Leveraging Joint Spectral and Spatial Learning with MAMBA for Multichannel Speech Enhancement

Wenze Ren[1], Haibin Wu[1], Yi-Cheng Lin[1], Xuanjun Chen[1], Rong Chao[24], Kuo-Hsuan Hung[34]
You-Jin Li[14], Wen-Yuan Ting[4], Hsin-Min Wang[4], Yu Tsao[4]
[1]*Graduate Institute of Communication Engineering, National Taiwan University*
[2]*Department of Computer Science and Information Engineering, National Taiwan University*
[3]*Department of Biomedical Engineering, National Taiwan University* [4]*Academia Sinica, Taiwan*
r11942166@ntu.edu.tw, yu.tsao@citi.sinica.edu.tw

***Abstract*—In multichannel speech enhancement, effectively capturing spatial and spectral information across different microphones is crucial for noise reduction. Traditional methods, such as CNN or LSTM, attempt to model the temporal dynamics of full-band and sub-band spectral and spatial features. However, these approaches face limitations in fully modeling complex temporal dependencies, especially in dynamic acoustic environments. To overcome these challenges, we modify the current advanced model McNet by introducing an improved version of Mamba, a state-space model, and further propose MCMamba. MCMamba has been completely reengineered to integrate full-band and narrow-band spatial information with sub-band and full-band spectral features, providing a more comprehensive approach to modeling spatial and spectral information. Our experimental results demonstrate that MCMamba significantly improves the modeling of spatial and spectral features in multichannel speech enhancement, outperforming McNet and achieving very promising performance on the CHiME-3 dataset. Additionally, we find that Mamba performs exceptionally well in modeling spectral information.**

***Index Terms*—multichannel speech enhancement, spatial, spectral, state space model, Mamba**

I. Introduction

Multichannel speech enhancement (SE) is a critical and challenging research topic in speech signal processing. In addition to the spectral information from each channel, the spatial information captured by microphone arrays is an important cue for multichannel SE. Compared to single channel methods [1–5], multichannel techniques have the potential to significantly enhance speech quality by effectively incorporating both spatial and spectral information. Spatial information refers to the difference in sound signals captured by multiple microphones, which helps determine the location of the sound source. This enables the SE systems to distinguish between speech and noise based on direction. Spectral information, on the other hand, refers to the frequency content of the signal, helping separate speech from noise by analyzing different frequency bands. Combining both spatial and spectral data improves the performance of SE.

Recently, there are models specifically designed to process spatial and spectral information across multiple channels. Notably, this advancement is exemplified by mechanisms such as the channel attention approach introduced in [6]. In addition, specific neural networks are designed to specifically target the spatial distinctions between narrow-band speech and noise [7, 8], further optimizing spatial information processing. The spectral domain is another crucial feature for differentiating clean speech from noise. Researchers have demonstrated that both full-band and sub-band regions contain valuable spectral information [9]. The state-of-the-art performance is achieved by modeling the spatial information in narrow-band and full-band frequencies while effectively combining sub-band and full-band features [10].

Mamba [11], a recent state-space model with a selection mechanism, has demonstrated high performance in single-channel speech separation [12, 13] and SE [14]. Building on this, SpatialNet introduces a variant of Mamba tailored for long-term streaming multichannel SE [15], particularly suited for scenarios involving stationary and moving speakers. However, SpatialNet focuses primarily on utilizing spatial information. Research exploring the joint use of equally important spatial and spectral features in state-space models for multichannel SE remains unexplored.

This study proposes a novel multichannel SE system based on Mamba, which aims to effectively capture spatial and spectral information. We develop causal and non-causal versions of the system to meet practical application requirements. The specially designed Mamba module is dedicated to modeling spatial and spectral characteristics. Extensive experiments show that on the CHiME-3 dataset, our method achieves **state-of-the-art** performance, outperforming McNet [10] across all SE metrics. A careful ablation study in controlled settings reveals that Mamba is particularly advantageous in spectral modeling.

