Arabian Journal for Science and Engineering (2024) 49:12679–12691
https://doi.org/10.1007/s13369-024-08785-1

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

U-NET: A Supervised Approach for Monaural Source Separation

Samiul Basir[1] · Md. Nahid Hossain[1] · Md. Shakhawat Hosen[1] · Md. Sadek Ali[2,3] · Zainab Riaz[3] · Md. Shohidul Islam[1,3]

Received: 26 August 2023 / Accepted: 23 January 2024 / Published online: 26 February 2024

Abstract

Separating speech is a challenging area of research, especially when trying to separate the desired source from its combination. Deep learning has arisen as a promising solution, surpassing traditional methods. While prior research has mainly focused on the magnitude, log-magnitude, or a combination of the magnitude and phase portions, a new approach using the Short-time Fourier Transform (STFT), and a deep Convolutional Neural Network named U-NET has been proposed. This method, unlike others, considers both the real and imaginary components for decomposition. During the training stage, the mixed time-domain signal undergoes a transformation into a frequency-domain signal by using STFT, producing a mixed complex spectrogram. The spectrogram's real and imaginary parts are then divided and combined into a single matrix. The newly formed matrix is fed through U-NET to extract the source components. The same process is repeated at testing. The resulting concatenated matrix for the mixed test signal is passed through the saved model to generate two enhanced concatenated matrices for each source. These matrices are then transformed back into time-domain signals using inverse STFT by extracting the magnitude and phase. The proposed approach has been evaluated using the GRID audio visual corpuses, with results showing improved quality and intelligibility compared to the existing methods, as demonstrated by objective measurement metrics.

Keywords Speech separation · Source separation · Short-time Fourier transform (STFT) · U-NET

List of Symbols

p, P (small & capital)	Variable
p (small bold)	Vector
P (capital bold)	Matrix
P (capital bold italic)	Function
τ	Time frame index
f	Frequency bi index
R	It determines the real part of the complex matrix
I	It determines the imaginary part of the complex matrix
AI	Artificial Intelligence
CASSM	CASSM-based SS method
CDAE	CDAE-based SS method
CNN	Convolutional neural network
Conv-TasNet	Conv-TasNet-based SS method
DNN	Deep neural network
DSP	Digital signal processing
fwsegSNR	Average frequency weighted segmental SNR
GPU	Graphics processing unit
HASPI	Hearing aid's speech perception index
HASQI	Hearing aid's speech quality index
ISTFT	Inverse short-time Fourier transform

✉ Md. Shohidul Islam
shohid7@cse.iu.ac.bd

Samiul Basir
samiulbasir.cse@yahoo.com

Md. Nahid Hossain
nahid.cse.iu24@gmail.com

Md. Shakhawat Hosen
shakhawat.cse@yahoo.com

Md. Sadek Ali
sadek@ice.iu.ac.bd

Zainab Riaz
zriaz@hkcoche.org

[1] Department of Computer Science and Engineering, Islamic University, Kushtia 7003, Bangladesh

[2] Department of Information and Communication Technology, Islamic University, Kushtia 7003, Bangladesh

[3] Hong Kong Centre for Cerebro-Cardiovascular Health Engineering (COCHE), Hong Kong, China

LSTM	Long short-term memory
MP	Multilayer perceptron
MSE	Mean squared error
NMF	Non-negative matrix factorization
NMF-DNN	NMF-DNN-based SS method
PESQ	Perceptual evaluation of speech quality
RELU	Rectified linear units
RNN	Recurrent neural network
SCSS	Single-channel source separation
SDR	Source to distortion ratio
SE	Speech enhancement
SNR	Signal-to-noise ratio
SS	Speech separation, source separation
STFT	Short-time Fourier transform
STOI	Short-time objective intelligibility
VAT-SNet	VAT-SNet-based SS method
ULSTM	ULSTM-based SS method

1 Introduction

In recent years, the field of audio signal processing has witnessed significant advancements, particularly in the area of source separation (SS). Single-channel source separation (SCSS), also known as monaural source separation, refers to the process of separating individual sound sources from a mixed audio signal, typically captured by a single microphone or channel. It has become a highly desirable and challenging task in various applications, such as music production, speech enhancement (SE), audio transcription, and immersive audio systems [1–4]. Numerous potential benefits exist for the segregation of mixed speech. In contemporary speech processing, the role of SS is becoming increasingly crucial, demanding a growing number of devices to effectively perform this task.

While humans can effortlessly separate speech, constructing an automated system that emulates the human auditory system proves to be exceptionally challenging. Consequently, the pursuit of developing effective automatic SS systems has consistently been a significant focus in speech processing research. Conventionally, SS methods relied on the assumption of having multiple microphones or channels to exploit spatial information. However, in many real-world scenarios, such as live concert recordings, teleconferencing, or historical audio restoration, the availability of multiple channels is limited or nonexistent. This limitation prompted the development of SCSS [5–7] techniques that aim to recover individual sound sources from a monoaural mixture.

Due to the increasing fascination with SS, several conventional SCSS models have been suggested, taking into account various factors such as phase, magnitude, frequency, energy, and the spectrogram of the speech signal. A notable success in separating individual speakers has been achieved through the use of factorial hidden Markov models (HMMs) [8]. Moreover, researchers are increasingly utilizing nonnegative matrix factorization (NMF), a collection of methods in multivariate analysis that involves decomposing a matrix into two other nonnegative matrices based on their components and weights to separate source signals in SCSS [9].

However, these conventional approaches often face limitations when dealing with complex acoustic environments, overlapping sources, and nonstationary signals. To overcome these challenges, researchers have turned to deep learning (DL) algorithms [10, 11] and architectures to develop data-driven approaches for SS and achieving unprecedented performance improvements. SCSS focuses on learning a mapping function that estimates individual source signals from mixed audio inputs using a training dataset consisting of paired mixtures and their corresponding source signals [12].

In the context of audio SS, the Short-Time Fourier Transform (STFT) [13] is widely used to analyze and manipulate the audio signals. STFT represents a signal in the time-frequency domain, decomposing it into a series of spectral components. Each component is characterized by its magnitude and phase information, which provide valuable cues for separating the sources. In traditional as well as many DL approaches, the magnitude spectrogram has received significant attention and has been the main focal point for SS. However, phase information has also been recognized as an important factor in performance.

In this study, we propose an approach for SCSS using U-NET that considers both the real and imaginary parts of the complex spectrum generated by the STFT. As a result, the phase component should also be noteworthy in terms of its magnitude. Our method aims to leverage the benefits of DL and exploit the additional information contained in the complex spectrum to enhance separation performance. We have designed a modified U-NET architecture that can effectively handle the complex input features and learn to extract individual sources from the mixed audio signal.

The rest of the article is organized as follows: Sect. 2 provides a comprehensive overview of relative research in this domain, focusing on the evolution of deep learning techniques. Section 3 presents the U-NET architecture in detail, elucidating its key components and highlighting the reasons behind its suitability for audio SS. Section 4 presents the proposed methodologies, describing the architectural choices, proposed algorithm, training, and evaluation procedures. Section 5 showcases the outcomes of the experiments conducted and the subsequent analysis by encompassing both the dataset employed in this study as well as the evaluation metrics utilized to gauge the performance. Finally, Sect. 6

concludes the article by summarizing the key findings and outlining future research directions.

2 Relative Research

For supervised SS, there are two different categories of learning models: (1) methods that are traditional, like processes based on models and voice improvement techniques; and (2) innovative methods based on DNN. As a consequence of the speech production process, the input characteristics and desired outcomes of the SS process display an apparent spatiotemporal structure. Deep models are ideal for modeling due to these characteristics.

In speech separation, numerous deep models are actively deployed. Sun et al. [14] devised a two-stage method employing two DNN-based algorithms to tackle the difficulty of current speech separation systems' performance. The authors of [15] created new training aims in addition to current magnitude training objectives, utilizing neural network approaches to adjust for target phase in order to attain higher separation performance.

In order to understand the temporal characteristics of geographic data, Zhou et al. [16] developed a separation system based on RNN with LSTM. The statistical properties of noise are not constrained in supervised speech separation, and it is not essential to know the spatial orientation of the sound sources. It offers certain benefits and a bright future for study when used in monaural, nonstationary, or in cases of poor SNR [17, 18].

