IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 August 2023, accepted 5 September 2023, date of publication 11 September 2023,
date of current version 19 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3313972

RESEARCH ARTICLE

A Robust Hybrid Neural Network Architecture for Blind Source Separation of Speech Signals Exploiting Deep Learning

SAM ANSARI[1], KHAWLA A. ALNAJJAR[1], (Member, IEEE), TAREK KHATER[1], SOLIMAN MAHMOUD[1,2], (Senior Member, IEEE), AND ABIR HUSSAIN[1,3], (Senior Member, IEEE)

[1]Department of Electrical Engineering, University of Sharjah, Sharjah, United Arab Emirates
[2]University of Khorfakkan, Khor Fakkan, United Arab Emirates
[3]School of Computer Science and Mathematics, Faculty of Engineering, Liverpool John Moores University, L3 3AF Liverpool, U.K.

Corresponding author: Abir Hussain (abir.hussain@sharjah.ac.ae)

ABSTRACT In the contemporary era, blind source separation has emerged as a highly appealing and significant research topic within the field of signal processing. The imperative for the integration of blind source separation techniques within the context of beyond fifth-generation and sixth-generation networks arises from the increasing demand for reliable and efficient communication systems that can effectively handle the challenges posed by high-density networks, dynamic interference environments, and the coexistence of diverse signal sources, thereby enabling enhanced signal extraction and separation for improved system performance. Particularly, audio processing presents a critical domain where the challenge lies in effectively handling files containing a mixture of human speech, silence, and music. Addressing this challenge, speech separation systems can be regarded as a specialized form of human speech recognition or audio signal classification systems that are leveraged to separate, identify, or delineate segments of audio signals encompassing human speech. In various applications such as volume reduction, quality enhancement, detection, and identification, the need arises to separate human speech by eliminating silence, music, or environmental noise from the audio signals. Consequently, the development of robust methods for accurate and efficient speech separation holds paramount importance in optimizing audio signal processing tasks. This study proposes a novel three-way neural network architecture that incorporates transfer learning, a pre-trained dual-path recurrent neural network, and a transformer. In addition to learning the time series associated with audio signals, this network possesses the unique capability of direct context-awareness for modeling the speech sequence within the transformer framework. A comprehensive array of simulations is meticulously conducted to evaluate the performance of the proposed model, which is benchmarked with seven prominent state-of-the-art deep learning-based architectures. The results obtained from these evaluations demonstrate notable advancements in multiple objective metrics. Specifically, our proposed solution showcases an average improvement of 4.60% in terms of short-time objective intelligibility, 14.84% in source-to-distortion ratio, and 9.87% in scale-invariant signal-to-noise ratio. These extraordinary advancements surpass those achieved by the nearest rival, namely the dual-path recurrent neural network time-domain audio separation network, firmly establishing the superiority of our proposed model's performance.

INDEX TERMS Artificial intelligence, blind source separation, deep learning, dual-path recurrent neural network, transfer learning, time-domain audio separation network.

I. INTRODUCTION

In the realm of wireless communication systems, the advent of beyond fifth-generation (B5G) and sixth-generation (6G)

The associate editor coordinating the review of this manuscript and approving it for publication was Felix Albu.

networks brings forth new challenges and opportunities. The need to support ultra-high data rates, massive connectivity, low latency, improved spectral efficiency, enhanced quality-of-service (QOS), and diverse services necessitates the exploration of advanced signal processing techniques [1], [2]. Blind source separation (BSS) emerges as a promising approach to address these challenges by leveraging the capability to separate mixed source signals and extract valuable information from complex mixtures [3], [4].

BSS refers to the process of recovering the source signals from an observed mixture without any prior knowledge of the mixing algorithm, or the source signals themselves [5]. The observed mixture can be either single-channel in nature [6] or involve multiple channels [7]. In scenarios where the number of observed channels is fewer than the number of sources, such as in the case of musical audio, the separation problem becomes underdetermined. Therefore, incorporating prior knowledge about the original signals becomes crucial for enhancing the efficacy of the separation process.

Television, radio, internet, and satellite channels serve as conduits for a plethora of information on a daily basis, encompassing diverse and valuable content [8]. In conjunction with this transmitted data, various supplementary components, including music, noise, and other elements, coexist [9]. Nevertheless, the significance of these supplementary components varies depending on the target audience. Consequently, the demand for systems capable of distinguishing between trivial sources and signals devoid of value and those of substantial importance becomes evident [10]. Consequently, it is imperative to extract the desired and beneficial signals while filtering out inconsequential sources that lack relevance or value to a specific audience, such as ambient noise, background music, vocal performances, or guitar solos. Moreover, the development of an efficient system to segregate insignificant data from noteworthy content assumes significance in terms of reducing storage volume [11].

In the sphere of telecommunications networks, an essential undertaking involves minimizing the amount of data transmitted by users. To address the volume reduction or capacity enhancement, it is imperative for the system to identify and eliminate silence frames present within speech frames [12]. Another scenario arises when multiple individuals utilize various devices located in diverse geographic locations to engage in simultaneous conversations. In such cases, the speeches from these devices are merged and transmitted to a designated receiver node. The primary objective is to differentiate and recover the individual speeches by leveraging the available perceptual data, i.e., BSS of audio files [13]. As evident from the observations, the development of a robust framework capable of effectively separating speech and music has the potential to yield substantial benefits across numerous lucrative applications [14], [15].

Due to the significant importance of human-computer interaction and communication in the new millennium, recent research efforts have increasingly focused on advanced multi-microphone signal processing solutions aimed at enhancing speech understanding in challenging environments. One prominent signal processing technique in this context is BSS. Various BSS algorithms and architectures are investigated, considering their potential to improve interference management, channel estimation, beamforming, and resource allocation in these next-generation wireless networks [4], [15], [16], [17], [18], [19], [20].

This study investigates the simultaneous recovery of signals in a reverberant or echoic environment using two or more microphones. In this scenario, each microphone captures the direct contributions from individual sources as well as multiple reflections of the original signals at varying propagation delays, resulting in complex compositions of the source signals. This study proposes an efficient BSS framework that achieves accurate separation of the signals and explores its application in the context of B5G and 6G networks. This work presents a novel hybrid model extensively harnessing the potential of deep learning techniques to achieve highly effective source signal separation.

The integration of advanced BSS techniques within B5G and 6G communication networks marks a pivotal stride toward enhancing signal quality, interference management, and resource optimization. In this context, the study presents a significant contribution through the introduction of a robust hybrid neural network architecture for BSS of audio/speech signals, adeptly harnessing the capabilities of deep learning. This architecture, characterized by its innovative tripartite structure involving transfer learning, a pre-trained dual-path recurrent neural network (DPRNN), and a transformer, showcases a versatile approach to untangle mixed audio/speech sources in complex and dynamic communication environments. The utilization of this architecture aligns seamlessly with the imperatives of B5G and 6G networks, which necessitate the seamless extraction of relevant and reliable information from diverse and often convoluted signal mixtures. The proposed architecture's adaptability to diverse scenarios, its capacity to accommodate a wide array of mixed sources, and its resilience against adversarial perturbations exemplify its viability for real-world deployment within these networks. Consequently, the seamless amalgamation of BSS techniques, epitomized by the novel hybrid neural network architecture, presents an avenue for addressing the intricate demands posed by B5G and 6G networks, ultimately culminating in improved communication fidelity, spectral efficiency, and quality of experience.

In light of the above, the main contributions of this work can be delineated as follows.

- Proposal of a three-way neural network-based BSS framework incorporating transfer learning, a pre-trained DPRNN, and a transformer for accurate signal separation, exploring BSS application in B5G and 6G networks to enhance signal processing capabilities and enable advanced functionalities, and contributing to the evolution of wireless communication systems mitigating interference and noise.

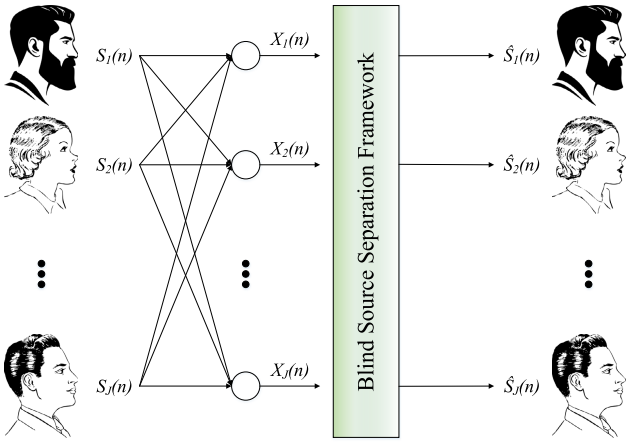


FIGURE 1. A visual representation of mixture signals and the implementation of a BSS framework.

- Development of divide and conquer strategies to address the challenges posed by long sequential input, including the partitioning of input into smaller chunks and the iterative application of intra- and inter-chunk operations.
- Implementation of skip connection to provide an additional gradient path for improved model convergence, along with the utilization of a convolutional network (decoder) for the separation of the mixed signal.

The rest of this paper follows a structured organization as outlined below. Section II delves into the related background of BSS and conducts a comprehensive analysis of existing literature and research in BSS, identifying gaps and limitations in current approaches. Section III details the architecture and components of the novel model or framework proposed in this research work. Section IV presents the experimental setup, methodology, and evaluation results, demonstrating the effectiveness and efficiency of the proposed model. Finally, Section V summarizes the key findings, implications, and provides avenues for future research in the field of BSS.

II. RELATED BACKGROUND AND LITERATURE REVIEW

This section is bifurcated into two distinct segments. The initial part presents a comprehensive background on BSS, elucidating its fundamental concepts and principles. Subsequently, this section diligently undertakes an exhaustive literature review, meticulously examining and synthesizing the existing body of scholarly works in BSS domain. Numerous research articles have been meticulously examined to gain a comprehensive understanding of the topic and become acquainted with the state-of-the-art techniques. A concise overview and discussion of the existing BSS models are provided, offering valuable insights into the latest advancements and methodologies in the field. To provide an illustrative example for better comprehension of the concepts, Figure 1 demonstrates the structure of a BSS framework.

It is worth highlighting that mixing systems can encompass various types, including linear and non-linear mixing,

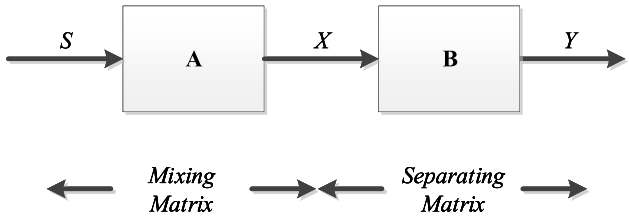


FIGURE 2. The linear space schematic diagram of BSS.

momentary mixing, convolutional, stationary, time-varying, and combinations involving noise in the signals. In the context of BSS, certain assumptions and limitations arise. One key assumption is the statistical independence of the components that need to be estimated (i.e., the sources). Additionally, it is preferred that the sources to be estimated do not follow a Gaussian distribution, with the exception of at most one source having such a distribution. For the sake of simplicity, it is assumed that the composition matrix is square and invertible, facilitating the separation process. To furnish an all-encompassing comprehension of the intricate procedure underlying the retrieval and acquisition of the original signal sources (S), a meticulously delineated schematic diagram is showcased in Figure 2.

The mathematical formulation of the mixing system can be expressed as follows [21], [22]:

$$X = AS, \tag{1}$$

where $S = [S_1, S_2, \ldots, S_M]^T$ represent a vector containing the source signals, and $X = [X_1, X_2, \ldots X_N]^T$ denote a vector comprising the signals obtained by the combination of the mixing matrix A (i.e., observed signals). In BSS, the objective is to estimate the matrix B such that $Y = [Y_1, Y_2, \ldots, Y_N]^T$, which are the reconstructed/estimated signals, closely resemble the original input signals while maintaining statistical independence. Assuming the independence of the input signals and linear mixing processes (which can also accommodate nonlinearity), the precise mathematical formulation to determine the output is defined by [21] and [22]

$$Y = BX, \tag{2}$$

where B denotes the separating matrix derived through the utilization of a mathematical algorithm along with iterative and artificial intelligence (AI)-driven methodologies.

In (1) and (2), the number of input sources is equivalent to the number of sensors (X_S). However, assuming the existence of P linear combinations of m sources, the unknown matrix A takes on dimensions $M \times P$ and can be represented as $X_1, X_2, \ldots, X_P$. One of the underlying assumptions employed in this context is sparsity. Alternatively, when accounting for the assumption of non-linearity, the precise mathematical

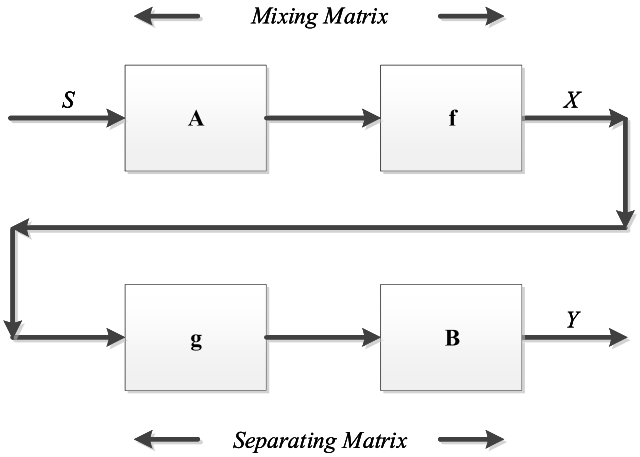


FIGURE 3. The diagrammatic representation of BSS within a nonlinear space.

expression employed is determined by [21], [22], and [23]

$$X = f\,[AS]\,, \tag{3a}$$
$$g = f^{(-1)}, \tag{3b}$$
$$Y = Bg\,[X]\,, \tag{3c}$$

where f signifies a non-linear mapping that is invertible, while g represents the corresponding inverse function that needs to be determined initially. The objective of finding the transformation function g is to ensure statistical independence among the components of $\boldsymbol{Y}$. The schematic diagram illustrating the nonlinear space is presented in Figure 3.

Various researchers have proposed systems and algorithms to achieve the separation of multiple audio sources [24]. The methods for signal separation can be categorized into single-channel algorithms and multi-channel algorithms [6], [7]. In single-channel algorithms, only one mixed output signal is available for processing. These methods primarily focus on separating a specific source signal. They leverage the characteristics and assumptions inherent in the nature of the source signals. By utilizing the existing features and statistical properties of the signals, discriminative algorithms can be implemented. On the other hand, multi-channel signals are composed of multiple sources that exhibit correlation or similarity across channels. In the processing of multi-channel signals, the sources in different channels influence each other. BSS is a well-known technique employed in multi-channel scenarios, aiming to reconstruct the individual source signals.

