Interspeech 2024
1-5 September 2024, Kos, Greece

DNN-based monaural speech enhancement using alternate analysis windows for phase and magnitude modification

Xi Liu, John H.L. Hansen[1]

[1]Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, USA

xi.liu@utdallas.edu, john.hansen@utdallas.edu

Abstract

In recent decades, considerable research has been devoted to speech enhancement leveraging the short-term Fourier transform (STFT) analysis. As speech processing technology evolves, the significance of phase information in enhancing speech intelligibility becomes more noticeable. Typically, the Hanning window has been widely employed as analysis window in STFT. In this study, we propose the Chebyshev window for phase analysis, and the Hanning window for magnitude analysis. Next, we introduce a novel cepstral domain enhancement approach designed to robustly reinforce the harmonic structure of speech. The performance of our model is evaluated using the DNS challenge test set as well as the naturalistic APOLLO Fearless Steps evaluation set. Experimental results demonstrate that the Chebyshev-based phase solution outperforms the Hanning option for in phase-aware speech enhancement. Furthermore, the incorporation of quefrency emphasis proves effective in enhancing overall speech quality.

Index Terms: speech enhancement, cepstral analysis, analysis window, deep neural network

1. Introduction

Monaural speech enhancement is one of the most challenging task in the field of acoustic signal processing. It involves using advanced algorithms to extract or reconstruct clean speech from noise-corrupted observations. Effective speech enhancement significantly improves speech quality and intelligibility by suppressing noise while preserving the target speech[1]. Consequently, Speech enhancement technology is extensively used in wireless communications, hearing aids[2, 3] and other intelligent voice-enabled speech technologies.

Traditional speech enhancement methods, such as Wiener filtering method [4], spectral subtraction [5], and MMSE-based method[6], excel in handling stationary background noises but face performance degradation with non-stationary noises. To tackle this issue, Cohen et al. proposed a sophisticated system[7] that enhances the capability of traditional speech enhancement by suppressing non-stationary noises. However, some algorithms designed for this purpose introduce artificial noises. To address this, efforts have been made to solve this issue[8]. Notably, these methods primarily focus on magnitude denoising, often neglecting the crucial role of phase estimation. Yet, previous studies [9, 10] emphasize the importance of phase modification in improving speech quality and intelligibility. Paliwal et al. [11] conducted a comprehensive study demonstrating the significant impact of the phase spectrum on analysis-modifiation-synthesis (AMS) based speech enhancement. Their work indicates that combining noisy magnitude with either clean or compensated noisy phase notably enhances speech quality. In addition, the use of a Chebyshev analysis window with a lower dynamic range further augments performance in this regard.

In recent years, data-driven speech enhancement methods using deep neural networks (DNNs) have been widely studied, demonstrating significant improvement under various noise conditions [12]. By utilizing advanced machine learning technology and training strategy, DNNs can either directly map the clean spectrogram[13] or estimate the complex-valued time-frequency(TF) mask[14]. While real-value time-frequency domain methods have advantages in modeling the magnitude of each frequency band, they still have difficulty reconstructing the phase of the target signal. Considering that phase has a critical impact on perceived speech quality, particularly under low signal-to-noise ratio conditions[15], most TF domain methods currently adopt complex-valued frequency features to implicitly estimate phase information[16]. Another mainstream approach aims to estimate clean speech waveform from noisy speech waveform in the time domain using neural networks[17] which can avoid the decoupling of magnitude and the phase information in the frequency domain. To take advantage of both domain features, Tang et al. proposed a cross-domain framework [18] that jointly uses the spectrum and waveform.

