), the proposed framework with 3 correctors (i.e., N=3) has improved PESQ by 0.187 and SI-SDR by 0.896 dB. This improvement can validate the effectiveness of the proposed framework. When N=4, the slight drop can be attributed to the following reason. Correction is mainly reflected in the repair of low-energy details of the spectrum. When N=3, each corrector has already captured the acoustic structure information to a great extent. Therefore, redundant modification (i.e., N=4) is detrimental to spectrum reconstruction. It is noteworthy that different PEMs result in different pre-enhancement terms, consequently causing variations in the maximum value of N. Hence, leveraging the post-processing style of our framework, an appropriate number of correctors can be applied to different spectrum estimation methods. Furthermore, we increase the number of convolutional channels in DPCRN to be comparable in parameters with the VoiCor with 3 correctors, denoted as PEM (DPCRN large). The experimental result shows that blindly increasing the parameters of model is not advisable. Starting from the underlying articulatory mechanism, reasonable parametric modeling is helpful to achieve better performance with fewer additional parameters.

We also compare using only $\mathcal{L}_{SI-SNR}$ as the loss function by removing auditory constraint loss function $\mathcal{L}_{PMSQE}$ (i.e., -AC). One can see that the auditory and articulatory attributes effectively reinforce each other within the established framework. This is also closely related to the aforementioned characteristics of human speech. Further removing ASE (i.e., -ASE) during the correction process results in an obvious drop in objective metrics. This directly confirms the structural features

extracted by ASE effectively assist in calibrating the spectrum. In other words, the prerequisite for RIC to make reasonable correction is that correction should be consistent with articulatory mechanism.

Table 1: *Ablation study on our 10 hours test dataset.*

	N	Para. (M)	PESQ	SI-SDR (dB)	STOI (%)
VoiCor [DPCRN] w/ N correctors w/o AC	0	1.64	2.771	20.389	94.40
	1	1.97	2.935	21.219	95.02
	2	2.25	2.941	21.262	95.14
	3	2.54	2.958	21.285	95.06
	4	2.82	2.909	21.113	94.92
PEM (DPCRN large)	0	2.95	2.774	20.380	94.39
VoiCor [DPCRN]	3	2.54	3.072	21.520	95.36
- AC	3	2.54	2.958	21.285	95.06
- ASE	3	2.47	2.687	19.943	93.96

5.2. Generalizability of VoiCor

Besides DPCRN, we additionally select two spectrum estimation strategies for further exploration, including the baseline method NSNet [33] from the 2020 DNS challenge and CRN [6]. As shown in Table 2, the generalizability of VoiCor enables it to better adapt to different PEMs. First, it is not surprising that CRN achieves greater improvement in objective metrics compared to NSNet. NSNet optimizes only the magnitude and retains the noise phase, while CRN estimates both the real and imaginary parts to enhance the magnitude and phase responses of noisy speech. Then, we focus on the improvement in objective metrics brought by VoiCor after applying the same number of correctors (N=4). Correcting NSNet (i.e., VoiCor [NSNet]) has improved PESQ by 0.555, SI-SDR by 3.437 dB and STOI by 3.50 %, while correcting CRN (i.e., VoiCor [CRN]) has improved PESQ by 0.302, SI-SDR by 1.009 dB and STOI by 0.92 %. This can be attributed to the fact that the unoptimized noise phase in NSNet, which provides VoiCor with greater room for improvement. On the other hand, CRN provides VoiCor with finer correction space by offering a more stable pre-enhancement term. So in terms of the final upper bound, one can see that VoiCor [CRN] is higher.

Table 2: *Generalizability experiments for different PEMs on our 10 hours test dataset.*

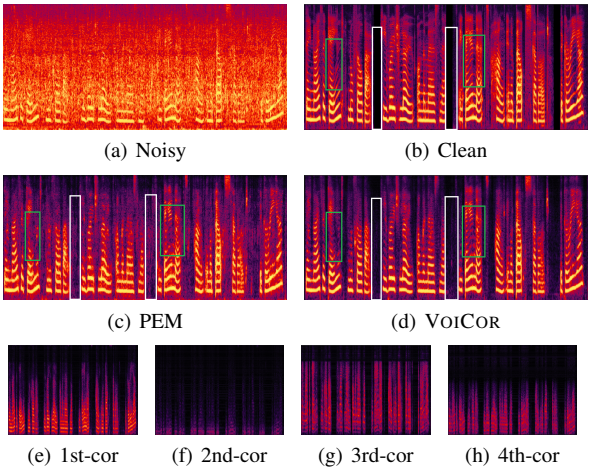
Model	Para. (M)	PESQ	SI-SDR (dB)	STOI (%)
NSNet	2.79	2.274	17.023	91.06
VoiCor [NSNet]	3.97	2.829	20.460	94.56
CRN	1.70	2.732	20.295	94.31
VoiCor [CRN]	2.88	3.034	21.304	95.23

5.3. Visual Analysis

To conduct a more in-depth analysis of the specific implication of correction, we visualize the output of each module in VoiCor [CRN] with 4 correctors, as shown in Fig. 2. Despite the noticeable improvement in the output of PEM, there are still issues related to noise residue and speech degradation (colorful boxes in Fig. 2(c)). Fig. 2(d) shows that VoiCor not only removes the residual noise in the white box but also enhances the harmonic structure of speech in the green box, actively correcting the acoustic structure. The output of each corrector not only presents sparse comb-like details of the spectrum, but also systematic harmonic patterns in different frequency bands.