II. Methodology

This section first presents the Mamba framework and its two novel variants: Uni-Mamba and Bi-Mamba in Fig. 1. Subsequently, we introduce the innovative MCMamba model, which builds upon and extends the capabilities of the Mamba and Bi-Mamba architectures.

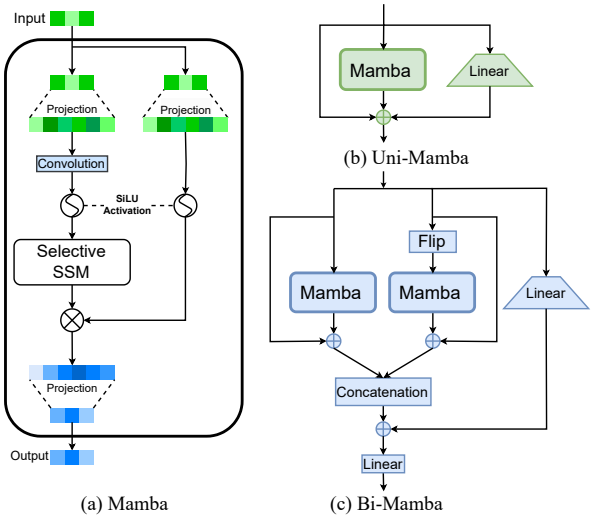


Fig. 1: (a) The basic model architecture of Mamba, (b) The proposed Uni-directional Mamba block, (c) The proposed Bi-directional Mamba block.

A. Core Mamba Architecture

Fig. 1(a) illustrates the core architecture of Mamba. Mamba integrates the H3 architecture [16] and a gated multilayer perceptron block into a stacked structure. Its main components include three linear layers, a deep convolutional network, and a selective State-Space Model (SSM). Given an input X and an output Y, the system can be described by the following mathematical formulation:

$$Y' \leftarrow \text{SSM}\left(\text{SiLU}\left(\text{Conv}\left(\text{Linear}(X)\right)\right)\right) \odot \text{SiLU}\left(\text{Linear}(X)\right),$$

$$Y \leftarrow \text{Linear}(Y').$$

The selective SSM, which is the core processing unit in Mamba, maps an input x to an output y through a higher-dimensional latent state h, described by the following equations:

$$h_n = \bar{A}h_{n-1} + \bar{B}x_n,$$

$$y_n = Ch_n,$$

where $\bar{A}$ and $\bar{B}$ represent discretized state matrices. The discretization process transforms continuous parameters (Δ, A, B) into their discrete counterparts $(\bar{A}, \bar{B})$, enabling the model to process discrete-time audio signals effectively [11].

The Structured SSM [17] offers a compelling alternative to convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for handling long-range dependencies in sequential data [18]. Building on this foundation, the Mamba model incorporates a novel input selection mechanism, effectively merging the strengths of CNNs and RNNs, significantly enhancing its capacity to model discrete data within the SSM framework. This mechanism allows the model to dynamically parameterize SSM components based on the input, selectively focusing on or ignoring specific parts of the input sequence. Consequently, Mamba excels at efficiently processing complex and extended sequences.

B. Uni-directional and bi-directional Mamba

To address the distinct requirements of real-time and offline multichannel SE, we propose two variants of the Mamba model: Uni-Mamba in Fig. 1(b) and Bi-Mamba in Fig. 1(c).

Uni-Mamba is a causal model specifically designed for real-time applications, where only the past and present information is processed. This characteristic makes Uni-Mamba particularly well-suited for real-time multichannel SE tasks, ensuring efficient performance under strict latency constraints. A key innovation is to **modify the traditional residual connection**. Unlike SPmamba [12] and SEmamba [14], which combine Mamba with a traditional residual structure, we add an additional learnable linear residual path. This provides the model with an additional representation that complements other paths. The linear residual connection of Uni-Mamba has an input with the same dimension as the output. This modification introduces greater flexibility and enhances the model's capacity to capture more complex representations and dependencies within the speech data, ultimately improving the performance in multichannel SE as shown in the experiments.

