# Single-Channel Speech Enhancement in Spherical-Mapped Short-Time Spectral Domain

Yu Morinaga, Naoto Kotake, Iori Hashimoto, Suehiro Shimauchi, and Shigeaki Aoki
Kanazawa Institute of Technology, Japan
E-mail: c6401263@st.kanazawa-it.ac.jp

***Abstract*— A single-channel speech enhancement method is proposed, which performs feature extraction and mask processing in spherical-mapped short-time spectral (SMSTS) domain. The features in the SMSTS domain are represented as three-dimensional direction cosines, obtained by mapping complex-valued short-time spectral components onto a three-dimensional unit sphere. The direction cosines have continuous properties, in contrast to the inherently discontinuous nature of conventional short-time spectral phase. Simulation results demonstrate that the proposed method outperforms conventional speech enhancement approaches, which use real and imaginary components in the complex-valued short-time spectral domain as features, in terms of scale-invariant signal-to-distortion ratio (SI-SDR) and short-time objective intelligibility (STOI) metrics.**

## I. Introduction

Speech-based applications, such as cellular communications, voice assistants, hearing aids, and teleconferencing systems, play crucial roles in daily life. A key technical challenge is mitigating quality degradation caused by noise that corrupts speech during recordings in noisy environments. Speech enhancement or noise suppression techniques aimed at reducing noise and enhancing speech clarity in noisy environments have been investigated over the years. Speech enhancement methods can be realized using both single-channel and multi-channel audio inputs. Multi-channel [1] approaches leverage spatial differences between target sound and noise sources through beamforming with multiple microphones, while single-channel [2] approaches use statistical properties of time-frequency characteristics through single-microphone information. Improving single-channel speech enhancement remains critical, as it requires only a single microphone and can be seamlessly integrated into existing legacy single-microphone systems.

Most single-channel speech enhancement methods transform the input speech signal into a short-time spectral domain based on short-time Fourier transform (STFT). In this domain, the desired speech features are estimated and enhanced. The enhanced output signal is reconstructed using inverse STFT (ISTFT). The short-time spectrum of the input speech signal obtained by STFT has complex values and is typically represented as an amplitude spectrum and a phase spectrum via polar coordinate transformation. The phase spectrum is often regarded as having minimal impact on sound quality enhancement, leading many methods to focus primarily on the amplitude spectrum [3], which is more visually interpretable and easier to model. However, recently, the importance of estimating the phase spectrum has been reaffirmed [4]–[6]. The primary challenge in estimating the phase spectrum lies in its discontinuous and complex nature, which makes it difficult to model directly using the same techniques employed for the amplitude spectrum. Therefore, some deep-learning methods have been proposed to estimate the complex spectrum as a pair of real and imaginary parts [7]–[13], instead of directly estimating the phase.

Given the continuity of the real and imaginary parts of the complex spectrum, estimating these components is more practical than directly estimating the discontinuous phases. However, this approach may not effectively exploit the intuitively accessible features of the amplitude spectrum. To address this issue, a feature representation method was proposed that maps the short-time spectrum in a two-dimensional complex plane onto a three-dimensional sphere, where the direction angles relative to the axes of the three-dimensional coordinates are used as feature values [14]. In this method, unlike conventional phase angles, each direction angle remains continuous.

In this study, we propose a single-channel speech enhancement method that performs feature extraction and mask processing in spherical-mapped short-time spectral (SMSTS) domain [14]. According to the evaluation results [14] based on nonlinear correlations calculated using the maximal information coefficient (MIC) [15], both the real and imaginary spectra tend to be highly correlated with both the amplitude and phase spectra. In contrast, the direction angle spectra in the SMSTS domain exhibit lower correlation with either amplitude or phase spectra, respectively. Therefore, the feature representation based on the direction angles is expected to capture the intuitive characteristics of amplitude and phase more effectively than the real and imaginary representations while avoiding the discontinuity of the phase.

## II. Speech Enhancement with Complex-Domain Ratio Masking

This section describes an example of a conventional speech enhancement system based on a complex ideal ratio mask (cIRM) estimation [7], as shown in Fig. 1. A complex spectrogram $\mathbb{S}_a$ of size $M \times K$ is obtained by applying STFT to a noisy speech signal $a(n) = c(n) + v(n)$, where $M$ and $K$ denote the number of frames and frequency bins, respectively, $c(n)$ and $v(n)$ are a clean speech and noise signals, and $n$ indicates discrete time index. Each element of

This work was partially supported by JSPS KAKENHI Grant Number JP22K12102.

