

Performance Evaluation of DenseNet-121 and Bi-LSTM with Attention Mechanism based Vocal Signal Separation

Arfana Chowdhary

*DSP Laboratory, Department of Electronics, University of Jammu
Jammu-180006, India*

arfaniaqua@gmail.com

Parveen Kumar Lehana

*DSP Laboratory, Department of Electronics, University of Jammu
Jammu-180006, India*

pklehana@gmail.com

Corresponding Author: Parveen Kumar Lehana

Copyright © 2026 Arfana Chowdhary and Parveen Kumar Lehana. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Audio source separation is the process of separating a mixed audio signal into its individual sources. This task remains difficult because of significant overlap in time and frequency. The presence of noise and fluctuating signal characteristics further deteriorate the task. Traditional signal processing methods rely on strong statistical assumptions, which often limit their effectiveness in real-world acoustic signal processing. To overcome some of these challenges, this paper presents a supervised attention-based deep learning approach for separating single-channel audio signals. This approach directly predicts the individual source magnitude spectrograms without employing traditional time-frequency masking methods. The approach is based on fixed-length segments of audio signals, which are represented at a constant rate and mapped to magnitude spectrograms using short-time Fourier transform (STFT). A structured preprocessing employed in the algorithm ensure normalisation, cropping, and temporal alignment, maintains constant input data size during training and testing phases. The separation approach employs a DenseNet-121 network that uses hierarchical spectral feature extraction. Thereby enabling the effective extraction of complex harmonic and formant characteristics. To predict time-frequency regions containing salient perceptual significance, circular spatial attention mechanism has been employed for enabling the network to concentrate on significant acoustic elements. The temporal characteristics and contextual parameters from spectrum frames are captured using a bidirectional long short-term memory (Bi-LSTM) network and multi-head self-attention. The combined feature representations are processed through fully connected layers of the DenseNet-121 network to estimate the magnitude spectrograms of the constituent sources. The setup allows for end-to-end learning of nonlinear source mappings. The network is trained using scale-invariant signal-to-distortion ratio (SI-SDR) as loss function and the trained network is evaluated using objective metrics, such as short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ). The experimental results demonstrate satisfactory performance of separation, intelligibility, and perceptual quality across different mixing ratios. The approach establishes a low complexity, scalable, and interpretable solution to realistic problems in source separation.

Keywords: Audio source separation, Short-Time Fourier Transform (STFT), Deep learning, Spectrogram estimation, Multi-head attention, DenseNet-121, Circular attention, Bi-LSTM, Speech enhancement, PESQ, STOI.

1. INTRODUCTION

Audio source separation separates a mixed audio signal into its constituent sources like vocal from the instrumental background and individual speakers from a mixed signal of voices from different subjects. This is essentially a fundamental problem in contemporary audio signal processing that has captured the interest of researchers working in the field of acoustics due to its wide applicability to music production, telecommunications, assistive hearing devices, speech recognition, human-computer interaction, robotics, intelligent acoustic environments, etc. [1, 2]. In real-world scenarios, audio mixtures contain multiple overlapping sources that interact in complex temporal and spectral domains making the problem of separation a difficult task. Accurate separation of sources with high quality is an essentially unstructured problem and hence, practically difficult if not impossible. Environmental noise, reverberation, recording artefacts, and non-stationary properties of the signals further degrade perceptual quality and hinder the reliable extraction of constituent sources [3]. Traditional approaches to audio separation largely rely on statistical and spectral analysis techniques. The last few decades have witnessed significant investigations of traditional approaches such as Non-negative Matrix Factorisation (NMF), Independent Component Analysis (ICA), Wiener filtering, spectral subtraction, and beamforming [4–6]. These approaches, generally assume strong premises such as the statistical independence of sources, linear mixing models, and noise stationarity. Therefore, these techniques pose the problem of spectra overlapping and noise [7]. ICA is not expected to provide a good quality of the separated signals as speech and music signals are harmonically related. Also, there is a problem in NMF as multiple sources occupy overlapping frequency bands, leading to unsatisfactory separation of the signal components [5]. Historically, frequency-domain approaches have dominated audio separation investigations due to their mathematical robustness and ease of illustration. Transform operations like Short-Time Fourier Transform (STFT) and wavelet transforms make it possible for signals to be represented in the time-frequency domain, where prominent spectral patterns become evident and filterable [8, 9]. These methods are mostly used for many classical systems for speech amplification, noise inhibition, and music transcription [10, 11]. However, there are inherent limitations to frequency domain processing. The phase information is usually neglected, resulting in reconstruction artefacts. Spectral leakage and resolution trade-offs also exist in time-frequency representations. These limitations affect the accuracy of transient's separation for highly non-stationary and highly time-varying signals [11, 12]. Frequency-domain processing is associated with mathematical overhead, which is undesirable in real-time applications such as live audio processing and usable listening devices [7]. The lack of success of traditional signal processing techniques has forced a paradigm shift to machine learning and deep learning models. Compared to traditional model-based approaches, traditional machine learning approaches have the advantage of using separation mappings without unique analytical assumptions [13, 14]. This new data-centric approach has also been made possible by the fast-evolving developments in computational power, audio datasets, and optimisation techniques. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have achieved exceptional capabilities for modelling complex non-linear relationships between mixed and clean signals [2]. CNNs are especially effective in estimating local spectral patterns from

spectrograms, treating them as two-dimensional images and extracting ordered features by using convolution filters [15, 16]. However, CNN-based models alone are limited in their ability to model long-range temporal dependencies. Audio signals, especially speech and music, are inherently sequential, and their semantic content is distributed across extended time intervals. Temporal coherence plays a crucial role in distinguishing between sources, particularly in scenarios involving overlapping speech or sustained musical instruments [17]. To address this limitation, RNNs, and especially Long Short-Term Memory (LSTM) networks, have been widely adopted. LSTM adopt a gated memory system, which permits them to retain relevant information across long ranges for better temporal behaviour compared to quality feed-forward networks [18]. One step ahead, Bidirectional Long Short-Term Memory (Bi-LSTM) allows information to flow from both forward and backward enabling them to capture more contextual information, especially in sequential data flow, empowers the model to leverage existing and future context [19].

Although powerful models, CNNs and Bi-LSTMs also face problems in handling complex acoustic patterns. The homogeneity of time–frequency features lead to skewed representations when the neighbouring or noisy regions of the input features are much larger than the crucial signal constituents. This finding affects the selection of attention mechanisms in deep learning architectures. Attention enables neural networks to dynamically allocate different weights to different parts of the input, focusing on informative segments and rejecting irrelevant or noisy parts [20, 21]. This mechanism is much like the human auditory system, which naturally preferentially focuses on salient sound samples and rejects peripheral noise. The self-attention mechanism, widely used in transformer architectures, is particularly well-suited to model global dependencies in the context of sequential data. In the case of separation of speech and music, self-attention helps the model to recognise relational aspects between different distant frames, improving the discrimination of overlapping segments [21]. By enabling several concurrent attention operations, multi-head attention goes one step further and teach each head to focus on distinct acoustic characteristics, such as pitch, timbre, phonetic structure, or rhythmic patterns. Richer feature selection and robustness in noisy environment are the advantages of networks using these combinations of advanced architectures [21–23]. These developments are also contributing developments in time-domain signal processing for audio signal separation. Time-domain learning avoids spectral modifications because separation is performed directly in the waveform, unlike the frequency-domain approach. The Time-Domain Audio Separation Network (Tas-Net), which proved the feasibility of training a neural network to separate speech from mixed acoustic signals, is a key breakthrough for this field [24]. Tas-Net employs a convolution encoder to learn a latent representation of the signal, followed by a temporal convolutional network that estimates source-specific masks, and a decoder that reconstructs the separated signals.

