



Exploiting Bispectral Features for Single-Channel Speech Enhancement

Venkatesh Parvathala, Ramesh Gundluru, Sreekanth Sankala, Sri Rama Murty Kodukula

Speech Information Processing Lab, Department of Electrical Engineering, Indian Institute of Technology Hyderabad, India

ee22resch01005@iith.ac.in, ee22m24p000001@iith.ac.in, ee20resch11011@iith.ac.in,
ksrm@ee.iith.ac.in

Abstract

Deep neural network-based speech enhancement (SE) algorithms often utilize an encoder - dual-path processing module - decoder architecture. The encoder, consisting of pointwise convolutions and a dilated densenet (DDN), maps time-frequency bins to a higher-dimensional space to better capture signal and noise characteristics but significantly adds to the computational complexity. In this work, we propose a computationally efficient modification by replacing the DDN block within the encoder with bispectral features, which effectively capture frequency correlations and phase information. We demonstrate the effectiveness of the method by incorporating it in two SE models, CMGAN and MP-SENet, and evaluating on two datasets, DNS-2020 and VoiceBank+DEMAND. Our experiments show that the proposed modification achieves similar or better performance while reducing the multiply-accumulate operations by 23.87% in CMGAN and 17.45% in MP-SENet, demonstrating its computational efficiency.

Index Terms: Speech Enhancement, Bispectrum, Densenet, Dual-Path Processing Module

1. Introduction

The objective of speech enhancement (SE) is to improve the quality and intelligibility of speech that has been degraded by additive background noise [1]. SE plays a critical role in various applications, including voice communication, voice assistants, hearing aids, and cochlear implants [2–4]. However, SE presents significant challenges due to the varying nature of noise across different environments [5, 6]. With the emergence of deep neural networks (DNN), substantial advancements have been made in overcoming these challenges, leading to notable improvements in SE performance.

DNN based algorithms can be broadly categorized into time-domain and frequency-domain approaches. Time-domain methods operate directly on raw samples using an encoder-separator-decoder architecture. The encoder maps the input signal to a feature space, the separator estimates a mask, and the decoder reconstructs the time-domain signal [7–11]. While these methods learn effective data-driven transformations, they are often less interpretable and impose high computational demands, as smaller filter and stride sizes are required to achieve optimal performance. In contrast, frequency-domain algorithms employ a fixed Fourier transform, estimating either a mask or the target spectra directly [12–16]. Although the Fourier transform representations are interpretable, it may not always serve as the optimal transformation for SE.

Recently, a new approach has gained attention, utilizing encoder - dual-path processing module (DPPM) - decoder architectures operating on fixed Fourier transform representa-

tions [17–21]. In this framework, the encoder is employed to extract efficient embeddings for each time-frequency bin, while the DPPM aggregates both temporal and spectral context within the learned embedding space. The output of the DPPM is further processed by the decoder to estimate the short-time Fourier transform (STFT) representations of clean speech. This method leverages the strengths of both learned and fixed transformations, enhancing the overall performance.

Although the encoder - DPPM - decoder based architectures achieve superior enhancement quality, their computational complexity is also higher, limiting their usage in resource constrained applications. Among the other components, the encoder, typically consisting of pointwise convolutions and a dilated densenet (DDN), contribute significantly to the model's overall complexity. Pointwise convolutions project the input into a higher-dimensional space, while the DDN is responsible for multi-scale feature learning [22]. This design expands the receptive field but multiplies computational operations due to repeated feature reuse. For instance, in CMGAN's architecture [19], the encoder accounts for 25% of total MACs and 14% of total parameters, with the DDN alone consuming 97% of the encoder's MACs and 95% of its parameters. In this work, we investigate whether a deterministic transformation can replace the learnable DDN, which could be replaceable given the DPPM's ability to aggregate longer contextual information. In this regard, we propose substituting the DDN with bispectral features, with the aim of reducing computational complexity while maintaining performance.