Researchers have made significant contributions in developing a wide range of systems and algorithms aimed at signal separation. However, despite these advancements, there are persistent challenges that impede the achievement of precise and timely separation of all signals [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36]. Extensive research efforts have been devoted to the field of BSS, leading to the proposal of numerous techniques that utilize a range of existing methods [37], [38], [39], [40], [41]. Interference management is a critical aspect in B5G and 6G networks. By employing BSS algorithms, it becomes possible to separate and mitigate interference sources, enhancing the overall system performance. This section explores various BSS techniques, such as independent component analysis (ICA), non-negative matrix factorization (NMF), and sparse component analysis (SCA), and their applicability in interference cancellation and suppression.

Effective resource allocation plays a vital role in optimizing the utilization of network resources and achieving efficient communication in B5G and 6G networks. BSS techniques offer the potential to enhance resource allocation algorithms by separating individual source signals and allocating resources based on their specific characteristics. This section explores the application of BSS in resource allocation optimization, considering aspects such as power allocation, bandwidth allocation, and user scheduling.

The BSS problem has been extensively addressed through the introduction and evaluation of various methods [23], [27], [30], [31], [35], [36], [42], [43], [44], [45], [46], [47], [48], which can be broadly classified into two main categories: mathematical methods and AI algorithms. The objective of these methods is to determine the optimal coefficients for the separation matrix, aiming to minimize the interdependence among the estimated sources. Both mathematical methods and AI algorithms have their strengths and limitations. Mathematical methods often rely on specific assumptions about the sources and mixing process, while AI algorithms offer more flexibility and adaptability but may require a large amount of training data. Researchers continue to explore and develop new BSS methods that combine the strengths of both mathematical and AI approaches.

A. MATHEMATICAL-BASED METHODS

Mathematical-based methods aim to estimate the separating matrix coefficients while minimizing the interdependence among the estimated sources. Examples of mathematical methods include ICA [49], SCA [50], NMF [51], [52], joint diagonalization (JD) [53], second-order statistics (SOS) [54], and time-frequency analysis (TFA) [55] methods. It is important to note that each technique has its own strengths and limitations, and their suitability varies depending on the specific application and characteristics of the source signals. These algorithms find applications in diverse fields, including speech separation, image processing, electroencephalogram (EEG) analysis, bioinformatics, finance, radar signal processing, and wireless communication.

ICA is widely recognized as the prevailing and most effective approach to address the challenge of blind signal separation. The fundamental objective of ICA is to identify signal components that exhibit the highest degree of statistical independence. These methods primarily rely on higher-order statistical properties and aim to establish a linear representation of non-Gaussian data, where the resultant components possess either complete statistical independence or, at the

TABLE 1. A comparative analysis of prominent methods in the BSS problem.

Tech.	Advantages	Disadvantages	Applications
ICA [49]	Effective for linear and nonlinear mixtures	Requires the number of sources to be known	Speech separation, image processing, and EEG analysis
NMF [51]	Intuitive interpretation of results	Prone to local minima in the optimization process	Audio source separation and image analysis
SCA [50]	Can handle non-Gaussian sources	Sensitivity to the choice of sparsity parameter	Speech separation, bioinformatics, and finance
JD [53]	Suitable for instantaneous mixtures	Limited performance for convolutive mixtures	EEG analysis and array signal processing
SOS [54]	Simple and computationally efficient	Limited performance for non-Gaussian mixtures	Radar signal processing and wireless communication
TFA [55]	Effective for time-varying mixtures	Challenging to handle source collisions	Audio signal separation and biomedical signal analysis

very least, a significant level of independence. Within the discussed methods, the maximum likelihood (ML) approach serves as a fundamental pillar for mutual information (MI) estimation. Additionally, the linear ICA presents a distinctive and singular solution approach [56]. It is important to note that these methods demonstrate optimal performance when the number of sources is equal to or less than the number of observations, while they may not be suitable for underdetermined problems where the number of sources exceeds the number of available observations. In contrast to non-linear ICA, which poses a challenging problem with infinite solutions and lacks a straightforward relationship between these solutions, linear ICA methods have garnered significant attention in separation algorithms. This preference for linear ICA methods stems from their ability to offer more tractable solutions and better interpretability in the context of BSS [57], [58]. In order to provide a comprehensive overview of the different techniques in BSS, Table 1 summarizes their respective advantages, disadvantages, and applications.

B. ARTIFICIAL INTELLIGENCE (AI)-BASED METHODS

AI-based BSS algorithms utilize neural networks, i.e., classical or shallow models, evolutionary algorithms, and deep learning architectures to learn the optimal separation coefficients and minimize the statistical dependency among the estimated components [27], [59], [60], [61], [62], [63], [64], [65], [66]. The utilization of neural networks in the BSS methodology plays a crucial role in reducing the statistical dependency among the estimated components. The effectiveness of this reduction heavily relies on the distribution function of the source signals, as the availability of such information facilitates accurate estimation. Evolutionary algorithms, on the other hand, leverage the application of a well-defined fitness function to guide the convergence of the initial population towards a population that achieves minimal interdependence among the estimated signals.

An overview of the particle swarm optimization (PSO) algorithm in the context of BSS is presented in Li et al. [59]. The authors highlight certain limitations of the standard PSO approach, including low accuracy and a tendency to get trapped in premature convergence. To address these shortcomings, the authors propose an enhanced PSO algorithm that incorporates adaptive adjustment of the inertia weight. This modification aims to improve the performance of BSS. The proposed algorithm is compared against the FastICA algorithm, which is a fast fixed-point algorithm commonly used in BSS, as well as conventional PSO techniques. The experimental evaluation is conducted in a noisy environment, and the correlation coefficient matrix is employed as the evaluation metric. The results indicate that the proposed enhanced PSO algorithm exhibits robustness against noise and significantly enhances the accuracy of BSS compared to both FastICA and conventional PSO methods.

Khalfa et al. [60] introduce a method called high exploration particle swarm optimization (HEPSO) for the purpose of separating signal sources from a given set of observations. This method represents an enhanced version of the PSO algorithm, incorporating two additional operators: the genetic algorithm (GA) and the artificial bee colony (ABC) mechanism. In their proposed HEPSO method, the GA and ABC frameworks are utilized to update the speed and position of particles within the optimization process. The authors employ kurtosis and MI as fitness functions to guide the optimization process. These fitness functions allow the proposed model to search for the appropriate transform/mixing matrix that can effectively separate the signal sources. To evaluate the performance of the HEPSO method, the authors conduct simulations using three test datasets.

The work in [27] proposes a BSS technique that addresses the issue of slow convergence rate by incorporating ABC optimization and kurtosis. Unlike some existing methods, their approach does not rely on any specific assumptions about the source signals. The algorithm presented in the study utilizes adaptive function values to select iterative updates

and step sizes. By incorporating the kurtosis objective function, the proposed model can be applied to signal distributions without any restriction. The model of Wang et al. is not constrained by assumptions about the source signals and exhibits improved performance compared to alternative approaches, as verified through extensive simulations. To evaluate the performance of their technique, the authors conduct various simulations.

In the work conducted by Kumar and Jayanthi [61], the effectiveness of FastICA for BSS in determined or over-determined instantaneous mixture signals is explored. The study focuses on investigating different contrast functions within the FastICA algorithm, which serve as nonlinear measurements of the independence between the estimated sources and the mixture signals. Specifically, the research aims to identify highly efficient contrast functions for analyzing signals in noisy environments. Notably, the contrast functions examined in FastICA include negentropy, ML, and kurtosis. To evaluate the proposed model, both real-time recorded mixture signals and synthetic instantaneous mixtures are utilized. The performance assessment of the contrast functions is based on several metrics, including source-to-interference ratio (SIR), signal-to-distortion ratio (SDR), signal-to-artifact ratio (SAR), and computational complexity. The simulation results obtained in noisy environments indicate that the ML contrast function demonstrates superior performance compared to the other contrast functions analyzed. The finding suggests that ML is particularly effective for BSS using the FastICA algorithm.

In the study conducted by Liu et al. [62], the main objective is to enhance the steady-state performance and convergence speed of BSS methods. The authors propose a novel approach that optimizes the performance of neural networks-based BSS by addressing the loss function utilized in the BSS method. The proposed model employs neural networks and the ML estimation approach. The neural network architecture incorporates a bias term, which contributes to improving the steady-state performance. Additionally, L2 regularization terms are introduced to the loss function to handle the weights and biases, further enhancing the model's performance. To accelerate the training process, a new optimization model is developed, featuring a dual acceleration strategy that aids in gradient descent. This strategy significantly improves the convergence rate of the algorithm. The presented model's performance is extensively evaluated through various simulations, considering scenarios with and without prior knowledge of the mixing systems and source signals. The authors conclude by stating that their technique is well-suited for engineering applications, underscoring its practical relevance and applicability.

Addressing the challenge of separating a singing voice from its musical accompaniment, Lin et al. [63] present a BSS technique that utilizes a unique neural network architecture based on pixel-wise image classification. Their model employs a pretraining stage of a Convolutional neural network (CNN) using cross-entropy loss, which functions as an autoencoder on singing voice spectrograms. The target output label in the CNN is trained using the ideal binary mask (IBM). By utilizing pixel-wise classification, the model predicts the label of sound sources, thereby eliminating the need for common pre- and postprocessing tasks typically associated with BSS methods. During the training phase, the objective is to minimize the error between the predicted and target labels by leveraging the cross-entropy loss. By converting the BSS problem into a pixel-wise classification task, the approach eliminates the requirement for postprocessing techniques such as the Wiener filter. To evaluate the performance of the proposed model, the authors employ various datasets and models.

Laugs et al. [64] investigate the influence of mixed audio on emotion recognition in music and speech using both a random forest model and a deep neural network (DNN). The random forest algorithm employed in the study is utilized to rank the features relevant to speech and music emotion recognition. By analyzing the significance of these features, the algorithm provides insights into the differences between the models and features used for each sound type. The speech DNN architecture consists of 512 neurons distributed across three hidden layers, with a dropout rate of 0.5 applied to mitigate overfitting. Rectified linear units (ReLU) activation functions are used in the hidden layer neurons, while the output layer utilizes the softmax activation function. The model is implemented and evaluated on six datasets. Through the presentation of simulation results, the paper asserts that their BSS model achieves higher accuracy in music and speech emotion recognition compared to alternative approaches. These results suggest that the proposed BSS model outperforms other methods in accurately identifying and distinguishing emotions in music and speech.

The study of [65] highlights the capability of machine learning algorithms, specifically convolutional time-domain audio separation network (Conv-TasNet) and deep extractor for music sources (Demucs), to discriminate between two interfering signals (such as speech and music) without prior knowledge of the mixture operation. The Demucs algorithm is a waveform-to-waveform model that exhibits a higher decoding capacity compared to the Conv-TasNet model, leveraging the same technique as the audio generation algorithm. Conversely, Conv-TasNet is a fully convolutional time-domain audio separation technique. The selected algorithms are evaluated based on their ability to achieve high-quality and precise signal separation while considering lower time complexity, indicating higher execution time efficiency. Four specific scenarios (music-child, music-male, music-conversation, and music-female) are defined to conduct experiments and assess the performance of the chosen models. Evaluation metrics employed to assess the results include R-squared, mean absolute error (MAE), root mean square error (RMSE), and scores from the music information retrieval evaluation (mir_eval) system, which is a Python library specifically designed for music-related evaluation tasks. The accuracy of the selected models is computed

using RMSE and MAE criteria, which involve calculating the absolute values and average magnitude of the errors between the observed and predicted data. Overall, the study demonstrates the effectiveness of Conv-TasNet and Demucs in discriminating between speech and music signals without prior information, with each algorithm displaying strengths and trade-offs in terms of signal separation quality, execution time, and computational complexity.

Issa et al. [66] address the problem of mixed speech signals that can occur when speech signals are converted and transferred to computers. This interference or mixture may arise from other speech sources or environmental noises. One common example is the cocktail party problem, where multiple people speaking simultaneously result in a mixture of different speech signals. To overcome the challenge, BSS techniques are employed to extract the desired audio signals from the mixture. The authors propose a novel BSS framework that utilizes deep recurrent neural networks (DRNN) equipped with bi-directional long short-term memory (BLSTM). The presented algorithm aims to separate audio signals from a monaural mixed signal that includes both male and female speech. To perform the separation, two types of time-frequency (TF) masks are estimated: the ideal ratio mask (IRM) and the optimal ratio mask (ORM). These masks help in determining the importance or relevance of different components in the TF domain, enabling the separation of desired audio signals from the mixture.

It is important to acknowledge that shallow machine learning models are typically characterized by their simplicity and lower parameter count when compared to deep learning models. Although they may not attain the same level of performance as deep learning techniques, they can still prove to be effective in specific scenarios and serve as a reliable baseline for audio BSS tasks. Table 2 indexes an arrangement of information, including descriptions, advantages, and disadvantages, that juxtaposes various AI-based approaches in the context of audio BSS.

Despite the significant advancements they bring to the field, existing BSS methods are not exempt from certain limitations and challenges. One notable limitation lies in handling complex audio scenes with overlapping sources. In such scenarios, the separation algorithms often struggle to accurately separate individual sources, leading to incomplete or distorted results. Moreover, the current approaches heavily rely on statistical assumptions and spectral analysis, which may not capture all the intricacies and nuances of real-world audio signals. Additionally, many existing techniques assume that the number of sources is known in advance, making them less effective in scenarios where the number of sources is unknown or variable. Innovative approaches need to be devised to enhance the accuracy and completeness of source separation in complex audio scenes. Furthermore, computational complexity remains a concern, as some separation algorithms require extensive processing power, hindering real-time applications. Lastly, the evaluation and benchmarking of BSS algorithms can be subjective and inconsistent, as different metrics and datasets may yield conflicting results. Accordingly, the concerted effort to address these shortcomings is crucial in driving the development of advanced, reliable, robust, and adaptive BSS methods that can effectively meet the demands of today's technological landscape, i.e., B5G and 6G.

In contrast to alternative methodologies, this work presents a novel proposition involving the incorporation of a CONV-1D encoder network to extract valuable insights from a composite signal. Moreover, it involves the segmentation of the derived information into smaller units, thereby yielding enhanced operational effectiveness. Furthermore, the suggested framework encompasses the utilization of a pre-trained DPRNN integrated network, accompanied by the incorporation of transfer learning within the primary processing unit. This amalgamation leverages the respective advantages of both approaches, thereby facilitating improved signal separation and processing capabilities.

III. PROPOSED MODEL

Signal separation to extract valuable and meaningful information has garnered significant interest among researchers due to its fundamental importance. The selection of an appropriate BSS algorithm for the attainment of a desired output holds significant importance across numerous contexts. Within the scope of this research endeavor, our aim is to accomplish this objective through the utilization of deep learning algorithms.