Also, the time domain signal separation avoids phase reconstruction and significantly reduces reversion, making it particularly suitable for real-time applications. Building on Tas-Net, Conv-Tas-Net introduced swollen convolutions to expand the neutral field without increasing computational complexity, enabling the model to capture long-range temporal dependencies more [25]. More recent models like Dual-Path Recurrent Neural Networks have continued to improve performance by the inclusion of local and global temporal information [26]. These models integrate signals at combined temporal opinions and capture fine-grained transient details at the same time as long-term surrounding structure, resulting in very high performance for real-time speech enhancement, hearing aids, and communication systems. Beyond audio processing, deep learning-based signal separation has been widely applied in biomedical engineering, telecommunications, and environmental

monitoring. Time-domain approaches using neural networks have been shown as effective in separating Electroencephalogram (EEG), Electrocardiogram (ECG), and Electromyography (EMG) signals in biomedical signal analysis for accurate diagnosis of neurological and cardiac disorders [27, 28]. In telecommunications, machine learning-based signal separation has been reported to achieve better performance in multi-user detection, channel estimation, and interference alleviation than classical linear equalisation techniques [29]. But despite their numerous advantages, deep learning-based separation systems present many challenges. Arithmetically, this limitation is particularly critical for training systems as they demand vast datasets and substantial hardware resources [13]. Robustness in highly noisy environments remains an area of concern, as models trained on limited datasets may not effectively generalise [30]. Furthermore, logical justification is a critical concern since deep neural networks operate as black boxes. The result is that in certain areas critical to safety, such as a healthcare and assistive technology, the experts are able to interpret the results, how they are dependent on input data [31]. Modern audio separation systems are moving from traditional assumption-based analysis to flexible, data-driven deep learning frameworks. Convolution architectures excel at finding spectral features. At the same time, recurrent networks address temporality, and attention mechanisms are also contributing for pin pointing the most important segments in the signals.

Based on the above insights, for the present research, DenseNet-121 is selected for its ability to extract deep hierarchical features from spectrograms with efficient gradient flow and feature reuse, especially useful for modelling complex audio patterns. To enhance the reliability of temporal contextual parameter, a Bi-LSTM network is incorporated, enabling the model to learn dependencies from both past and future frames, which is essential for audio signals. Additionally, the integration of multi-head self-attention mechanisms allows the model to focus on the most relevant time–frequency regions, improving separation accuracy and perceptual quality. The detail of the technique employed is explained in the following section. Results and discussion are presented in **Section 3**. This is followed by Conclusions and future scope.

2. MATERIAL AND METHOD

The proposed system using supervised deep learning approach to separate vocal signals from the given mixture is shown in **FIGURE 1**. The input signal applied to the system consists of a mixture of vocal and instrumental signals. The system may be explained using two-phases i.e. training and testing, both utilising similar pre-processing and feature extraction steps. For the investigations, the acoustic material was collected using experiment specific recordings and from online resources. For the recording of vocal signals, three male and three female speakers were used. All speakers were Hindi speaking and in the age group of 20-25 years. A total of 100 sentences from each speaker were recorded 16 kHz sampling frequency and 16-bit quantization using RØDE condenser microphone (cardioid polar pattern, 20 Hz–20 kHz frequency response, and low self-noise), ensuring high-quality audio capture. Four types of instrumental sounds piano, drum, guitar, and flute were downloaded from the reliable online resources. In total 600 vocal and four instrumental signals were finalized. The vocal and instrumental signals were mixed in the ratios ranging from 0.1 to 0.8, resulting in 1,53,600 overall samples. The samples were randomly divided into training, validation, and testing subsets as per the ratio 70: 20: 10.

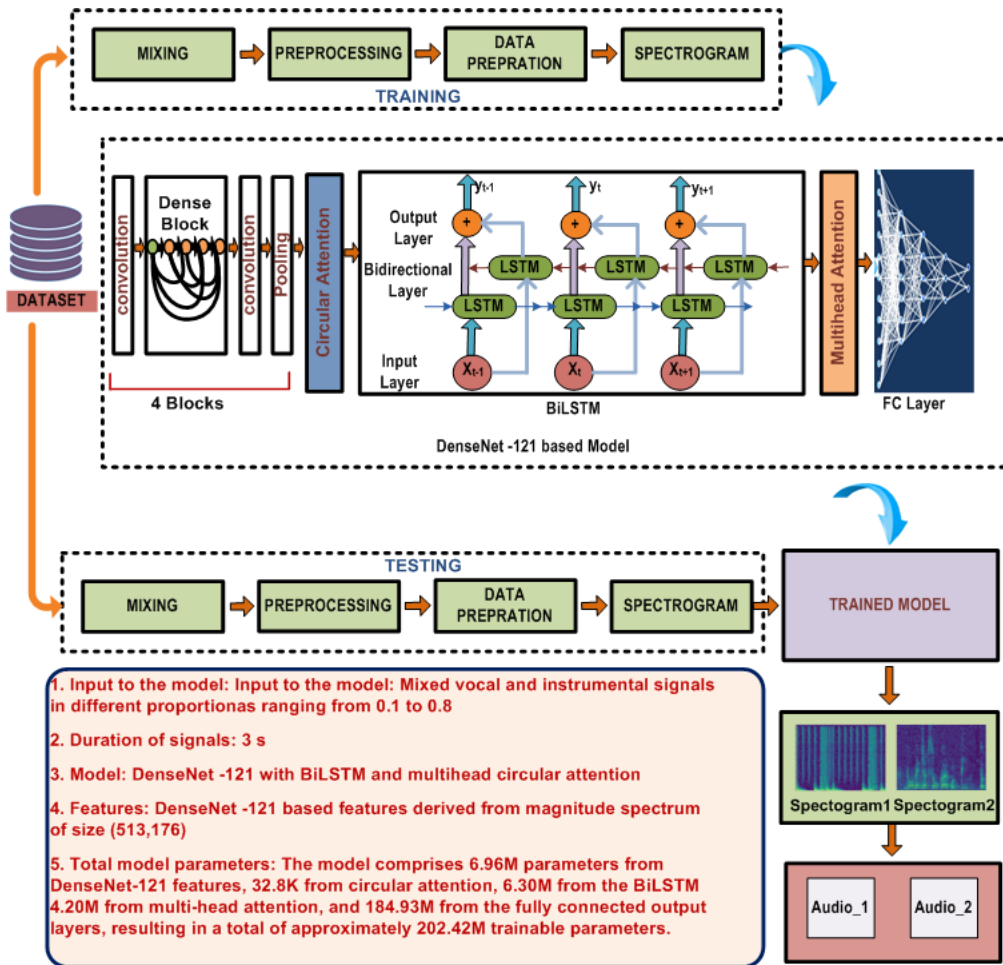


Figure 1: Block diagram of proposed framework for training and testing for signal separation and reconstruction from spectrogram comprising of DenseNet-121, Bi-LSTM, and attention mechanism.

In preprocessing, the duration of each sample was restricted to 3s and amplitude was normalized to reduce sample to sample drastic energy variations. Normalisation of signal amplitude also necessary for numerical stability. The normalised samples are transformed into magnitude spectrograms using STFT with a fixed window length (1024 samples) and hop size (256 samples). The STFT provides localised time–frequency representation that preserves both signals spectral content and its temporal evolution at fixed dimension of (513×176) . Only the magnitude component was used for learning, while the phase was stored and reused later for waveform reconstruction. DenseNet-121 based model along with circular attention, Bi-LSTM, multi-head attention, and a fully connected layer has been used for the training and testing of the system. The model comprises of 6.92 million parameters from DenseNet-121, 32.8 k from circular attention, 6.30 million from Bi-LSTM, 4.20 million from multi-head attention, and 184.93 million from fully connected output layer. The complete architecture comprises of approximately 202.42 million trainable parameters. Unlike conventional approaches that use fixed mixing conditions, the proposed methodology incorporates wide

range of systematic variable mixing strategy to simulate diverse real-world interference scenarios. This enables more comprehensive evaluation of the model robustness and generalization capability. Furthermore, the integration of DenseNet-121, Bi-LSTM, and attention mechanisms is expected to effectively capture spectral, temporal, and contextual dependencies, providing a task-specific enhancement over existing architectures.

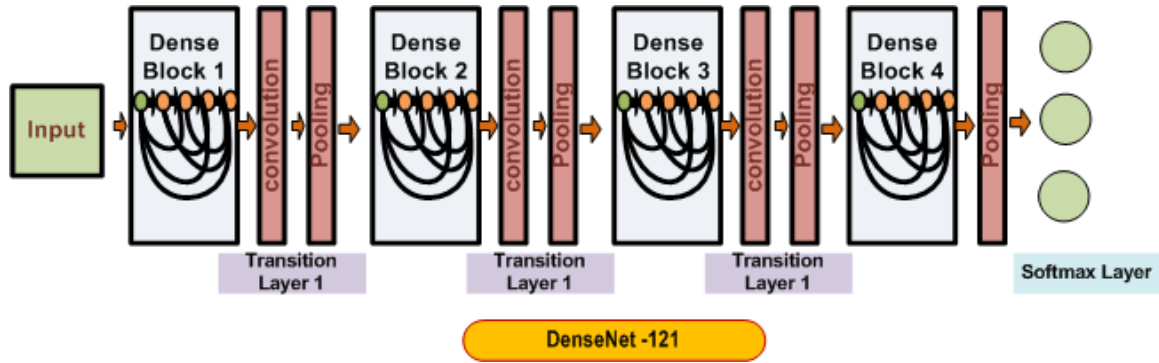


Figure 2: Architecture of DenseNet-121 for feature extraction compresses of four dense blocks and connectivity with convolution and pooling layer for deep learning.