The bispectrum offers three key advantages that motivate its integration in a DNN framework: (1) it captures phase information, unlike conventional second-order spectra, (2) it models third-order frequency correlations, and (3) it is computationally efficient compared to the DDN. Moreover, the bispectrum is zero for symmetric density functions, such as Gaussian noise processes, and it remains flat for white noise processes with nonsymmetric density functions [23–25]. Previous works have taken advantage of these properties to suppress additive Gaussian noise [26–28]. In our approach, we extract bispectral features to assist the DNN in leveraging phase information and higher-order frequency relations. Experimental results show that replacing the DDN with bispectral features achieves similar or better enhancement quality while reducing the overall computational complexity.

2. Proposed Method

Figure 1(a) shows a typical frequency-domain encoder-DPPM-decoder architecture. In this section, we provide a detailed explanation of the key components within each module.

As depicted in Figure 1(b), the encoder is composed of

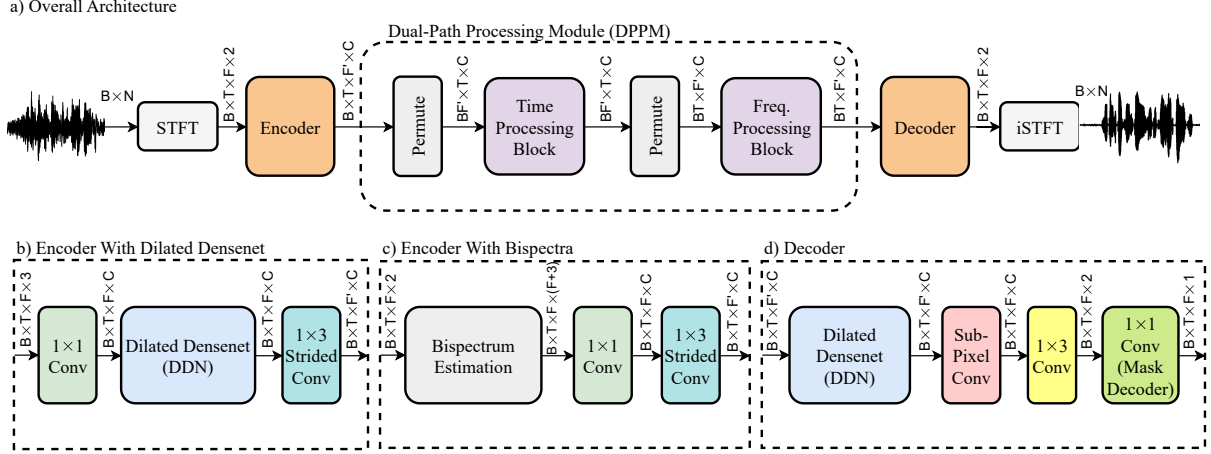


Figure 1: Overall network architecture

three blocks. A pointwise (1×1) convolution block to project the input feature maps to an intermediate feature space of dimension C . It is followed by a dilated densenet (DDN), comprising four dilated convolution blocks interconnected through dense residual connections. Finally, a 1×3 strided convolution block is used to downsample the frequency dimension to F' , thereby reducing the complexity of subsequent blocks. Each convolution block comprises a convolution layer, an instance normalization and a PReLU activation. The input to the encoder is formed by concatenating the magnitude, real, and imaginary components of the STFT representation.

The DPPM employs two distinct stages to effectively capture temporal and spectral dependencies. In the first stage, the encoder feature maps are reshaped to $X \in \mathbb{R}^{B F' \times T \times C}$ and processed using a sequence processing module. Here, B refers to batch size, F' is the reduced frequency dimension, and T is the total number of frames. The conformer is typically used as a sequence processing module to effectively capture both global and local contexts. In the second stage, the output of previous stage is reshaped to $X \in \mathbb{R}^{B T \times F' \times C}$ and processed using another conformer module. Multiple DPPM blocks are stacked to effectively model the temporal and spectral dependencies.