This research focuses on the application of deep learning approaches as domain-based methods for BSS. The term domain refers to the information set that deep learning approaches utilize for learning purposes. This set encompasses the underlying structure of the target source(s), the mixture signal(s), and the interdependencies between the sources and the mixture. Domain learning can be categorized as either supervised or unsupervised. Supervised learning aims to efficiently acquire knowledge of the mixing and unmixing processes. However, one particular challenge that arises in this context is the generalization of the learned denoising process to signals that lie outside the observed range, i.e., the ability to apply the process to invisible audio mixtures. Another challenge pertains to the limitations imposed by the available computational resources or the size of the domain used for learning.

Additionally, there is a need to interpret and comprehend deep learning methods for audio BSS that rely on domain information. This challenge offers an alternative perspective for evaluating deep learning approaches beyond the conventional criteria used to assess their learning capabilities. In the case of unsupervised learning, the objective is to grasp the underlying signal structure of the source(s) and mixture, enabling the subsequent unmixing of the target source(s). Two challenges emerge in this scenario. First, it is crucial to learn representations of audio signals that are well-suited for isolating the target source(s). Second, there is a need to interpret the representations of the learned signals. The

TABLE 2. Comparative analysis of AI-based approaches for audio BSS.

Technique	Description	Advantages	Disadvantages
ICA [67]	Separates sources based on statistical independence by assuming non-Gaussian distributions for the sources.	Simple and widely used method.	Assumes statistical independence and may struggle with correlated sources.
NMF [68]	Decomposes the audio spectrogram into a non-negative matrix product, separating sources based on their spectral patterns.	Intuitive interpretation and good performance in some scenarios.	Requires appropriate initialization and may struggle with complex spectra.
SVM [69]	Uses a hyperplane to separate sources based on the extracted features from the mixed audio.	Effective in binary classification scenarios.	May struggle with high-dimensional audio data and require feature engineering.
KNN [70]	Assigns sources to the nearest neighbors in the feature space, based on their similarity to the mixed audio.	Simple and easy to implement.	Performance may degrade with high-dimensional audio data and large datasets.
Decision Trees [71]	Builds a tree-like model to separate sources by making decisions based on the audio features and their thresholds.	Easy to interpret and visualize the separation process.	Prone to overfitting and may struggle with complex decision boundaries.
Random Forests [72]	Ensembles multiple decision trees to improve the separation performance and reduce overfitting.	Robust performance and handles high-dimensional data well.	Computationally expensive and may be challenging to interpret.
DNN [73]	Utilizes deep learning models, such as deep autoencoders or recurrent neural networks, to learn the mapping from mixed audio to individual sources.	Can handle complex dependencies and provide high-quality separation.	Requires a large amount of training data and computational resources.
CNN [74]	Applies convolutional layers to extract spatial and spectral features for source separation.	Effective for processing spectrogram-like inputs.	Limited temporal modeling capabilities.
RNN [75]	Utilizes recurrent connections to model temporal dependencies in the audio signal for source separation.	Suitable for handling sequential data and long-term dependencies.	May suffer from vanishing/exploding gradient problems.
GAN [76]	Employs a generative model and a discriminative model to compete against each other, enhancing the quality of source separation.	Can generate realistic and high-quality separated sources.	Training can be unstable and requires careful tuning.
Sparse Coding [77]	Represents the mixed audio as a linear combination of basis functions, promoting sparsity to separate sources.	Can handle complex mixtures and perform well in sparse scenarios.	Sensitive to the choice of sparsity regularization parameters.

proposed research work is presented considering the relevant challenges within the process of BSS.

The workflow outlined for the proposed system is structured as follows.

- Distilling deep nonlinear neural networks: Presenting an algorithm designed to distill deep nonlinear neural networks, focusing on the understanding of how DNNs effectively separate audio sources in the frequency domain. The algorithm is explained in detail, highlighting its key components and techniques used.
- Data-driven filter operators and skip-filtering connections: Providing experimental evidence to demonstrate that DNNs have the ability to learn data-driven filter operators. Furthermore, the potential for enhancing these operators using skip-filtering connections is explored. Reviewing a simple technique employed in advanced audio BSS approaches, shedding light on its efficacy and benefits.
- Efficient neural architecture for vocal separation and harmonic/percussive source separation (HPSS): Focusing on introducing an efficient neural architecture that utilizes band-pass filter connections. This architecture demonstrates promising results in vocal separation and HPSS. Discussing the design and implementation of this

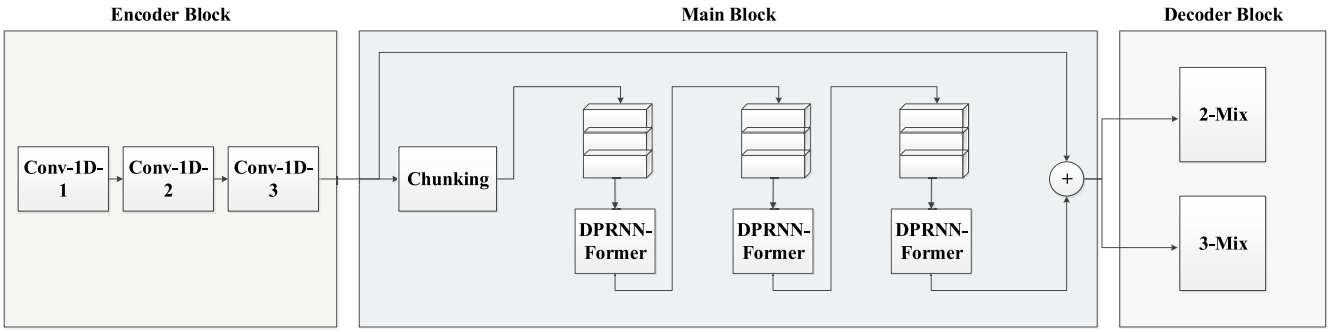


FIGURE 4. The schematic diagram illustrating the architecture of the proposed network.

architecture, highlighting its competitive performance in these specific tasks.

- Reparameterization scheme for interpretable signal representation: Presenting a reparameterization scheme for the decoding functions of deep autoencoder networks (DAEs). This scheme enables the computation of an interpretable signal representation suitable for source isolation. Explaining the reparameterization technique and discuss its implications for obtaining meaningful and actionable insights from the separated audio sources.

Within the scope of this research, the remarkable attributes inherent in the transformer framework are harnessed to achieve direct context awareness. The proposed architecture for this study is presented in Figure 4. The algorithm employed in this study encompasses several key steps. First, the mixed signal is inputted into the encoder network, i.e., CONV-1D. After extracting pertinent information, the data is divided into smaller segments to facilitate faster processing exploiting the engineering concept of divide and conquer. Subsequently, the information proceeds through the primary processing block, which integrates a pre-trained DPRNN and transfer learning. By combining the capabilities of transformers in handling long-term dependencies with the establishment of automatic regression to capture crucial local dependency information, the incorporation of a stepwise distributed transformer and a two-way DPRNN process significantly enhances network flexibility. Moreover, the integration of skip connection, which provides an alternate gradient path during backpropagation, proves beneficial for model convergence, as empirically confirmed. Finally, the separated mixed signal is obtained through the utilization of a convolutional network, i.e., decoder. In the proposed approach, the assimilation of a recurrent neural network (RNN) into the original transformer, devoid of positional encoding, facilitated the acquisition of sequential order information pertaining to speech sequences. In what fellows the three parts of our proposed system will be thoroughly explained.

A. ENCODER BLOCK

The purpose of the audio encoder in this context is to process audio data and extract relevant features to improve classification accuracy. The audio encoder module is composed of three 1D convolutional layers. Its input is the raw waveform of the audio, and it employs a series of 1D convolutional operations to extract meaningful features from the raw audio signal. Each convolutional block comprises a convolutional operation followed by batch normalization, ReLU activation, and max pooling operations. Each convolutional layer has its own kernel size and filters. The pseudocode provided in Algorithm 1 outlines the systematic process of the encoder block, allowing for the conversion of input data into an encoded representation, i.e., extraction of appropriate features.

Algorithm 1 Algorithm for the Encoder Block

Input: Raw audio waveform
Output: Extracted audio features

function AudioEncoder(RawAudio)
  *Conv*1 ← 1D convolutional layer with 1×3 filter
  *Conv*2 ← 1D convolutional layer with 1×5 filter
  *Conv*3 ← 1D convolutional layer with 1×7 filter
  *Block*1 ← Conv1 → Batch Normalization → ReLU → Max Pooling
  *Block*2 ← Conv2 → Batch Normalization → ReLU → Max Pooling
  *Block*3 ← Conv3 → Batch Normalization → ReLU → Max Pooling
  Features ← RawAudio
  Features ← Block1(*Features*)
  Features ← Block2(*Features*)
  Features ← Block3(*Features*)
  return *Features*
end function

B. MAIN PROCESSING BLOCK

Due to the extensive time steps in complex speech data, it is impractical for the model to process the entire sequence simultaneously. To address this, a chunking operation, i.e., divide and conquer, is employed. In the chunking operation, the latent representation, denoted as $z \in \mathbb{R}^{N \times T}$, derived from the encoder output, is partitioned into blocks of length K with a hop size of P. Consequently, the sequence is divided into

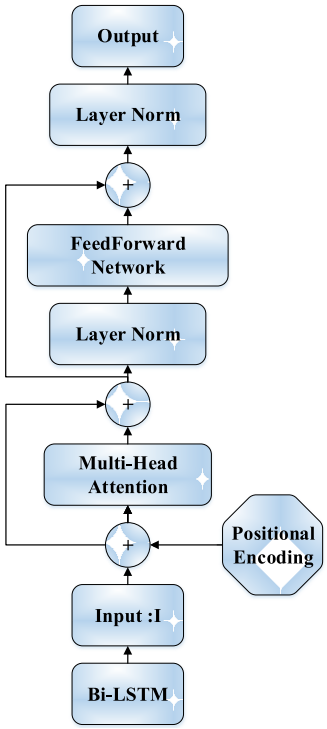


FIGURE 5. The schematic diagram of the DPRNN-Former model.

smaller sections, similar to chunking paper. Subsequently, all these sections are combined into a three-dimensional tensor denoted as $\boldsymbol{v} \in \mathbb{R}^{N\times K\times R}$. The resulting tensor $\boldsymbol{v}$ allows for more manageable processing of the speech data by the model, enabling effective analysis and manipulation of the information contained within the individual sections.

Following the chunking operation, the resulting piece-wise/chunked output $\boldsymbol{v} \in \mathbb{R}^{N\times K\times R}$ is fed into the DPRNN module, which consists of m bidirectional long short-term memory (Bi-LSTM) blocks. A pre-trained DPRNN is employed to mitigate the training error within the proprietary network. A pre-trained model denotes a stored model or network that has been constructed and trained by another individual, leveraging a substantial dataset to address a comparable problem. Figure 5 shows the schematic diagram of the DPRNN-Former model. The odd blocks, denoted as B_{2i-1}, where $i = 1, \ldots, {}^{m}/_{2}$, apply Bi-LSTM operations along the time-dependent dimension with a size of R. These blocks focus on capturing temporal dependencies within each segment. On the other hand, the even blocks, denoted as B_{2i}, are applied along the segmentation dimension with a size of K. These blocks analyze the relationships between the different segments. Intuitively, the DPRNN models the 3D tensor $\boldsymbol{v}$ in the second dimension, effectively capturing local information within small segments. Simultaneously, it utilizes the third dimension to capture the relationships between each segment, facilitating interactions within and between the blocks. The output of the DPRNN, denoted as $\boldsymbol{u} \in \mathbb{R}^{N\times K\times R}$, is then forwarded as input to the transformer module, which continues

Algorithm 2 Algorithm for the Main Processing Block

Input: Latent representation $z \in \mathbb{R}^{N\times T}$, representing the feature extracted by Algorithm 1

Output: Transformed output $\boldsymbol{u} \in \mathbb{R}^{N\times K\times R}$

function MainProcessing(Algorithm 1 Output)
  Perform chunking operation on z with block length K and hop size P to obtain $\boldsymbol{v} \in \mathbb{R}^{N\times K\times R}$
  Apply DPRNN and transfer learning on $\boldsymbol{v}$ with m bidirectional LSTM blocks
  for $i = 1$ to m **do**
    if i is odd **then**
      Apply Bi-LSTM operations along the time-dependent dimension with size R to B_{2i-1}
    else
      Apply Bi-LSTM operations along the segmentation dimension with size K to B_{2i}
    end if
  end for
  Obtain output $\boldsymbol{u}$ from the DPRNN-Transfer Learning
  Apply transformer module on $\boldsymbol{u}$ for further processing
  Output final transformed output
end function

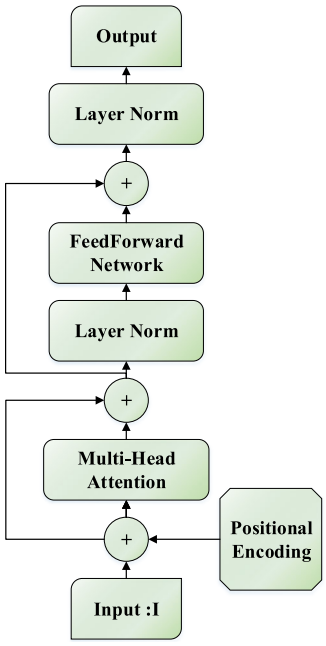


FIGURE 6. Main transformer architecture.

the processing of the data. Algorithm 2 presents a detailed outline in pseudocode of the systematic procedure employed by the main processing block. In addition, the general architecture of the main transformer encoder, as illustrated in Figure 6, can be formulated as follows:

$$MHA = MultiHeadAttention\,(I)\,, \tag{4a}$$

$$LN = LayerNorm\,(I + MHA)\,, \tag{4b}$$

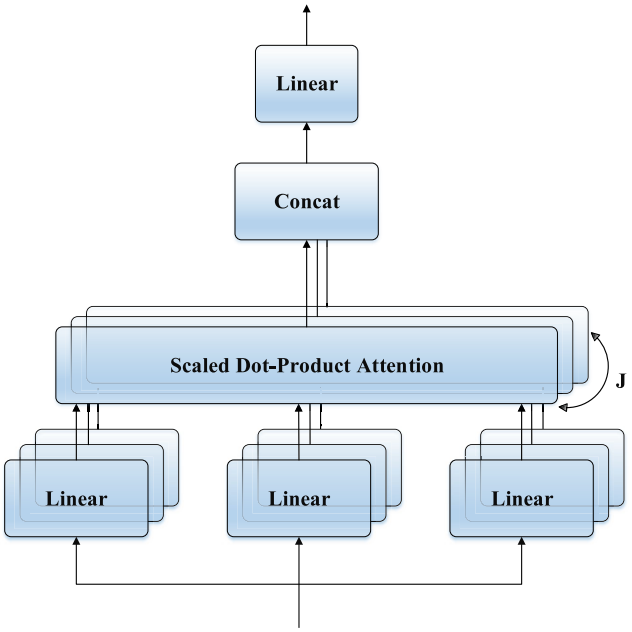


FIGURE 7. Multi-head attention architecture.