The Deep Recurrent Neural Network (DRNN) is a deep learning model frequently used in speech separation. It excels in using Markov models to identify the hidden states of RNN units like LSTM [19] and GRU (Gated Recurrent Unit, GRU) [20] in SS. Some past information will still be preserved from the previous concealed state; however, the magnitude spectra of mixed speech have a prolonged duration, causing loss in sequence analysis, impacting both the separation of mixed speech and the accuracy of speech prediction.

CNN has been commonly used in DL since Lecun et al. [21] first presented it in 1998. CNN clearly has advantages in 2-D signal processing, and applications like picture recognition have shown off its impressive modeling abilities. CNN is currently being used for SS and has outperformed speech separation systems based on DNN in terms of removal efficiency under identical circumstances.

[22] introduces a method for SCSS using deep, fully convolutional denoising autoencoders (CDAEs). Trained to extract specific sources from mixtures, CDAEs performs well deep feedforward neural networks in SS. They learn unique spectral–temporal patterns for successfully isolating the sources in mixed signals. Additionally, it explores the use of spectral masks to scale the mixed signal based on each source's contribution, ensuring an accurate estimation of the mixed signal's sources.

To address the problem of time-frequency masking, Luo et al. [23] developed Conv-TasNet, a network for SS in the time domain that utilizes fully convolutional techniques. Its impressive modeling abilities have been shown in applications like picture recognition. To mitigate the disparity in accuracy measures such as hit rate, error rate, and classification accuracy, Wang et al. [24] modified the loss function of CNN.

[25] suggests a system that addresses challenges like oversmoothing and incomplete separation in SCSS by integrating time-frequency non-negative matrix factorization (TFNMF) and deep neural networks with sigmoid-based normalization (SNDNN). TFNMF is utilized for feature extraction, and the resulting classified features are transformed into softmax.

The paper [26] introduces VAT-SNet, a time-domain music separation model that directly utilizes music waveform data as input. VAT-SNet enhances the network structure of Conv-TasNet by preserving deep acoustic features through sample-level convolution in both the encoder and decoder. Additionally, it incorporates vocal and accompaniment embeddings from an auxiliary network to enhance the purity of the separation, aligning with the principles of independent component analysis (ICA) and providing a mathematical model for the separation process.

UFLSTM [27] is a deep learning model for speech enhancement (SE). UFLSTM utilizes adaptive power law transformation to redistribute energy, maintaining constant total energy in speech signals for improved intelligibility and quality, incorporating residual connections to prevent gradient decay, and adjusting the forget gate using an attention process.

Although conventional and separation models based on DNN have shown impressive results, they all have a few flaws. Using CNN each element may absorb local features without learning global characteristics in order to benefit from the spatial connectivity of the input data, and in the process of feature extraction, localized features are initially identified and then combined to create more comprehensive features at higher levels. Using weight sharing can improve the speed of the model by reducing the number of parameters that need to be computed for each neuron.

Various feature maps that can recognize the same type of feature in various locations and partly assure the invariance of displacement and distortion may be produced by combining a number of convolution filters. As a result, this study provides a CNN-based approach to alleviate the issue of mixed-language speakers' loss of extended sequence information. Our model may boost the speech separation impact by concentrating on the timing sequence stage, which offers the highest contribution, and by partially solving the difficulty of the temporal model's short memory.

3 U-NET Architecture

In order to extract the features of the desired source from the mixed coefficients, we employed the U-NET architecture. Figure 1 presents a pictorial representation of the network structure, comprising two main components: a contracting path on the left side and an expansive path to the right. The contracting path adheres to the typical architecture of a convolutional network. The structure involves the iterative utilization of two sets of 3×3 convolutions. Subsequently, a Leaky rectified linear unit (LeakyReLU) and a 2×2 max pooling operation with a stride of 2 are applied for downsampling. During each downsampling stage, the quantity of feature channels is increased twofold.

Each iteration in the extensive trajectory involves enlarging the feature map through upsampling, followed by a 2×2 convolution that reduces the number of feature channels by half. This is followed by combining the enlarged feature map with the corresponding cropped feature map from the contracting trajectory. Cropping is essential to address the removal of border pixel elements during convolutions at each step. At the last layer, a 1×1 convolution is employed to transform each 16-component feature vector into the specified number of classes. Altogether, the network comprises 24 convolutional layers. To ensure the output segmentation map can be seamlessly tiled, it is crucial to choose the input tile size in such a way that all 2×2 max-pooling operations are performed on a layer with both x- and y-dimensions being even.

Huber loss is a robust alternative to mean squared error (MSE) loss, which is commonly affected by outliers and sensitivity issues. By balancing between quadratic loss for small errors and linear loss for larger errors, Huber loss effectively addresses these challenges and improves model performance. Huber loss combines squared loss for minor errors and absolute loss for significant errors. By incorporating a parameter called delta (δ), the loss function determines the threshold at which the transition occurs from quadratic to linear. When errors are smaller than δ, the loss function resembles MSE, while for errors exceeding δ, it behaves similarly to MAE. Mathematically, this loss function is represented as per Eq. (1), where y denotes the actual or desired value, y' signifies

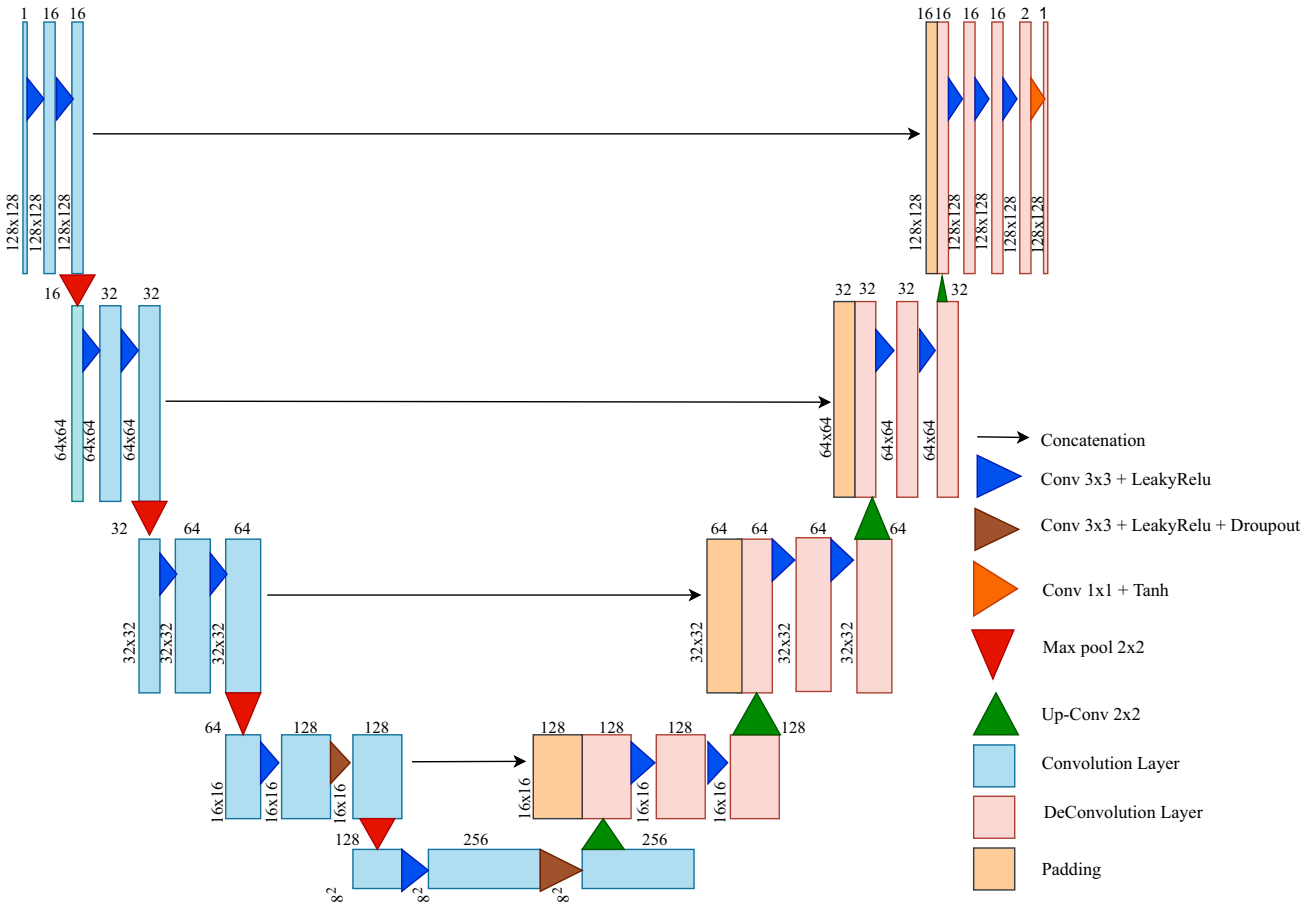

Fig. 1 U-NET Architecture

the predicted value, and δ represents the threshold parameter.

$$L(y, y') = \begin{cases} \frac{1}{2}(y - y')^2, & if|y - y'| \leq \delta \\ \delta|y - y'| - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases} \quad (1)$$

The networks' parameters were randomly initialized, amounting to a total of 1,941,093. They underwent training using backpropagation and the Adam optimizer with a learning rate of 0.001, employing the default settings for all other parameters.