Inspired by these studies, we propose a novel approach that explores the effectiveness of using alternate (mismatched) analysis windows for phase and magnitude estimation in neural network-based speech enhancement solution[1]. Specifically, we use a Hanning window for noisy magnitude extraction, which is proven effective for magnitude spectrum. Simultaneously, a Chebyshev analysis window is applied for noisy phase extraction. The motivation of using Chebyshev window will be further discussed in Section 2.3. The separated phase and magnitude are processed independently using a dual-path parallel convolutional encoder-decoder structure. In addition, we propose employing cepstral enhancement for magnitude denoising, since emphasizing pitch information in time-quefrency(TQ) domain is crucial for speech harmonic reconstruction. After separate processing, the complex-valued spectrum is easily reconstructed by multiplying the two parts. Our contributions can be summarized as follows: 1) We propose a novel cepstral domain (magnitude) and spectral domain (phase) fusion speech enhancement framework, demonstrating superior performance compared to single-domain approaches. 2) The adoption of a Chebyshev analysis window for DNN-based clean phase estimation is introduced, showcasing superior performance compared to the Hanning analysis window.

[1]This work was supported in part by NSF CCRI: Community Resource Project under Grant 2016725 and in part by University of Texas at Dallas – Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen

10.21437/Interspeech.2024-2244

2. Proposed method

2.1. AMS based speech enhancement

In time domain, the noisy speech can be formulated as:

$$y(n) = s(n) * h(n) + d(n) \tag{1}$$

where $y(n)$ is noisy signal captured by single omni-directional microphone, $s(n)$ is the clean speech, $h(n)$ is the room impulse response(RIR), $d(n)$ is the additive noise and $*$ denotes the convolution operation. The first step of AMS process is employing short-term Fourier transform (STFT) for frequency analysis. The time-frequency(TF) domain noisy representation is given by:

$$Y(k,n) = \sum_{m} y(m)w(n-m)e^{-j2\pi km/L} \tag{2}$$

where k and n are frequency and time indices respectively, and $w(n)$ is the analysis window with a length of L. Hence, the TF domain representation of equation1 can be derived as:

$$Y(k,n) = S(k,n)H(k,n) + D(k,n) \tag{3}$$

In this study, we aim to remove $D(k,n)$ and preserve $H(k,n)$ by only cancelling the additive noise. Eventually, the last step of AMS is to reconstruct time domain speech by using inverse STFT and overlap and add(OLA) synthesis method.

2.2. Cepstral analysis

While time and frequency domain methods have received significant attention in speech enhancement, the potential of cepstral domain speech enhancement has been relatively overlooked. The concept of cepstrum was initially introduced by Bogert et al. [19] to analyze the periodicity of the spectrum. In 1991, Sorensen et al. [20] first adopted a multi-layer neural network to suppress noise in the cepstral domain for noise-robust speech recognition. A standard cepstral analysis and its inverse process can be found in [21].

In the cepstral domain, the excitation and spectral envelope exhibit contrasting distributions along the quefrency axis, as illustrated in Figure 1 right plots. Moreover, the distribution of noise in the cepstral domain also differs from the location of speech component, as noise usuallt lacks the similar 'source-filter' pattern observed in speech. For instance, in Figure 1, pitch contour of speech still remains visible in the time-quefrency (TQ) domain while most of speech harmonic is masked by noise in time-frequency (TF) domain. Thus, we believe cepstrum is potentially a better feature than spectrum for speech enhancement. Prior works, such as Liu et al. [22], have employed an in-place cepstral processing block to effectively restore the speech harmonic structure that is masked by noise, demonstrating the considerable potential of cepstral modification.