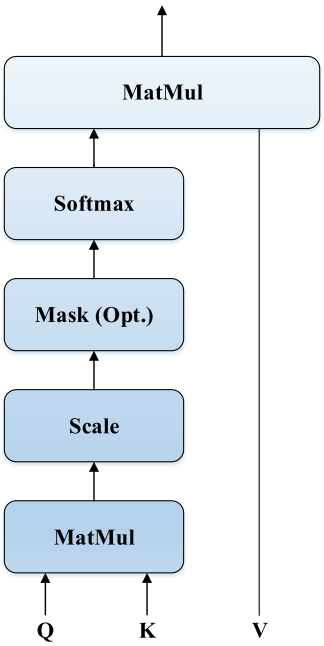


FIGURE 8. Scaled dot-product attention process.

$$FFN = FeedForward\,(LN)\,, \tag{4c}$$
$$Output = LayerNorm\,(LN + FFN)\,. \tag{4d}$$

Multi-head attention has been empirically validated as an effective self-attention mechanism for capturing and modeling long-term dependencies, as illustrated in Figure 7. The computational procedure of multi-head attention can be obtained by

$$head_j = A\left(Q_j, K_j, V_j\right), \quad j \in [1, J], \tag{5a}$$
$$\textbf{\textit{Multihead}} = concat\,(head_1, \ldots, head_J)\,W^{\circ}, \tag{5b}$$

where A is the scaled dot-product attention operation, and W° is a linear projection. In the context of multi-head attention, where the query, key, and value inputs are segregated into J distinct heads, the scaled dot-product attention operation, depicted in Figure 8, can be expressed as:

$$A(Q_j, K_j, V_j) = softmax\left(\frac{\left(Q_j^T K_j\right)}{\sqrt{\frac{D}{J}}}\right) V_i, \quad j \in [1, J], \tag{6}$$

where the softmax function is employed to convert the linear value into class probabilities. This function ensures that the output values are normalized and represent the probabilities associated with each class.

C. DECODER BLOCK

The decoder part of the model takes the outputs from the main processor and generates the final output or outcome. It can be a simple output layer in the case of classification tasks, a generative model for tasks like image generation, or any other architecture suitable for the task. The decoder unit acts as the final component of a transfer learning model, responsible for transforming the intermediate representations obtained from the main processor into meaningful and actionable output. Its primary purpose is to map the learned features to the desired output space of the target task. By utilizing the knowledge acquired from the source task, the decoder unit plays a critical role in adapting the model's knowledge to the new task, facilitating effective knowledge transfer. The decoder unit operates by employing various architectures and techniques, depending on the specific task and data domain. Its functionality can be broadly categorized into two main types: discriminative and generative.

a: DISCRIMINATIVE DECODERS

Discriminative decoders are commonly used in transfer learning scenarios involving classification, regression, or any task where a direct mapping to a specific output is required. These decoders typically consist of fully connected layers or softmax layers that take the encoded features as input and produce class probabilities, regression values, or any relevant outputs. They serve to bridge the gap between the transferred knowledge and the desired target task output, enabling effective utilization of the learned representations for accurate predictions.

b: GENERATIVE DECODERS

Generative decoders are employed in transfer learning settings where the goal is to generate new samples or to model the underlying distribution of the target data. These decoders can take various forms such as CNNs or RNNs, variational autoencoders (VAEs), or generative adversarial networks (GANs). Generative decoders learn to generate new instances that resemble the target task's data distribution by leveraging

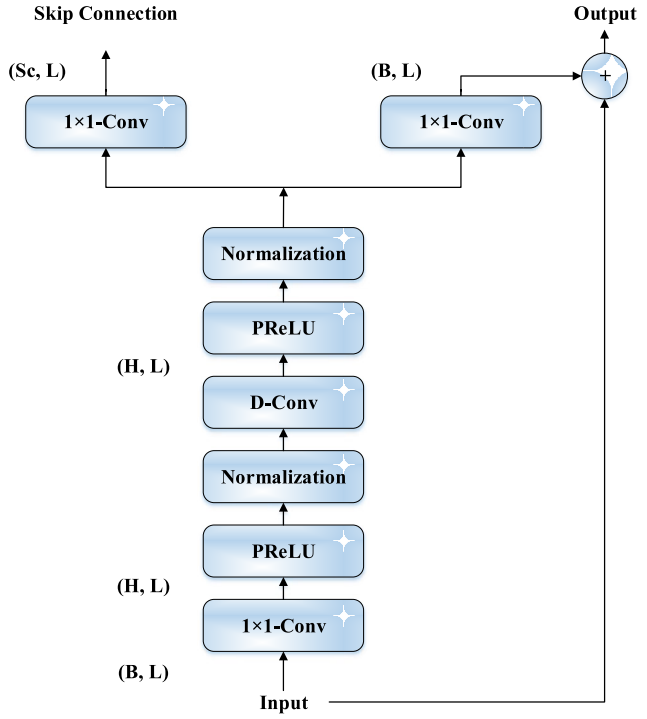


FIGURE 9. Decoder architecture.

the latent representations acquired from the source task. This allows for data augmentation, sample synthesis, or even unsupervised learning in scenarios where labeled target data is scarce.

During transfer learning, the decoder is typically fine-tuned or updated along with the main processor to align the model's output with the specific requirements of the target task. Fine-tuning enables the decoder to learn task-specific nuances and optimize its parameters to improve performance on the target task. This process is especially important when the source and target tasks have significant differences, ensuring that the transferred knowledge is appropriately adapted to the new context. The decoder module is responsible for reconstructing the source waveforms based on the masked features. Figure 9 illustrates the structure of each 1D convolutional block used in this process. The design of the 1D convolutional blocks incorporates both a residual path and a skip connection. Besides, Algorithm 3 provides a detailed depiction in pseudocode of the sequential operations employed by the decoder block.

In this design, the output of one convolutional block serves as the input to the subsequent block, creating a sequential flow of information. This sequential pathway ensures that the residual path of each block serves as the input to the next block, allowing the propagation of information through the network. Additionally, a skip connection is established, allowing the direct transfer of information across all blocks. This skip connection facilitates the propagation of information and gradients throughout the network, enhancing the

Algorithm 3 Algorithm for the Decoder Block

Input: Transformed output $\boldsymbol{u} \in \mathbb{R}^{N\times K\times R}$, representing the output extracted by Algorithm 2
Output: Estimated/Separated source signals

```
function Decoder(Main Processor Outputs)
    # Decoder
        ClassProbabilities, SeparatedSignals ← Decoder(MainProcessorOutputs)
    # Perform further processing or analysis on ClassProbabilities and SeparatedSignals
    return ClassProbabilities, SeparatedSignals
     Input ← Main Processor Outputs
    # Transfer Learning
     PretrainedModel ← LoadPretrainedModel()
     PretrainedFeatures ← PretrainedModel(Input)
    # DPRNN Module
        DPRNNOutputs ← DPRNNModule(PretrainedFeatures)
    # Transformer Module
        TransformerOutputs ← TransformerModule(DPRNNOutputs)
    # Classification
        ClassProbabilities ← ClassificationModule(TransformerOutputs)
    # Source Separation
        SeparatedSignals ← SourceSeparationModule(TransformerOutputs)
    # Decoder Block
    1 × 1-Convolution ← Input
     PReLU ← 1 × 1-Convolution
     Normalization ← PReLU
     D-Conv ← Normalization
     PReLU ← D-Conv
     Normalization ← PReLU
    1 × 1-Convolution ← Normalization
     SkipConnection(Output) ← 1 × 1-Convolution
        SeparatedSignals(Output) ← 1 × 1-Convolution(Output)
    return ClassProbabilities, SeparatedSignals
end function
```

overall learning and reconstruction capabilities of the decoder module. The mathematical formulation of the procedure can be expressed as:

$$D-Conv\,(Y,K) = Concat\left(y_j \circledast k_j\right),\quad j=1,\ldots N, \tag{7a}$$

$$S-Conv\,(Y,K,L) = D-Conv\,(Y,K) \circledast L, \tag{7b}$$

where $Y \in \mathbb{R}^{G\times M}$ represent the input to the $S-Conv$ operation, G corresponds to the number of heads and M represents the feature dimension. Moreover, $K \in \mathbb{R}^{G\times P}$ denotes the convolution kernel with a size of P, and $y_j \in \mathbb{R}^{1\times M}$ and $k_j \in \mathbb{R}^{1\times M}$ are individual rows extracted from the Y and K matrices, respectively. $L \in \mathbb{R}^{G\times H\times 1}$ corresponds to the

convolution kernel with a size of 1 in the three dimension. The symbol ⊛ represents the convolution operation.

To further clarify, the $D-Conv$ operation convolutes each row of the Y input with the corresponding row of the K matrix. In this operation, the convolution is performed between individual rows of Y and K. Additionally, the 1×1-Conv block linearly transforms the feature space. When comparing the depth-resolvable convolution with the standard convolution using a kernel size of $\hat{K} \in \mathbb{R}^{G\times H\times P}$, it is worth noting that the depth-resolvable convolution requires fewer parameters. Specifically, it involves $G \times P + G \times H$ parameters, whereas the standard convolution has $G \times H \times P$ parameters. This reduction in parameters becomes more significant when H is much larger than P, i.e., $H \gg P$, approximately by a factor of $(H \times P)/(H + P) \approx P$. Therefore, in scenarios where the dimensionality H of the kernel is considerably larger than the spatial dimension P, employing the depth-resolvable convolution can help reduce the overall model size by a factor approximately equal to P.

After the 1×1-Conv block and the D-Conv block, additional operations are incorporated to enhance the model's capabilities. Specifically, a nonlinear activation function is applied following the 1×1-Conv block, and a normalization operation is performed after the D-Conv block. The nonlinear activation function used after the 1×1-Conv block is the parametric rectified linear unit (PReLU) [78], which is given by

$$PReLU(x) = \begin{cases} x, & \text{if } x \geq 0 \\ ax, & \text{otherwise} \end{cases} \tag{8}$$

where $a \in \mathbb{R}$ is a trainable scalar that controls the negative bias of the rectifier. When considering the condition of causality in the network, the choice of normalization method can have an impact. In the case of a non-causal configuration, it has been observed through experimentation that global layer normalization (gLN) surpasses other normalization methods in terms of performance. In gLN, the feature is normalized across both the channel and time dimensions. This means that normalization is applied to each feature independently, considering its distribution along the channel dimension, i.e., across different channels, as well as its distribution along the time dimension, i.e., across different time steps. The normalization across both the channel and time dimensions are computed by

$$gLN(F) = \frac{F - E[F]}{\sqrt{Var[F] + \varepsilon}} \odot \gamma + \beta, \tag{9a}$$

$$E[F] = \frac{1}{NT} \sum_{NT} F, \tag{9b}$$

$$Var[F] = \frac{1}{NT} \sum_{NT} (F - E[F])^2, \tag{9c}$$

where $F \in \mathbb{R}^{N\times T}$ is the feature, $\beta, \gamma \in \mathbb{R}^{N\times 1}$ are trainable parameters, and ε is a small constant for numerical stability. This formulation is identical to the standard layer normalization applied in computer vision models, where the channel and time dimensions correspond to the width and height dimensions in an image [79]. In the causal configuration, gLN cannot be applied because it relies on future values of the signal at each time step. In contrast, a cumulative layer normalization (cLN) operation is developed in [80] to perform stepwise normalization on the causal system, and the mathematical equations are given by

$$cLN(F) = \frac{f - E[f_{t\leq k}]}{\sqrt{Var[f_{t\leq k}] + \varepsilon}} \odot \gamma + \beta, \tag{10a}$$

$$E[F] = \frac{1}{NT} \sum_{NT} f_{t\leq k}, \tag{10b}$$

$$Var\left[f_{t\leq k}\right] = \frac{1}{NT} \sum_{NT} \left(f_{t\leq k} - E\left[f_{t\leq k}\right]\right)^2, \tag{10c}$$

where $f_k \in \mathbb{R}^{N\times 1}$ is the k-th frame of the complete feature $F, f_{(t\leq k)} \in \mathbb{R}^{N\times K}$ corresponds to the k feature of the frame $[f_1, f_2, \ldots, f_k]$. The trainable parameters β and $\gamma \in \mathbb{R}^{N\times 1}$ are applied uniformly to all frames. To ensure that the decoupling module remains invariant to the input scale, a chosen normalization method is applied to the output of the encoder w before it is passed to the decoupling module.

The separation module begins with the introduction of a 1×1-convolutional linear block, which serves as a bottleneck layer. This block not only determines the number of input channels but also functions as the remaining path for subsequent convolutional blocks. For instance, in the case where the linear bottleneck layer comprises B channels, a 1D convolutional block with H channels and a kernel size of P would require a kernel size denoted as $O \in \mathbb{R}^{B\times H\times 1}$ for the first 1×1-convolutional block, and $K \in \mathbb{R}^{H\times P}$ for the initial D-convolutional block. Subsequently, the kernel size in the remaining paths should be represented as $L_{Rs} \in \mathbb{R}^{H\times B\times 1}$. It is worth noting that the number of output channels in the hop connection path may deviate from B. For this purpose, the size of the cores in that particular path can be denoted as $L_{Sc} \in \mathbb{R}^{H\times Sc\times 1}$. These design choices in the separation module contribute to its effectiveness and enable efficient information flow between the various convolutional blocks.

IV. SIMULATION RESULTS

In this section, the results of the proposed algorithm are analyzed. First, the dataset and considered parameters are examined, followed by a benchmark of the results of the proposed algorithm with the state-of-the-art work. To assess the efficacy of BSS in B5G and 6G networks, extensive simulations and performance evaluations are conducted. The section presents the experimental setup, performance metrics, and comparative analysis of BSS-enabled solutions against conventional approaches.

This work implements and benchmarks eight distinct deep learning-based BSS frameworks, which encompass deep clustering (DeepClustering) [7], [78], [81], fixed attractor deep affinity network (FixedAttractorDANet) [82], [83], long short-term memory time-domain audio separation network

(LSTMTasNet) [84], convolutional time-domain audio separation network (ConvTasNet) [80], DPRNN time-domain audio separation network (DPRNNTasNet) [85], dual-path transformer network (DPTNet) [86], separation transformer (SepFormer) [87], and the proposed technique. The results demonstrate the potential of BSS in improving interference management, channel estimation accuracy, beamforming performance, and resource allocation efficiency.

A. DATASET

The Wall Street Journal (WSJ0) dataset, which is extensively employed in research related to automatic speech recognition, is characterized by read English speech from 101 speakers. Originally recorded at a sampling rate of 16 kHz, the individual speeches in the WSJ0 collection exhibit considerable variation in length. While shorter sentences or sentence fragments may span only a few seconds, longer texts can extend beyond a minute. However, in the realm of speech and audio processing, it is customary to partition these recordings into smaller, fixed-length segments for the purposes of analysis or model training. The duration of these chunks typically ranges from hundreds of milliseconds to a few seconds.