4 Proposed Method

This section outlines the proposed SCSS technique and provides details about the substances it utilizes. In the context of audio or time-series data, the signal is represented by the STFT as a complex matrix, where each element corresponds to a specific frequency and time bin index. The real part signifies the magnitude or intensity of the frequency component, while the imaginary part encodes phase information. Unlike most SS systems that focus solely on the magnitude of the STFT, neglecting the phase component, this article combines STFT with U-NET, a deep CNN, taking both the real and imaginary components into consideration. The utilization of both components during U-NET training enables the model to effectively capture complex-valued frequency information in the input data.

It is important to note that no approach is universally superior, and trade-offs exist. The associated trade-offs were that the utilization of U-NET for SS introduced computational complexities, and its performance was contingent upon the quantity and quality of available data. Notably, there were associated risks of overfitting, especially when confronted with limited data, potentially limiting the model's interpretability. Furthermore, the implementation of U-NET demanded substantial computational resources and prolonged training times. Achieving robust generalization across diverse acoustic environments posed a significant challenge. Therefore, a pivotal aspect in this methodology involved striking a balance between U-NET's model complexity and the specific requirements of the application.

However, the trade-offs of the proposed method stated earlier here include a breakdown of potential reasons for better performance. Unlike others, incorporating both the real and imaginary components together in the model yielded a comprehensive representation of the audio signal, capturing both the amplitude and phase details. This refined representation enhanced accuracy, especially in scenarios involving overlapping speech. Besides, preserving phase information was crucial for maintaining temporal attributes, leading to more natural and intelligible speech output. The end-to-end learning approach streamlined training, allowing the model to autonomously learn relevant features and promoting better generalization across speakers and acoustic environments. Furthermore, supervised learning with labeled data enhanced adaptability to diverse acoustic environments, increasing robustness in real-world scenarios. U-NET efficiency and hardware acceleration allowed real-time audio processing, crucial for low-latency applications like live streaming and interactive platforms. The proposed SS method has two stages, the training stage and the testing stage, which are depicted in Fig. 2.

Algorithm 1 Algorithm for the training and testing stages of the proposed method

Training Stage

Input: Training sets $\mathbf{m}(t)$ and $\mathbf{p}(t)$

Output: $\mathbf{b}_{\mathbf{M}_{RI}}$ and $\mathbf{W}_{\mathbf{M}_{RI}RI}$

Step 1: Compute the complex spectrogram $m1_{RI(\tau,f)}$, $m2_{RI(\tau,f)}$,...., $mn_{RI(\tau,f)}$ and $p1_{RI(\tau,f)}$, $p2_{RI(\tau,f)}$,...., $pn_{RI(\tau,f)}$ by applying *STFT*.

Step 2: Separate the real $m1_{R(\tau,f)}$, $m2_{R(\tau,f)}$,...., $mn_{R(\tau,f)}$ and $p1_{R(\tau,f)}$, $p2_{R(\tau,f)}$,...., $pn_{R(\tau,f)}$ and imaginary $m1_{I(\tau,f)}$, $m2_{I(\tau,f)}$,...., $mn_{I(\tau,f)}$ and $p1_{I(\tau,f)}$, $p2_{I(\tau,f)}$,...., $pn_{I(\tau,f)}$ components of the complex spectrogram.

Step 3: Concatenate both the real and imaginary portions of the complex spectrogram to make a single matrix $\mathbf{M}_{RI}^{\text{Train}}$ and $\mathbf{P}_{RI}^{\text{Train}}$ for both ends.

Step 4: Train the network with the generated concatenated matrices.

Step 5: Determine the bias $\mathbf{b}_{\mathbf{M}_{RI}}$ and weight matrices $\mathbf{W}_{\mathbf{M}_{RI}}$ and update the network.

Testing Stage

Input: Testing set $\mathbf{m}(t)$

Output: Estimated the male and female sources $\mathbf{p}'(t)$ and $\mathbf{q}'(t)$

Step 1: Compute the complex spectrogram $m1_{RI(\tau,f)}$, $m2_{RI(\tau,f)}$,...., $mn_{RI(\tau,f)}$ by applying *STFT*.

Step 2: Separate the real $m1_{R(\tau,f)}$, $m2_{R(\tau,f)}$,...., $mn_{R(\tau,f)}$ and imaginary $m1_{I(\tau,f)}$, $m2_{I(\tau,f)}$,...., $mn_{I(\tau,f)}$ components of the complex spectrogram.

Step 3: Concatenate both the real and imaginary portions of the complex spectrogram to make a single matrix $\mathbf{M}_{RI}^{\text{Test}}$.

Step 4: From the best training weights of the network, $\mathbf{M}_{RI}^{\text{Test}}$ generated the $\mathbf{P}_{RI}^{\text{EnhancedMatrix}}$ matrix for the first source.

Step 5: Subtract $\mathbf{P}_{RI}^{\text{EnhancedMatrix}}$ from $\mathbf{M}_{RI}^{\text{Test}}$ to generate the second $\mathbf{P}_{RI}^{\text{EnhancedMatrix}}$ matrix.

Step 6: Separate both the real and imaginary parts from the first and second sources.

Step 7: The re-complex matrices $\mathbf{P}^{\text{recmplx}}$ and $\mathbf{Q}^{\text{recmplx}}$ are generated by incorporating the real and imaginary parts with a complex number.

Step 8: The magnitude and phase parts are extracted from the re-complex series for both sources.

Step 9: Obtain the estimated sources $\mathbf{p}'(t)$ and $\mathbf{q}'(t)$ by applying *ISTFT*.

4.1 Training Stage

During the training phase, we think about a signal $\mathbf{m}(t)$ called the mixed, consisting of two different sources $\mathbf{p}(t)$ and $\mathbf{q}(t)$, respectively. $\mathbf{m}(t)$ is utilized here as an input signal, and $\mathbf{p}(t)$ is the corresponding label. The ***STFT*** processes both mixed

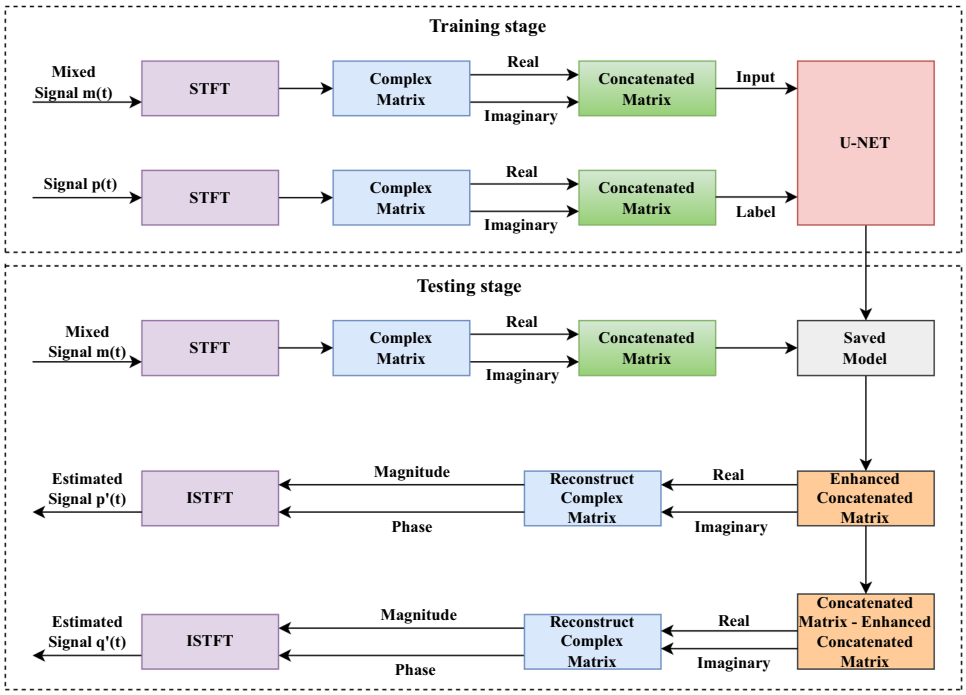

Fig. 2 Block diagram of the proposed SS approach

and labeled signals to calculate the complex spectrograms $\mathbf{M}_{(\tau,f)}$ and $\mathbf{P}_{(\tau,f)}$. These are denoted in Eqs. (2) and (3), with τ and f indicating the time and frequency bin indices, respectively.