979-8-3315-7206-8/25/$31.00 

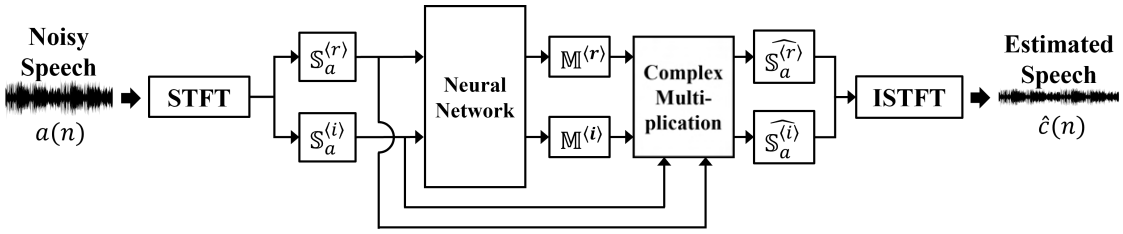



Fig. 1: System configuration with complex-domain ratio masking.

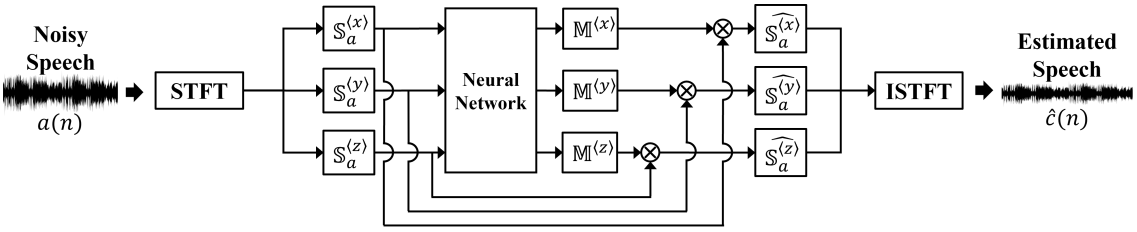



Fig. 2: System configuration with SMSTS-domain ratio masking.

$\mathbb{S}_a$ is represented as $S_{m,k}$, corresponding to the $m$-th frame and the $k$-th frequency bin, i.e.,

$$\mathbb{S}_a = [S_{m,k}], \tag{1}$$
$$= \mathbb{S}_a^{<r>} + j\mathbb{S}_a^{<i>}, \tag{2}$$
$$= \left[S_{m,k}^{<r>} + jS_{m,k}^{<i>}\right], \tag{3}$$

where $\mathbb{S}_a^{<r>}$ and $\mathbb{S}_a^{<i>}$ denote the real and imaginary parts of $\mathbb{S}_a$, and $S_{m,k}^{<r>}$ and $S_{m,k}^{<i>}$ denote their elements. By applying $\mathbb{S}_a^{<r>}$ and $\mathbb{S}_a^{<i>}$ to a neural network as inputs, a complex mask

$$\mathbb{M} = \mathbb{M}^{<r>} + j\mathbb{M}^{<i>}, \tag{4}$$

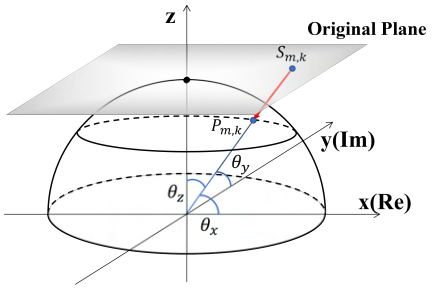
where

$$[M_{m,k}] = \left[M_{m,k}^{<r>} + jS_{m,k}^{<i>}\right], \tag{5}$$

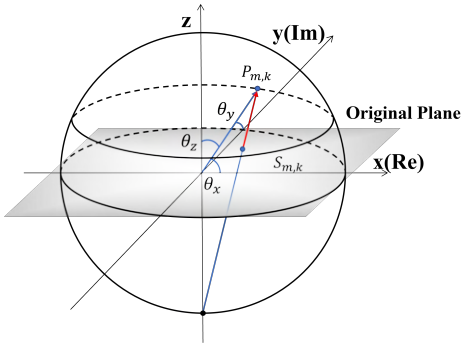
in element-wise form, is estimated. Subsequently, a masked spectrogram is obtained by element-wise complex multiplication,

$$\tilde{\mathbb{S}}_a = \left[(S_{m,k}^{<r>} + jS_{m,k}^{<i>}) \cdot (M_{m,k}^{<r>} + jM_{m,k}^{<i>})\right]. \tag{6}$$

By applying ISTFT to $\tilde{\mathbb{S}}_a$, the clean speech estimate $\hat{c}(n)$ is reconstructed.