The WSJ0-2mix and WSJ0-3mix datasets have gained significant popularity as valuable resources for investigating speech separation, a prominent challenge within the domain of audio signal processing. WSJ0-2mix is a derived subset of the WSJ0 corpus [78], where a deliberate mixing process has been applied to generate a dataset comprising speech mixtures from two distinct speakers. Consequently, the task of separating these individual speakers becomes notably difficult. Building upon the WSJ0-2mix dataset, WSJ0-3mix takes a similar approach but introduces an additional complexity by including three speakers in each audio sample. Consequently, this extended dataset presents an even greater challenge for source separation algorithms due to the necessity of disentangling three overlapping speakers. These datasets were developed as part of the research conducted in [81].

For the creation of the WSJ0-2mix and WSJ0-3mix datasets, the researchers followed a specific procedure. They randomly selected pairs of utterances from different speakers within the WSJ0 dataset and combined them to form two- or three-speaker mixes. Prior to mixing, the original utterances were appropriately scaled to ensure a rough equivalence in speaker power within the resulting mixture. The mixes were generated across a range of SNR, spanning from −5 dB to 5 dB. To facilitate the training and evaluation of separation algorithms, the researchers also provided accompanying IBMs for each speaker in both datasets. These masks serve as ground truth references, aiding in the separation of the mixed speech signals. Each dataset comprises three distinct subsets: a training set, a validation set, and a test set. These subsets enable researchers to effectively train, validate, and assess the performance of speech separation algorithms on these datasets.

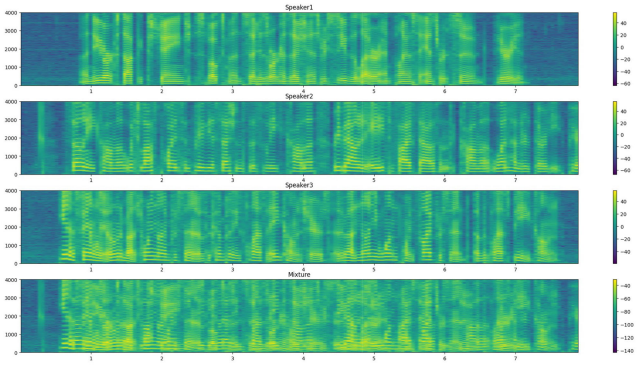


FIGURE 10. The spectrograms of the original files and the test mixture files.

In order to assess the performance of the proposed network, a mixed file is generated using the following procedure. First, three audio files are selected from the Carnegie Mellon University Arctic (CMU Arctic) dataset [88], each recorded by a different speaker. To reduce the size of the test and evaluation files, the samples are resampled at a rate of 8000 Hz. The selected samples are then mixed using the provided application programming interface (API), and their relevant spectrograms are shown in Figure 10.

This study focuses on the development and evaluation of a BSS framework for the task of separating randomly mixed signals. The primary objective of the framework is to achieve accurate and robust separation and estimation of individual sources without prior knowledge of the mixing matrix or the characteristics of the source signals. For the purpose of this study, we denote each source as Speaker 1, Speaker 2, and Speaker 3. The proposed BSS framework is thoroughly examined in the context of separating mixtures containing two or three source signals, corresponding to speakers in our analysis. By evaluating the model's performance in scenarios involving multiple sources, we assess its effectiveness in simultaneously disentangling and recovering individual sources within the mixed audio signals. Throughout this study, our investigation revolves around the core challenge of source separation, where the framework is expected to successfully separate and accurately estimate the underlying sources while being agnostic to the specific mixing configuration and the characteristics of the individual speakers.

B. BENCHMARKING CRITERIA

In this section, the graphical and statistical tools required for comparing and evaluating the outputs of the proposed BSS model are discussed. In order to accurately assess the performance of audio source separation approaches, researchers have dedicated efforts to develop appropriate evaluation methods. Objective evaluation of audio source separation approaches primarily relies on the calculation of criteria that measure the extent to which energy is separated among the estimated source components. Key measures used in this context include the SDR, SIR, and SAR. These measures are

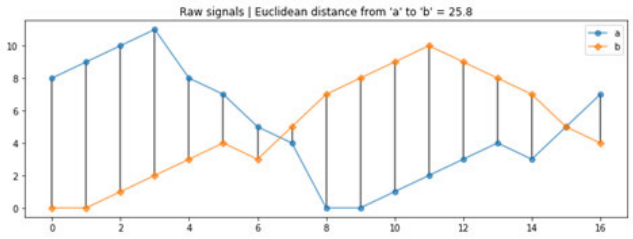


FIGURE 11. Euclidean distance example of two vectors ***a*** and ***b***.

widely recognized and serve as standard evaluation metrics for assessing source isolation quality. On the other hand, subjective evaluation of audio source separation approaches focuses on the perceptual assessment of the quality of the separated/estimated source(s).

1) COMPARISON OF SIGNALS IN THE TIME DOMAIN

When comparing time series signals, such as signal a and signal b, determining if they are the same or similar requires defining a notion of similarity. There are various approaches to tackle this task, and the choice of a suitable comparison function depends on the specific characteristics and goals of the analysis. One simple approach, is to compare each value in signal a to the corresponding value in signal b. However, this approach may not be sufficient for capturing the overall similarity of the signals, especially when dealing with noisy or time-shifted data. The consideration of the following techniques allows for the definition of a more comprehensive comparison function:

a: EUCLIDEAN DISTANCE

The Euclidean distance between two points p and q is defined as the length of the line segment connecting them. In the context of comparing signals $\boldsymbol{a}$ and $\boldsymbol{b}$, one approach is to iterate through the arrays and calculate the Euclidean distance between each corresponding pair of points as follows:

$$D_E(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{n} (a_i - b_a)^2}. \tag{11}$$

The Euclidean distance between the two signals is calculated by treating them as multidimensional vectors. This approach measures the overall difference between the signal values but does not consider any temporal dependencies. Figure 11 serves as an illustrative example to provide a clearer explanation of the Euclidean distance.

b: CROSS-CORRELATION

Cross-correlation is a similarity measure utilized to identify the maximum overlap between two signals by sliding one signal over the other. It is closely related to the concept of convolution. Cross-correlation finds extensive application in various fields, including pattern recognition, computer vision, etc. It is often employed to search for a shorter, known signal within a longer signal.

a) Scale-Invariant Signal-to-Noise Ratio (SI-SNR): is a metric that provides a measure of the quality of a signal in the presence of noise, while accounting for the scale or magnitude of the signal. It is particularly useful when comparing signals of different amplitudes or scales. The SI-SNR can be calculated by [84]

$$SI-SNR = 10\log_{10} \frac{\left\| s_{t\arg et} \right\|^2}{\left\| e_{noise} \right\|^2}, \tag{12a}$$

$$s_{target} = \frac{\langle \hat{s}, s \rangle}{\|s\|^2}, \tag{12b}$$

$$e_{noise} = \hat{s} - s_{target}. \tag{12c}$$

b) Signal-to-Distortion (SDR) Ratio: is a measure used to assess the quality of a signal after it has been distorted or corrupted by noise, interference, or other factors. SDR is particularly relevant in audio signal processing and source separation applications, and is expressed by [89]

$$SDR = 10\log_{10} \frac{\left\| s_{target} + e_{interf} \right\|^2}{\left\| e_{interf} + e_{noise} + e_{artif} \right\|^2}, \tag{13a}$$

$$s_{target} = P_{sj}\hat{s}_J, \tag{13b}$$

$$e_{interf} = P_s\hat{s}_J - P_{sj}\hat{s}_J, \tag{13c}$$

$$e_{noise} = P_{s,n}\hat{s}_J - P_s\hat{s}_J, \tag{13d}$$

$$e_{artif} = \hat{s}_J - P_{s,n}\hat{s}_J. \tag{13e}$$

c) Short-Time Objective Intelligibility (STOI): is a metric used to assess the intelligibility or understandability of speech signals. STOI measures the similarity between the clean or reference speech signal and a degraded or processed speech signal, taking into account the effects of noise, distortion, and other factors [90]. The STOI metric is based on the short-time magnitude spectrum of the speech signals. It calculates a similarity index between the clean and degraded signals, representing the degree to which the degraded signal retains the intelligibility of the clean reference signal. STOI values range from 0 to 1, where a value of 1 indicates perfect intelligibility, meaning the degraded signal is indistinguishable from the clean reference signal, while a value of 0 indicates complete loss of intelligibility. STOI is calculated as [91], (14a) and (14b), as shown at the bottom of the next page, where $X_j(n)$ denotes the norm of the j^{th} one-third octave band, $Y'_j(n)$ represents the normalized and clipped TF-unit, N represents the number of consecutive TF-units from both $X_j(n)$ and $Y_j(n)$, $l \in \mathcal{M} = \{(m-N+1), (m-N+2), \ldots, m-1, m\}$, J indicates the number of one-third octave bands, and M is the total number of frames. Moreover, STOI can be mathematically formulated as

$$STOI = \left(\prod_{m=1}^{M} G(m) \right)^{\frac{1}{M}}, \tag{15a}$$

$$G(m) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M}\sum_{m=1}^{M}\frac{M(k,m)}{D(K,m)}, \tag{15b}$$

where $D(k,m) = \|X(k,m) - Y(k,m)\|_2$, $M(k,m)$ denotes the IBM indicating clean speech dominance, $X(k,m)$ represents the magnitude spectrum of the clean reference signal, $Y(k,m)$ indicates the magnitude spectrum of the degraded signal, K is the total number of frequency bins, and M is the total number of frames.

C. SIMULATION RESULTS

This Section presents a discussion on the comparison of algorithms and the evaluation using the correlation coefficient criterion. All the simulations in this study are performed exploiting hardware with the specifications outlined as follows: The central processing unit (CPU) is an Intel(R) Xeon(R) CPU running at a clock speed of 2.20GHz. The system has a random access memory (RAM) capacity of 12.7GB. For graphical processing, it is equipped with an NVIDIA Tesla T4 graphics processing unit (GPU). In terms of storage, the system has a disk capacity of 107.7GB. Table 3 examines the correlation coefficients of three parameters between the original and estimated signals, providing a comprehensive comparison across various algorithms. Furthermore, Figures 12 through 14 present the graphical representation of the results obtained using the selected techniques. The analyses focus on evaluating the relationship and consistency between the original and estimated signals using correlation coefficients as a measure of performance.

DeepClustering achieves a mean STOI score of 0.73, mean SDR of -0.23, and mean SI-SNR of -1.11. For Speaker 1, it obtains a reasonable STOI score of 0.86, SDR of 4.10, and SI-SNR of 3.46. However, it performs poorly for Speaker 2 and Speaker 3, with negative SDR and SI-SNR scores. Overall, DeepClustering shows limited effectiveness in separating audio sources, as indicated by the negative mean SDR and SI-SNR values.

FixedAttractorDANet demonstrates improved performance compared to DeepClustering. It achieves a mean STOI score of 0.72, mean SDR of 2.42, and mean SI-SNR of 1.79. Notably, it exhibits better separation results for all three speakers compared to DeepClustering, with positive SDR and SI-SNR scores. The improvement percentage can be calculated by comparing the mean scores of FixedAttractorDANet with DeepClustering.

LSTMTasNet achieves the mean STOI score of 0.82, mean SDR of 5.30, and mean SI-SNR of 4.72. It consistently outperforms the previous two algorithms for all three speakers, achieving positive SDR and SI-SNR scores. Compared to FixedAttractorDANet, LSTMTasNet shows improvements in mean STOI, SDR, and SI-SNR by 13.89%, 119.01%, and 163.69%, respectively.

ConvTasNet performs reasonably well with a mean STOI score of 0.80, mean SDR of 3.41, and mean SI-SNR of 2.66. It demonstrates better results compared to DeepClustering for all three speakers, but falls short in performance when compared to FixedAttractorDANet and LSTMTasNet. Nonetheless, it shows improvements over DeepClustering, with positive SDR and SI-SNR scores and higher mean scores.

DPRNNTasNet surpasses ConvTasNet in performance, achieving a mean STOI score of 0.87, mean SDR of 7.48, and mean SI-SNR of 7.20. It outperforms all previous algorithms in terms of mean STOI, SDR, and SI-SNR scores. Compared to ConvTasNet, DPRNNTasNet exhibits improvements in mean STOI, SDR, and SI-SNR by 8.75%, 119.35%, and 170.67%, respectively.

DPTNet performs well, achieving a mean STOI score of 0.84, mean SDR of 6.32, and mean SI-SNR of 5.76. It shows improvements over ConvTasNet, but falls slightly behind DPRNNTasNet in terms of mean scores. The improvement percentages relative to ConvTasNet are 5.00%, 85.34%, and 116.54% for mean STOI, SDR, and SI-SNR, respectively.

SepFormer exhibits an average performance, with a mean STOI score of 0.76, mean SDR of 4.25, and mean SI-SNR of 3.74. It achieves lower scores compared to ConvTasNet and DPTNet. The improvement percentages relative to ConvTasNet are −5.00%, 24.63%, and 40.60% for mean STOI, SDR, and SI-SNR, respectively.

The proposed model demonstrates the highest performance in terms of mean STOI, SDR, and SI-SNR scores. It achieves a mean STOI score of 0.91, mean SDR of 8.59, and mean SI-SNR of 7.91. The proposed model consistently outperforms all other algorithms, including DPRNNTasNet, which was previously the top-performing algorithm. Compared to DPRNNTasNet, the proposed model shows improvements in mean STOI, SDR, and SI-SNR by 4.60%, 14.84%, and 9.87%, respectively.

$$d_j(m) = \frac{\sum_m \left(X_j(n) - \frac{1}{N}\sum_l X_j(l)\right)\left(Y'_j(n) - \frac{1}{N}\sum_l Y'(l)\right)}{\sqrt{\sum_n \left(X_j(n) - \frac{1}{N}\sum_l X_j(l)\right)^2 \sum_n \left(Y'_j(n) - \frac{1}{N}\sum_l Y'(l)\right)^2}}, \tag{14a}$$

$$d = \frac{1}{JM}\sum_{j,m} d_j(m), \tag{14b}$$

TABLE 3. Comparative evaluation of SI-SNR, SDR, and STOI metrics for main and separated signals across various algorithms.