$$\mathbf{M}_{(\tau,f)} = \mathbf{M}_{R(\tau,f)} + \mathbf{M}_{I(\tau,f)}i \tag{2}$$

$$\mathbf{P}_{(\tau,f)} = \mathbf{P}_{R(\tau,f)} + \mathbf{P}_{I(\tau,f)}i \tag{3}$$

The concatenated forms of the real and imaginary components for both $\mathbf{M}_{RI}^{\text{Train}}$ and $\mathbf{P}_{RI}^{\text{Train}}$ matrices are then forwarded into the U-NET model. The network model next decomposes the $\mathbf{M}_{RI}^{\text{Train}}$ matrix into its bias and weight matrices as per Eq. (4), where the terms $\mathbf{W}_{\mathbf{M}_{RI}}$ and $\mathbf{b}_{\mathbf{M}_{RI}}$ represent the weight and bias matrices corresponding to the mixed source, and g represents the nonlinear activation function.

$$\mathbf{M}_{RI}^{\text{Train}} \approx g(\mathbf{W}_{\mathbf{M}_{RI}} + \mathbf{b}_{\mathbf{M}_{RI}}) \tag{4}$$

Initially, the bias and weight metrics are assigned to zero and random values, respectively. The weighted matrix $\mathbf{W}_{\mathbf{M}_{RI}}$ and the bias metrics $\mathbf{b}_{\mathbf{M}_{RI}}$ were updated continuously by minimizing the cost between $\mathbf{M}_{RI}^{\text{Train}}$ and $\mathbf{P}_{RI}^{\text{Train}}$ using Eq. (5) with the help of Eqs. (6) and (7), where α is called learning rate. During training, the model was saved, and after completing the training, the best bias and weights were fixed.

$$\mathbf{M}_{RI}(\text{Error}) = \mathbf{M}_{RI}(\text{Label Output}) - \mathbf{M}_{RI}(\text{Predicted Output}) \tag{5}$$

$$\mathbf{W}_{\mathbf{M}_{RI}}(\text{New}) = \mathbf{W}_{\mathbf{M}_{RI}}(\text{Old}) - \alpha \frac{\partial \mathbf{M}_{RI}(\text{Error})}{\partial \mathbf{W}_{\mathbf{M}_{RI}}(\text{Old})} \tag{6}$$

$$\mathbf{b}_{\mathbf{M}_{RI}}(\text{New}) = \mathbf{b}_{\mathbf{M}_{RI}}(\text{Old}) - \alpha \frac{\partial \mathbf{M}_{RI}(\text{Error})}{\partial \mathbf{b}_{\mathbf{M}_{RI}}(\text{Old})} \tag{7}$$

4.2 Testing Stage

During the testing phase, the signal $\mathbf{m}(t)$ in Eq. (8), which is a combination or mixture of the signals $\mathbf{p}(t)$ and $\mathbf{q}(t)$, undergoes ***STFT*** to generate the complex spectrogram.

$$\mathbf{M}_{(\tau,f)} = \mathbf{M}_{R(\tau,f)} + \mathbf{M}_{I(\tau,f)}i \tag{8}$$

From the complex spectrogram of the mixed signal, the real and imaginary components were separated and concatenated to construct $\mathbf{M}_{RI}^{\text{Test}}$, which is passed through the U-NET saved model. The model then generated the enhanced concatenated matrices $\mathbf{P}_{RI}^{E}$ for the first source. To compute the enhanced concatenation matrix $\mathbf{Q}_{RI}^{E}$ for the second source, we subtract $\mathbf{P}_{RI}^{E}$ from $\mathbf{M}_{RI}^{\text{Test}}$ as per Eq. (9).

$$\mathbf{Q}_{RI}^{E} = \mathbf{M}_{RI}^{\text{Test}} - \mathbf{P}_{RI}^{E} \tag{9}$$

From the initial estimation of the first enhanced concatenated matrix $\mathbf{P}_{RI}^{E}$, the real and imaginary components were separated once again to reconstruct a complex matrix $\mathbf{P}^{\text{recmplx}}$ with the help of following Eq. (10).

$$\mathbf{P}^{\text{recmplx}} = \mathbf{P}_{R}^{E} + \mathbf{P}_{I}^{E}i \tag{10}$$

Similarly, the real and imaginary components were separated from the female enhanced concatenated matrix to

reconstruct another complex matrix $\mathbf{Q}^{\text{recmplx}}$ for the female source as per Eq. (11).

$$\mathbf{Q}^{\text{recmplx}} = \mathbf{Q}_R^E + \mathbf{Q}_I^E i \tag{11}$$

From the reconstructed complex matrix $\mathbf{P}^{\text{recmplx}}$, the magnitude and phase components $\mathbf{P}_{\text{Emag}}$ and $\mathbf{P}_{\text{Ephase}}$ were generated for the first source, respectively, with the aid of following Eq. (12).

$$\begin{aligned} \mathbf{P}_{\text{Emag}} &= \text{magnitude}(\mathbf{P}^{\text{recmplx}}) \\ \mathbf{P}_{\text{Ephase}} &= \text{phase}(\mathbf{P}^{\text{recmplx}}) \end{aligned} \tag{12}$$

The magnitude and phase components $\mathbf{Q}_{\text{Emag}}$ and $\mathbf{Q}_{\text{Ephase}}$ for the other source were extracted from the reconstruct complex matrix $\mathbf{Q}^{\text{recmplx}}$ as per Eq. (13).

$$\begin{aligned} \mathbf{Q}_{\text{Emag}} &= \text{magnitude}(\mathbf{Q}^{\text{recmplx}}) \\ \mathbf{Q}_{\text{Ephase}} &= \text{phase}(\mathbf{Q}^{\text{recmplx}}) \end{aligned} \tag{13}$$

As input for the first source, the newly generated enhanced magnitude and enhanced phase as per Eq. (12) were fed into the inverse STFT. The inverse STFT then transforms it into a time-domain signal, and we get the first estimated source as per Eq. (14). Similarly, the inverse STFT in Eq. (15), after getting the enhanced magnitude and enhanced phase as per Eq. (13), generated the second source as well.

$$\mathbf{p}'(t) = \boldsymbol{ISTFT}(\mathbf{P}_{\text{Emag}} \times \mathbf{P}_{\text{Ephase}}) \tag{14}$$

$$\mathbf{q}'(t) = \boldsymbol{ISTFT}(\mathbf{Q}_{\text{Emag}} \times \mathbf{Q}_{\text{Ephase}}) \tag{15}$$

5 Results and Discussion

This section offers experimental findings and discussions. Initially, a brief overview of the experiment's design and evaluation methods will be given, followed by a discussion of the metrics used to measure the results. Third, we examine how the join features compare to the single-domain techniques with regard to the SDR, SIR, fwsegSNR, STOI, and HASQI scores. Fourth, we compared the general effectiveness of our suggested approach to the CDAE, Conv-TasNet, CASSM, NMF-DNN, VAT-SNet, and ULSTM techniques in terms of PESQ, STOI, fwseqSNR, and SDR, SIR, and SAR. To the end, the time domain waveform and spectrogram of the clear, mixed, and segregated male and female sounds were provided.

5.1 Experimental Setup

To assess the efficiency of the suggested approach, we compare the proposed model with CDAE [22], Conv-TasNet [23], CASSM [24], NMF-DNN [25], VAT-SNet [26], and ULSTM [27]. In this system, we collect the signal speech from GRID audio visual corpuses [28], which were used for training as well as testing data. There are 1000 utterances spoken by thirty-four speakers (eighteen male and sixteen female). We concatenate sentences all together for each speaker. For the opposite gender speech separation, to form an experimental group, six male and six female speakers' utterances are exploited here. Each training signal lasts for about 25 min, and each test signal lasts for around 60 s. These signals are sampled at 8000 Hz. Like the speech-noise scenario, we consider female as noise and male as the speech signal. We mixed the female source with the male at -10, -5, 0, 5, and 10 dB.