The detail of the process of DenseNet-121 used for feature extraction in the proposed system is shown in **FIGURE 2**. The main advantage of using model is that it has already been trained on large-scale image data. It has a built-in dense layer, inter connections for feature reuse, and enhanced gradient flow for high accuracy of the classification. The magnitude spectrogram was first adapted to match the input requirements of the DenseNet-121 backbone by replicating the single-channel spectrogram into three channels and resizing it to 224×224 .

Such a structure allows the model to learn highly discriminative hierarchical representations while maintaining computational efficiency [32]. The network produces high-level feature maps of size $1024 \times 7 \times 7$, capturing complex spectral patterns associated with vocal harmonies and instrumental textures. Transfer learning from visual domains to audio spectrograms has been shown to significantly improve convergence speed and generalisation performance in audio processing tasks. The convolutional layers capture local spectral patterns such as harmonics and formant structure, while deeper layers learn more abstract representations related to source identity and temporal modulation.

Detail of circular spatial attention mechanism is shown in **FIGURE 3**, which is applied for extracting the most informative time–frequency regions subsiding the spectral structures and irrelevant background components using eight attention heads. By computing attention weights across all the time steps, the model selectively focuses on relevant spectro-temporal patterns and improves the representation of overlapping sources. So, by applying attention multiplicatively to the feature maps, the model dynamically focuses on regions that contribute most to source discrimination. The circular attention maps also provide interpretability by allowing visualisation of spectral regions prioritised by the model during separation, which is particularly valuable for understanding and validating system behaviour in complex acoustic environments [33]. While convolutional layers are effective at capturing spatial structure, they are limited in modelling long-term temporal depen-

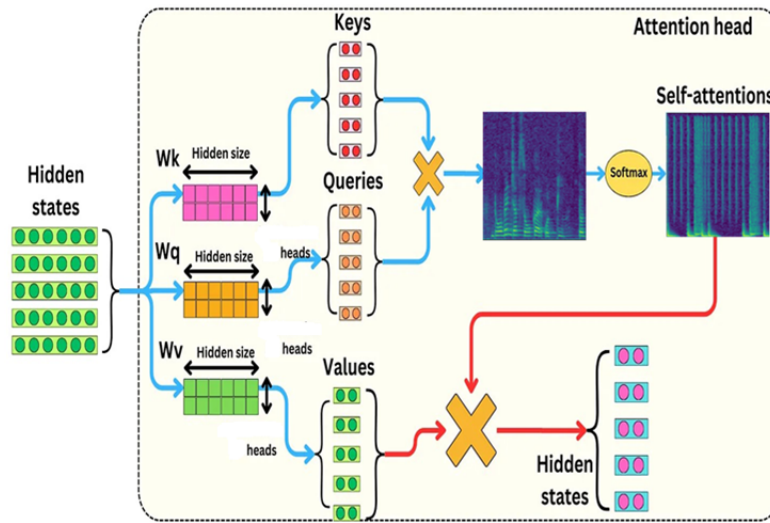


Figure 3: Architecture of multi-head self-Attention mechanism applied to hidden state for processing parallel heads for refine representation.

dencies. To overcome this limitation, the system employs a Bi-LSTM network that processes the feature sequences in both forward and backward directions.

The attended features are reshaped into a temporal sequence of length 49 (7×7) with a feature dimension of 1024 and passed to a Bi-LSTM network as shown in **FIGURE 4**. The Bi-LSTM consists of 512 hidden units in each direction. The gate values in LSTMs help in selecting desired features while filtering out unwanted noise and fluctuations. To improve temporal discrimination, the Bi-LSTM output is passed through a multi-head self-attention layer. Self-attention computes pairwise relationships between all time frames, allowing the model to capture global dependencies and contextual relevance across the entire sequence. Multi-head attention improves this paradigm by allowing concurrent attention to a spectrum of features like pitch, rhythm, speaker identity, and harmonic structure. The diversity in attention representations results in richer feature learning and improved classification performance, especially in scripts with severe spectral overlaps. Self-attention mechanism outperforms classical recurrent architectures because they can model long-range dependencies in a single layer and avoid inclination decay [34].

The representational layer obtained from the attention layers are mapped directly to the assessed magnitude spectrograms of the target sources using fully connected layers. Two parallel output branches are used to predict the vocal and instrumental components independently. The system is trained using SI-SDR of the predicted sources with respect to the actual target values. The entire network is optimised end-to-end using the Adam optimiser, which adapts learning rates dynamically and accelerates convergence [35]. Early stopping is employed to prevent over fitting and ensure generalisation to unseen data.

Finally, two fully connected layers directly regress the magnitude spectrograms of the vocal and instrumental sources from the global feature representation. Each output layer maps a 1024-dimensional feature vector to a 513×176 spectrogram. After training, testing of the trained model is carried using

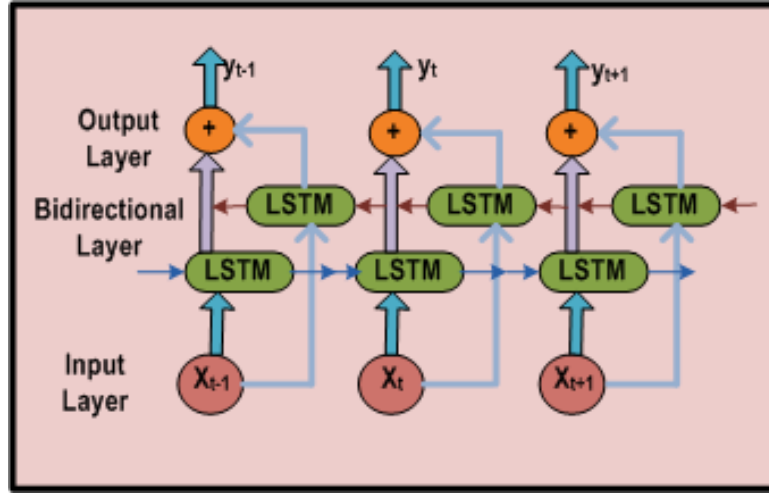


Figure 4: Architecture of Bi-LSTM capturing past and future dependencies to generate time step wise output from hidden states.

the mixtures from the testing subset. The trained model predicts the separated spectrograms from each audio component. These estimated spectrograms are combined with the mixture phase and converted back to time-domain audio signals using the inverse STFT. Although phase estimation remains a challenging problem, reusing the mixture phase has been shown to produce perceptually acceptable reconstructions in many practical systems. Objective evaluation is performed using PESQ to assess listening quality and STOI to evaluate speech intelligibility. Together, these metrics provide a comprehensive assessment of both technical accuracy and perceptual realism.

PESQ measures perceptual speech quality by modelling the human auditory system and calculating a disturbance measure between the reference signal $s(t)$ and the estimated signal $\hat{s}(t)$. Mathematically, PESQ can be expressed as

$$\text{PESQ} = a_0 + a_1 D_{\text{sym}} + a_2 D_{\text{asym}} \quad (1)$$

Regression coefficient (a_0, a_1, a_2) obtained during PESQ calibration. During large-scale subjective listening test results a_0 is a constant term providing an objective score with human perception. The a_1 and a_2 are relative weights of different distortions. D_{sym} represents the symmetric disturbance, meaning distortions that speech quality equally. Mainly, this involves additive noise, coloration, and those aspects related to both time and frequency. These are seen to be contained between the reference signal and the degraded speech signal. The meaning of D_{asym} is “asymmetrical disturbance”. It is used to describe the perceptions with the occurrence of asymmetrical disturbance. Where the visibility of audible speech quality is more dominant in one particular direction then in the other. It usually comprises the missing speech components that could be found within the artefacts. It is suggested that the auditory system is more sensitive to missing speech components than to noisy components. Secondly, it is proposed that the PESQ scores, which are determined, are the linear combination of the computed disturbances along with regression coefficients. Speech Intelligibility using STOI calculates the correlation among clean signals along with clean and short-time temporal cover.