(a) Hemisphere mapping.




(b) Whole-sphere mapping.

Fig. 3: Concept of spherical mappings.

## III. Speech Enhancement with SMSTS-Domain Ratio Masking

In this section, we propose a novel speech enhancement approach that masks unwanted components in spherical-mapped short-time spectral (SMSTS) domain, whose configuration is shown in Fig. 2.

### A. Spherical-Mapping of Short-Time Spectral Components

The concept of spherical mapping of short-time spectrum [14] is introduced here to address both the $2\pi$-phase jumps caused by phase wrapping and the $\pi$-phase jumps owing to amplitude zero crossing. There are two types of mapping schemes: hemisphere and whole-sphere mappings.

Figure 3(a) shows the hemisphere mapping scheme. The unit hemisphere of radius 1 is placed in the three-dimensional $xyz$ space, where the original complex spectral component $S_{m,k}$ is on the plane at $z = 1$. Subsequently, $S_{m,k}$ is mapped onto the hemisphere as the point $P_{m,k}$ that intersects the line connecting the spectral component and the origin in $xyz$ space. The cosine components of the direction angles $\theta_{m,k}^{<x>}$, $\theta_{m,k}^{<y>}$ and $\theta_{m,k}^{<z>}$ are

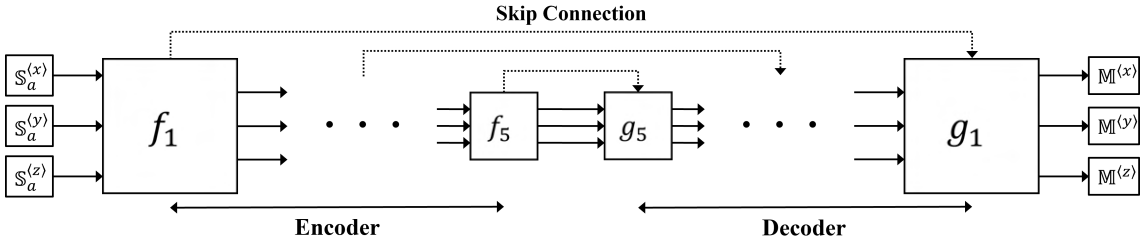



Fig. 4: U-Net structure.

calculated as follows:

$$\cos\theta_{m,k}^{<x>} = \frac{S_{m,k}^{<r>}}{\sqrt{(S_{m,k}^{<r>})^2 + (S_{m,k}^{<i>})^2 + 1}}, \tag{7}$$

$$\cos\theta_{m,k}^{<y>} = \frac{S_{m,k}^{<i>}}{\sqrt{(S_{m,k}^{<r>})^2 + (S_{m,k}^{<i>})^2 + 1}}, \tag{8}$$

$$\cos\theta_{m,k}^{<z>} = \frac{1}{\sqrt{(S_{m,k}^{<r>})^2 + (S_{m,k}^{<i>})^2 + 1}}, \tag{9}$$

where $\theta_{m,k}^{<x>} \in [0, \pi]$, $\theta_{m,k}^{<y>} \in [0, \pi]$, and $\theta_{m,k}^{<z>} \in [0, \pi/2]$.

Similarly, Fig. 3(b) shows the whole-sphere mapping scheme. Here, the original complex spectral component $S_{m,k}$ is on the plane at $z = 0$. Subsequently, $S_{m,k}$ is mapped to the unit whole-sphere as the point $P_{m,k}$ intersecting the line connecting the spectral component and the point $(0, 0, -1)$. The cosine components of the direction angles $\theta_{m,k}^{<x>}$, $\theta_{m,k}^{<y>}$ and $\theta_{m,k}^{<z>}$ are calculated as follows:

$$\cos\theta_{m,k}^{<x>} = \frac{2S_{m,k}^{<r>}}{(S_{m,k}^{<r>})^2 + (S_{m,k}^{<i>})^2 + 1}, \tag{10}$$

$$\cos\theta_{m,k}^{<y>} = \frac{2S_{m,k}^{<i>}}{(S_{m,k}^{<r>})^2 + (S_{m,k}^{<i>})^2 + 1}, \tag{11}$$

$$\cos\theta_{m,k}^{<z>} = \frac{1 - (S_{m,k}^{<r>})^2 - (S_{m,k}^{<i>})^2}{(S_{m,k}^{<r>})^2 + (S_{m,k}^{<i>})^2 + 1}, \tag{12}$$

where $\theta_{m,k}^{<x>} \in [0, \pi]$, $\theta_{m,k}^{<y>} \in [0, \pi]$, and $\theta_{m,k}^{<z>} \in [0, \pi]$.

