

Research Paper

**Audio Strips Network (ASNet) and Amalgamated Audio Features (A2F):
A Synergistic Approach for Audio Source Separation**Shunmugalingam Parvathi SAKTHIDEVI*, Chandrasekar DIVYA^{id}*Centre for Information Technology and Engineering, Manonmaniam Sundaranar University
Tirunelveli, India**Corresponding Author: spsakthidevi2000@gmail.com*Received May 29, 2025; revised December 10, 2025; accepted April 25, 2026;
available online April 29, 2026; version of record September 7, 2026; published issue September 23, 2026.*

Audio source separation refers to the procedure of decomposing a mixed audio signal into its constituent components. This technique enables numerous applications, including creative music production, educational tools, karaoke, transcription, and music analysis. Despite the recent success of deep learning-based source separation techniques, these techniques often do not perform very accurately and do not provide high-quality separation of sources when many contain complex combinations in their mixtures. Source separation techniques generally rely on temporal or spectral features for analysis, which does not fully capture the complex dynamics of audio signals. To address these limitations, this paper proposes the amalgamated audio features (A2F), a hybrid representation combining temporal and spectral features. Then, the audio strip network (ASNet) is proposed, a novel framework designed to achieve clean and precise separation of individual audio sources with enhanced performance. ASNet utilizes A2F to separate sources more effectively. The model is trained and evaluated on the MUSDB, DSD100, and MUSDB18-HQ datasets, a benchmark for music source separation, and its standard measures, such as the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR), are used to examine performance. ASNet achieves enhanced separation performance with SDR values of 12.63 for drums, 11.42 for vocal, 12.01 for bass, and 11.14 for other, and SIR values of 9.57 for drums, 9.61 for vocal, 9.66 for bass, and 9.67 for other. This advancement benefits musicians through high-quality remixing and creativity while aiding researchers in improving deep learning and hybrid audio processing models.

Keywords: feature extraction, amalgamated audio features (A2F), source extraction, audio strip network (ASNet).



Copyright © 2026 The Author(s).
This work is licensed under the Creative Commons Attribution 4.0 International CC BY 4.0
(<https://creativecommons.org/licenses/by/4.0/>).

1. INTRODUCTION

Audio signal processing (ZÖLZER, 2022) has expanded greatly in recent years, particularly in the domain of audio source separation: the process of extracting different individual constituent sources from a combined auditory mixture. Source separation is crucial in many areas, including music production (HUBER *et al.*, 2023), speech transcription (O'CONNELL, KOWAL, 2009), audio enhancement (CHUANG *et al.*, 2022) and restoration, hearing aids (SANDERS *et al.*, 2021), and forensic audio analysis (RENUKADEVI *et al.*, 2024). Even though some advancements have been made, creating realistic source separation remains difficult due to the fact that audio signals are often inherently complex and often exhibit overlapping frequencies, transient characteristics, dense harmonics, environmental noise, and reverberation, all of which can impair model performance.

Advances in the field of deep learning (DL) (ISSA, AL-IRHAYM, 2021), have transformed the source separation landscape by introducing powerful, data-driven approaches capable of learning complex patterns directly from raw or transformed audio inputs. Various techniques have achieved state-of-the-art performance in source separation.

ration tasks. For example, convolutional neural network (CNN) such as U-Net (DENG *et al.*, 2024), Conv-TasNet (DÉFOSSEZ *et al.*, 2021), SCNet (TONG *et al.*, 2024), DTTNet (CHEN *et al.*, 2024), and DenseNet (TAKAHASHI, MITSUFUJI, 2020) have shown great success in modeling spectral and spatial correlations. At the same time, recurrent neural networks (RNNs) (SUN *et al.*, 2021; LUO, YU, 2023), incorporating gated linear unit (GLU) (TONG *et al.*, 2024; DÉFOSSEZ *et al.*, 2019), and bidirectional long short-term memory (Bi-LSTM) networks (CHEN *et al.*, 2024; DENG *et al.*, 2024; DÉFOSSEZ *et al.*, 2021), have demonstrated efficacy in capturing long-range temporal dependencies within audio streams. While CNNs effectively capture spatial and spectral characteristics, RNNs are capable of modeling the temporal dependencies of audio segments while assigning increased computational weight to relevant acoustic regions. Nevertheless, both CNNs and RNNs face limitations, particularly when audio signals are highly overlapping or contaminated with noise.