Given clean and estimated short-time temporal envelopes $X_{j,m}$ and $\hat{X}_{j,m}$ frequency band and M^{th} time segment, STOI is defined as

$$\text{STOI} = \frac{1}{JM} \sum_{j=1}^J \sum_{m=1}^M \frac{(X_{j,m} - \mu_{j,m})^T (\hat{X}_{j,m} - \hat{\mu}_{j,m})}{\|X_{j,m} - \mu_{j,m}\|_2 \|\hat{X}_{j,m} - \hat{\mu}_{j,m}\|_2} \quad (2)$$

Whereas J is the number of frequency bands, M is the number of time segments, $\mu_{j,m}$ and $\hat{\mu}_{j,m}$ denote the mean values of the corresponding envelope segments. STOI values range from 0 to 1, with higher values indicating improved speech intelligibility. Together, PESQ and STOI provide a comprehensive quantitative evaluation of the proposed system, capturing both perceptual quality and intelligibility of the separated vocal and instrumental signals. On the other hand, SI-SDR measures how close an estimated (processed) signal is to the original clean signal while ignoring differences in signal scale (amplitude). This makes it more reliable than traditional SDR when the output signal has different loudness. It has been used as the inverse loss function in the training. The speech reconstruction process in the proposed framework is performed in the time–frequency domain: the neural network estimates the magnitude spectrograms of the target sources, while the mixture’s phase information is reused for waveform synthesis. Given a time-domain mixture signal $x(t)$, the STFT is applied to obtain its complex spectrogram.

$$X(f, \tau) = \sum_t x(t) \omega(t - \tau) e^{-j2\pi f t} \quad (3)$$

where $\omega(\cdot)$ denotes the analysis window, f is the frequency bin index, and τ represents the time frame. The complex spectrogram is decomposed into magnitude and phase components as

$$X(f, \tau) = \sqrt{R\{X\}^2 + I\{X\}^2}, \quad \angle X(f, \tau) = \tan^{-1} \left(\frac{I\{X\}}{R\{X\}} \right) \quad (4)$$

In the implementation, the magnitude spectrogram is normalised and used as the learning target, while the phase is preserved for later reconstruction.

The deep network directly estimates the source-specific magnitude spectrograms $\hat{\mathbf{S}}_k(f, \tau)$ for each source $k \in \{1, 2\}$ (vocal and instrumental) using a DenseNet-121 backbone followed by circular attention, Bi-LSTM, and multi-head self-attention. Unlike masking-based approaches, no explicit time–frequency mask is estimated. Instead, the model learns a nonlinear mapping.

$$\hat{\mathbf{S}}_k = F_\theta(|\mathbf{X}|) \quad (5)$$

where $F_\theta(\cdot)$ represents the trained neural network with parameters θ . The loss function used during training is the MSE between the predicted and reference magnitudes, defined as

$$\text{MSE} = \sum_{k=1}^2 \left\| \hat{\mathbf{S}}_k(f, \tau) - \mathbf{S}_k(f, \tau) \right\|_2^2 \quad (6)$$

which enforces accurate spectral reconstruction of both sources across all time–frequency bins.

For waveform synthesis, the estimated magnitude spectrogram $\hat{\mathbf{S}}_k(f, \tau)$ is combined with the mixture phase $\angle \mathbf{X}(f, \tau)$ to form the reconstructed complex spectrogram.

$$\mathbf{X}_k(f, \tau) = \hat{\mathbf{S}}_k(f, \tau) \quad (7)$$

which enforces accurate spectral reconstruction of both sources across all time–frequency bins.

For waveform synthesis, the estimated magnitude spectrogram $\widehat{\mathbf{S}}_k(\mathbf{f}, \tau)$ is combined with the mixture phase $\angle \mathbf{X}(f, \tau)$ to form the reconstructed complex spectrogram.

$$\widehat{\mathbf{X}}_k(f, \tau) = \widehat{\mathbf{S}}_k(f, \tau) e^{j\angle \mathbf{X}(f, \tau)} \quad (8)$$

The time-domain signal is then obtained by applying the inverse STFT.

$$\widehat{s}_k(t) = \sum_{\tau, f} \widehat{\mathbf{X}}_k(f, \tau) e^{j2\pi f t} \omega(t - \tau), \quad (9)$$

Although phase estimation remains a challenging problem, reuse of the mixture phase has been shown to yield perceptually acceptable reconstructions when magnitude estimates are accurate, particularly for speech and music separation tasks.

3. RESULTS AND DISCUSSION

The performance of the proposed attention-based source separation structure is evaluated through extensive analysis of both training dynamics and testing behaviour under varying mixture conditions. The objective of this evaluation is to understand how spectrogram-domain optimisation translates into waveform and audio level separation quality and how attention mechanisms evolve with training to support robust and perceptually meaningful source reconstruction. **FIGURE 5** shows the training plot for SI-SDR, PESQ, STOI versus epoch on the x-axis and loss on the y-axis. The training epoch and the mean squared error minimising difference between the predicted and ground-truth magnitude spectrograms at that epoch. For each epoch, the estimated spectrogram is combined with the mixture phase and transformed back to the time domain audio signal using inverse STFT to obtain the reconstructed waveform, and the separation quality is then evaluated on this waveform using the SI-SDR. Also, during the early epochs, the significant decline in the mean squared error results in a sharp improvement in separation performance. This shows that there is a significant improvement in waveform reconstruction error reduction, corresponding to the model learning the major vocal patterns and greatly diminishing instrumental influences. However, as the model progresses through subsequent epochs, the rate of improvement gradually slows. But as seen from the graph above, the smooth rate of improvement of the metrics pertaining to perceptual evaluation and intelligibility scores emphasises the significance of the evaluation to the information contained within the temporal envelope, as opposed to the error contained within the energy values. By changing the learning rate, smooth changes in loss occur in the spectrogram domain which affects the perceptual quality. Although minor changes are evident across the following stages of improvement, they are due to fine-tuning of perception. We have employed early stopping to prevent the model from continuing to learn, as evidenced by the constant rate of convergence towards minimal improvement. Therefore, a balance is achieved between the accuracy of the spectrogram and the waveform separation, while keeping the perceptual level minimal. STOI measures the cover correlation based on intelligibility normalisation the correlation concerned with the intelligibility of the speech. Learning rate scheduling is evident as step sizes are minimised. It can be seen that as the loss is minimised further on, moderate amounts of increase are observed on the PESQ plots. Changes seen in the PESQ are minimal, indicating of fine-tuning of the model and perception rather than an unstable model.

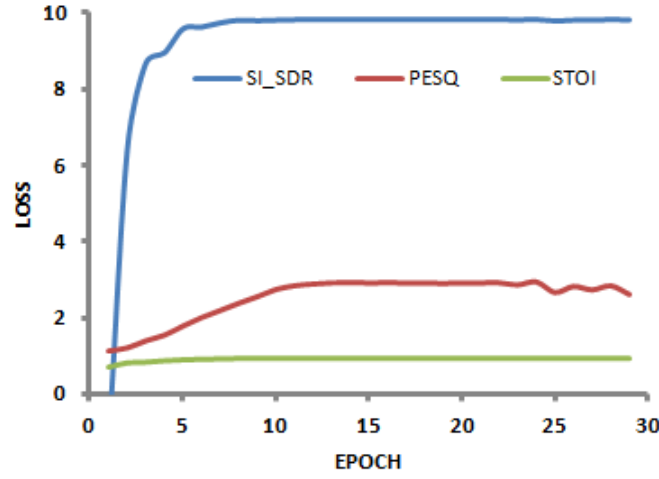


Figure 5: Training loss and evaluation metrics (SI-SDR, PESQ, STOI)

Early stopping after validation set loss and all evaluation metrics have converged ensures that the chosen model best balances the trade-off of spectrogram accuracy against waveform-level perceptual quality.