In this work, we directly estimate the clean cepstrum from the noisy observation for clean magnitude estimation. It should be noted that we don't use complex cepstrum in this contribution due to the phase wrapping issue occurred during complex cepstral analysis. Therefore, the proposed method adopts minimum-phase reconstruction method by using real-valued cepstrum and an appropriate window. A real-valued cepstral analysis is derived by applying real-valued logarithm to the magnitude of the Fourier transform of speech:

$$S_{cepstrum} = real(IFFT\{ln(|FFT\{s(t)\}|)\}) \tag{4}$$

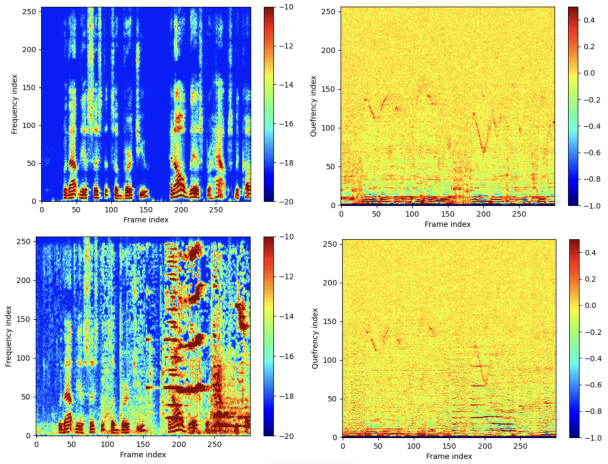

Figure 1: *Clean and noisy speech in different domain. Top left: clean speech in TF domain; Top right: clean speech in TQ domain; bottom left: noisy speech in TF domain; bottom right: noisy speech in TQ domain*

$$window(n) = \begin{cases} 1 & if \quad n = 0, n = \frac{N}{2} - 1 \\ 2 & if \quad 0 < n < \frac{N}{2} - 1 \\ 0 & if \quad \frac{N}{2} - 1 < n < N - 1 \end{cases} \tag{5}$$

Applying the window function in equation (5) on real cepstrum $S_{cepstrum}$ yields the minimum-phase speech component in cepstral space. The minimum-phase component contains the equivalent energy of the speech spectrum[23].

2.3. The effect of analysis window

With the development of speech processing, researchers realized phase also plays a crucial role in recovering distorted speech, particularly in low signal-to-noise ratio cases [9]. The choice of using different analysis windows for STFT has been extensively studied in recent decades. In most AMS framework-based speech enhancement approaches, such as [6], the Hamming or Hanning window is commonly employed as the analysis window for magnitude spectral enhancement, demonstrating satisfactory performance. Conversely, numbers of works [24, 25, 26] have shown that using a short-length Hamming window for phase spectral estimation can lead to poor speech quality. The 'phase-only' reconstructed speech becomes intelligible when the window length exceeds 1 second. However, employing such long window lengths is impractical for speech processing, as speech is not stationary in such extended duration. Therefore, Paliwa et al. [11] thoroughly investigated the suitable window types for both magnitude and phase spectrum estimation. Experimental results indicate that when clean reference speech is unavailable, using the Chebyshev window for phase extraction and the Hamming window for magnitude extraction can significantly improve the quality of reconstructed speech compared to solely using the Hamming window. Based on these findings, we are inspired to use the Chebyshev window for phase-aware speech enhancement.

2.4. The DNN-based dual path structure

To separately estimate the magnitude and phase spectrum, we employed a dual-path parallel deep neural network structure, as