The decoder further processes the feature maps to estimate the final output. Different network architectures use different decoder configurations. For example, the CMGAN [19] employs two independent decoders: a mask decoder and a complex decoder. The mask decoder is responsible for estimating a mask on the magnitude, while the complex decoder predicts the residual real and imaginary components. Each decoder comprises a DDN module, followed by a sub-pixel convolution to upsample the frequency dimension back to F . A final convolution layer reduces the channel dimensions to 1 for the mask decoder and 2 for the complex decoder. For the mask decoder, the final mask is generated using a pointwise convolution, followed by a PReLU activation function. The masked magnitude is combined with the noisy phase and added to the output of the complex decoder to generate the final real and imaginary components. Finally, an inverse STFT (iSTFT) is used to compute time-domain signal from the estimated real and imaginary components.

The encoder plays an important role in learning efficient embedding vector. In the following subsection, we conduct an analysis of the encoder's feature maps.

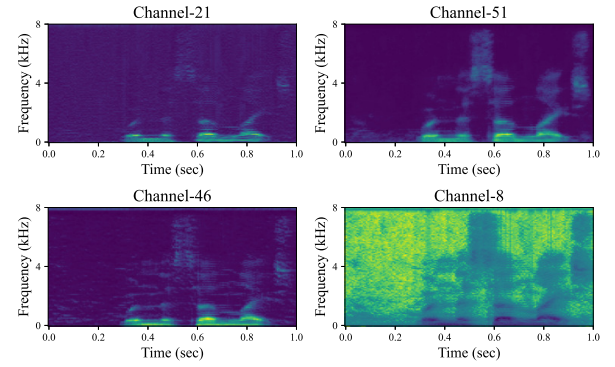


Figure 2: Output feature maps of the CMGAN's encoder for a male utterance (p232_006) from VoiceBank+DEMAND testset

2.1. Encoder Feature Maps

Figure 2 illustrates some of the output feature maps generated by the learned CMGAN's encoder. As shown in the figure, each feature map learns different characteristics of the speech signal. For instance, channel-21 captures formant transitions in the lower-frequency region, where signal-to-noise ratio (SNR) is higher, while channels-51, and 46 capture formant transitions at progressively higher frequencies with improved resolution. On the other hand, channel-8 focuses on lower SNR regions. Although the encoder learns interpretable feature maps, given that the DPPM's ability to effectively capture long-term temporal and spectral context, the longer contextual aggregation is redundant in the encoder. This observation motivates the replacement of the DDN with bispectral features, which can capture the necessary frequency relations and phase information. In the following subsection, we discuss the proposed encoder with bispectral features.

2.2. Encoder with Bispectral Features

The bispectra is defined as the two-dimensional Fourier transform of its third-order cumulant [23]. For a zero mean, stationary random process $g[n]$, the third-order cumulant is defined as

$$c_3[l, m] = \mathbb{E} [g[n]g[n+l]g[n+m]]$$

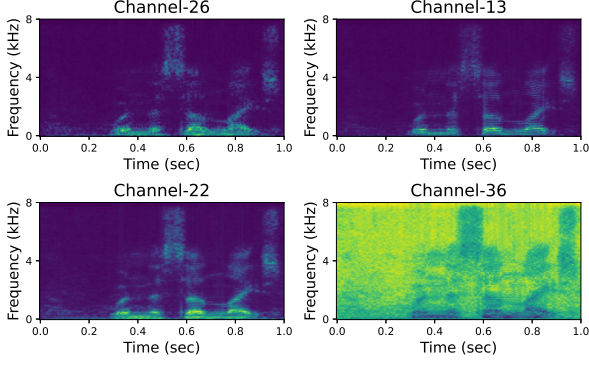


Figure 3: Output feature maps of the CMGAN's encoder with the bispectral features for a male utterance (p232_006) from VoiceBank+DEMAND test set