For obtaining the reconstructed signal, the estimated spectrogram is combined with the mixture phase and transformed back to the time domains using inverse short-time Fourier transform to obtain the reconstructed signal. Let n denote the epoch number, $s(t)$ the reference signal, $\hat{s}(t, n)$ the corresponding separated signal, and $\alpha(n)$ the normalization factor, the quality of the separated signal is measured in terms of the SI-SDR using (10).

$$\text{SI-SDR}(n) = 10 \log_{10} \left(\frac{\|\alpha(n) s(t)\|^2}{\|\alpha(n) s(t) - \hat{s}(t, n)\|^2} \right), \text{ where } \alpha(n) = \frac{\langle \hat{s}(t, n), s(t) \rangle}{\|s(t)\|^2} \quad (10)$$

SI-SDR quantifies the similarity between the reconstructed waveform and the desired signal, regardless of scale; thus, it may be a powerful measure of fidelity at the waveform level.

The spectrogram representation signifies that the separation performance under different training condition in the early training phase that trade-off in **FIGURE 6**, approximately between epochs 5 and 25, a rapid decrease in mean square error is observed, which corresponds to a steep increase in SI-SDR. This shows that the model learns quickly that the main source of interference is being suppressed and that the estimated waveform energy aligns with that of the original signal. It is now that the entire learning process can be defined by the alignment of the entire spectrum. The model captures the prominent characteristics of the human voice and eliminates the instrument's large feature. This is because the changes that the model is showing are not as drastic as in the earlier stages, the error correction does not show as prominently. This is not unusual in the world of deep learning. Unlike the SI-SDR, the overall performance of the model, as shown by the metrics used in the PESQ and the STOI, shows consistent and smoother trends. The overall qualitative process that is revealed by the spectrogram shows these trends. After five epochs, the training mixture with a vocal-to-instrumental ratio of 0.2:0.8 is largely dominated by instrumental energy. On one hand, At 5 epochs the mixture spectrogram clearly shows significant vertical harmonic and rhythmic structures across the spectrum. On the other hand, the vocal content forms faint patterns in the

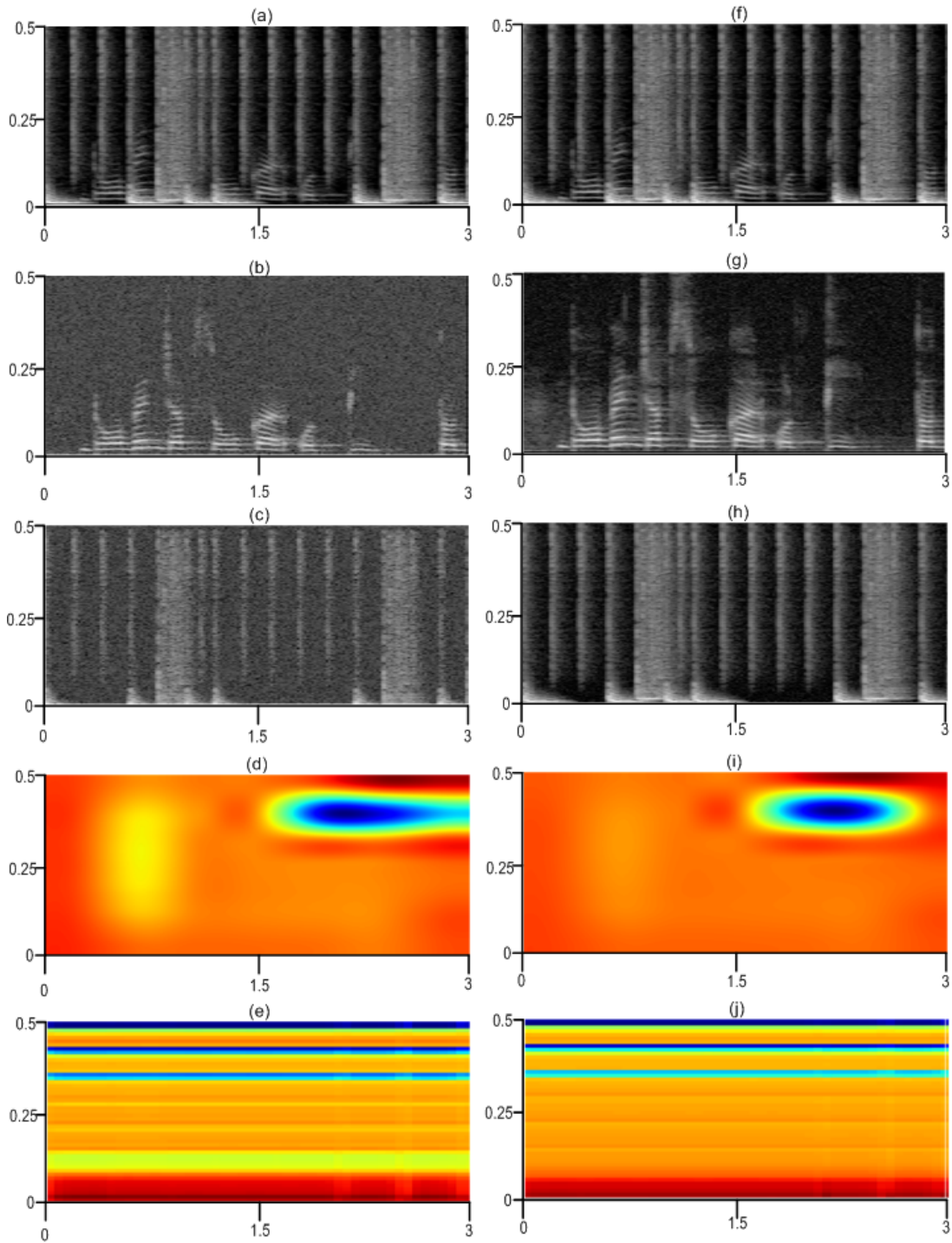


Figure 6: Effect of number of epochs on the quality of the separated components during training phase of the proposed system. First column shows the plots for 5 epochs and second column for 25 epochs. First row is for mixtures, second for separated vocals, third for separated instrumental, fourth for circular attention, and last row for average multi-head attention.