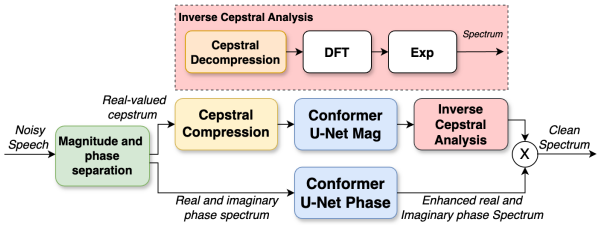

Figure 2: *System diagram of proposed dual-path structure*

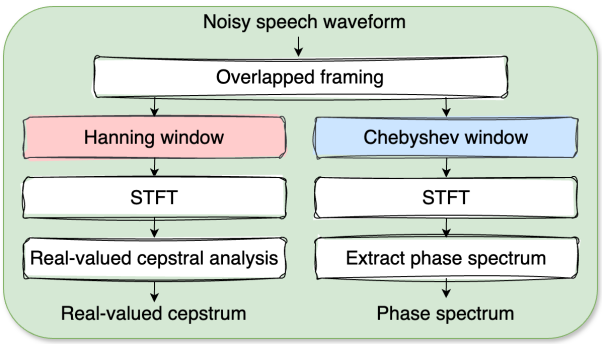

Figure 3: *Magnitude and phase separation process*

depicted in Figure 2. In the system diagram, the Conformer U-Net blocks constitute the DNN structure, and the rest processing blocks are signal processing methods. The upper branch in Figure 2 is designed to process the real-valued speech magnitude in the cepstral space, and the lower branch is intended for complex-valued phase estimation. The final denoised speech spectrum can be calculated by multiplying the outputs from the two branches. Figure 3 illustrates how the noisy magnitude and phase representations are extracted by applying different window functions.

$$S_{ceptrum}^{cpr} = \begin{cases} |S_{ceptrum}|^{\alpha} & if \quad S_{ceptrum} \geq 0 \\ -|S_{ceptrum}|^{\alpha} & if \quad S_{ceptrum} < 0 \end{cases} \tag{6}$$

In the cepstral domain, the energy of the speech envelope representation is much higher than the pitch peak. This is primarily because the energy decays exponentially along the quefrency axis [27], causing the pitch peak in the high quefrency region to have a very small value. To enhance the DNN's performance for better estimation, we employ a power-law cepstrum compression method (equation (6)), inspired by similar spectrum compression techniques [28]. This method aims to narrow the dynamic range and normalize the energy distribution.

We employed two distinct Conformer U-Net structures for real-valued and complex-valued features, as illustrated in Figure 4. The Conformer U-Net comprises a convolutional encoder-decoder structure and residual connections. The key difference between these two networks is that the Conformer U-Net for the phase has two decoders for the real and imaginary parts. To achieve simultaneous phase and magnitude estimation, we utilized a joint loss function [16] for this task:

$$L_R = \frac{1}{T \times F} \sum_{t,f} \left| |S_{real}(t,f)|^{\beta} - |\widetilde{S}_{real}(t,f)|^{\beta} \right|^2, \tag{7}$$

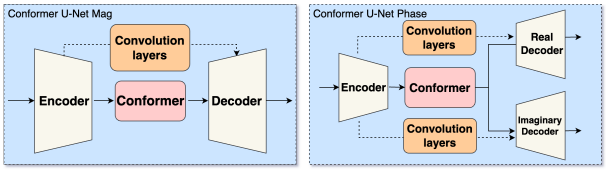

Figure 4: *U-Net structure*

$$L_I = \frac{1}{T \times F} \sum_{t,f} \left| |S_{imag}(t,f)|^{\beta} - |\widetilde{S}_{imag}(t,f)|^{\beta} \right|^2, \tag{8}$$

$$L_{mag} = \frac{1}{T \times F} \sum_{t,f} \left| |S(t,f)|^{\beta} - |\widetilde{S}(t,f)|^{\beta} \right|^2. \tag{9}$$

The joint loss function is:

$$L_{joint} = \gamma L_{mag} + L_R + L_I, \tag{10}$$

Here, L_R and L_I represent the real and imaginary spectrum losses, L_{mag} is the magnitude loss function, and γ is a weight factor. n this work, we set the weight factor γ and compression factor β to 5 and 0.3 respectively.

3. Experiments

3.1. Dataset

In this study, we utilized two datasets for training and evaluation: the Deep Noise Suppression (DNS) challenge[29] dataset and the APOLLO Fearless Steps dataset[30]. The training set consists of 562.72 hours of clean reference speech in English and 181 hours of noise. Additionally, to enhance system robustness in reverberant environments, 75% of the clean speech is convolved with Room Impulse Responses (RIR) randomly selected from OpenSLR26[31] and OpenSLR28[31] to simulate reverberation. The proposed system incorporates an on-the-fly data augmentation method, ensuring that all noisy data used in the training process is synthesized by randomly applying different Signal-to-Noise Ratios (SNR) ranging from -5 to 20 dB.