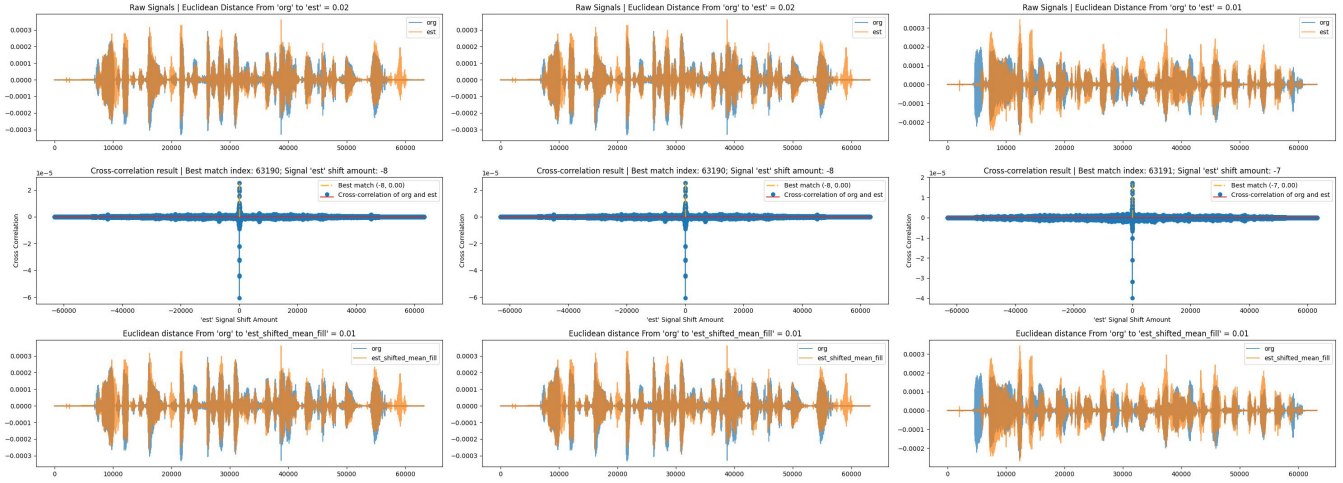
#	Algorithms	Sources	SI-SNR	SDR	STOI	#	Algorithms	Sources	SI-SNR	SDR	STOI
1	Deep-Clustering [7], [78], [81]	Speaker 1	3.46	4.10	0.86	5	DPRNN-TasNet [85]	Speaker 1	8.60	10.10	0.92
		Speaker 2	-2.96	-1.78	0.62			Speaker 2	6.20	6.51	0.79
		Speaker 3	-3.84	-3.02	0.70			Speaker 3	6.80	5.82	0.90
		Mean	-1.11	-0.23	0.73			Mean	7.20	7.48	0.87
2	Fixed-Attractor-DANet [82], [83]	Speaker 1	3.69	4.39	0.84	6	DPTNet [86]	Speaker 1	7.72	8.34	0.92
		Speaker 2	0.31	1.25	0.63			Speaker 2	4.43	4.95	0.77
		Speaker 3	1.38	1.84	0.70			Speaker 3	5.13	5.68	0.85
		Mean	1.79	2.42	0.72			Mean	5.76	6.32	0.84
3	LSTM-TasNet [84]	Speaker 1	9.27	9.81	0.95	7	Sep-Former [87]	Speaker 1	4.49	5.09	0.78
		Speaker 2	1.70	2.40	0.72			Speaker 2	3.91	4.41	0.72
		Speaker 3	3.19	3.69	0.80			Speaker 3	2.81	3.25	0.77
		Mean	4.72	5.30	0.82			Mean	3.74	4.25	0.76
4	Conv-TasNet [80]	Speaker 1	8.00	8.45	0.88	8	Proposed Model	Speaker 1	9.62	10.25	0.95
		Speaker 2	-0.09	1.06	0.73			Speaker 2	6.25	8.02	0.86
		Speaker 3	0.07	0.72	0.79			Speaker 3	7.85	7.49	0.93
		Mean	2.66	3.41	0.80			Mean	7.91	8.59	0.91

Furthermore, Table 4 concisely provides an enlightening explanation of the computational complexities associated with different BSS models. These complexities are elegantly expressed using big O notation and runtime. The notations employed in the table bear significance in concisely representing key dimensions and attributes of the models operations. In this context, I indicates the number of iterations, T symbolizes the sequence length, capturing the temporal extent of the input data, N holds significance as it signifies the count of input features, illuminating the dimensionality of the input space. Moreover, C assumes the role of representing the number of channels, indicating the diversity of information streams within the data. The parameters L, H, W, and M collectively delineate critical aspects of the data, where L denotes length, H signifies height, W encapsulates width, and M characterizes the number of masks. Furthermore, the variables m, K, and R respectively denote the quantities of blocks, chunk length, and number of chunks.

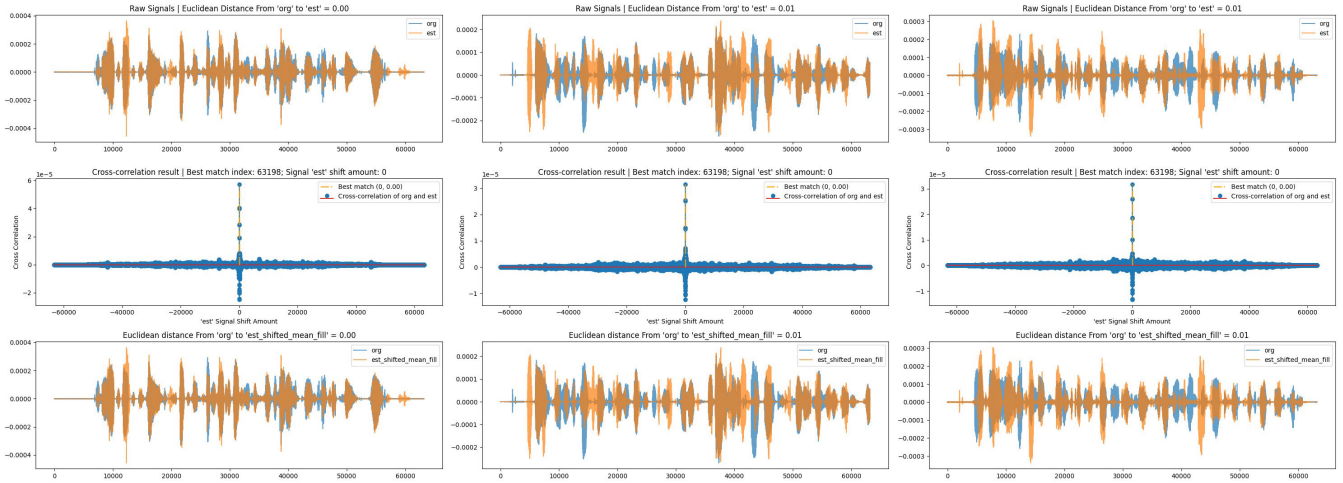
TABLE 4. Computational complexities of BSS models in terms of big O notation and runtime.

Model	Computational Complexity	Big O Notation	Runtime (s)
Deep Clustering	Moderate, highly dependent on iterations and clustering algorithm	$\geq O(I \times N)$	12.34
Fixed-Attractor-DANet	Moderate	$O(N \times M \times H \times W \times C^2)$	16.29
LSTM-TasNet	Moderate to High	$O(T \times N \times C^2 \times L)$	22.07
Conv-TasNet	Moderate	$O(T \times N \times C \times L^2)$	18.56
DPRNN-TasNet	Moderate to High	$O(T \times N \times C^2 \times L)$	27.07
DPTNet	Moderate to High	$O(T \times N^2 \times C^2 \times H)$	28.43
Sep-Former	High	$O(T \times N^2 \times C^2 \times L)$	34.39
Proposed Model	Moderate to High	$O(m \times N \times K \times R \times C^2)$	26.33

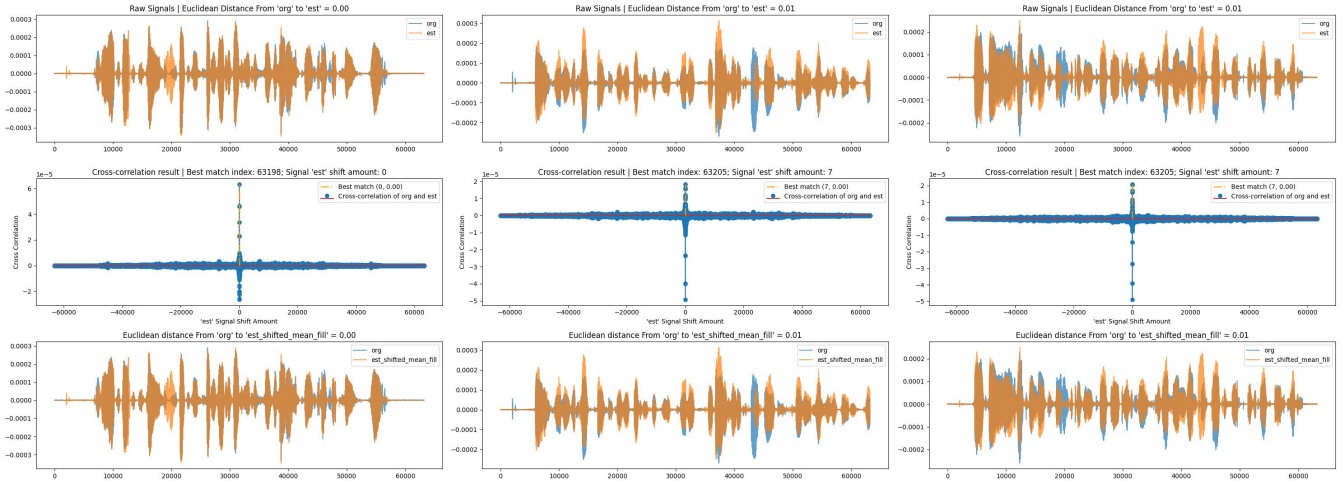
In the landscape of BSS models, an examination of computational complexities reveals intriguing insights into their distinctive characteristics. DeepClustering, a methodology involving iterative clustering algorithms such as k-means applied to embedded feature representations, offers a versatile approach to untangle mixed sources. The computational complexity of DeepClustering hinges on factors like iteration count, cluster numbers, and embedding dimensions, rendering its scalability contingent upon dataset scale. Moving forward, the FixedAttractorDANet model embraces deep neural networks as its foundation, navigating forward and backward passes to process information. Its computational demand, while moderate due to network architecture, adapts with network depth and input dimensions. LSTMTasNet, by leveraging LSTM layers for temporal dependencies, introduces higher complexity due to its recurrent nature. The sequential matrix multiplications and non-linear activations amplify computational requirements, particularly within LSTM layers. On the other hand, ConvTasNet's proficiency



(a) DeepClustering v. 1. (b) DeepClustering v. 2. (c) DeepClustering v. 3.

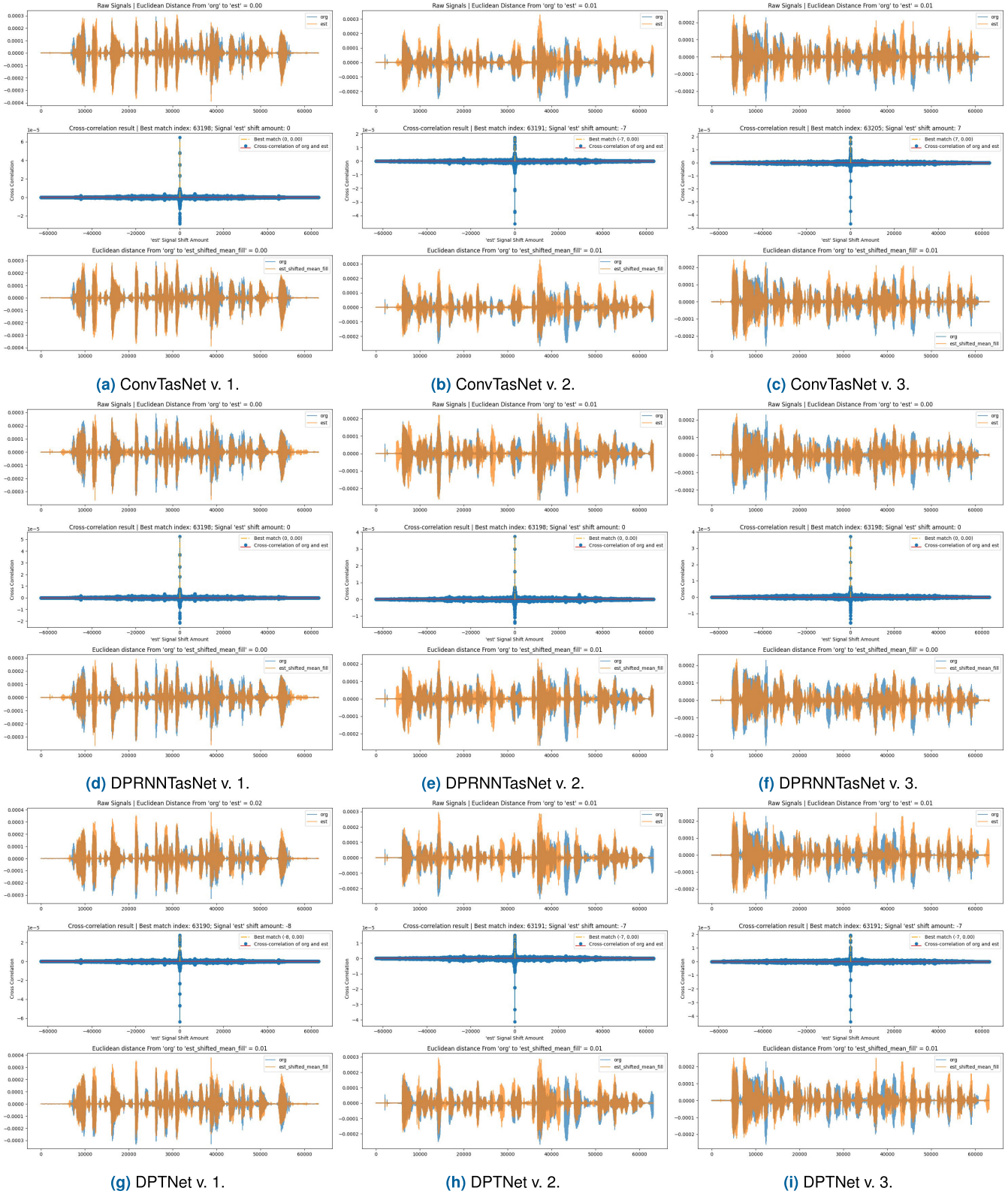


(d) FixedAttractorDANet v. 1. (e) FixedAttractorDANet v. 2. (f) FixedAttractorDANet v. 3.



(g) LSTMTasNet v. 1. (h) LSTMTasNet v. 2. (i) LSTMTasNet v. 3.

FIGURE 12. Comparative analysis of correlation coefficients and Euclidean distances scrutinizing the congruence between original signals (the main, i.e., org) and estimated signals (separated signals, i.e., est) across multiple algorithms, including DeepClustering, FixedAttractorDANet, and LSTMTasNet networks.



(a) ConvTasNet v. 1. **(b)** ConvTasNet v. 2. **(c)** ConvTasNet v. 3.