Bi-Mamba, in contrast, is a non-causal model designed for offline processing. Like Uni-Mamba, the strategy it adopts is to project the input of the Mamba module to a larger dimension (twice the original) and then form a linear residual connection. The architecture utilizes two Mamba blocks: one processes the input in its original direction, while the other handles the flipped input. The outputs from both directions are combined and refined through a final linear layer. By leveraging past and future information, Bi-Mamba performs better in SE tasks.

C. Proposed Method: MCMamba

In this work, we present MCMamba in Fig. 2, a multi-channel SE network specifically designed to process temporal and spectral cues through four dedicated modules: full-band spatial, narrow-band spatial, sub-band spectral, and full-band spectral. While inspired by the McNet framework, MCMamba introduces a novel variant of Mamba to further enhance its capabilities. Each module is carefully structured to progressively refine the features extracted from the preceding stage, ensuring that essential information is preserved throughout the process. MCMamba offers flexibility based on latency requirements, utilizing Bi-Mamba for non-causal (offline) processing and Uni-Mamba for causal (real-time) processing.

1) spatial modeling: Spatial feature modeling in MCMamba comprises two distinct modules: the full-band spatial module (Fig. 2(a)) and the narrow-band spatial module (Fig. 2(b)). Each module is responsible for extracting spatial features across different frequency ranges in multichannel speech, ensuring comprehensive spatial representation across the spectrum. The spatial characteristics across the entire frequency band are captured by processing the multichannel Short-Time Fourier Transform (STFT) coefficients along the

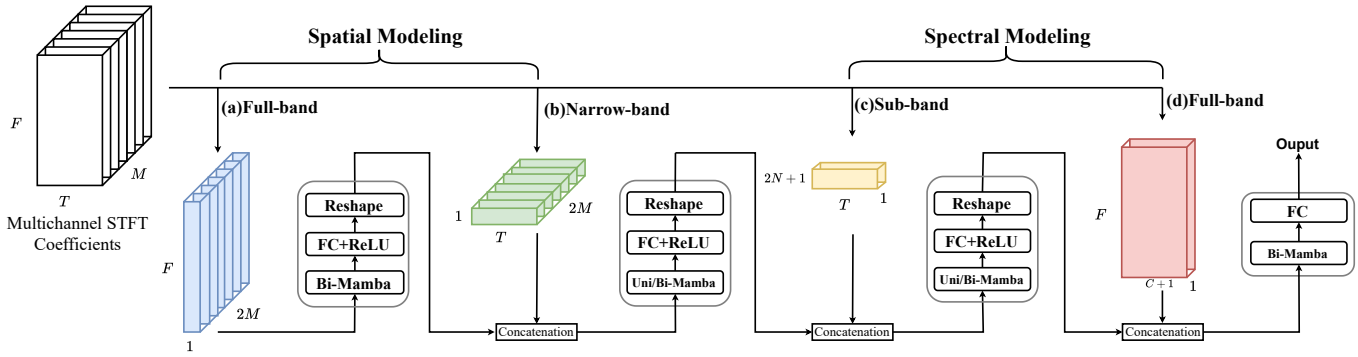


Fig. 2: Diagram of the proposed MCMamba architecture.

frequency axis, effectively modeling the spatial correlations across all frequencies. The spatial information of the input signal at each time-frequency point (t, f) is represented by concatenating the real (Re) and imaginary (Im) parts of STFT coefficients from all M channels:

$$x^1(t,f) = [\text{Re}(X_1(t,f)), \text{Im}(X_1(t,f)), \ldots, \text{Re}(X_M(t,f)), \text{Im}(X_M(t,f))]^T. \quad (1)$$

This allows learning of spatial cue correlations across frequencies, such as the inter-channel phase difference (IPD). The component does not capture temporal dependencies (always in a causal manner); therefore, the Bi-Mamba module extracts the full-band spatial features. Fully connected (FC) layers with ReLU activation are used to project the output to the desired dimension, after which the data is reshaped for further spatial processing.