For objective evaluation, we employed a non-blind synthetic test set to assess our model and compare its performance with state-of-the-art (SOTA) methods. This synthetic test set was released as part of the 2020 DNS Challenge and comprises 12 classes of noise, with Signal-to-Noise Ratios (SNR) ranging from 0dB to 25dB. Additionally, real recording data from the APOLLO Fearless Step test set[2] was used for evaluation in a real-world environment. The APOLLO data was upsampled to 16kHz before evaluation, as the original sampling rate was 8kHz.

3.2. Experimental setups

The proposed framework employs a 32ms window length with a 16ms hop length for STFT, resulting in 257 frequency/quefrency bins in each frame for 16kHz sampling rate data. In this work, the dynamic range of the Chebyshev window is set to 30 dB based on previous studies. The AdamW optimizer is used with an initial learning rate of 0.001. Additionally, we utilize the Xavier initializer to initialize the values of weights in each layer. The channel numbers of convolution layers in encoders are 32, 32, 64, 64, 96, 256, with (5,3) kernel

[2]https://fearless-steps.github.io/ChallengePhase2/

Table 1: *Objective evaluation on 2020 DNS challenge test set*

Methods	Feature Domain	Without reverb		
		WB-PESQ	NB-PESQ	STOI
Noisy		1.582	2.265	91.52
NS-Net[32]	F	2.145	2.873	94.47
DTLN[33]	F	-	3.040	94.76
DCCRN−E[34]	F	-	3.266	-
ConvTasNet[17]	T	2.730	-	-
FullSubNet[35]	F	2.777	3.305	**96.11**
Proposed	C+F	**2.840**	**3.338**	96.05

Table 2: *Ablation study on 2020 DNS challenge test set*

Methods	Without reverb		
	WB-PESQ	NB-PESQ	STOI
Spec match	2.610	3.171	94.61
Spec mismatch	2.802	3.322	95.38
Fusion match	2.815	3.306	95.74
Fusion mismatch	**2.840**	**3.338**	**96.05**
−Cep compress	2.722	3.229	94.67

and (2,1) stride size in the shape of (frequency/quefrency, time). The decoder consists of transpose convolution layers, which are the reverse version of convolution layers. The convolution layers in the skip path have a (3,3) kernel size with (1,1) stride and zero padding. All convolution layers in the encoder and decoder are followed by 2D-batchnorm and PReLU activation. The 2-layer Conformer structure has 4 attention heads, 256 FFN dimensions, and a 15 kernel size for depth-wise convolution.

To illustrate the effectiveness of cepstral domain enhancement and the Chebyshev window, we conducted several ablation studies on the DNS challenge test set, as shown in Table 2. In general, the methods are divided into spectral domain methods (Spec) and domain fusion (Fusion) methods. The spectral domain methods directly use spectral magnitude as input for 'Conformer U-Net Mag' instead of cepstrum. The term 'Match' indicates that the analysis windows for magnitude and phase are the same (Hanning window), and 'mismatch' means the window functions are different for magnitude (Hanning) and phase (Chebyshev). Additionally, importance of cepstrum compression is also evaluated in this study.

3.3. Results and analysis

In Table 1, we compared our method with other top-ranked systems in the DNS challenge. To ensure fairness, we directly used the scores reported in their papers; the character '-' denotes an unreported score. The results indicate that our method outperforms the other compared systems in terms of wide-band perceptual evaluation of speech quality (WB-PESQ) score[36] and narrow-band PESQ (NB-PESQ) score.