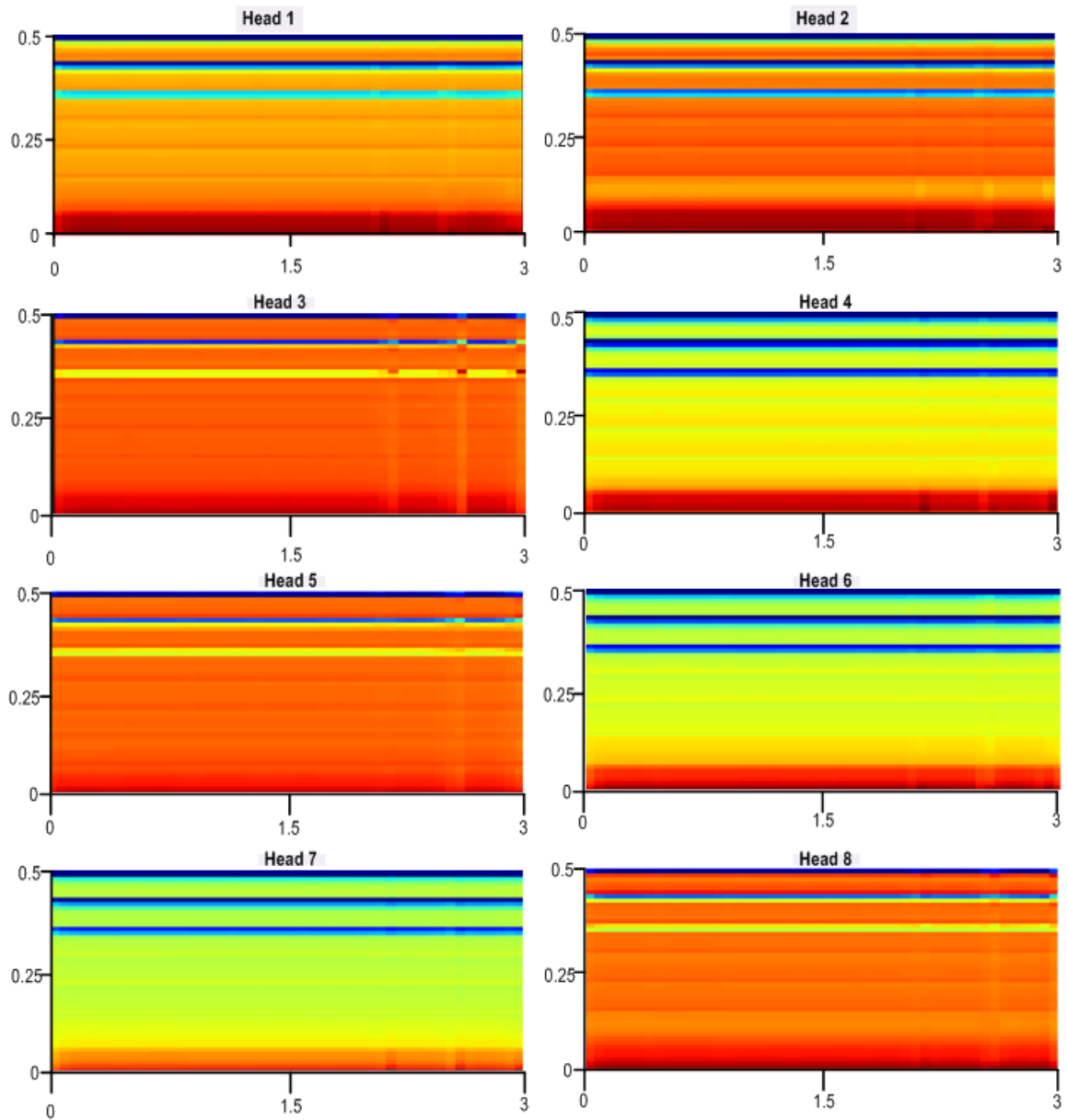


Figure 7: The figure shows output of all the eight heads after training.

low to mid-frequency range of the spectrogram. However, the separated vocal spectrogram shows fragmented harmonic structures, along with significant instrumental leakage at this point in the output, suggesting that the model has not fully learned the characteristics of the vocal content. The output of the separated instrumental sound is quite similar to that of the mixture itself, suggesting the model dominance over a stronger sound. The circular attention map is wide and diffuse, showing weak temporal selectivity. In contrast, the multi-head attention map displays fairly uniform patterns, suggesting that attention heads are not yet specialized. Quantitatively, this stage achieves SI-SDR= 9.59 dB, PESQ = 1.80, and STOI = 0.90, which means it has reasonable intelligibility but moderate

distortion. After twenty-five epochs, the training mixture with a ratio of 0.8:0.2 ratio result in a better situation in which vocal energy takes precedence. From the separated vocal spectrogram, it is clear that the formants are well-defined and that there is minimal noise. But at the same time, the instrumental output is sparse and clean, indicating that the interference has been well suppressed and that it originated from another source. The circular attention map is very sharply focused. The multi-head attention exhibits organised highlights on specific frequencies, reflecting stable and targeted attention learning. These are quantified as SI-SDR = 9.80 dB, PESQ = 2.67, and STOI = 0.93, confirming that perceptual quality and intelligibility have been improved. Together, these observations show that increasing the number of training epochs has allowed the model to evolve from coarse spectral alignment to fine-grained perceptual optimisation. **FIGURE 7**, shows how the eight attention heads in the multi-head attention mechanism focus on different aspects of the same time–frequency input. Although all heads receive identical input, their attention patterns differ, indicating that each head learns to extract a specific type of information. Heads are given greater focus on the lower portions of the representation, where high energy is usually heard, for instance, in presiding speech harmonics or fundamental musical tones. Such heads therefore only focus on the strongest and most energetic parts of the signal to provide the model with the main structure of the audio data. Other heads scatter their attention more evenly over wider spaces, meaning they capture higher-level circumstantial context rather than localised particularities. Such heads may be involved in modelling long-term secular patterns of speech or sustained musical textures. Some heads exhibit intermediate behaviour, focussing on some frequency bands but still varying over time, indicating sensitivity to passage.

The spectrograms shown in **FIGURE 8** are testing results in which about the testing results the signal on the left side shows the mixture consisting of a very weak vocal component (10%) and a dominant instrumental component (80%), resulting in strong broadband and vertical harmonic structures across the time–frequency plane. The multi-head attention shows organised emphasis on specific frequencies, indicating stable and targeted attention learning. This is also reflected quantitatively by SI-SDR = 9.80 dB, PESQ = 2.67, and STOI = 0.93, confirming improved perceptual quality and intelligibility. These observations together not only show that an increased number of training epochs can help the model evolve from coarse spectral alignment to fine-grained perceptual optimisation. It is also evident that the value of the STOI measure, which is 0.92, shows that intelligibility is present. Additionally, it is evident that the PESQ score is 2.67. This implies that cognitive quality is moderate. Regarding the same, it was observed that there is a strong rhythmic pattern in the separated spectrogram. It is also apparent that this shows that strong reconstruction of a strong source. It is also evident that circular patterns were presented due to attention. This shows that the areas are small. It was also observed that global multi-head attention is focused on the frequency-based layer patterns. This implies that it uses both local and global context to recover intelligible speech, especially during extreme masking.

Regarding vocal output, this appears to be close to the mixture and mainly comprises clean, sharp, and continuous formant structures, while a high SI-SDR of 24.40 dB and an STOI of 0.99 indicate perfect reconstruction. The circular attention map is localised at the centre around the prominent vocal regions, and multi-head attention displays well-organised horizontal frequency bands, indicating mature global context modelling. The high of 3.72 for the separated speech clearly conform of natural, undistorted sound. A comparison of the two test examples forms the harsh conditions of the operating regime of this speech separation model. In the unfavourable case, where 0.1 vocal and 0.8 instrumental signals are mixed, the energy of the signal under consideration is much lower

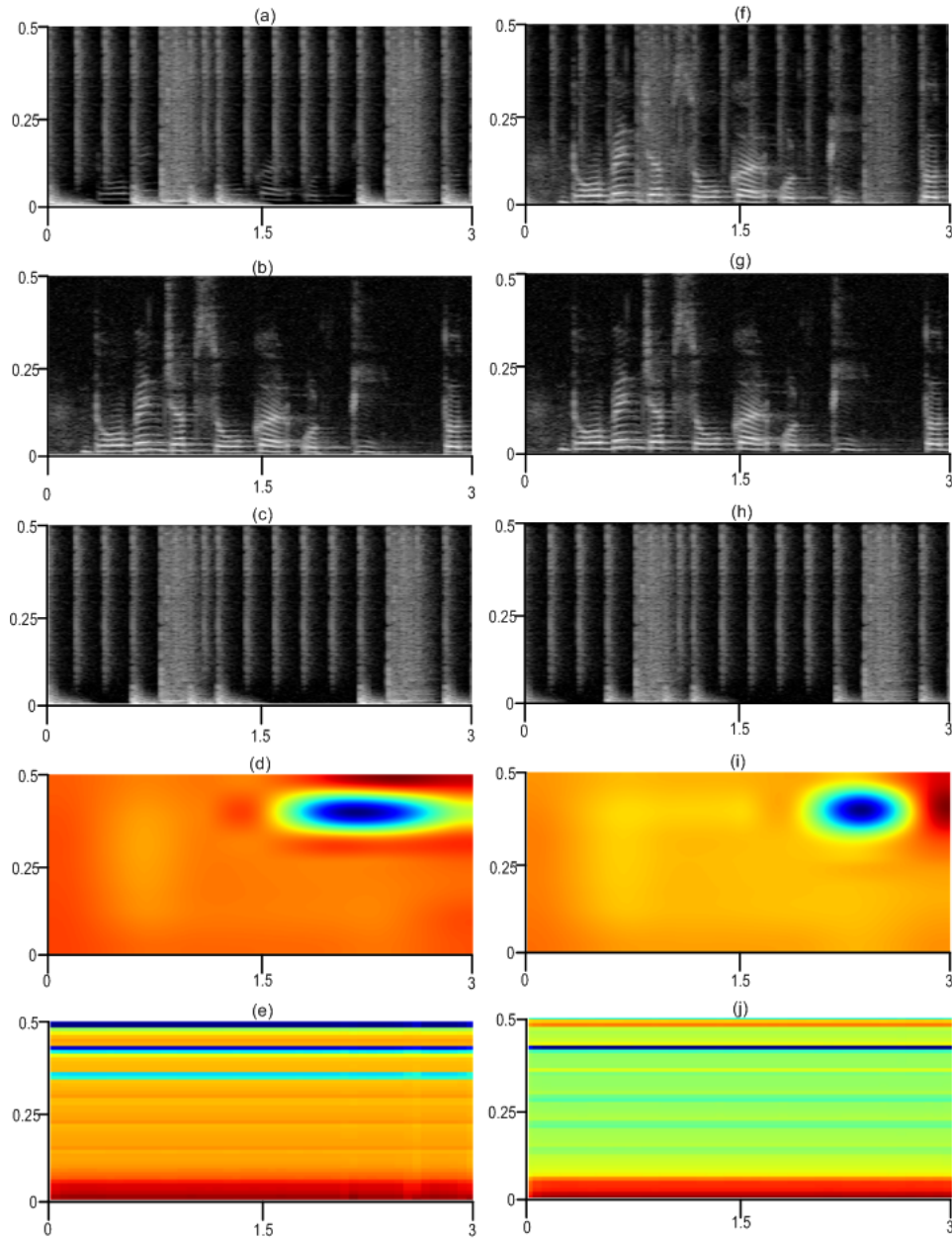


Figure 8: Effect of signal strengths (mixing Ratio 1 and Ratio 2) on the quality of the separated components during testing phase of the proposed system. First column shows the plots for mixing ratio set (0.8, 0.1) and second column for (0.1, 0.8). First row is for mixtures, second for separated vocals, third for separated instrumental, fourth for circular attention, and last row for average multiread attention.