The bispectrum of this process is given by

$$\mathcal{B}(\omega_1, \omega_2) = \sum_l \sum_m c_3[l, m] e^{-j(\omega_1 l + \omega_2 m)}$$

The above moment based computation of bispectra is not suitable for speech as it is a nonstationary signal. Instead, we rely on estimating the bispectrum using non-parametric methods [24]. In particular, the direct bispectrum estimator for a discrete real signal $s[n]$, based on the periodogram is incorporated for each frame as follows [25, 26]

$$\hat{\mathcal{B}}[t, f_1, f_2] = S[t, f_1] S[t, f_2] S^*[t, f_1 + f_2] \quad (1)$$

where $S[t, f]$ is the STFT of the signal $s[n]$ defined as

$$S[t, f] = \sum_{p=0}^{P-1} s[p + tR] w[p] e^{-j \frac{2\pi p f}{F}}, \quad f = 0, 1, \dots, F-1$$

where $w[n]$ is a window of length P and shift R , t and f represents time and frequency indices, respectively. F is the DFT length and let $\bar{F}$ denotes the total number of non-redundant frequency bins considering conjugate symmetry. The sample bispectra also, by definition, has redundant regions due to symmetry [24]. The principal region of bispectrum is bounded by the triangle $f_1 + f_2 \leq \bar{F}$, $0 \leq f_2 \leq f_1$. The resulting bispectrum in this region is a complex quantity with dimension $\hat{\mathcal{B}}[t, f_1, f_2] \in \mathbb{C}^{T \times \bar{F} \times \frac{\bar{F}}{2}}$. We concatenate real and imaginary parts resulting in the final bispectral features $\hat{\mathcal{B}}[t, f_1, f_2] \in \mathbb{R}^{T \times \bar{F} \times \bar{F}}$. We then replace the DDN in the encoder block with the estimated bispectral features as shown in Figure 1(c). First, the sample bispectra is concatenated with magnitude, real, and imaginary components of STFT and normalized using instance normalization. A 1×1 convolution reduces the channel dimension from $(\bar{F} + 3)$ to C , followed by a strided convolution to downsample the frequency dimension to F' . The output of strided convolution is passed through the subsequent modules as shown in Figure 1(a).

A subset of the feature maps of the encoder with bispectral features, closely related to the feature maps of the encoder with DDN, are shown in Figure 3. A clear correspondence between the feature maps of encoder with DDN and their bispectral counterparts highlights the effectiveness of bispectral features in modeling the phase and higher-order frequency relations and it's potential to replace the DDN block within the encoder.

3. Experimental Deatails

3.1. Datasets

We train and evaluate our proposed method on two datasets. The first is the DNS-2020 dataset, which contains a large set of clean and noise samples collected from various sources [29]. Over 500 hours of clean samples from 2150 speakers are available with 180 hours of noise clips from 150 classes. For training, we prepare 500 hours of clean and noisy signal pairs by mixing a randomly chosen noise clip with clean speech at SNRs ranging from -5 to 20 dB. All the models are tested on 150 synthetic, non-reverberant, non-blind test pairs from the challenge.

The second dataset is the VoiceBank+DEMAND dataset, which is a publicly available small-scale dataset [30]. The dataset contains clean utterances from 28 speakers for training and 2 speakers for testing, with around 400 utterances of each speaker from the VoiceBank corpus [31]. Parallel noisy utterances are synthesized by mixing the clean speech with real noise recordings from the DEMAND database [32]. For training, eight types of real noises and two artificially generated noises – speech-shaped noise and babble noise – are used. Testing employs five unseen real noise types. The noisy utterances are synthesized at signal-to-noise ratios (SNRs) of 0 dB, 5 dB, 10 dB, and 15 dB for training, and at SNRs of 2.5 dB, 7.5 dB, 12.5 dB, and 17.5 dB for testing. All the utterances are downsampled to 16 kHz.