In addition, there are many existing approaches that are computationally intensive, requiring great computational resources for training and inference, which limit their use in resource-constrained scenarios. Noise and reverberation further complicate these separation difficulties, undermining the efficacy of many current approaches. To address these challenges, a novel methodology is proposed that combines the strengths of DL with advanced feature extraction techniques. This methodology introduces hybrid features that integrate temporal and spectral cues, enabling a more accurate representation of the complex dynamics present in musical mixtures. Built upon this feature set, the proposed hybrid DL model achieves higher SDR and SIR scores across benchmark datasets, demonstrating superior separation quality and robustness compared to existing models. Its relevance is further reflected in practical applications such as music production, remixing, and audio processing research, where high-quality, efficient source separation is essential.

2. BACKGROUND ANALYSIS

This section reviews existing DL methodologies and feature extraction techniques, while addressing the challenges associated with these approaches. The gaps in current methods are identified to establish the foundation for the proposed framework.

2.1. ANALYSIS OF FEATURE EXTRACTION

Mel frequency cepstral coefficients (MFCCs) (HAMZA *et al.*, 2022) mimic human hearing by utilizing the fast Fourier transform (FFT) (RAJABY, SAYEDI, 2022), pre-emphasis, framing, a hamming window, Mel filterbanks, log-energy calculation, and the discrete cosine transform (DCT) to extract speech and audio features. However, they face challenges with computational efficiency and noise robustness. While effective under controlled conditions, MFCCs struggle with real-world noise, edge-device computation, and dynamic speech variations (REZAUL *et al.*, 2024). Similarly, cochlear implant strategies comparing the Crystallis method (128-point FFT) and spectral feature extraction (SFE) (512-point FFT + interpolation), show that while SFE improves spectral resolution it lacks validation in noise robustness, efficiency, and generalizability. Future work should explore DL integration and lightweight SFE variants for clinical scalability, as long-term acclimatization and real-world performance remain unassessed (ZHANG *et al.*, 2025). Furthermore, the 1-transistor-1-memristor (1T1R) neuromorphic system mimics biological neurons for audio feature extraction, achieving an accuracy range from 85.38 % to 95.91 % on the spiking Heidelberg digits (SHD) dataset but lacks real-world noise robustness testing. Future work should improve complementary metal-oxide-semiconductor (CMOS) compatibility, scalability beyond memristors, and generalization to diverse acoustic environments (WU *et al.*, 2024).

2.2. PERFORMANCE OF DL APPROACHES

A denoising autoencoder with U-Net and Bi-LSTM (DAEUBL) is applied to denoise multi-level random telegraph signals (RTS) (DENG *et al.*, 2024) contaminated with mixed noise. Unlike single-noise-focused models, the DAEUBL generalizes across various noise profiles and signal lengths via unified training and adaptive segmentation. However, it lacks real-world benchmark validation and a computational efficiency analysis, limiting

scalability insights and practical deployment trade-offs for complex RTS denoising (DENG *et al.*, 2024). Similarly, the sparse conformer network (SCNet) applies the short-time Fourier transform (STFT) and sparse downsampling (SD) to compress binaural audio spectrograms for source separation. It uses group normalization (GroupNorm), gated linear units (GLUs), and dual-path RNN to model sequence dependencies. Despite its success, research gaps for SCNet include the need for better high-frequency modeling, lower complexity, and improved generalization to unseen data (TONG *et al.*, 2024).