5.2 Evaluation Metrics

The performances of the separated utterances are evaluated through the SDR [29], SIR [29], SAR [29], fwsegSNR [30], STOI [31], PESQ [32], HASPI [33], and HASQI [34] scores. The SDR value, which is a measure of overall speech quality, is calculated as the ratio of the strength of the input signal to the power of the difference between the input and reconstructed signals. Performance restoration is governed by higher SDR scores. Along with SDR, SIR also detects errors brought on by source separation process failures to eliminate the interfering signal. Better separation quality is indicated by a higher SIR value. Comparing the separated speech to comparable clean speech allows for the evaluation of PESQ, which results in scores between -0.5 and 4.5, with a greater number indicating better quality. A higher STOI value allows for more intelligibility. Short-time temporal wrappers, with a score ranging from 0 to 1, are correlated with clean and separated speech. The intelligibility of the collected signal was evaluated by fwsegSNR, and the greater the value, the better the performance. The HASQI and HASPI are instruments designed to measure how well hearing-impaired people and hearing-unimpaired people perceive sound. Higher scores, which range from 0 to 1, are related to greater sound quality and understandability.

5.3 The Impact of Single Over Join Features

The source signals are characterized by being brief, unchanging, and infrequent. The transformation of the signal into the time-frequency domain using STFT resulted in the generation of its complex spectra, which were used for speech separation techniques. There are certain methods that have been described that solely consider the magnitude part of a complex spectra, ignoring the real and imagined components. In this contrast, the real and imaginary portions are individually evaluated, even the magnitude section is evaluated separately, and the real and imaginary portions are evaluated

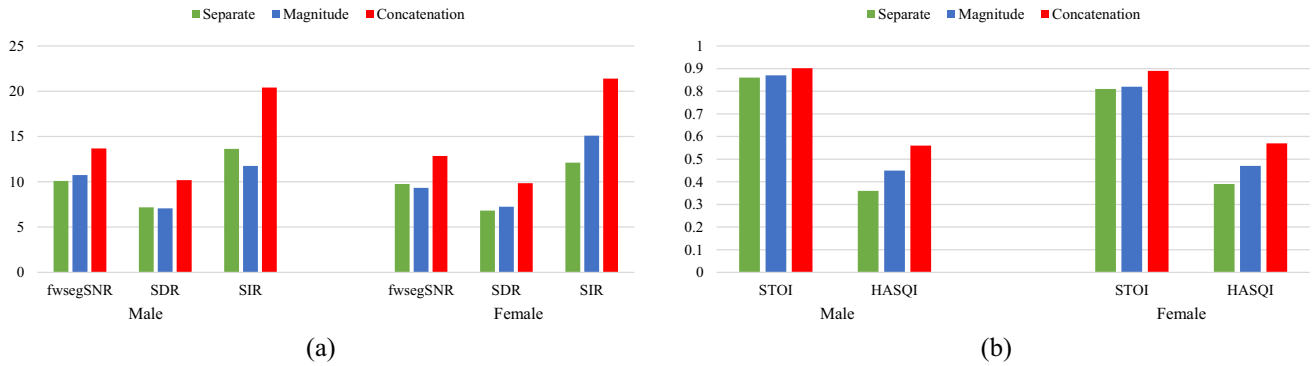

Fig. 3 Comparison of **a** SDR, SIR, fwsegSNR, **b** HASQI and STOI of single over join feature

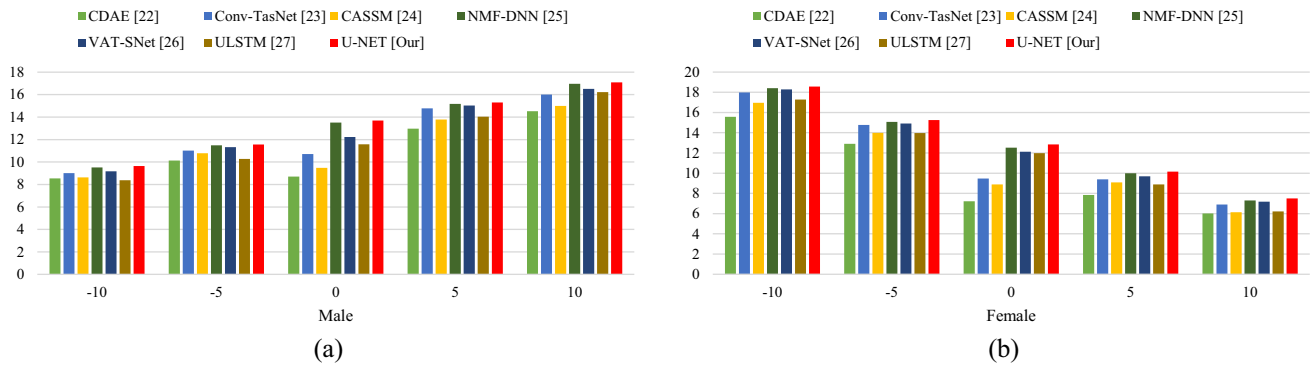

Fig. 4 Comparison of fwsegSNR for **a** male, **b** female source, respectively

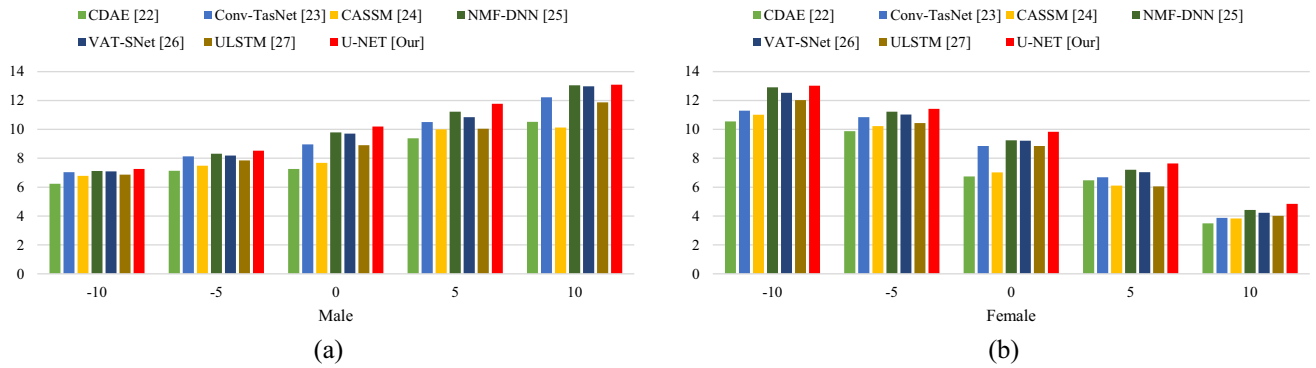

Fig. 5 Comparison of SDR for **a** male, **b** female source, respectively

jointly. SDR, SIR, fwsegSNR, HASQI, and STOI measurement techniques are compared in Fig. 3. As we can see from the figures that the method which uses the real and imaginary portions together outperforms than others. As a result, in the suggested technique, we examine the real and imaginary portions simultaneously, which improves a SCSS's quality and intangibility.

5.4 Overall Performance of the Proposed Algorithm

In Fig. 4, the fwsegSNR performance of the proposed model is compared with that of current models. Based on the following graphs, it appears that the suggested model outperforms the other current techniques in all circumstances. Our strategy boosts fwsegSNR by 9.65% for −10 SNR than the presented approaches, 11.56% for −5 SNR, 13.69% for 0 SNR, for 5 SNR 15.31% and 17.09% for 10 SNR to separate male sources. Similarly, our proposed approach gained 18.56%, 15.26%, 12.85%, 10.16%, 7.51% for −10 SNR, −5 SNR, 0 SNR, 5 SNR, and 10 SNR, respectively, for female source separation.