Moreover, based on the results of the ablation studies in Table 2, we observe that using a mismatched window combination yields superior performance compared to using matched window functions. This suggests that employing the Chebyshev window for noisy phase not only enhances speech quality [11] but also improves the performance of DNN-based phase-aware speech enhancement. Figure 5 illustrates that using a mismatched window can reduce part of noise, but applying it in clean speech can cause speech distortion. Consequently, we use matched window function for clean speech in trainning process. The speech is synthesized using the Hanning window in Figure 5. By comparing the evaluation results of the proposed method (Fusion mismatch) and the spectral domain method (Spec mismatch), we can conclude that cepstrum is potentially a better feature than spectrum for magnitude estimation. When we eliminate cepstral compression in the proposed method, the neural network performance and speech quality degrade. This demonstrates that pre-processing for normalization is important to deep neural network training and modeling.

Table 3: *Evaluation on APOLLO Fearless Steps test set*

Methods	OVRL ↑	SIG ↑	BAK ↑
Noisy	1.88	2.50	2.11
DTLN	2.06	2.42	3.55
Proposed	**2.17**	**2.72**	**4.11**

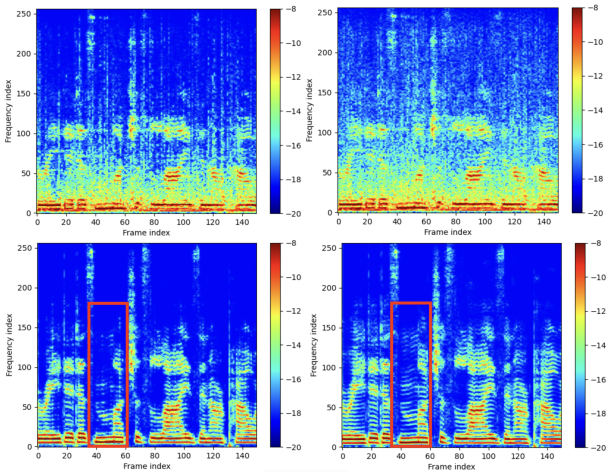

Figure 5: *Spectrogram with matched and mismatched window: top left is noisy speech with mismatched window; top right is noisy speech with matched window; bottom left shows clean speech with mismatched window and bottom right is clean speech with matched window*

To showcase the performance of the proposed system on real recording data, we also evaluate our method on the Fearless Step APOLLO dataset in terms of the DNS Mean Opinion Score (MOS), a non-intrusive objective evaluation metric [37]. The results in Table 3 demonstrate that our method exhibits better performance on the naturalistic low-quality dataset, where OVRL, SIG, and BAK represent overall effect, speech signal, and background noise.

4. Conclusions

In this paper, we introduced a novel approach for speech enhancement, leveraging mismatched window functions and feature fusion. More precisely, we proposed a two-branch structure to independently handle speech magnitude features extracted by the Hanning window and phase features extracted by the Chebyshev window. Experimental results demonstrate the superiority of the Chebyshev window over the Hanning window in phase-aware speech enhancement. And We also proved noise suppression in cepstral space further enhances performance. In future research, we aim to use a single encoder for different domain features and extend this work to simultaneous denoising and dereverberation tasks. Audio samples are available online[3].

[3]Audio Samples: https://winkee520.github.io/CSFNet/

5. References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. USA: CRC Press, Inc., 2013.

[2] M. Nursadul and H. John, "Speech enhancement for cochlear implant recipients using deep complex convolution transformer with frequency transformation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 32, pp. 2616–2629, 05 2024.

[3] H. J. Shekar RCMC, "A convolutional neural network-based framework for analysis and assessment of non-linguistic sound classification and enhancement for normal hearing and cochlear implant listeners." *Journal of the Acoustical Society of America*, vol. 152(5), pp. 2720–2734, 11 2022.

[4] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transa on Acoustics, Speech, and Signal Proc*, vol. 26, no. 3, pp. 197–210, 1978.

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transa on Acoustics, Speech, and Signal Proc*, vol. 27, no. 2, pp. 113–120, 1979.

[6] 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.