than the interference energy. This results in a low SIR. Mathematically, the SI-SDR equation yield a high error term, thereby yielding a low SI-SDR. In the favourable case, where 0.8 vocal and 0.1 instrumental signals are mixed, the numerator of the equation for SI-SDR has a high value, the error

term is low, and this results in an extremely high value for SI-SDR. This theoretical behaviour is reflected visually in cleaner spectrograms and quantitatively in near ideal STOI and PESQ values.

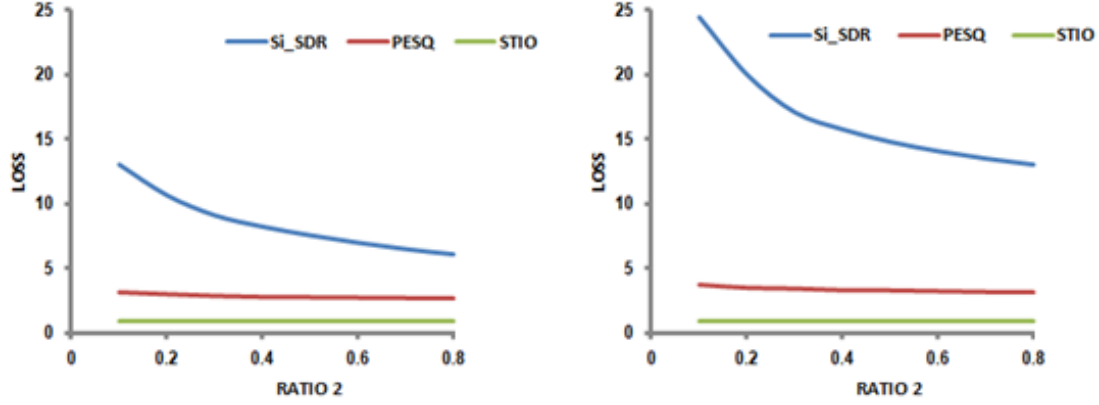


Figure 9: The figure illustrates the effect of signal strengths (mixing ratios) on the signal quality parameters (SI-SDR, PESQ, and STOI). Left plot is for RATIO 1 = 0.1, and RATIO 2 varied from 0.1 to 0.8. Right plot is for RATIO 1 = 0.8, and RATIO 2 varied from 0.1 to 0.8.

The effect of mixing ratios for the two constituent signals on their quality is shown in **FIGURE 9**. When R1 is fixed at 0.1 and R2 is varied from 0 to 0.8, SI-SDR decreases sharply from 13.09 dB to 6.16 dB, PESQ drops from 3.17 to 2.67, and STOI decreases slightly from 0.95 to 0.92. This indicates that increasing interference primarily degrades signal fidelity while intelligibility remains relatively robust. Conversely, when R1 is fixed at 0.8, SI-SDR remains significantly higher across all ratios, decreasing from 24.40 dB to 13.09 dB, while PESQ and STOI remain consistently high.

Table 1: Comparative analysis of performance of the proposed model with recently reported models.

Model Type	SI-SDR (dB)
CNN/Conv-TasNet [25]	10-15
LSTM/DPRNN [36]	15-20
Attention/Transformer [37]	18-24
Attention-Bi-LSTM [Proposed]	9.8

TABLE 1 compares the performance of the proposed model with recently reported models. While the proposed Attention-Bi-LSTM framework achieves an SI-SDR of 9.8 dB, numerically even lower than the results reported for large-scale Conv-TasNet, DPRN, Transformer-based architectures, this cannot be considered a metric comparison under these circumstances. The proposed model operates in the spectrogram domain and performs direct magnitude estimation on fixed-duration input data, while waveform reconstruction is performed using the mixture phase. This formulation poses more strict conditions on separation compared to purely time-domain models that jointly optimise magnitude and phase or make use of oracle phase information, for which SI-SDR gains

are inherently easier to obtain. Furthermore, the evaluation in this work includes extremely adverse mixtures, which contribute only around 10–20% of the total mixture energy from the vocal part. Under such low signal-to-interference conditions, SI-SDR inevitably saturates owing to unavoidable residual instrumental leakage, even when the isolated vocal signal remains perceptually clear. This is further in agreement with the high STOI values, approximately 0.90 - 0.99, and the derived PESQ scores, which indicate that the model is indeed proficient in maintaining intelligibility and perceptual structure, to the extent that this may be a higher priority than the actual separation of energies. Although some of the cited SI-SDRs fall within the best range, between 18 and 24 dB, this is, analogous to time-domain networks, which are well-suited for large multi-speaker speech data sets and are tuned with waveform-level losses, and great additional computational resources are made available under the computational power economy. These usually perform better in better environments and are not easily comparable to music-vocal separation, since these devices show higher complexity but do not allow any numerical comparisons. The performance of the proposed Attention-Bi-LSTM model depicted consistent trends rather than just high performance under ideal conditions. This will improve the perception of the region related to vocals by employing circular spatial attention and multi-head self-attention mechanisms; thus, both techniques can be given the freedom to improve only the relevant region for the vocal, thereby the robust intelligibility of the outputs produced by the spectrogram. This is to ensure that, when we bring it into practice in cases such as speech and music enhancement and processing, this method is considered very appropriate. Moreover, clarity turns is more relevant in this case as well. If we are considering the use of the mixture phase reconstruction system, and our system is considered as the mixture phase reconstruction system without special consideration of the energy of separation, it can be considered as achieved and to exhibit satisfactory behaviour within our desired SI-SDR range of 9-10 dB.

Further improvements, including phase-aware modelling, perceptual loss functions and hybrid time-frequency decoding, will enable the proposed architecture to achieve SI-SDR improvements that not only retain the existing strengths but also align with the objectives of the current work. This discovery also supported another theoretical finding regarding the theory of deep separation systems, which asserts that intelligibility is more robust to distortion than actual reconstruction accuracy and thus achieves optimal perception when the source in question is dominant. For this, it has been observed that, as confirmed with computational results of training and testing of this suggested method, it can effectively learn discriminative models of spectro-temporal representations and attention, and can cope with unfavourable situations. Also, as more epochs are given, more structural attention, less interference leakage, and more purity of the spectrogram will be achieved, as has been confirmed with the testing, securing high intelligibility and quality of perception even with severe masking, thus proving this system satisfactory and efficient, considering it an effective model for separation of the sound source, as has also been confirmed with theory.

It may also be noted that adapting a pre-trained DenseNet-121 image classification backbone requires resizing spectrograms to a fixed $224 \times 224 \times 3$ format, which can lead to a loss of fine-grained time-frequency resolution and may affect harmonic structure preservation. The use of DenseNet-121 was primarily motivated by its strong feature extraction capability, dense connectivity, and parameter efficiency, which enable stable training even with limited data. However, image-based architectures are not inherently optimized for audio signal characteristics. To address this concern, informal experiments using native audio architectures U-Net, BS-RoFormer, and DemucsV have also been conducted. The results indicate that, under the same dataset and experimental conditions, these models achieve performance comparable to our approach. This suggests that

the dataset complexity—particularly the use of synthetically generated mixtures with challenging mixing ratios—plays a significant role in limiting overall SI-SDR performance across architectures.