We demonstrated that in Fig. 5, the proposed model's SDR achieves much superior outcomes compared to the alternatives, notably CDAE, Conv-TasNet, CASSM, NMF-DNN,

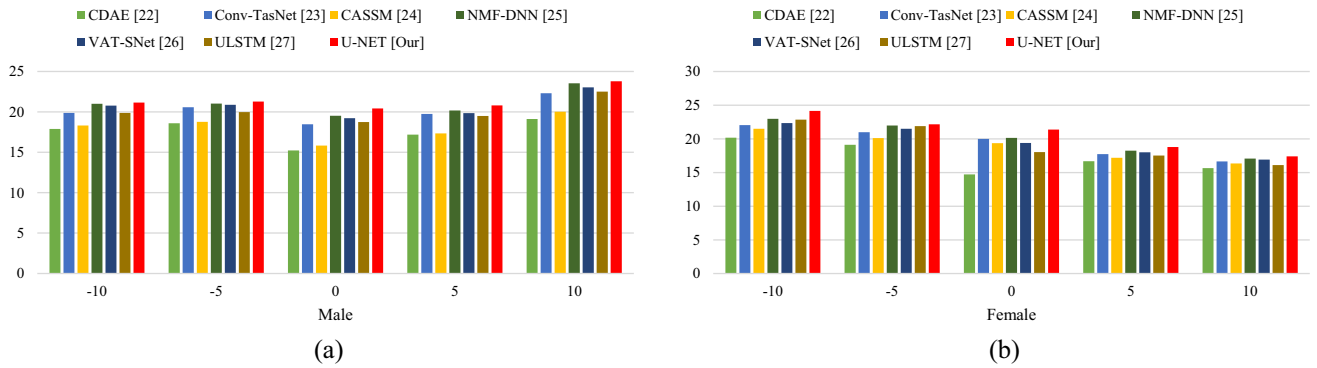

Fig. 6 Comparison of SIR for **a** male, **b** female source, respectively

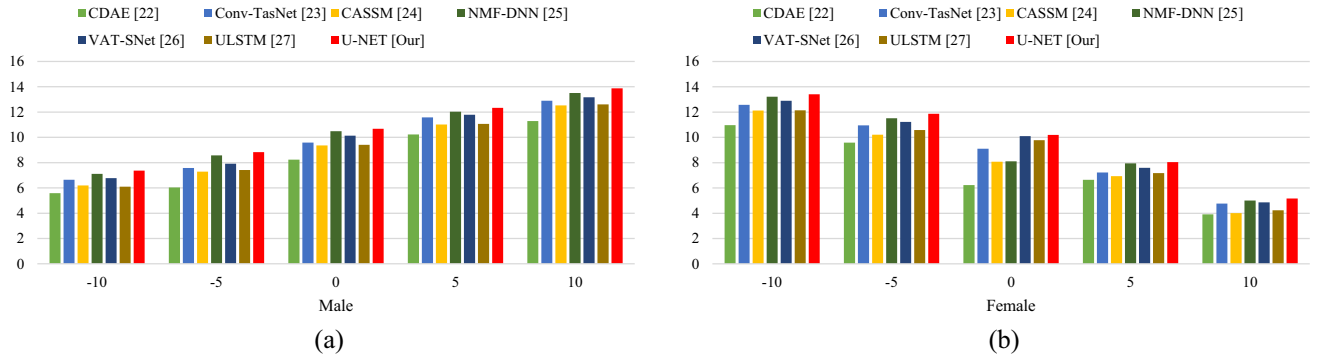

Fig. 7 Comparison of SAR for **a** male, **b** female source, respectively

VAT-SNet, and ULSTM for both male and female gender. The suggested model's SDR values are greater than the previous models in all circumstances of separation.

The suggested models increase SDR for 7.26 dB for − 10 SNR, 8.53 dB for − 5 SNR, 10.19 dB for 0 SNR, 11.78 dB for 5 SNR, and 13.10 dB for 10 SNR to separate the male sources. Accordingly, 13.02 dB, 11.43 dB, 9.84 dB, 7.63 dB, and 4.84 dB for -10 SNR, − 5 SNR, 0 SNR, 5 SNR, and 10 SNR, respectively, separate the female sources. Similarly in **Fig.** 6 SIR values for predicted signals get higher than the current models, as seen by this figure.

From **Fig.** 7, we examined that our proposed approach performed in a better manner in terms of source to artifacts ratio (SAR) for both of the male and female sources than the other methods stated in this article.

Tables 1, 2, 3, and 4 compare the suggested technique's performance in terms of PESQ and STOI to those of other current approaches. Our suggested technique improves PESQ scores 2.25 for − 10 dB, 2.40 for − 5 dB, 2.63 for 0 dB, 2.81 for 5 dB, and 2.98 for 10 dB for separating the male source, over the methods existing for comparisons. Likewise, a separate female source achieved 3.23, 2.98, 2.70, 2.35, and 1.97 for − 10 dB, − 5 dB, 0 dB, 5 dB, and 10 dB, respectively. Further, the table demonstrates that expected signals have a higher STOI performance than do models that are already in use.

Table 1 Comparison of PESQ scores for the male source with six different approaches

Methods	− 10	− 5	0	5	10
CDAE [22]	1.98	2.03	2.10	2.23	2.52
Conv-TasNet [23]	2.10	2.19	2.41	2.51	2.81
CASSM [24]	2.01	2.07	2.33	2.46	2.77
NMF-DNN [25]	2.15	2.31	2.56	2.73	2.93
VAT-SNet [26]	2.19	2.23	2.44	2.64	2.84
ULSTM [27]	2.11	2.13	2.37	2.41	2.57
U-NET [Our]	**2.25**	**2.40**	**2.63**	**2.81**	**2.98**

Bold indicate all the approaches, both existing and proposed values. The purpose of using bold text is to highlight and differentiate both the existing approaches and our proposed approach

Table 2 Comparison of PESQ scores for the female source with six different approaches

Methods	− 10	− 5	0	5	10
CDAE [22]	2.57	2.37	1.99	2.01	1.57
Conv-TasNet [23]	3.13	2.88	2.53	2.21	1.85
CASSM [24]	2.87	2.57	2.42	2.11	1.59
NMF-DNN [25]	3.17	2.89	2.63	2.29	1.91
VAT-SNet [26]	3.15	2.89	2.61	2.23	1.89
ULSTM [27]	3.02	2.58	2.57	2.19	1.88
U-NET [Our]	**3.23**	**2.98**	**2.70**	**2.35**	**1.97**

Table 3 Comparison of STOI scores for the male source with six different approaches

Methods	− 10	− 5	0	5	10
CDAE [22]	0.63	0.71	0.81	0.83	0.86
Conv-TasNet [23]	0.79	0.83	0.89	0.91	0.91
CASSM [24]	0.75	0.78	0.81	0.83	0.86
NMF-DNN [25]	0.79	0.84	0.89	0.92	0.94
VAT-SNet [26]	0.78	0.84	0.88	0.91	0.92
ULSTM [27]	0.76	0.83	0.87	0.88	0.90
U-NET [Our]	**0.81**	**0.86**	**0.90**	**0.93**	**0.95**

Bold indicate all the approaches, both existing and proposed values. The purpose of using bold text is to highlight and differentiate both the existing approaches and our proposed approach

Table 4 Comparison of STOI scores for the female source with six different approaches

Methods	− 10	− 5	0	5	10
CDAE [22]	0.89	0.88	0.83	0.76	0.68
Conv-TasNet [23]	0.92	0.92	0.88	0.80	0.70
CASSM [24]	0.89	0.89	0.84	0.76	0.69
NMF-DNN [25]	0.93	0.91	0.88	0.80	0.71
VAT-SNet [26]	0.92	0.92	0.87	0.79	0.70
ULSTM [27]	0.91	0.90	0.87	0.77	0.68
U-NET [Our]	**0.96**	**0.93**	**0.89**	**0.82**	**0.72**

Bold indicate all the approaches, both existing and proposed values. The purpose of using bold text is to highlight and differentiate both the existing approaches and our proposed approach

Table 5 Comparison of HASPI values for the male source with six different approaches

Methods	− 10	− 5	0	5	10
CDAE [22]	0.9985	0.9993	0.9995	0.9994	0.9995
Conv-TasNet [23]	0.9989	0.9995	0.9997	0.9998	0.9998
CASSM [24]	0.9986	0.9994	0.9995	0.9994	0.9995
NMF-DNN [25]	0.9990	0.9996	0.9998	0.9999	0.9999
VAT-SNet [26]	0.9989	0.9995	0.9997	0.9998	0.9999
ULSTM [27]	0.9988	0.9994	0.9995	0.9995	0.9996
U-NET [Our]	**0.9990**	**0.9996**	**0.9998**	**0.9999**	**0.9999**