[7] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[8] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori snr estimation approach based on selective cepstro-temporal smoothing," in *IEEE ICASSP 2008*, 2008, pp. 4897–4900.

[9] B. J. Shannon and K. K. Paliwal, "Role of phase estimation in speech enhancement," in *Proc. Interspeech 2006*, 2006, pp. paper 1330–Tue3FoP.4.

[10] E. Loweimi, S. M. Ahadi, and S. Loveymi, "On the importance of phase and magnitude spectra in speech enhancement," in *2011 19th Iranian Conf on Electr Engineering*, 2011, pp. 1–6.

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

[12] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 10, p. 1702–1726, oct 2018.

[13] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *IEEE ICASSP*, 2019, pp. 6865–6869.

[14] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE Transa on Audio, Speech, and Lang Proc*, vol. 24, no. 3, pp. 483–492, 2016.

[15] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel source separation," in *Proc. Interspeech 2012*, 2012, pp. 1548–1551.

[16] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *AAAI Conf on Artificial Intel*, 2019, pp. 9458–9465.

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

[18] C. Tang, C. Luo, Z. Zhao, W. Xie, and W. Zeng, "Joint time-frequency and time domain learning for speech enhancement," in *International Joint Conference on Artificial Intelligence*, 2020, pp. 3816–3822.

[19] B. P. Bogert, "The quefrency analysis of time series for echoes : cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the Symposium on Time Series Analysis*, 1963, pp. 209–243.

[20] H. Sorensen, "A cepstral noise reduction multi-layer neural network," in *IEEE ICASSP*, 1991, pp. 933–936 vol.2.

[21] W. Jiang, Z. Liu, K. Yu, and F. Wen, "Speech enhancement with neural homomorphic synthesis," in *IEEE ICASSP*, 2022, pp. 376–380.

[22] J. Liu and X. Zhang, "Iccrn: Inplace cepstral convolutional recurrent neural network for monaural speech enhancement," in *IEEE ICASSP 2023*, 2023, pp. 1–5.

[23] E. Loweimi, J. Barker, and T. Hain, "Source-filter separation of speech signal in the phase domain," in *ISCA Interspeech 2015*, 2015, pp. 598–602.

[24] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants1audiofiles available. see http://www.elsevier.nl/locate/specom.1," *Speech Communication*, vol. 22, no. 4, pp. 403–417, 1997.

[25] A. Oppenheim, J. Lim, G. Kopec, and S. Pohlig, "Phase in speech and pictures," in *IEEE ICASSP-79. IEEE Intern Conf on Acoustics, Speech, and Signal Proc*, 1979, pp. 632–637.

[26] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529–541, 1981.

[27] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. USA: Prentice Hall Press, 2010.

[28] S. Lv, Y. Fu, M. Xing, J. Sun, L. Xie, J. Huang, Y. Wang, and T. Yu, "S-dccrn: Super wide band dccrn with learnable complex feature for speech enhancement," in *IEEE ICASSP*, 2021, pp. 7767–7771.

[29] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *IEEE ICASSP*, 2021, pp. 6623–6627.

[30] J. H. L. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *Interspeech*, 2018, pp. 2758–2762.

[31] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE ICASSP*, 2017, pp. 5220–5224.

[32] Y. Xia, S. Braun, C. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *IEEE ICASSP*, 2020, pp. 871–875.

[33] N. L. Westhausen and B. T. Meyer, "Dual-Signal Transformation LSTM Network for Real-Time Noise Suppression," *ISCA Interspeech 2020*, pp. 2477–2481, 2020.

[34] 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 *ISCA INTERSPEECH*, 08 2020, pp. 2472–2476.

[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," *IEEE ICASSP 2021*, pp. 6633–6637, 2021.

[36] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," *IEEE ICASSP 2001*, vol. 2, pp. 749–752 vol.2, 2001.

[37] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *IEEE ICASSP*, pp. 6493–6497, 2021.