3.2. Network Configuration

We demonstrate the effectiveness of the bispectral features by incorporating them in two state-of-the-art architectures, CMGAN [19] and MP-SENet [20]. While the encoder and DPPM blocks remain same for both the architectures, the structure of decoder is changed. CMGAN uses two independent decoders, a mask decoder and a complex decoder. The mask decoder estimates a magnitude mask, while the complex decoder predicts the residual real and imaginary components. On the other hand, MP-SENet replaces the complex decoder with a phase decoder to explicitly estimate the phase. In addition to the time-domain, frequency-domain and metric discriminator based loss functions used in CMGAN, the MP-SENet uses anti-wrapping phase loss to accurately estimate the phase.

Power-law compressed STFT magnitude, real, and imaginary parts are used as input to the CMGAN, while magnitude and phase are used for MP-SENet. The STFT is computed with 25 ms window length and 6.25 ms shift. Both the architectures incorporate 4 DPPM blocks, and the number of channels C are set to 64. Other hyperparameter details can be found in [19] and [20] for CMGAN and MP-SENet, respectively. The CMGAN models are trained for 120 epochs on VoiceBank+DEMAND and 30 epochs on DNS-2020 dataset. The MP-SENet is trained for 80 epochs on VoiceBank+DEMAND dataset.

3.3. Evaluation Metrics

We evaluate all the models with six commonly used objective quality and intelligibility measures: perceptual evaluation of speech quality (PESQ) [33], short-time objective intelligibility (STOI) [34], three composite measures [35] (CSIG, CBAK, COVL), and segmental SNR [36]. For all metrics, higher values correspond to better performance. In addition to the quality and intelligibility measures, we assess the computational complexity of each model by measuring the number of trainable parameters and multiply-accumulate operations (MACs) required to enhance a one-second noisy signal.

Table 1: Comparison with state-of-the-art methods on VoiceBank+DEMAND testset

Method	PESQ	STOI (%)	CSIG	CBAK	COVL	SSNR (dB)	MACs (G)	Parameters (M)
Noisy	1.97	92	3.35	2.44	2.63	1.68	-	-
TSTNN [10]	2.96	95	4.33	3.53	3.67	9.70	20.28	0.92
DEMUCS [9]	3.07	95	4.31	3.40	3.63	-	7.72	33.53
MetricGAN+ [14]	3.15	-	4.14	3.16	3.64	-	0.12	1.89
DPT-FSNet [18]	3.33	96	4.58	3.72	4.00	-	46.36	0.88
CMGAN [19]	3.41	96	4.63	3.94	4.12	11.10	31.67	1.83
MP-SENet [20]	3.60	96	4.81	3.99	4.34	-	42.98	2.26
CMGAN* ¹	3.37	96	4.56	3.86	4.06	10.41	31.67	1.83
CMGAN* - w/o DDN	3.35	96	4.60	3.84	4.06	10.32	23.68	1.59
CMGAN* - Bispectra	3.40	96	4.62	3.90	4.11	10.74	24.11	1.60
MP-SENet* ²	3.55	96	4.76	3.96	4.27	10.71	42.98	2.26
MP-SENet* - w/o DDN	3.51	96	4.74	3.94	4.23	10.64	35.05	2.02
MP-SENet* - Bispectra	3.58	96	4.75	3.99	4.28	11.01	35.48	2.03