In the dilated dense block (D2Block) architecture, each feature map is processed with varying dilation factors to ensure better input coverage without blind spots. The deep dual-path densely connected network (D3Net) extends this architecture by stacking D2Blocks with a repeated dilation pattern, allowing flexible feature transformation across resolutions and improving computational efficiency with channel reduction mechanisms. However, certain shortcomings persist, including the potential for aliasing artifacts, increased model complexity due to dense connections, and challenges in maintaining effective receptive field coverage across layers (TAKAHASHI, MITSUFUJI, 2020). Furthermore, the dual-PATH TFC-TDF U-Net (DTTNet) uses time-distributed fully connected (TDF) layers, time-frequency convolutions (TFC), and an improved dual-path module (IDPM) for music source separation. GroupNorm and Bi-LSTM enhance sequence modeling across time and frequency. While its real-world generalization has been tested on patterns such as ‘wah-wah’ guitar filters and vocal chops beyond the MUSDB18-HQ benchmark, primary challenges remain. These challenges include better fine-tuning for distorted audio patterns and more robust separation across unseen acoustic environments (CHEN *et al.*, 2024).

Finally, DL architectures such as Demucs and the convolutional time domain audio separation network (Conv-TasNet) employ convolutional encoders/decoders, GLU and Bi-LSTM to achieve complete source separation in musical mixtures. Improvements, such as weight rescaling at initialization, the shift trick for time-equivariance, and sinc-based resampling, address performance bottlenecks. Nonetheless, key difficulties involve maintaining phase consistency, managing varying loudness levels across entire tracks, preserving time-shift invariance, and optimizing memory usage when working with these large DL models (DÉFOSSEZ *et al.*, 2021).

3. METHODOLOGY

This section reveals the innovative DL framework designed to address key challenges in music source separation, featuring two core components: A2F for effective feature extraction and ASNet for robust source separation. Each component presents innovative techniques to overcome the limitations of previous methods. The overall workflow of the system is displayed in Fig. 1 and starts with raw audio inputs, which next undergo preprocessing steps such as loading, normalizing, applying data augmentation and converting to generate spectrograms. Once the features are extracted via A2F, the representations are processed through ASNet and then postprocessed to reconstruct the final separated audio signals.

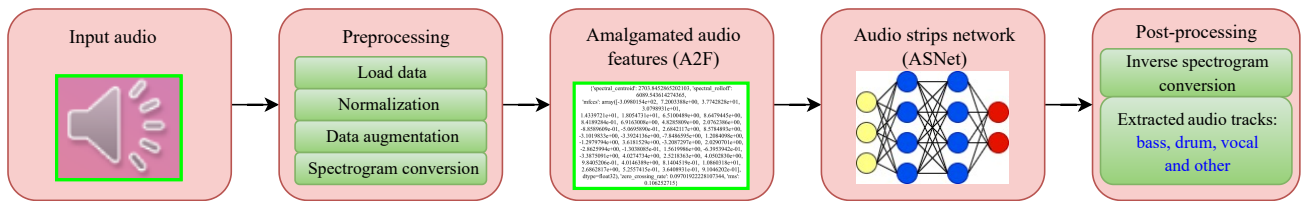


FIG. 1. End-to-end architecture of the proposed framework.

3.1. DATASET

To evaluate and train the proposed model for audio source separation, three benchmark datasets were employed: MUSDB, DSD100, and MUSDB18-HQ. The model was trained independently on each dataset to assess its separation performance and generalization ability across diverse musical compositions, genres, and recording environments. Figure 2 shows accuracy and loss curves for training and validation across the proposed ASNet model evaluated on the MUSDB18-HQ, DSD100, and MUSDB datasets. It includes accuracy vs. epochs and loss