The narrow-band spatial module (Fig. 2(b)) captures the spatial information within each frequency band and models its temporal evolution. The narrow-band process below will be executed several times in parallel for each narrow band. The concatenation of the original multichannel STFT coefficients and the output from the full-band spatial module are passed to the narrow-band module. Different frequencies are processed independently but share the same network parameters. The temporal sequence at one frequency can be formulated as follows:

$$x^2(f) = [x(1,f), x(2,f), \ldots, x(T,f)]. \quad (2)$$

By employing either the Bi-Mamba (non-causal) or Uni-Mamba (causal) module, followed by a FC layer to refine spatial characteristics, MCMamba effectively divides spatial modeling into full-band and narrow-band components. This approach ensures that frequency and time domain correlations across channels are fully leveraged, resulting in a more accurate separation of speech from noise.

2) spectral modeling: Spectrum modeling in MCMamba is carried out through the sub-band and full-band modules, each designed to process the frequency and spectral information of the reference channel. The sub-band spectral module (Fig. 2(c)) is tasked with capturing local spectral patterns and signal smoothness, both of which are critical for differentiating speech from noise. This module processes the STFT amplitude of the reference channel, namely $|X_r(t,f)|$. The data from adjacent frequencies are interconnected to utilize spectral information more effectively:

$$x^3(t,f) = [|X_r(t,f-N)|, \ldots, |X_r(t,f+N)|], \quad (3)$$

where N is the number of neighbors considered. Depending on whether real-time or offline processing is required, either Bi-Mamba or Uni-Mamba is employed to capture the partial spectral dynamics over time, followed by a FC layer with ReLU activation for further refinement.

The full-band spectral module (Fig. 2(d)) integrates information across the entire frequency range to model a broad-spectrum pattern. It operates along the frequency axis, combining multiple consecutive frames to capture temporal dependencies within the broader spectral context. The model gains access to more comprehensive information by connecting these context frames:

$$x^4(t,f) = [|X_r(t-C,f)|, \ldots, |X_r(t,f)|], \quad (4)$$

where C is the number of context frames considered. By utilizing both sub-band and full-band modules, MCMamba effectively captures local and global spectral cues, ensuring that spectral modeling complements spatial information to enhance the overall accuracy of speech reconstruction.

III. Experiments

A. Experimental setup

1) Dataset: We used the simulated dataset from the 3rd CHiME Challenge, comprising 7,138, 1,640, and 1,320 utterances for training, development, and testing. The data, sampled at 16 kHz using a six-microphone tablet, includes multichannel noise from four environments: bus, cafe, pedestrian area, and street intersection. The official CHiME-3 Matlab script was used to generate the dataset. We dynamically generated samples with random noise clips and varied SNR between -5 and 10 dB to enhance training. Speech signals were simulated by adding delays to single-channel utterances without rever-

TABLE I: Speech enhancement performance comparison. †Cited from other original papers, *Cited from McNet. Some papers don't provide causal results.

Method	NB-PESQ	WB-PESQ	STOI	SDR
Performance of Non-Causal (Offline) speech enhancement				
Noisy	1.82	1.27	87.0	7.5
MNMF Beamforming†[19]	–	–	94.0	16.2
Oracle MVDR*	2.49	1.94	97.0	17.3
CA Dense U-net†[6]	–	2.44	–	18.6
Narrow-band Net*[7]	2.74	2.13	95.0	16.6
FT-JNF*[20]	3.17	2.48	96.2	17.7
McNet*[10]	3.38	2.73	97.6	19.6
MVDR-Embedded U-Net†[21]	–	2.64	97.4	20.6
MCMamba	**3.49**	**2.98**	**98.2**	**20.8**
Performance of Causal (Online) speech enhancement				
Noisy	1.82	1.27	87.0	7.5
Narrow-band Net*[7]	2.70	2.15	94.7	16.0
FT-JNF*[20]	2.80	2.23	95.4	16.9
McNet*[10]	3.29	2.67	97.2	19.0
MCMamba	**3.41**	**2.89**	**97.6**	**19.8**