(d) DPRNNTasNet v. 1. **(e)** DPRNNTasNet v. 2. **(f)** DPRNNTasNet v. 3.

(g) DPTNet v. 1. **(h)** DPTNet v. 2. **(i)** DPTNet v. 3.

FIGURE 13. Comparative analysis of correlation coefficients and Euclidean distances scrutinizing the congruence between original signals (the main, i.e., org) and estimated signals (separated signals, i.e., est) across multiple algorithms, including ConvTasNet, DPRNNTasNet, and DPTNet networks.

in processing time-domain signals stems from its utilization of convolutional layers, yielding comparatively lower complexity than recurrent layers. The complexity scales with network depth and kernel sizes. Introducing parallel and recurrent processing, DPRNNTasNet adopts a DPRNN architecture. This parallelism optimizes computational efficiency,

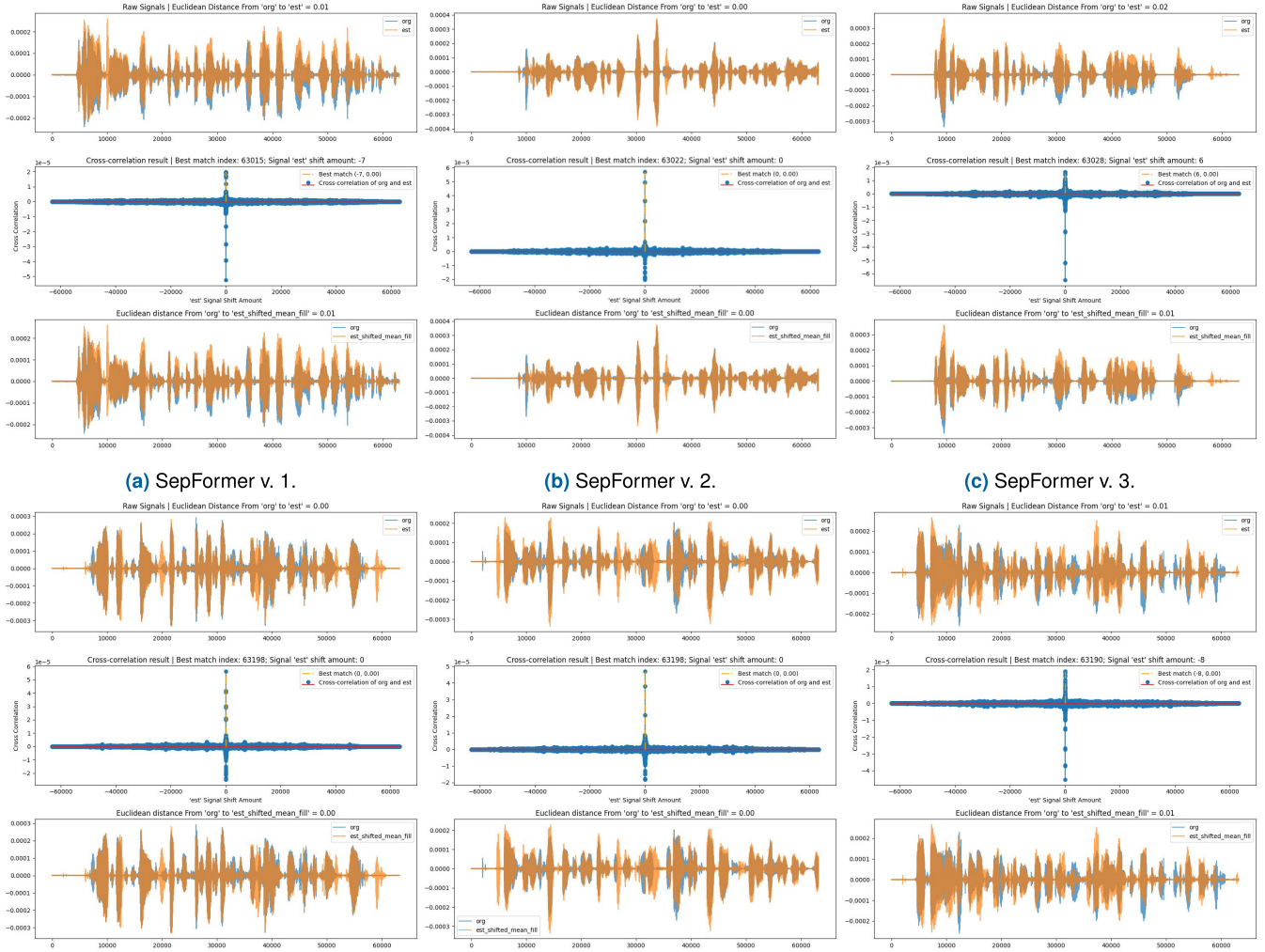


FIGURE 14. **Comparative analysis of correlation coefficients and Euclidean distances scrutinizing the congruence between original signals (the main, i.e., org) and estimated signals (separated signals, i.e., est) across the SepFormer network and the Proposed model.**

positioning its complexity as moderate in comparison to intricate models. Meanwhile, DPTNet's intricacy is dictated by architecture and components; the inclusion of attention mechanisms or transformers can elevate its complexity due to the quadratic time complexity arising from self-attention operations. Finally, SepFormer's transformer-based design bestows it with elevated computational complexity. Multi-head self-attention and feedforward layers, inherent to transformers, demand significant computation, particularly when handling lengthy input sequences. Amidst this panorama of complexities, the proposed model emerges as a vanguard, adorning a three-way neural network architecture entwining transfer learning, pre-trained DPRNN, and transformers. This amalgamation seamlessly navigates through the intricacies of audio/speech signal separation, positioning itself as a potent contender in addressing contemporary challenges.

In summary, the evaluation results highlight the effectiveness of different algorithms in separating audio sources. LSTMTasNet, DPTNet, DPRNNTasNet, and the proposed model consistently demonstrate superior performance compared to other algorithms, with the proposed model ultimately achieving the highest mean scores. It is important to note that the improvement percentages are calculated by comparing the mean scores of each algorithm with the baseline algorithm. These improvements indicate the advancements made by each algorithm in terms of audio/speech separation quality and highlight the progression within the field.

In the landscape of advanced communication paradigms, characterized by the imminent advent of B5G and 6G networks, the pursuit of efficient and intelligent signal processing techniques takes precedence. Addressing this demand, our work introduces a pioneering BSS model that holds profound implications for the augmentation of B5G and 6G communication systems. The novelty of our approach resides in the strategic fusion of three distinctive neural network components: transfer learning, a pre-trained DPRNN, and a transformer architecture. The synergistic amalgamation of these components engenders a multi-faceted framework

capable of disentangling complex mixed signals encountered in the dynamic and heterogeneous communication environments envisioned by B5G and 6G networks.

The incorporation of transfer learning enables the model to leverage pre-existing knowledge and adapt to context-specific scenarios, while the pre-trained DPRNN contributes temporal context preservation and signal continuity. The transformative inclusion of a transformer architecture further empowers the model with attention mechanisms that discern salient features across signals. This holistic architecture, poised at the intersection of deep learning and signal processing, not only advances the efficacy of BSS techniques but also resonates deeply with the multifaceted demands of B5G and 6G communications. By enabling the robust extraction of pristine source signals from complex mixtures, the proposed model intricately aligns with the imperatives of spectral efficiency, interference mitigation, and adaptive resource allocation that define the trajectory of B5G and 6G networks. As such, this work contributes significantly to the evolving landscape of next-generation communications by offering a sophisticated and adaptable BSS framework tailored to the demands of B5G and 6G communication systems.

It is worth mentioning that the evaluation metrics alone may not provide a complete assessment of audio separation algorithms. Other factors, such as computational complexity, real-time processing capabilities, and subjective listening tests, should also be considered when determining the practical suitability of an algorithm for specific applications. Overall, the provided analyses offer insights into the performance of different audio/speech separation algorithms and their improvements over the baseline algorithm. The field of audio source separation continues to evolve, and these advancements contribute to enhancing the quality of separated audio sources and opening up new possibilities in various audio-related applications.

V. CONCLUSION AND FUTURE DIRECTION

In conclusion, this paper highlights the significance of BSS in enhancing signal processing capabilities in B5G and 6G. In the context of future wireless communication systems such as B5G and 6G, the process of transmitting a signal from the transmitter side and subsequently receiving it at the receiver side introduces the possibility of signal contamination due to undesired components in the transmission channel. This article presents a novel algorithm aimed at restoring the original signals from such contamination. This study introduces a novel three-way neural network architecture that combines transfer learning, a pre-trained DPRNN, and a transformer model. The proposed algorithm outperforms all the benchmarked techniques in terms of SI-SNR, SDR, and STOI metrics, showcasing the highest improvement percentages across the board compared to that of selected algorithms. In particular, the proposed technique achieves an average SI-SNR of 7.91, which is a 9.87% improvement compared to DPRNNTasNet, the nearest competitor. In terms of SDR, the proposed algorithm achieves 8.59, showing a remarkable 14.84% enhancement compared to that of DPRNNTasNet. Additionally, the proposed algorithm achieves an STOI of 0.91, indicating a substantial 10.98% and 4.60% increase over LSTMTasNet and DPRNNTasNet, respectively.

The effectiveness of the proposed algorithm in tackling real-world challenges, such as complex acoustic environments characterized by noise and reverberation, is clearly evident. The robustness of the framework enables its applicability in practical domains such as speech enhancement, audio transcription, and audio-visual processing, particularly within the realm of B5G and 6G technologies. Future work in audio BSS could focus on exploring hybrid approaches that combine the strengths of different algorithms to further improve separation performance. Additionally, the development of novel evaluation metrics that capture additional aspects of audio quality and perceptual attributes could provide a more comprehensive assessment of separation algorithms.

REFERENCES

[1] B. A. Adoum, K. Zoukalne, M. S. Idriss, A. M. Ali, A. Moungache, and M. Y. Khayal, "A comprehensive survey of candidate waveforms for 5G, beyond 5G and 6G wireless communication systems," *Open J. Appl. Sci.*, vol. 13, no. 1, pp. 136–161, 2023.

[2] U. Gustavsson, P. Frenger, C. Fager, T. Eriksson, H. Zirath, F. Dielacher, C. Studer, A. Parssinen, R. Correia, and J. N. Matos, "Implementation challenges and opportunities in beyond-5G and 6G communication," *IEEE J. Microw.*, vol. 1, no. 1, pp. 86–100, Jan. 2021.

[3] K. Xie, K. Jiang, and Q. Yang, "Multi-channel underdetermined blind source separation for recorded audio mixture signals using an unmanned aerial vehicle," *IET Commun.*, vol. 15, no. 10, pp. 1412–1422, Jun. 2021.

[4] W. Zhang, A. Tait, C. Huang, T. Ferreira de Lima, S. Bilodeau, E. C. Blow, A. Jha, B. J. Shastri, and P. Prucnal, "Broadband physical layer cognitive radio with an integrated photonic processor for blind source separation," *Nature Commun.*, vol. 14, no. 1, p. 1107, Feb. 2023.

[5] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Cambridge, MA, USA: Academic Press, 2010.

[6] P. He, T. She, W. Li, and W. Yuan, "Single channel blind source separation on the instantaneous mixed signal of multiple dynamic sources," *Mech. Syst. Signal Process.*, vol. 113, pp. 22–35, Dec. 2018.

[7] L. Drude, D. Hasenklever, and R. Haeb-Umbach, "Unsupervised training of a deep clustering model for multichannel blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 695–699.

[8] D.-S. Huang and W. Jiang, "A general CPL-AdS methodology for fixing dynamic parameters in dual environments," *IEEE Trans. Syst., Man, Cybern., B*, vol. 42, no. 5, pp. 1489–1500, Oct. 2012.

[9] L. Shang, D.-S. Huang, C.-H. Zheng, and Z.-L. Sun, "Noise removal using a novel non-negative sparse coding shrinkage technique," *Neurocomputing*, vol. 69, pp. 874–877, Mar. 2006.

[10] K. B. Bhangale and M. Kothandaraman, "Survey of deep learning paradigms for speech processing," *Wireless Pers. Commun.*, vol. 125, pp. 1913–1949, Jul. 2022.

[11] C. Li, L. Zhu, Z. Luo, Z. Zhang, and Y. Yang, "Effective methods and performance analysis on data transmission security with blind source separation in space-based AIS," *China Commun.*, vol. 19, no. 4, pp. 154–165, Apr. 2022.

[12] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: State of the art and research challenges," *Multimedia Tools Appl.*, vol. 80, pp. 23745–23812, Jan. 2021.

[13] J. J. C. Sheeja and B. Sankaragomathi, "CNN-QTLBO: An optimal blind source separation and blind dereverberation scheme using lightweight CNN-QTLBO and PCDP-LDA for speech mixtures," *Signal, Image Video Process.*, vol. 16, no. 5, pp. 1323–1331, Jul. 2022.

[14] H. Ma, X. Zheng, X. Wu, L. Yu, and P. Xiang, "A blind separation algorithm for underdetermined convolutional mixed communication signals based on time-frequency soft mask," *Phys. Commun.*, vol. 53, Aug. 2022, Art. no. 101747.

[15] S. Soni, R. N. Yadav, and L. Gupta, "State-of-the-art analysis of deep learning-based monaural speech source separation techniques," *IEEE Access*, vol. 11, pp. 4242–4269, 2023.

[16] Z. Luo, C. Li, and L. Zhu, "A comprehensive survey on blind source separation for wireless adaptive processing: Principles, perspectives, challenges and new research directions," *IEEE Access*, vol. 6, pp. 66685–66708, 2018.

[17] W. Cui, S. Guo, L. Ren, and Y. Yu, "Underdetermined blind source separation for linear instantaneous mixing system in the non-cooperative wireless communication," *Phys. Commun.*, vol. 45, Apr. 2021, Art. no. 101255.

[18] M. E. Fouda, C. Shen, and A. E. Eltawil, "Blind source separation for full-duplex systems: Potential and challenges," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1379–1389, 2021.

[19] J. He, H. Wu, X. Xiao, R. Bassoli, and F. H. P. Fitzek, "Functional split of in-network deep learning for 6G: A feasibility study," *IEEE Wireless Commun.*, vol. 29, no. 5, pp. 36–42, Oct. 2022.

[20] S. Ansari, K. A. Alnajjar, S. Mahmoud, R. Alabdan, H. Alzaabi, M. Alkaabi, and A. Hussain, "Blind source separation based on genetic algorithm-optimized multiuser kurtosis," in *Proc. 46th Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2023, pp. 164–171.

[21] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[22] J. F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Procces.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.

[23] C. Fevotte and J. Cardoso, "Maximum likelihood approach for blind audio source separation using time-frequency Gaussian source models," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2005, pp. 78–81.