The original complex spectral component $S_{m,k}$ can be reconstructed by employing the directional cosines $\cos\theta_{m,k}^{<x>}$, $\cos\theta_{m,k}^{<y>}$ and $\cos\theta_{m,k}^{<z>}$ as

$$S_{m,k} = \frac{\cos\theta_{m,k}^{<x>}}{\cos\theta_{m,k}^{<z>}} + j\frac{\cos\theta_{m,k}^{<y>}}{\cos\theta_{m,k}^{<z>}} \tag{13}$$

in the hemisphere case, and as

$$S_{m,k} = \frac{\cos\theta_{m,k}^{<x>}}{1 + \cos\theta_{m,k}^{<z>}} + j\frac{\cos\theta_{m,k}^{<y>}}{1 + \cos\theta_{m,k}^{<z>}}, \tag{14}$$

in the whole-sphere case.

### B. SMSTS-Domain Masking Procedure

The diagram in Fig. 2 illustrates a speech enhancement procedure based on the SMSTS-domain ratio masking. The three direction cosine spectrograms $\mathbb{S}_a^{<d>} = [\cos\theta_{m,k}^{<d>}]$, where $d \in \{x, y, z\}$, are calculated from each element of the original spectrogram $\mathbb{S}_a$, based on (7) to (9) for the hemisphere scenario, or on (10) to (12) for the whole-sphere scenario. By applying the three spectrograms $\mathbb{S}_a^{<d>}$ to a neural network as inputs, three mask spectrograms $\mathbb{M}^{<d>}$, whose elements are denoted as $M_{m,k}^{<d>}$, are estimated. Thus, the masked direction cosine spectrograms $\tilde{\mathbb{S}}_a^{<d>}$ are obtained as

$$\tilde{\mathbb{S}}_a^{<d>} = \left[\cos\theta_{m,k}^{<d>} \cdot M_{m,k}^{<d>}\right]. \tag{15}$$

Consequently, the masked complex spectrogram $\tilde{\mathbb{S}}_a$ is reconstructed by element-wisely applying either the hemisphere-mapping reconstruction formula (13) or the whole-sphere-mapping reconstruction formula (14) to $\tilde{\mathbb{S}}_a^{<d>}$ depending on the mapping scenario.

## IV. Simulations

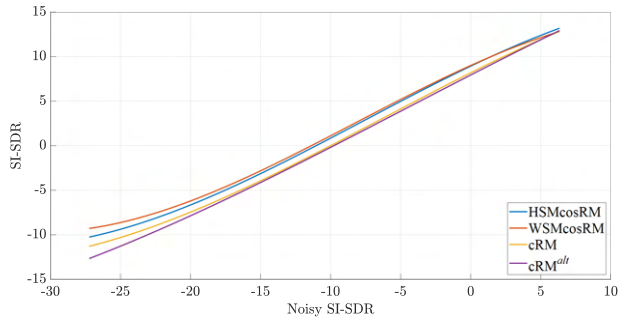
In this section, we present the simulation results to compare the performances of the conventional complex-domain and the proposed SMSTS-domain speech enhancement processing.

### A. Simulation Conditions

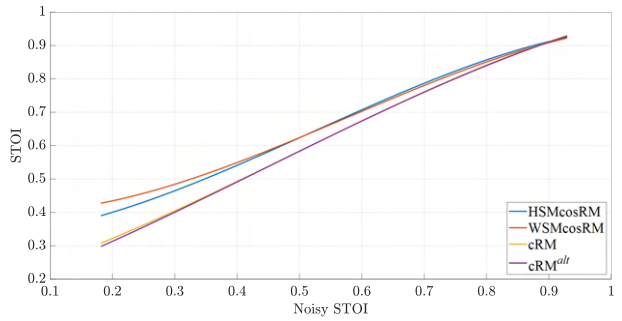
We compared the speech enhancement performances of two proposed methods and two conventional methods, i.e., HSMcosRM (hemisphere-mapped direction cosine ratio masking), WSMcosRM (whole-sphere-mapped direction cosine ratio masking), cRM (complex ratio masking as in (6)), and cRM$^{alt}$, where cRM$^{alt}$ is an alternative complex ratio masking with the following multiplication [7],

$$\tilde{\mathbb{S}}_a = \left[S_{m,k}^{<r>} \cdot M_{m,k}^{<r>} + jS_{m,k}^{<i>} \cdot M_{m,k}^{<i>}\right]. \tag{16}$$