TABLE II: Ablation study on modules for spatial and spectral modeling, where the spatial features of the narrow band in Fig. 2(b) and the spectral features of the sub-band Fig. 2(c) are time-dependent.

Spatial	Spectral	NB-PESQ	WB-PESQ	STOI	SDR
LSTM	LSTM	3.27	2.72	97.46	19.30
Mamba	LSTM	3.46	2.80	97.60	19.22
LSTM	Mamba	**3.52**	**3.02**	**98.10**	**20.50**

beration. A Hann window of 512 samples with a 256-sample step size was applied for processing.

2) Model configurations: The output sizes of the Mamba blocks (either Uni- or Bi-) for full-band spatial, narrow-band spatial, sub-band spectral, and full-band spectral modeling are 64, 64, 64, and 2. The corresponding hidden cell sizes are 128, 256, 384, and 128, respectively. The sub-band spectral model uses 6 adjacent frequency bands ($N = 3$), while the full-band spectral model uses 5 context frames ($C = 5$).

3) Training setting: The max epoch is 500, with the best model selected based on the validation set, with an initial learning rate of 0.001 and a decay rate of 0.992 per epoch, using the Adam optimizer. The 5th channel was selected as the reference channel, and the performance was evaluated using NB-PESQ and WB-PESQ [22], STOI [23], and SDR [24].

4) Baseline methods: We compare our method with the following multichannel SE techniques: MNMF Beamforming [19], which employs multichannel non-negative matrix factorization (MNMF) followed by beamforming for SE; Oracle MVDR, which utilizes the actual noise spatial covariance matrix for enhancement; Narrow-band Net [7], which uses two LSTMs to capture narrow-band spatial information; FT-JNF [20], an improved version of Narrow-band Net that more effectively leverages full-frequency information; McNet [10] (previous SOTA), a fusion network that incorporates LSTMs to model both spatial and spectral information; and MVDR-Embedding U-Net [21], which integrates the minimum variance distortionless response (MVDR) module into the U-Net architecture, combining beamforming with neural networks for enhanced speech quality.

B. Experimental results

1) Results and analysis: In Table I, we fist compare the SE performance of various non-causal models. Evidently, traditional signal processing techniques, such as the MVDR-Embedded U-Net combined with deep learning methods, often yield superior outcomes. Notably, FT-JNF, which effectively leverages spatial information, outperforms Narrow-band Net. McNet, which integrates spatial and spectral features, underscores the critical importance of utilizing spectral information in multichannel SE. Our proposed MCMamba model achieves the **best** performance, demonstrating the strength of Bi-Mamba in modeling spatial and spectral information across multiple channels.

Given the significance of real-time SE, we also compare the performance of causal models. Again, MCMamba delivers highly competitive results, **surpassing even all the baseline non-causal models**. This highlights the effectiveness of our Mamba module in efficiently exploiting spatial information across different bandwidths and spectral cues across frequency bands. Notably, despite the absence of future information, MCMamba achieves excellent enhancement performance, affirming the robustness of the proposed approach.