5.4. Comparison on Public Test Dataset

Compared to the outstanding solutions on the public test set of DNS-Challenge 2020, as shown in Table 3, our proposed



(a) Noisy (b) Clean (c) PEM (d) VoiCor (e) 1st-cor (f) 2nd-cor (g) 3rd-cor (h) 4th-cor

Figure 2: *Noisy and clean spectrograms are provided as reference. (c)-(h) are the output visualizations of each module and* VoiCor*'s final result, the i-th corrector is denoted as ith-cor.*

VoiCor using DPCRN as the backbone of PEM achieves competitive performance with minimal model parameters. The superiority of the proposed framework over the complex-spectrum methods (e.g., DCCRN/DCCRN+) is predictable. For the FullSubNet+ with 8.67 M parameters, VoiCor achieves better performance with only 29% of its parameters. For the HGCN, which only performs harmonic compensation, our solution shows comprehensive improvement in objective metrics. It is also superior to the approach that combines DNN with a speech production model (GARNNHS). Moreover, as for multiple objective metrics, our VoiCor outperforms an advanced method of hierarchical optimization in the complex spectrum (GaGNet), with 43% of its parameters.

Table 3: *System comparisons on DNS-Challenge 2020 test set without reverberation. "-" denotes no published result.*

Model	Para. (M)	PESQ_{WB}	PESQ_{NB}	SI-SDR (dB)	STOI (%)
Noisy	-	1.58	2.45	9.07	91.52
DCCRN [8]	3.70	-	3.27	-	-
DCCRN+ [34]	3.30	-	3.33	-	-
FullSubNet [35]	2.97	2.78	3.31	17.29	96.11
FullSubNet+ [36]	8.67	2.98	3.50	18.34	96.69
HGCN [11]	-	2.88	-	18.14	96.50
CARNNHS [14]	-	2.89	3.43	18.80	96.70
GaGNet [19]	5.94	3.17	3.56	18.91	97.13
VoiCor [DPCRN]	2.54	3.09	3.61	19.56	97.25

6. Conclusion

In this work, we proposed a universal framework named VoiCor to correct the acoustic structure of pre-enhanced speech. From a novel perspective of decoupling mutual information, we decomposed the framework into pre-enhancement and correction components. To recalibrate the sparse spectral details, we further decoupled the correction component into acoustic structure extractor and multiple correctors in a residual iterative manner inspired by the chain rule for information. Experiments reveal the superior performance of the proposed light-weight correction method.

7. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62176182.

8. References

[1] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech communication*, vol. 52, no. 5, pp. 450–475, 2010.

[2] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction wiener filter," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, 2006.

[3] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE transactions on speech and audio processing*, vol. 8, no. 5, pp. 497–507, 2000.

[4] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[5] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Speech denoising in the waveform domain with self-attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7867–7871.

[6] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6865–6869.

[7] X. Le, H. Chen, K. Chen, and J. Lu, "Dpcrn: Dual-path convolution recurrent network for single channel speech enhancement," in *Proc. Interspeech*, 2021, pp. 2811–2815.

[8] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[9] W. Jiang, Z. Liu, K. Yu, and F. Wen, "Speech enhancement with neural homomorphic synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 376–380.

[10] Z. Kong, W. Ping, A. Dantrey, and B. Catanzaro, "Cleanunet 2: A hybrid speech denoising model on waveform and spectrogram," in *Proc. Interspeech*, 2023, pp. 790–794.

[11] T. Wang, W. Zhu, Y. Gao, J. Feng, and S. Zhang, "Hgcn: Harmonic gated compensation network for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 371–375.

[12] 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*, vol. 34, no. 05, 2020, pp. 9458–9465.

[13] T. Wang, W. Zhu, Y. Gao, Y. Chen, J. Feng, and S. Zhang, "Harmonic gated compensation network plus for icassp 2022 dns challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 9286–9290.

[14] S. He, W. Rao, J. Liu, J. Chen, Y. Ju, X. Zhang, Y. Wang, and S. Shang, "Speech enhancement with intelligent neural homomorphic synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[15] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[16] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The journal of the acoustical society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[17] R. D. Kent and H. K. Vorperian, "Static measurements of vowel formant frequencies and bandwidths: A review," *Journal of communication disorders*, vol. 74, pp. 74–97, 2018.

[18] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1614–1623, 2008.

[19] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, p. 108499, 2022.

[20] A. Li, S. You, G. Yu, C. Zheng, and X. Li, "Taylor, can you hear me now? a taylor-unfolding framework for monaural speech enhancement," in *Proceedings of the International Joint Conferences on Artificial Intelligence. (IJCAI)*, 2022, pp. 4193–4200.

[21] 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 Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2156–2172, 2022.

[22] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 ieee information theory workshop (itw)*. IEEE, 2015, pp. 1–5.

[23] C. Zheng, H. Zhang, W. Liu, X. Luo, A. Li, X. Li, and B. C. Moore, "Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods," *Trends in Hearing*, vol. 27, p. 23312165231209913, 2023.

[24] T. Wang, W. Zhu, Y. Gao, S. Zhang, and J. Feng, "Harmonic attention for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2424–2436, 2023.

[25] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 46–50.

[26] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech large language models," in *International Conference on Learning Representations*, 2024.

[27] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016, pp. 545–549.

[28] J. M. Martin-Donas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal processing letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[29] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020, pp. 2492–2496.

[30] 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," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 2, 2001, pp. 749–752.

[31] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 626–630.

[32] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[33] Y. Xia, S. Braun, C. K. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 871–875.

[34] S. Lv, Y. Hu, S. Zhang, and L. Xie, "Dccrn+: Channel-wise subband dccrn with snr estimation for speech enhancement," in *Proc. Interspeech*, 2021, pp. 2816–2820.

[35] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6633–6637.

[36] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7857–7861.