Bold indicate all the approaches, both existing and proposed values. The purpose of using bold text is to highlight and differentiate both the existing approaches and our proposed approach

Table 6 Comparison of HASPI values for the female source with six different approaches

Methods	− 10	− 5	0	5	10
CDAE [22]	0.9994	0.9995	0.9993	0.9992	0.9989
Conv-TasNet [23]	0.9998	0.9997	0.9996	0.9994	0.9991
CASSM [24]	0.9995	0.9995	0.9994	0.9993	0.9990
NMF-DNN [25]	0.9999	0.9999	0.9997	0.9995	0.9992
VAT-SNet [26]	0.9998	0.9998	0.9997	0.9995	0.9992
ULSTM [27]	0.9997	0.9997	0.9995	0.9994	0.9989
U-NET [Our]	**0.9999**	**0.9999**	**0.9997**	**0.9995**	**0.9992**

Bold indicate all the approaches, both existing and proposed values. The purpose of using bold text is to highlight and differentiate both the existing approaches and our proposed approach

Table 7 Comparison of HASQI values for the male source with six different approaches

Methods	− 10	− 5	0	5	10
CDAE [22]	0.4873	0.5652	0.6744	0.7433	0.8014
Conv-TasNet [23]	0.4957	0.6047	0.6931	0.7587	0.8189
CASSM [24]	0.4889	0.5663	0.6788	0.7477	0.8083
NMF-DNN [25]	0.5003	0.6099	0.6978	0.7641	0.8259
VAT-SNet [26]	0.4983	0.6074	0.6953	0.7588	0.8193
ULSTM [27]	0.4909	0.6059	0.6917	0.7599	0.8197
U-NET [Our]	**0.5021**	**0.6123**	**0.7002**	**0.7688**	**0.8289**

Bold indicate all the approaches, both existing and proposed values. The purpose of using bold text is to highlight and differentiate both the existing approaches and our proposed approach

Table 8 Comparison of HASQI values for the female source with six different approaches

Methods	− 10	− 5	0	5	10
CDAE [22]	0.7339	0.7125	0.6439	0.6133	0.5146
Conv-TasNet [23]	0.7303	0.7111	0.6419	0.6259	0.5163
CASSM [24]	0.7375	0.7179	0.6498	0.6181	0.5177
NMF-DNN [25]	0.7323	0.7273	0.6531	0.6269	0.5207
VAT-SNet [26]	0.7309	0.71123	0.6441	0.6275	0.5174
ULSTM [27]	0.7307	0.7237	0.6509	0.6229	0.5191
U-NET [Our]	**0.7459**	**0.7256**	**0.6595**	**0.6312**	**0.5216**

Bold indicate all the approaches, both existing and proposed values. The purpose of using bold text is to highlight and differentiate both the existing approaches and our proposed approach

Tables 5, 6, 7, and 8 show the HASPI and HASQI findings of several approaches, including CDAE, Conv-TasNet, CASSM, NMF-DNN, VAT-SNet, and ULSTM for male and female speech separation. Tables 5 and 6 show that U-NET produces higher HASPI values in all scenarios of separation. It can also be noted that in Tables 7 and 8 the HASQI findings of our approach outperform the other three techniques in all circumstances of separation.

5.5 Time-Domain and Spectrogram Representation

Time-domain and spectrogram representation offers distinct approaches to visualize and analyze signals, especially within the realm of signal processing. The time-domain representation depicts the signal's temporal evolution, offering insights into amplitude and serving as a valuable tool for comprehending temporal patterns and identifying specific events. On the

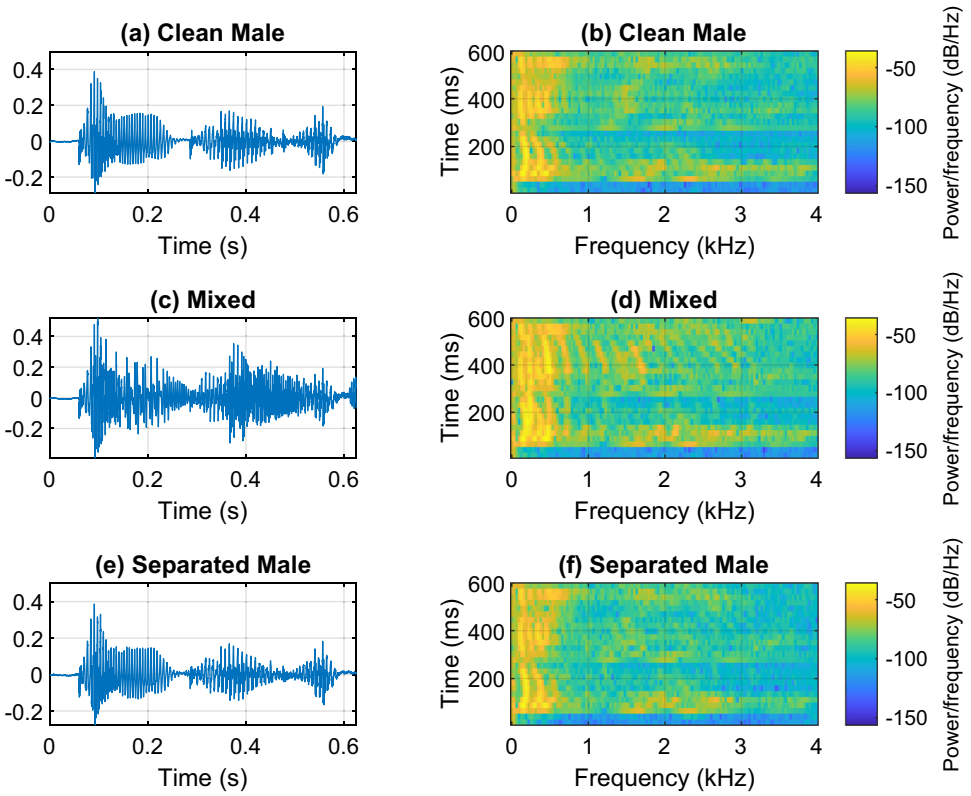

Fig. 8 **a** Waveform, **b** Spectrogram of Clean, **c** Waveform, **d** Spectrogram of Mixed and **e** Waveform, **f** Spectrogram of Separated male source, respectively

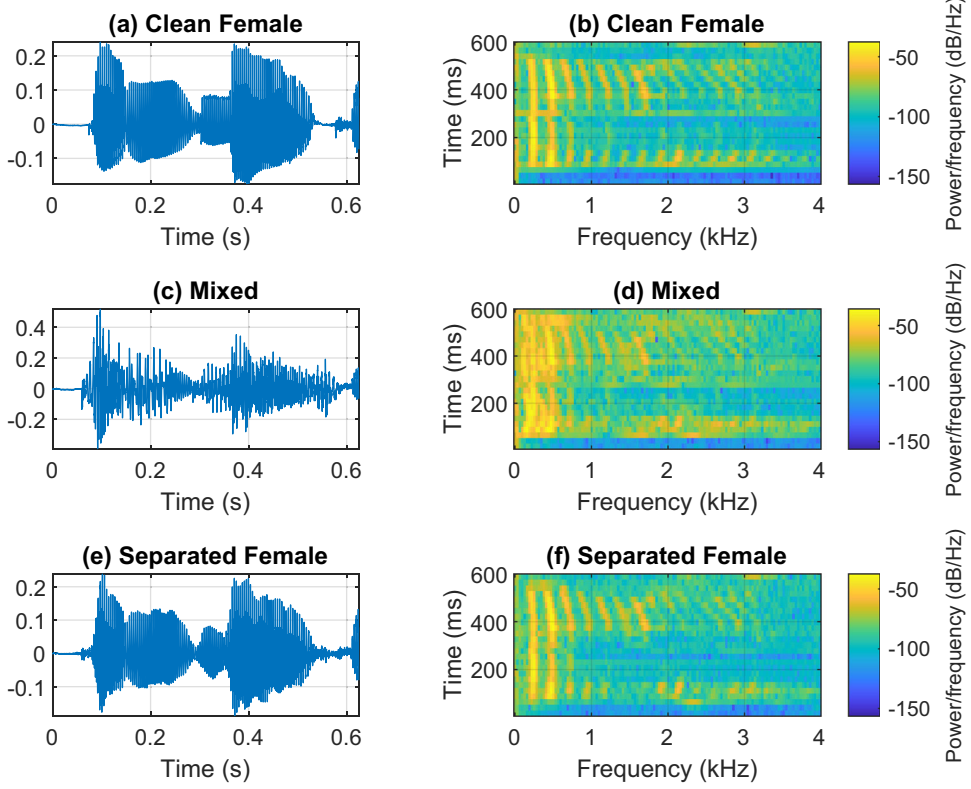

Fig. 9 **a** Waveform **b** Spectrogram of Clean, **c** Waveform **d** Spectrogram of Mixed and **e** Waveform **f** Spectrogram of Separated female source, respectively

other hand, a spectrogram serves as a graphical representation of a signal's frequency spectrum over time. It introduces an extra layer of information regarding frequency content over time, facilitating the examination of evolving spectral characteristics.