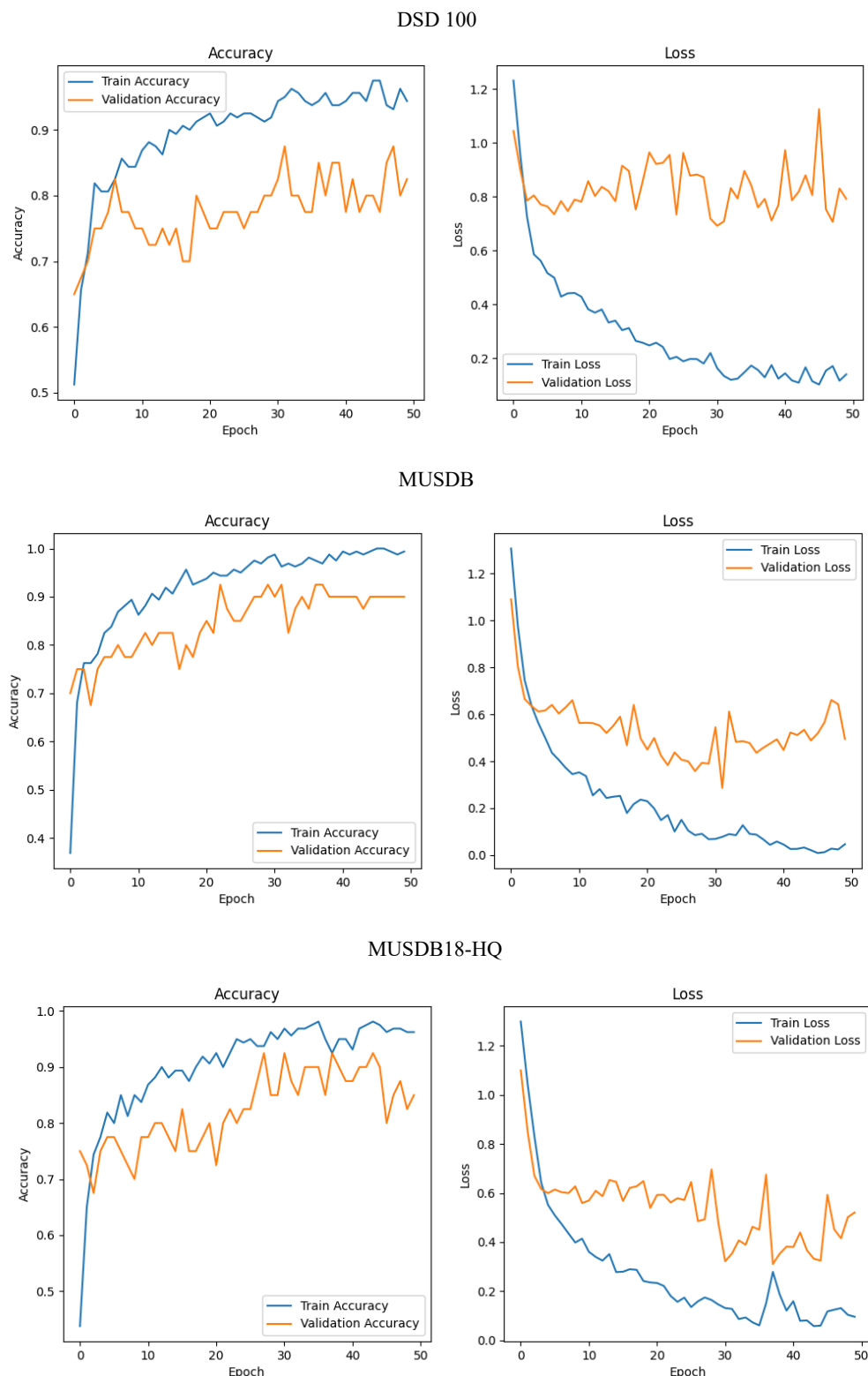


FIG. 2. Accuracy and loss visualization for training and validation across the various datasets.

vs. epochs plots, highlighting the convergence behavior, learning progress, and generalization capability of the ASNet.