For the mask estimation in both the complex- and SMSTS-domains, U-Net [16] was employed as a neural network in Figs. 1 and 2. Figure 4 shows a schematic diagram of the U-Net. Although this diagram reflects the SMSTS-domain implementation in terms of the input channel, the overall structure remains consistent with that of the complex-domain implementation. The encoder and decoder consisted of five symmetrical convolutional layers, interconnected by skip connections at the corresponding hierarchical level. The input data size for each channel at the top layer was $257 \times 512$, while the size at the bottom layer was $16 \times 16$. The learning rate was set to 0.003, batch size was set to 16, and number of epochs was set to 100.




(a) SI-SDR relationships: Unprocessed versus processed signals.




(b) STOI relationships: Unprocessed versus processed signals.

Fig. 5: Performance comparisons for SI-SDR and STOI scores (third-order regression curves fitted by 600 data clips).

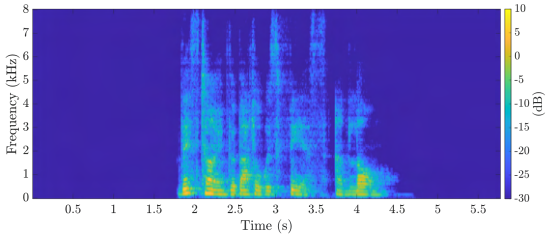
We utilized a portion of ICASSP 2023 Clarity Challenge dataset for training and evaluation. In this study, 6,000 pairs of clean and noisy speech data clips were randomly divided into 80 percent, 10 percent, and 10 percent for training, validation, and evaluation, respectively. The original audio was downsampled from 44.1 kHz to 16 kHz. The STFT frame size was 512 (32 ms), hop size was 128 (8 ms), and the window function was a Hann window of 512 sample length.

Scale-invariant signal-to-distortion ratio (SI-SDR) [17] and short-time objective intelligibility (STOI) [18] were used as evaluation metrics.

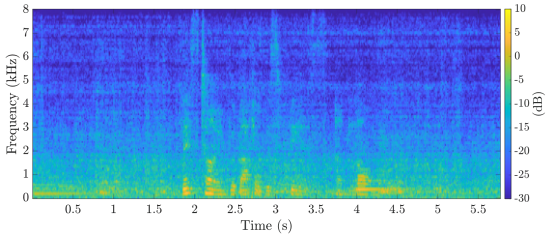
### *B. Results and Discussions*

Figure 5(a) illustrates the relationships between the SI-SDR of the noisy speech before processing (horizontal axis) and the SI-SDR of the speech after processing by each method (vertical axis). Figure 5(b) illustrates the relationships for the STOI scores in the same manner. Each curved line in the graph represents a third-order polynomial regression line, fitted to the results obtained from the 600 test data clips for each method. These figures demonstrate that the proposed two methods outperform the two conventional methods in both SI-SDR and STOI scores. Notably, the performance gap widened as the level of noise in unprocessed speech increased. When comparing the two proposed methods, WSMcosRM exhibited slightly better performance than HSMcosRM.

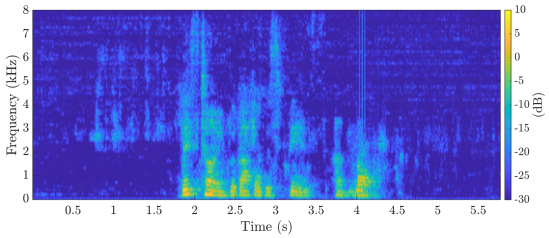
Figures 6(a) to 6(f) compare the amplitude spectrograms for an example of the simulation results. Figures 6(a) and 6(b) show the amplitude spectrograms of the target clean



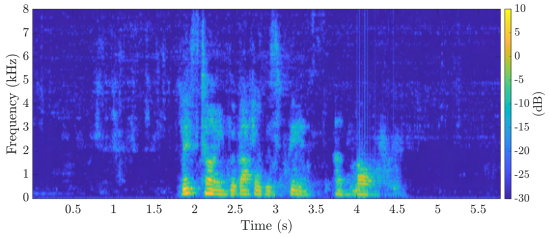

(a) Target speech.