* indicate the results reproduced with the official implementations³

4. Results & Analysis

4.1. Evaluation on VoiceBank+DEMAND dataset

Table 1 presents a comparative analysis of the proposed method and other models trained on the VoiceBank+DEMAND dataset. Among these models, DEMUCS and TSTNN operate in the time domain, while the others are frequency-domain models. The encoder-DPPM-decoder networks in the frequency domain – DPT-FSNet, CMGAN, and MP-SENet – achieve superior performance across all quality and intelligibility metrics. Notably, CMGAN and MP-SENet attain state-of-the-art performance across all metrics but at a significant computational cost. Within the network architecture, the DDN in the encoder contributes substantially to the model’s overall computational load due to its four layers of dilated convolutions with dense connections. Specifically, it accounts for 13.11% of the total parameters and 25.23% of the total MACs in CMGAN. As presented in Table 1, removing the DDN block from the encoder of CMGAN and MP-SENet has significantly reduced the MACs and parameters. However, this comes at the expense of a significant drop in performance across all metrics. In contrast, replacing the DDN with bispectral features has proven to be a more efficient and effective solution. This modification results in a model that not only achieves higher scores across all metrics, but also reduces the computational burden. In CMGAN, replacing the DDN with bispectral features has improved the PESQ to 3.40 and also reduced the MACs and parameters by 23.87% and 12.57%, respectively. Similar improvements in quality metrics are observed in MP-SENet with 17.45% reduction in MACs and 10.18% reduction in parameters, demonstrating its effectiveness across different architectures.

4.2. Evaluation on DNS Dataset

Table 2 summarizes the comparison of the proposed method with the other models trained on DNS-2020 challenge dataset. Similar to the VoiceBank+DEMAND dataset, the encoder-DPPM-decoder architectures in frequency-domain – FullSubNet, DPT-FSNet, CMGAN – proved to be more effective compared with the other methods. As presented in the Table 2, Removing the DDN has adversely affected all the performance measures. In contrast, proposed replacement of DDN with bispectral features has achieved similar performance compared to its DDN counterpart, demonstrating its applicability across the

Table 2: Comparison with state-of-the-art methods on DNS-2020 non-reverb testset

Method	PESQ	STOI (%)	SSNR	MACs (G)
Noisy	1.58	92	5.96	-
DEMUCS [9]	2.65	97	13.39	7.72
POCONet [13]	2.75	-	-	-
FullSubNet [17]	2.89	96	12.14	30.87
DPT-FSNet [18]	3.26	98	-	46.36
CMGAN	3.36	98	14.72	31.67
CMGAN - w/o DDN	3.29	97	14.67	23.68
CMGAN - Bispectra	3.35	98	14.87	24.11

datasets. Source codes for bispectral feature extraction and its incorporation in CMGAN and MP-SENet will be released upon publication of this work at <https://github.com/SIPLab-IITH/BiSpectral-SE/>

5. Conclusions

In this work, we have analyzed the role of the DDN block within the encoder of a frequency-domain encoder-DPPM-decoder architectures. The DDN block, through its use of dilated convolutions, aggregates long-range contextual information, while its dense connections facilitate multi-scale feature learning. However, given that the DPPM is designed to capture both long-term temporal and spectral context, the DDN block in the encoder becomes less effective. We take advantage of this by replacing the DDN with bispectral features, which effectively capture both higher-order frequency relations and phase information. We validate the proposed method across two state-of-the-art architectures (CMGAN and MP-SENet) and two benchmark datasets (VoiceBank+DEMAND and DNS-2020). Our experimental results demonstrate that incorporating bispectral features achieves comparable or better performance to DDN-based models, while significantly reducing computational complexity. In our future work, we aim to exploit the DPPM and decoder modules to reduce their computations.

¹<https://github.com/ruizhecao96/CMGAN>

²<https://github.com/yxlu-0102/MP-SENet>

³Note that the reproduced results do not exactly match with the original paper’s results. This issue has been discussed by other researchers as well at the corresponding repositories issue tracker, particularly <https://github.com/ruizhecao96/CMGAN/issues/37> and <https://github.com/yxlu-0102/MP-SENet/issues/15>