[24] S. Ansari, A. S. Alatrany, K. A. Alnajjar, T. Khater, S. A. Mahmoud, D. Al-Jumeily, and A. J. Hussain, "A survey of artificial intelligence approaches in blind source separation," *Neurocomputing*, 2023.

[25] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "Convolutive blind source separation methods," in *Springer Handbook of Speech Processing*. Berlin, Germany: Springer, 2008, pp. 1065–1094.

[26] J. Zi, D. Lv, J. Liu, X. Huang, W. Yao, M. Gao, R. Xi, and Y. Zhang, "Improved swarm intelligent blind source separation based on signal cross-correlation," *Sensors*, vol. 22, no. 1, p. 118, Dec. 2021.

[27] R. Wang, "Blind source separation based on adaptive artificial bee colony optimization and kurtosis," *Circuits, Syst., Signal Process.*, vol. 40, no. 7, pp. 3338–3354, Jul. 2021.

[28] H. M. Salman, A. K. M. Al-Qurabat, and A. A. R. Finjan, "Bigradient neural network-based quantum particle swarm optimization for blind source separation," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 10, no. 2, p. 355, Jun. 2021.

[29] M. Zhao, X. Yao, J. Wang, Y. Yan, X. Gao, and Y. Fan, "Single-channel blind source separation of spatial aliasing signal based on stacked-LSTM," *Sensors*, vol. 21, no. 14, p. 4844, Jul. 2021.

[30] J. A. Chambers, M. G. Jafari, and S. McLaughlin, "Variable step-size EASI algorithm for sequential blind source separation," *Electron. Lett.*, vol. 40, no. 6, pp. 393–394, 2004.

[31] N. Hassan and D. A. Ramli, "A comparative study of blind source separation for bioacoustics sounds based on FastICA, PCA and NMF," *Proc. Comput. Sci.*, vol. 126, pp. 363–372, Jan. 2018.

[32] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 11, pp. 1–12, Dec. 2003.

[33] J.-F. Cardoso, "Eigen-structure of the fourth-order cumulant tensor with application to the blind source separation problem," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1990, pp. 2655–2658.

[34] K. Matsuoka, M. Ohoya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Netw.*, vol. 8, no. 3, pp. 411–419, 1995.

[35] K. E. Hild, D. Erdogmus, and J. Principe, "Blind source separation using Renyi's mutual information," *IEEE Signal Process. Lett.*, vol. 8, no. 6, pp. 174–176, Jun. 2001.

[36] B. Ma and T. Zhang, "Single-channel blind source separation for vibration signals based on TVF-EMD and improved SCA," *IET Signal Process.*, vol. 14, no. 4, pp. 259–268, Jun. 2020.

[37] C. Jutten and J. Herault, "Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, no. 1, pp. 1–10, Jul. 1991.

[38] S. Haykin, Ed., *Unsupervised Adaptive Filtering, Blind Deconvolution*, vol. 2. Hoboken, NJ, USA: Wiley, Apr. 2001.

[39] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, nos. 4–5, pp. 411–430, Jun. 2000.

[40] A. Cichocki and S.-I. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. Hoboken, NJ, USA: Wiley, 2002.

[41] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*, vol. 615. Berlin, Germany: Springer, 2007.

[42] A. Hyvärinen, "Independent component analysis by minimization of mutual information," Dept. Comput. Sci. Eng., Lab. Comput. Inf. Sci., Helsinki Univ. Technol., Helsinki, Finland, Tech. Rep. A46, Aug. 1997.

[43] Z. Ding and T. Nguyen, "Stationary points of a kurtosis maximization algorithm for blind signal separation and antenna beamforming," *IEEE Trans. Signal Process.*, vol. 48, no. 6, pp. 1587–1596, Jun. 2000.

[44] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, Nov. 1995.

[45] L. B. Almeida, "MISEP—Linear and nonlinear ICA based on mutual information," *J. Mach. Learn. Res.*, vol. 4, pp. 1297–1318, Dec. 2003.

[46] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.

[47] V. Capdevielle, C. Serviere, and J. L. Lacoume, "Blind separation of wide-band sources in the frequency domain," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 2080–2083.

[48] J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Process. Lett.*, vol. 4, no. 4, pp. 112–114, Apr. 1997.

[49] S. N. Jain and C. Rai, "Blind source separation and ICA techniques: A review," *Int. J. Eng. Sci. Technol.*, vol. 4, no. 4, pp. 1490–1503, 2012.

[50] R. Gribonval and S. Lesage, "A survey of sparse component analysis for blind source separation: Principles, perspectives, and new challenges," in *Proc. 14th Eur. Symp. Artif. Neural Netw.*, 2006, pp. 323–330.

[51] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2610–2625, 2020.

[52] G. Zhou, Z. Yang, S. Xie, and J.-M. Yang, "Online blind source separation using incremental nonnegative matrix factorization with volume constraint," *IEEE Trans. Neural Netw.*, vol. 22, no. 4, pp. 550–560, Apr. 2011.

[53] N. Ito, R. Ikeshita, H. Sawada, and T. Nakatani, "A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel Wiener filter," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1950–1965, 2021.

[54] D. G. Fantinato, L. T. Duarte, Y. Deville, R. Attux, C. Jutten, and A. Neves, "A second-order statistics method for blind source separation in post-nonlinear mixtures," *Signal Process.*, vol. 155, pp. 63–72, Feb. 2019.

[55] S. M. Abdulla and J. Jayakumari, "Improving time-frequency sparsity for enhanced audio source separation in degenerate unmixing estimation technique algorithm," *J. Control Decis.*, vol. 9, no. 4, pp. 502–515, Oct. 2022.

[56] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[57] G. C. Marques, "Separation of nonlinear mixtures using pattern repulsion," in *Proc. Int. Workshop Independ. Compon. Anal. Blind Separat. Signals*, 1999, pp. 277–282.

[58] A. Hyvärinen and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *Neural Netw.*, vol. 12, no. 3, pp. 429–439, Apr. 1999.

[59] C. Li, Y. Jiang, F. Liu, and Y. Xiang, "Blind source separation algorithm based on improved particle swarm optimization under noisy condition," in *Proc. 2nd IEEE Adv. Inf. Manag., Communicates, Electronic Autom. Control Conf. (IMCEC)*, May 2018, pp. 398–401.

[60] A. Khalfa, N. Amardjia, E. Kenane, D. Chikouche, and A. Attia, "Blind audio source separation based on high exploration particle swarm optimization," *KSII Trans. Internet Inf. Syst.*, vol. 13, no. 5, pp. 2574–2587, May 2019, doi: 10.3837/tiis.2019.05.019.

[61] M. Kumar and V. E. Jayanthi, "Blind source separation using kurtosis, negentropy and maximum likelihood functions," *Int. J. Speech Technol.*, vol. 23, no. 1, pp. 13–21, Mar. 2020.

[62] S. Liu, B. Wang, and L. Zhang, "Blind source separation method based on neural network with bias term and maximum likelihood estimation criterion," *Sensors*, vol. 21, no. 3, p. 973, Feb. 2021.

[63] K. W. E. Lin, B. T. Balamurali, E. Koh, S. Lui, and D. Herremans, "Singing voice separation using a deep convolutional neural network trained by ideal binary mask and cross entropy," *Neural Comput. Appl.*, vol. 32, no. 4, pp. 1037–1050, Feb. 2020.
[64] C. Laugs, H. V. Koops, D. Odijk, H. Kaya, and A. Volk, "The influence of blind source separation on mixed audio speech and music emotion recognition," in *Proc. Companion Publication Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 67–71.
[65] A. Alghamdi, G. Healy, and H. Abdelhafez, "Real time blind audio source separation based on machine learning algorithms," in *Proc. 2nd Novel Intell. Lead. Emerg. Sci. Conf. (NILES)*, Oct. 2020, pp. 35–40.
[66] R. J. Issa and Y. F. Al-Irhaym, "Audio source separation using supervised deep neural network," *J. Phys., Conf.*, vol. 1879, no. 2, May 2021, Art. no. 022077.
[67] J. Agrawal, M. Gupta, and H. Garg, "Blind source separation in perspective of ICA algorithms: A review," in *Proc. Int. Conf. Comput. Intell. Sustain. Eng. Solutions (CISES)*, May 2022, pp. 78–85.
[68] D. Li, M. Wu, L. Yu, J. Han, and H. Zhang, "Single-channel blind source separation of underwater acoustic signals using improved NMF and FastICA," *Frontiers Mar. Sci.*, vol. 9, Jan. 2023, Art. no. 1097003.
[69] H. Abouzid and O. Chakkor, "Blind source separation approach for audio signals based on support vector machine classification," in *Proc. 2nd Int. Conf. Comput. Wireless Commun. Syst.*, Nov. 2017, pp. 1–6.
[70] P. R. Kunekar, K. S. Sadavarte, P. R. Khambad, R. B. Lokhande, and M. B. Kharat, "Audio feature extraction: Foreground and background audio separation using KNN algorithm," *Int. J. Sci. Res. Arch.*, vol. 9, no. 1, pp. 269–276, May 2023.
[71] A. C. Stasis, E. N. Loukis, S. A. Pavlopoulos, and D. Koutsouris, "Using decision tree algorithms as a basis for a heart sound diagnosis decision support system," in *Proc. 4th Int. IEEE EMBS Special Topic Conf. Inf. Technol. Appl. Biomed.,.*, 2003, pp. 354–357.
[72] C. Riday, S. Bhargava, R. H. R. Hahnloser, and S.-C. Liu, "Monaural source separation using a random forest classifier," in *Proc. Interspeech*, Sep. 2016, pp. 3344–3348.
[73] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1652–1664, Sep. 2016.
[74] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *Proc. Int. Conf. Latent Variable Anal. Signal Separat.* Cham, Switzerland: Springer, 2017, pp. 258–266.
[75] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. 16th Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sep. 2018, pp. 106–110.
[76] L. Li, H. Kameoka, and S. Makino, "Determined audio source separation with multichannel star generative adversarial network," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2020, pp. 1–6.
[77] Y. Xie, K. Xie, and S. Xie, "Underdetermined blind source separation of speech mixtures unifying dictionary learning and sparse representation," *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 12, pp. 3573–3583, Dec. 2021.
[78] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 61–65.
[79] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[80] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[81] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2016, pp. 31–35.
[82] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 246–250.
[83] L. Drude, T. von Neumann, and R. Haeb-Umbach, "Deep attractor networks for speaker re-identification and blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 11–15.
[84] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 696–700.
[85] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 46–50.
[86] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," 2020, *arXiv:2007.13975*.
[87] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 21–25.
[88] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Speech Synth. Workshop*, Pittsburgh, PA, USA, 2004, pp. 223–224.
[89] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[90] G. Cheng, L. Liao, H. Chen, and J. Lu, "Semi-blind source separation for nonlinear acoustic echo cancellation," *IEEE Signal Process. Lett.*, vol. 28, pp. 474–478, 2021.
[91] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4214–4217.

SAM ANSARI received the B.Sc. degree in telecommunication engineering from Canadian University Dubai, Dubai, United Arab Emirates, and the M.Sc. degree in electrical and computer engineering from Abu Dhabi University, Abu Dhabi, United Arab Emirates. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Sharjah, Sharjah, United Arab Emirates. His research interests include wireless communications, signal and image processing, molecular communication, artificial intelligence, machine learning, and renewable energy.

KHAWLA A. ALNAJJAR (Member, IEEE) received the B.S. degree in electrical engineering, communication track, from United Arab Emirates University (UAEU), Al-Ain, in 2008, the M.S. and P.E.E. degrees in electrical engineering from Columbia University, New York, in 2010 and 2012, respectively, and the Ph.D. degree in electrical and electronics engineering from the University of Canterbury, Christchurch, New Zealand, in 2015. She is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of Sharjah, United Arab Emirates. Her research interests include wireless communication systems, mathematical statistics, network information theory, and power grids. She has received more than 30 competitive awards for the successful studies and research during these ten years.

TAREK KHATER was born in TX, USA. He received the B.S. degree in biomedical engineering from Minia University, Egypt, in 2021. He is currently pursing the M.Sc. degree in biomedical engineering with the University of Sharjah with a thesis titled "Explainable artificial intelligence in the classification of breast cancer in young women." He is a Research and Teaching Assistant with the University of Sharjah. He has a number of publications in the research area of machine learning, explainable artificial intelligence, and their applications in biomedical engineering.

SOLIMAN MAHMOUD (Senior Member, IEEE) was born in Cairo, Egypt, in 1971. He received the B.Sc. (Hons.), M.Sc., and Ph.D. degrees from the Department of Electronics and Communications, Cairo University, Egypt, in 1994, 1996, and 1999, respectively. He is currently a Professor with the Department of Electrical Engineering, University of Sharjah, Sharjah, United Arab Emirates. He is a Distinguished Academic Leader in the field of electrical engineering, will be on leave from the University of Sharjah, and will assume the position of the Vice Chancellor for academic affairs with the University of Khorfakkan, Khor Fakkan, United Arab Emirates, effective from September 1, 2023. He supervised three Ph.D. students, 15 M.Sc. students, and more than 50 senior design projects. He is actively engaged in scholarly research work and has authored or coauthored more than 170 journals and conference publications since joining academia in 1996. He received ''The German-Egyptian Research Fund'' Grant, which has been used to finance the project ''Design of CMOS Field Programmable Analog Array and Its Applications.'' The project has been carried out in collaboration with the Ulm Microelectronics Institute, Ulm University, Germany. His articles received more than 2100 citations, and his Google Scholar H-index of 25 (his H-index from Scopus is 22). He has published six refereed research books. His research interests include mixed analog/digital integrated electronic circuit (IC) design, including mixed mode (voltage/current) analog circuits IC design, mixed (analog/digital) programmable CMOS electronics systems, biomedical circuits, field programmable analog arrays (FPAAs), and multi-standard wireless receiver design. In 2005, he received the Science Prize in Advanced Engineering Technology from the Academy of Scientific Research and Technology, Higher Ministry of Education, Cairo. He received the Distinguished Research Award from the University of Sharjah, from 2011 to 2012 and from 2014 to 2015.

ABIR HUSSAIN (Senior Member, IEEE) received the Ph.D. degree from The University of Manchester (UMIST), U.K., in 2000, with a thesis titled ''Polynomial neural networks for image and signal processing.'' She is currently a Professor of image and signal processing with the Electrical Engineering Department at the University of Sharjah, Sharjah, United Arab Emirates. She is also a Visiting Professor of machine learning with Liverpool John Moores University, U.K. She was involved with higher order and recurrent neural networks and their applications to e-health and medical image compression techniques. She has developed with her research students a number of recurrent neural network architectures. She is a Ph.D. supervisor and an external examiner for research degrees, including Ph.D. and M.Phil. students. She is one of the initiators and chairs of the development of e-Systems Engineering (DeSE) series. Her research interests include neural networks, signal prediction, telecommunication fraud detection, and image compression. She has published numerous refereed research papers in conferences and journals in the research areas.