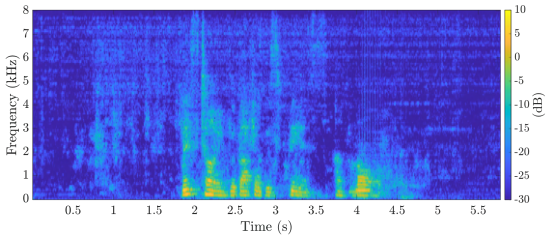

(b) Noisy speech.



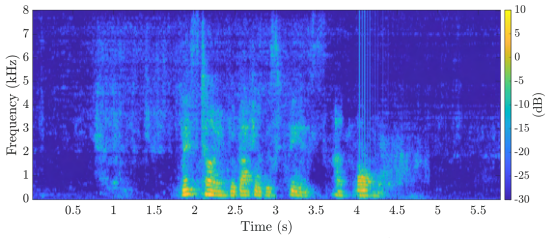

(c) Processed by HSMcosRM.




(d) Processed by WSMcosRM.




(e) Processed by cRM.




(f) Processed by cRM$^{alt}$.

Fig. 6: Amplitude Spectrograms.

speech, and the noisy speech, respectively. Figures 6(c) to 6(f) show the amplitude spectrograms obtained with HSMcosRM, WSMcosRM, cRM, and cRM$^{alt}$, respectively. From these figures, the proposed methods, HSMcosRM and WSMcosRM, consistently outperform the conventional methods, cRM and cRM$^{alt}$.

## V. Conclusions

In this study, we proposed a single-channel speech enhancement method that performs feature extraction and mask processing in the SMSTS domain. The features in the SMSTS domain are represented as three-dimensional direction cosines, which avoid discontinuities inherent in the original short-time spectral phase. Compared to conventional speech enhancement methods that use real and imaginary parts in the complex STFT domain as features, the proposed method outperformed in terms of SI-SDR and STOI metrics under our simulation settings. The results indicated that directional cosines in the SMSTS domain serve as effective features for enhancing speech quality.

## References


[1] G. Huang, J. R. Jensen, J. Chen, *et al.*, "Advances in Microphone Array Processing and Multichannel Speech Enhancement," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[2] P. Ochieng, "Deep Neural Network Techniques for Monaural Speech Enhancement and Separation: State of the Art Analysis," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 3651–3703, Dec. 2023.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[4] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-Sensitive and Recognition-Boosted Speech Separation Using Deep Recurrent Neural Networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[5] 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, Mar. 2015.

[6] 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, Apr. 2020, pp. 9458–9465.

[7] D. S. Williamson, Y. Wang, and D. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[8] A. Ephrat, I. Mosseri, O. Lang, *et al.*, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, Aug. 2018.

[9] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-Aware Speech Enhancement with Deep Complex U-Net," in *International Conference on Learning Representations*, Sep. 2018.

[10] K. Tan and D. Wang, "Complex Spectral Mapping with a Convolutional Recurrent Network for Monaural Speech Enhancement," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6865–6869.

[11] S. Zhao and B. Ma, "D2Former: A Fully Complex Dual-Path Dual-Decoder Conformer Network Using Joint Complex Masking and Complex Spectral Mapping for Monaural Speech Enhancement," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[12] Y. Hu, Y. Liu, S. Lv, *et al.*, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Interspeech 2020*, 2020, pp. 2472–2476.

[13] W. Chen, R. Yu, and Z. Ye, "Decoupling-Style Monaural Speech Enhancement with a Triple-Branch Cross-Domain Fusion Network," *Applied Acoustics*, vol. 217, p. 109 839, Feb. 2024.

[14] Y. Morinaga, N. Kotake, I. Hashimoto, S. Shimauchi, and S. Aoki, "Spherical Mapping of Short-time Spectral Components," in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 105–109.

[15] D. N. Reshef, Y. A. Reshef, H. K. Finucane, *et al.*, "Detecting Novel Associations in Large Data Sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 234–241.

[17] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-Baked or Well Done?" In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[18] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.