6. References

- [1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.
- [2] T. V. Ramabadran, J. P. Ashley, and M. J. McLaughlin, "Background noise suppression for speech enhancement and coding," in *1997 IEEE Workshop on Speech Coding for Telecommunications Proceedings*. IEEE, 1997, pp. 43–44.
- [3] L.-P. Yang and Q.-J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *The Journal of the Acoustical Society of America*, vol. 117, no. 3, 2005.
- [4] J. Gnanamanickam, Y. Natarajan, and S. Ramasamy, "A hybrid speech enhancement algorithm for voice assistance application," *Sensors*, vol. 21, p. 7025, 10 2021.
- [5] T. Virtanen, R. Singh, and B. Raj, *Techniques for noise robustness in automatic speech recognition*. John Wiley & Sons, 2012.
- [6] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.
- [7] 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, 2019.
- [8] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.
- [9] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020, pp. 3291–3295.
- [10] K. Wang, B. He, and W.-P. Zhu, "TSTNN: Two-Stage Transformer Based Neural Network for Speech Enhancement in the Time Domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
- [11] E. Kim and H. Seo, "SE-Conformer: Time-domain speech enhancement using conformer," in *Interspeech*, 2021, pp. 2736–2740.
- [12] A. Nicolson and K. K. Paliwal, "Masked multi-head self-attention for causal speech enhancement," *Speech Communication*, vol. 125, pp. 80–96, 2020.
- [13] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, K. Helwani, and A. Krishnaswamy, "PoCoNet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," in *Interspeech*, 2020, pp. 2487–2491.
- [14] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Interspeech*, 2021, pp. 201–205.
- [15] L. Zhang, M. Wang, Q. Zhang, X. Wang, and M. Liu, "PhaseDCN: A Phase-Enhanced Dual-Path Dilated Convolutional Network for Single-Channel Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2561–2574, 2021.
- [16] V. Parvathala, S. Andhavarapu, G. Pamisetty, and K. Murty, "Neural comb filtering using sliding window attention network for speech enhancement," *Circuits, Systems, and Signal Processing*, pp. 1–22, 2022.
- [17] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6633–6637.
- [18] F. Dang, H. Chen, and P. Zhang, "DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6857–6861.
- [19] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric gan for speech enhancement," in *Interspeech*, 2022, pp. 936–940.
- [20] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *Interspeech*, 2023, pp. 3834–3838.
- [21] R.-Y. Chao, W.-H. Cheng, M. L. Quatra, S. M. Siniscalchi, C.-H. H. Yang, S.-W. Fu, and Y. Tsao, "An investigation of incorporating mamba for speech enhancement," *IEEE Spoken Language Technology Workshop (SLT)*, pp. 302–308, 2024.
- [22] A. Antoniou, A. Słowik, E. J. Crowley, and A. Storkey, "Dilated densenets for relational reasoning," *arXiv preprint arXiv:1811.00410*, 2018.
- [23] A. W. Lohmann and B. Winitzer, "Triple correlations," *Proceedings of the IEEE*, vol. 72, no. 7, pp. 889–901, 1984.
- [24] D. A. Douglass, "Voice analysis using the bispectrum," *thesis*, 1994.
- [25] J. W. Fackrell, "Bispectral analysis of speech signals," *thesis*, 1997.
- [26] R. Fulchiero and A. S. Spanias, "Speech enhancement using the bispectrum," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. IEEE, 1993, pp. 488–491.
- [27] D. R. Green, "The utility of higher-order statistics in gaussian noise suppression," Ph.D. dissertation, Naval Postgraduate School, 2003.
- [28] T. Lavanya, P. Vijayalakshmi, K. Mrinalini, and T. Nagarajan, "Higher order statistics-driven magnitude and phase spectrum estimation for speech enhancement," *Computer Speech & Language*, vol. 87, p. 101639, 2024.
- [29] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matushevych, 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 *Interspeech*, 2020, pp. 2492–2496.
- [30] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in *SSW*, 2016, pp. 146–152.
- [31] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation, O-COCOSDA/CASLRE 2013*, 2013.
- [32] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, 2013.
- [33] 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 *IEEE international conference on acoustics, speech, and signal processing. Proceedings (ICASSP)*, vol. 2. IEEE, 2001, pp. 749–752.
- [34] 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 - Proceedings (ICASSP)*, 2010.
- [35] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, 2008.
- [36] J. H. L. Hansen and B. L. Pellom, "An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms," *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998.