2) Ablation study: Our ablation study further demonstrates Mamba's capabilities in spatial and spectral modeling. Since the LSTM architecture is known for exceptional temporal modeling, we compared its spatial and spectral modeling capabilities with that of the Mamba blocks we designed. Note that the models in Table II are causal, i.e., Uni-Mamba or LSTM is used to model temporal correlation. Bi-Mamba or BLSTM is only used to model frequency correlation. All three models in Table II are with similar parameter numbers. It serves as the baseline when modeling spatial and spectral features using LSTM. When Mamba is introduced for spectral or spatial modeling, all metrics improves dramatically. It is worth noting that when using Mamba for spectral modeling, the performance is significantly improved, especially for NB-PESQ and WB-PESQ, which are highly sensitive to spectral information. This demonstrates Mamba's advantages over LSTM in modeling spectral information, especially its superior ability to capture complex spectral dynamics.

IV. Conclusion

This paper proposes MCMamba, a novel model designed for multichannel speech enhancement, leveraging spatial and spectral information through Uni-Mamba and Bi-Mamba models for causal and non-causal processing. MCMamba outperforms the state-of-the-art baseline on the CHiME-3 dataset. Our experiments also show that the Mamba architecture significantly improves spectral modeling over traditional methods.

REFERENCES

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] 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. ICASSP*, 2021, pp. 6633–6637.

[3] 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," *arXiv preprint arXiv:2008.00264*, 2020.

[4] H. Wu, K. Tan, B. Xu, A. Kumar, and D. Wong, "Rethinking complex-valued deep neural networks for monaural speech enhancement," in *Proc. Interspeech*, 2023, pp. 3889–3893.

[5] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.

[6] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense U-Net for multichannel speech enhancement," in *Proc. ICASSP*, 2020, pp. 836–840.

[7] X. Li and R. Horaud, "Narrow-band deep filtering for multichannel speech enhancement," *arXiv preprint arXiv:1911.10791*, 2019.

[8] ——, "Multichannel speech enhancement based on time-frequency masking using subband long short-term memory," in *Proc. WASPAA*, 2019, pp. 298–302.

[9] F. Xiong, W. Chen, P. Wang, X. Li, and J. Feng, "Spectro-temporal subnet for real-time monaural speech denoising and dereverberation," in *Proc. Interspeech*, 2022, pp. 931–935.

[10] Y. Yang, C. Quan, and X. Li, "McNet: Fuse multiple cues for multichannel speech enhancement," in *Proc. ICASSP*, 2023, pp. 1–5.

[11] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[12] K. Li, G. Chen, R. Yang, and X. Hu, "SPMamba: State-space model is all you need in speech separation," *arXiv preprint arXiv:2404.02063*, 2024.

[13] X. Jiang, C. Han, and N. Mesgarani, "Dual-path Mamba: Short and long-term bidirectional selective structured state space models for speech separation," *arXiv preprint arXiv:2403.18257*, 2024.

[14] R. Chao, W.-H. Cheng, M. La Quatra, S. M. Siniscalchi, C.-H. H. Yang, S.-W. Fu, and Y. Tsao, "An investigation of incorporating mamba for speech enhancement," *arXiv preprint arXiv:2405.06573*, 2024.

[15] C. Quan and X. Li, "Multichannel long-term streaming neural speech enhancement for static and moving speakers," *IEEE Signal Processing Letters*, vol. 31, pp. 2295–2299, 2024.

[16] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, and C. Ré, "Hungry hungry hippos: Towards language modeling with state space models," *arXiv preprint arXiv:2212.14052*, 2022.

[17] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[18] A. Gu, K. Goel, A. Gupta, and C. Ré, "On the parameterization and initialization of diagonal state space models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 971–35 983, 2022.

[19] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel nmf-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 5, pp. 960–971, 2019.

[20] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 563–575, 2023.

[21] C.-H. Lee, K. Patel, C. Yang, Y. Shen, and H. Jin, "An MVDR-embedded U-Net beamformer for effective and robust multichannel speech enhancement," in *Proc. ICASSP*, 2024, pp. 8541–8545.

[22] 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," in *Proc. ICASSP*, 2001, pp. 749–752.

[23] C. Taal, R. 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, 2011.

[24] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.