4. CONCLUSIONS AND FUTURE SCOPE

In this paper, a supervised deep learning-based architecture for learning vocal signal is proposed. In the proposed structure, the techniques of circular spatial attention, Bi-LSTM, DenseNet-121, and multi-head self-attention have been effectively combined to directly learn the source magnitude spectrograms from the given mixtures of audio signals. In the proposed structure, the manner of learning the removal of spectrograms, temporal learning, and global learning is effectively incorporated. Therefore, the proposed structure is capable of effectively addressing even the most complex specialized acoustic layout. Further, from the experimental analysis presented here, it is clear that the proposed structure can learn to improve through a transformation that approximates better perceptual enhancement. The experimental analysis of the attention pattern learned by the proposed model further confirms improvements in attention to significant time-frequency regions. This would further enhance its accuracy, accessibility, and separability. Furthermore, an objective analysis of the model across metrics such as SI-SNR, PESQ, STOI, etc., suggests that the model performs significantly well, regardless of whether the mixed signal is good or bad. This would imply greater robustness, meaning it can produce high-quality output even when the Signal-to-Interference ratio is low. If one were to look at this model from a methodological perspective, one could say that taking the direct approach to spectrogram estimation would overcome difficulties encountered with conventional time-frequency masking methods. Thirdly, the structure also offers itself for expansion in different ways in the future. One of the most important improvements in this model is the addition of phase-awareness and multiplex-domain learning, which could further reduce reconstruction errors and improve waveform accuracy.

The model is also generalizable across multi-source scenarios that employ more than two speakers or instruments. Furthermore, the implementation of perceptual loss methods and adversarial learning could improve subjective audio quality in addition to mean-squared error optimization. The implementation of real-time compression and model compression would also be a positive trend to follow for field applications. Lastly, domain transformation using self-supervised or weakly supervised learning might yield stronger results in novel acoustic environments; hence, the system could be more robust and usable for real-world audio recordings, unimpeded by any factors.

References

- [1] Févotte C, Vincent E, Ozerov A. Single-Channel Audio Source Separation With Nmf: Divergences, Constraints and Algorithms. *Audio source separation*. Springer. 2018:1-24.
- [2] Purwins H, Li B, Virtanen T, Schlüter J, Chang SY, et al. Deep Learning for Audio Signal Processing. *IEEE J Sel Top Signal Process*. 2019;13:206-219.
- [3] Gannot S, Vincent E, Markovich-Golan S, Ozerov A. A Consolidated Perspective on Multi-Microphone Speech Enhancement and Source Separation. *IEEE/ACM Trans Audio Speech Lang Process*. 2017;25:692-730.

- [4] Hyvarinen A, Karhunen J, Oja E. Independent Component Analysis. New York: John Wiley Sons. 2001.
- [5] Smaragdis P, Fevotte C, Mysore GJ, Mohammadiha N, Hoffman M. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Process Mag.* 2014;31:66-75.
- [6] Ephraim Y, Van Trees HL. A Signal Subspace Approach for Speech Enhancement. *IEEE Trans Speech Audio Process.* 1995;3:251-266.
- [7] Wang D, Chen J. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE/ACM Trans Audio Speech Lang Process.* 2018;26:1702-1726.
- [8] Boashash B. Time-Frequency Signal Analysis and Processing. MA USA: Academic press. 2015.
- [9] Stanković L, Daković M, Thayaparan T. A Real-Time Time-Frequency Based Instantaneous Frequency Estimator. *Signal Process.* 2013;93:1392-1397.
- [10] Salamon J, Gómez E, Ellis DP, Richard G. Melody Extraction From Polyphonic Music Signals: Approaches Applications and Challenges. *IEEE Signal Process Mag.* 2014;31:118-134.
- [11] Prabhu KM. Window Functions and Their Applications in Signal Processing. 1st ed. CRC Press. 2014.
- [12] Pachori RB. Basics of Time-Frequency Analysis. In: Time-frequency analysis techniques and their applications. CRC Press. 2023:22.
- [13] Deng L, Yu D. Deep Learning: Methods and Applications. *Found Trends Signal Process.* 2014;7:197-387.
- [14] Sarker IH. Machine learning: Algorithms, real-world applications and research directions. *SN Comput Sci.* 2021 May;2(3):160.
- [15] Huang PS, Kim M, Hasegawa-Johnson M, Smaragdis P. Deep Learning for Monaural Speech Separation. In 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE. 2014:1562-1566.
- [16] Wu B. Singing Voice Separation Based on U-Net Neural Network and Its Application. *Appl Comput Eng.* 2024;75:185-192.
- [17] Yu D, Li J. Recent Progresses in Deep Learning Based Acoustic Models. *IEEE/CAA J Autom Sin.* 2017;4:396-409.
- [18] Hochreiter S, Schmidhuber J. Long Short-Term Memory. *Neural Comput.* 1997;9:1735-1780.
- [19] Yoshioka T, Erdogan H, Chen Z, Allea F. Multi-Microphone Neural Speech Separation for Far-Field Multi-Talker Speech Recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2018:5739-5743.
- [20] Zapotoczny M, Pietrzak P, Łańcucki A, Chorowski J. Lattice Generation in Attention-Based Speech Recognition Models. In *Proc Interspeech 2019.* 2019:2225-2229.

- [21] Subakan C, Ravanelli M, Cornell S, Bronzi M, Zhong J. Attention Is All You Need in Speech Separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. 2021:21-25.
- [22] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention Is All You Need. Adv Neural Inf Process Syst. 2017;30.
- [23] Sawhney T, Sharma A, Abrol P, Kumar Lehana P, Chaahat, et al. Fingerprint Matching for Noisy and Distorted Patterns Using a Siamese Network With RESNET50 and Multihead Attention. IEEE Access. 2025;13:88047-88064.
- [24] Luo Y, Mesgarani N. Tasnet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE. 2018:696-700.
- [25] Luo Y, Mesgarani N. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. IEEE/ACM Trans Audio Speech Lang Process. 2019;27:1256-1266.
- [26] Luo Y, Chen Z, Yoshioka T. Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Speech Separation. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020:46-50.
- [27] Negro F, Muceli S, Castronovo AM, Holobar A, Farina D. Multi-Channel Intramuscular and Surface Emg Decomposition by Convolutional Blind Source Separation. J Neural Eng. 2016;13:026027.
- [28] Rangayyan RM, Krishnan S. Biomedical signal analysis. 3rd ed. John Wiley & Sons. 2024.
- [29] Boutaba R, Salahuddin MA, Limam N, Ayoubi S, Shahriar N, et al. A Comprehensive Survey on Machine Learning for Networking: Evolution, Applications and Research Opportunities. J Internet Serv Appl. 2018;9:1-99.
- [30] Jiang X, Bian GB, Tian Z. Removal of Artifacts From EEG Signals: A Review. Sensors. 2019;19:987.
- [31] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, et al. MIMIC-III, a Freely Accessible Critical Care Database. Sci Data. 2016;3:160035.
- [32] Albelwi SA. Deep Architecture Based on Densenet-121 Model for Weather Image Recognition. Int J Adv Comput Sci Appl. 2022;13.
- [33] Zacarias-Morales N, Pancardo P, Hernández-Nolasco JA, Garcia-Constantino M. Attention-Inspired Neural Networks for Speech Processing: A Systematic Review. Symmetry. 2021;13:214.
- [34] Subakan C, Ravanelli M, Cornell S, Grondin F, Bronzi M. Exploring Self-Attention Mechanisms for Speech Separation. IEEE/ACM Trans Audio Speech Lang Process. 2023;31:2169-2180.
- [35] Y. Yang, “An improved Hilbert-Huang transform method for stochast signals,” *IET Conference Proceedings*, vol. 2024, no. 20, pp. 91–97, Dec. 2024, DOI: 10.1049/icp.2024.4087.
Yang Y. An improved Hilbert-Huang transform method for stochastic signals. In IET Conference Proceedings CP897. IET. 2024;2024:91-97.

- [36] Kothandaraman M, Ramalingam B, Sheng K, Verma A, Dhagat U, et al. Audio-Visual Source Separation With Localization and Individual Control. PLOS One. 2025;20: e0321856.
- [37] Li K, Chen G, Sang W, Luo Y, Chen Z, et al. Advances in Speech Separation: Techniques Challenges and Future Trends. 2025. ArXiv preprint: <https://arxiv.org/pdf/2508.10830>