Figure 8 depicts the time-domain and spectrogram representations of the clean, mixed, and separated signals for the male source. In this case, we chose a male that performed best, the corresponding mixed, and the estimated male signal. From the mentioned figures we see that our suggested approach segregated the male source from the mixed one in a pretty good way. In the similar fashion, we see from Fig. 9, the female source also separated from the mixed signal.

6 Conclusion

From the perspective of neural architecture, we developed U-NET, a convolutional neural network architecture that built on a few improvements in the original CNN design. The model architecture was created with two principles in mind. The initial concept was encoder connections, which use strides 2's max pooling layers to minimize data sizes. We must further repeat the convolutional layers, including a greater quantity of filters in the encoder block. The second idea is to employ a decoder block and its associated connections. As we move closer to the decoder, we observe that the quantity of filters in the convolutional layers begins to lessen, followed by a continual up-sampling in the subsequent layers at upmost. We can also see the use of skip connections to link the preceding outputs to the decoder blocks' layers. Using this network architecture to separate the intended sources, we get better performance in every SNR scenario. In comparison with the outcomes of the other approaches mentioned in this article, the quality and understandability of the separated speech signals are enhanced. The experimental results show that the proposed speech separation model outperforms the current models in terms of overall performance in assessments of the improvement in the separated speech signals using various evaluation methodologies. We intend to research other training and testing procedures in the future utilizing different deep neural networks.

Acknowledgements This work is supported by the "Image and Speech Signal Processing Lab," Department of Computer Science and Engineering, Islamic University, Kushtia-7003, Bangladesh.

Author Contributions SB was involved in conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft, writing—review and editing. MSH helped in writing—original draft and writing—review and editing. ZR, MSA contributed to writing—review and editing. MSI helped in writing—original draft, supervision, project administration.

Data Availability Data will be made available on reasonable request.

Declarations

Conflict of interest The authors declare no conflicts of interest regarding the publication of this paper.

References

1. Huang, P.-S.; Kim, M.; Hasegawa-Johnson, M.; Smaragdis, P.: Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(12), 2136–2147 (2015)
2. Rivet, B.; Wang, W.; Naqvi, S.M.; Chambers, J.A.: Audiovisual speech source separation: an overview of key methodologies. IEEE Signal Process. Mag. **31**(3), 125–134 (2014)
3. Khan, M.S.; Naqvi, S.M.; Wang, W.; Chambers, J.; et al.: Video-aided model-based source separation in real reverberant rooms. IEEE Trans. Audio Speech Lang. Process. **21**(9), 1900–1912 (2013)
4. Wu, B.; Li, K.; Yang, M.; Lee, C.-H.: A reverberation-time-aware approach to speech dereverberation based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(1), 102–111 (2016)
5. Demir, C.; Saraclar, M.; Cemgil, A.T.: Single-channel speech-music separation for robust ASR with mixture models. IEEE Trans. Audio Speech Lang. Process. **21**(4), 725–736 (2012)
6. Jiang, D.; He, Z.; Lin, Y.; Chen, Y.; Xu, L.: An improved unsupervised single-channel speech separation algorithm for processing speech sensor signals. Wirel. Commun. Mob. Comput. **2021**, 1–13 (2021)
7. Mowlaee, P.; Saeidi, R.; Christensen, M.G.; Tan, Z.-H.; Kinnunen, T.; Franti, P.; Jensen, S.H.: A joint approach for single-channel speaker identification and speech separation. IEEE Trans. Audio Speech Lang. Process. **20**(9), 2586–2601 (2012)
8. Muhsina, N.; Beegum, D.; Manjusree, S.; Lubaib, P.; Al Saheer, S.; Shenoy, A.J.: Signal enhancement of source separation techniques. In: 2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1–8 (2023). IEEE
9. Hossain, M.I.; Al Mahmud, T.H.; Islam, M.S.; Hossen, M.B.; Khan, R.; Ye, Z.: Dual transform based joint learning single channel speech separation using generative joint dictionary learning. Multimed. Tools Appl. **81**(20), 29321–29346 (2022)
10. Weng, C.; Yu, D.; Seltzer, M.L.; Droppo, J.: Deep neural networks for single-channel multi-talker speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(10), 1670–1679 (2015)
11. Wichern, G.; Antognini, J.; Flynn, M.; Zhu, L.R.; McQuinn, E.; Crow, D.; Manilow, E.; Roux, J.L.: Wham!: Extending speech separation to noisy environments. arXiv preprint arXiv:1907.01160 (2019)
12. Mayer, F.; Williamson, D.S.; Mowlaee, P.; Wang, D.: Impact of phase estimation on single-channel speech separation based on time-frequency masking. J. Acoust. Soc. Am. **141**(6), 4668–4679 (2017)
13. Wang, D.; Chen, J.: Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(10), 1702–1726 (2018)
14. Sun, Y.; Wang, W.; Chambers, J.; Naqvi, S.M.: Two-stage monaural source separation in reverberant room environments using deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(1), 125–139 (2018)

15. Wang, C.; Zhu, J.: Neural network based phase compensation methods on monaural speech separation. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 1384–1389 (2019). IEEE
16. Zhou, L.; Lu, S.; Zhong, Q.; Chen, Y.; Tang, Y.; Zhou, Y.: Binaural speech separation algorithm based on long and short time memory networks. Comput. Mater. Continua **63**(3), 1373–1386 (2020)
17. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
18. Weninger, F.; Hershey, J.R.; Le Roux, J.; Schuller, B.: Discriminatively trained recurrent neural networks for single-channel speech separation. In: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 577–581 (2014). IEEE
19. Hochreiter, S.; Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
20. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
21. Wang, Y.; Wang, D.: A deep neural network for time-domain signal reconstruction. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4390–4394 (2015). IEEE
22. Grais, E.M.; Plumbley, M.D.: Single channel audio source separation using convolutional denoising autoencoders. In: 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1265–1269 (2017). IEEE
23. Luo, Y.; Mesgarani, N.: Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(8), 1256–1266 (2019)
24. Yuan, C.-M.; Sun, X.-M.; Zhao, H.: Speech separation using convolutional neural network and attention mechanism. Discret. Dyn. Nat. Soc. **2020**, 1–10 (2020)
25. Koteswararao, Y.V.; Rama Rao, C.: Single channel source separation using time-frequency non-negative matrix factorization and sigmoid base normalization deep neural networks. Multidimens. Syst. Signal Process. **33**(3), 1023–1043 (2022)
26. Qiao, X.; Luo, M.; Shao, F.; Sui, Y.; Yin, X.; Sun, R.: Vat-snet: A convolutional music-separation network based on vocal and accompaniment time-domain features. Electronics **11**(24), 4078 (2022)
27. Saleem, N.; Khattak, M.I.; AlQahtani, S.A.; Jan, A.; Hussain, I.; Khan, M.N.; Dahshan, M.: U-shaped low-complexity type-2 fuzzy LSTM neural network for speech enhancement. IEEE Access **11**, 20814–20826 (2023)
28. Cooke, M.; Barker, J.; Cunningham, S.; Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**(5), 2421–2424 (2006)
29. Varshney, Y.V.; Abbasi, Z.A.; Abidi, M.R.; Farooq, O.: Frequency selection based separation of speech signals with reduced computational time using sparse NMF. Arch. Acoust. **42**(2), 287–295 (2017)
30. Vincent, E.; Gribonval, R.; Févotte, C.: Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006)
31. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J.: An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)
32. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P.: Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), vol. 2, pp. 749–752 (2001). IEEE
33. Kates, J.M.; Arehart, K.H.: The hearing-aid speech perception index (HASPI) version 2. Speech Commun. **131**, 35–46 (2021)
34. Kates, J.M.; Arehart, K.H.: The hearing-aid speech quality index (HASQI) version 2. J. Audio Eng. Soc. **62**(3), 99–117 (2014)