The MUSDB, DSD100, and MUSDB18-HQ (RAFII *et al.*, 2019) datasets consist of 166, 100, and 150 full-length music tracks, respectively. The MUSDB and MUSDB18-HQ provide approximately 10 hours of audio data each,

while DSD100 includes tracks ranging from 2 min and 22 s to 7 min and 20 s. These datasets cover a wide range of musical genres and include isolated stems for drums, bass, vocals, and other instruments. All stereophonic audio signals are captured at 44 100 Hz. Each dataset is partitioned into an 80 % split for training and a 20 % part for testing, using A2F audio features as the inputs, while the corresponding source categories (separated signals) function as the ground-truth labels. To maintain reproducibility and consistency, a random state of 42 is used during data splitting. The method is trained on a batch size of 32 over 50 epochs to preserve available memory and limit the computational burden.

For the MUSDB dataset, the training accuracy steadily rises, reaching approximately 0.95, while the validation accuracy also improves and saturates around 0.85, indicating effective learning with a slight generalization gap. The training loss consistently decreases throughout the epochs. However, the validation loss begins to show mild fluctuations after epoch 30. This suggests minor overfitting in the later stages of training. For the DSD100 dataset, the training accuracy steadily improves, reaching nearly 0.98, while the validation accuracy rises and stabilizes around 0.85 after epoch 30. Both training and validation loss decrease during the initial epochs. However, the validation loss becomes highly variable in the later stages. This implies some generalization instability, probably due to the complexity of the dataset. Finally, on the MUSDB18-HQ dataset, a training accuracy is at approximately 0.96, while the validation accuracy increases quickly and has very high confidence values between 0.85 and 0.90, which indicate strong generalization. As optimization continues, the training loss decreases continuously; however, the validation loss varies widely after epochs 25. This indicates possible overfitting, or the fact that the model can be highly sensitive to validation samples.

In summary, all three datasets have learning behavior characterized by increasing accuracy and decreasing loss. The model achieves consistent performance at high levels of validation accuracy for all three datasets. While variability exists, particularly for validation loss, this provides a good opportunity to further improve performance using techniques such as early stopping, regularization, or data augmentation. Among the datasets, MUSDB18-HQ seems to perform best with the highest level of stable validation accuracy; more importantly, the dataset's audio quality appears to have meaningful effects on the model's generalization ability.

3.2. PREPROCESSING

The audio preprocessing pipeline starts with reading the audio file (such as WAV or MP3), returning a tuple that contains the audio time series (1D if mono or 2D if stereo) and the sample rate, which is fixed at 48 000 samples per second. Then, the time-domain waveform is transformed using the STFT, resulting in the signal transformed to a complex-valued time-frequency representation. If a user provides an MP3 file, it must first be decoded into pulse-code modulation (PCM), as MP3 is a compressed frequency-domain format; this decoding step restores the actual time-domain waveform required for further analysis. The magnitude of this complex STFT output is calculated to obtain a 2D array of values (frequency bins $\times$ time frames), representing the signal's energy distribution across frequency over time. To better align with human auditory perception, the magnitude spectrogram is further converted to the decibel scale.

Normalization techniques are applied to ensure numerical stability and consistency across inputs. In the time domain, peak normalization scales waveform amplitudes to the range $[-1, 1]$. In the frequency domain, log-magnitude or decibel scaling is used to manage dynamic range and improve learning efficiency. Data augmentation is employed across both the time and frequency domains to enhance generalization and address the challenge of limited labeled data. In the time domain, augmentations such as time warping and pitch shifting modify the waveform directly, while corresponding spectral transformations are applied to the decibel spectrograms in the frequency domain. These preprocessing steps, spanning both waveform and spectral representations, enable robust feature extraction and improved model generalization. This dual-domain approach is necessary for optimal source separation performance. Figure 3a shows the time-domain waveform of the audio signal, representing amplitude variations over time, while Fig. 3b and Fig. 3c display the magnitude spectrogram and decibel spectrogram, respectively, capturing the frequency content of the signal across time with different scaling.

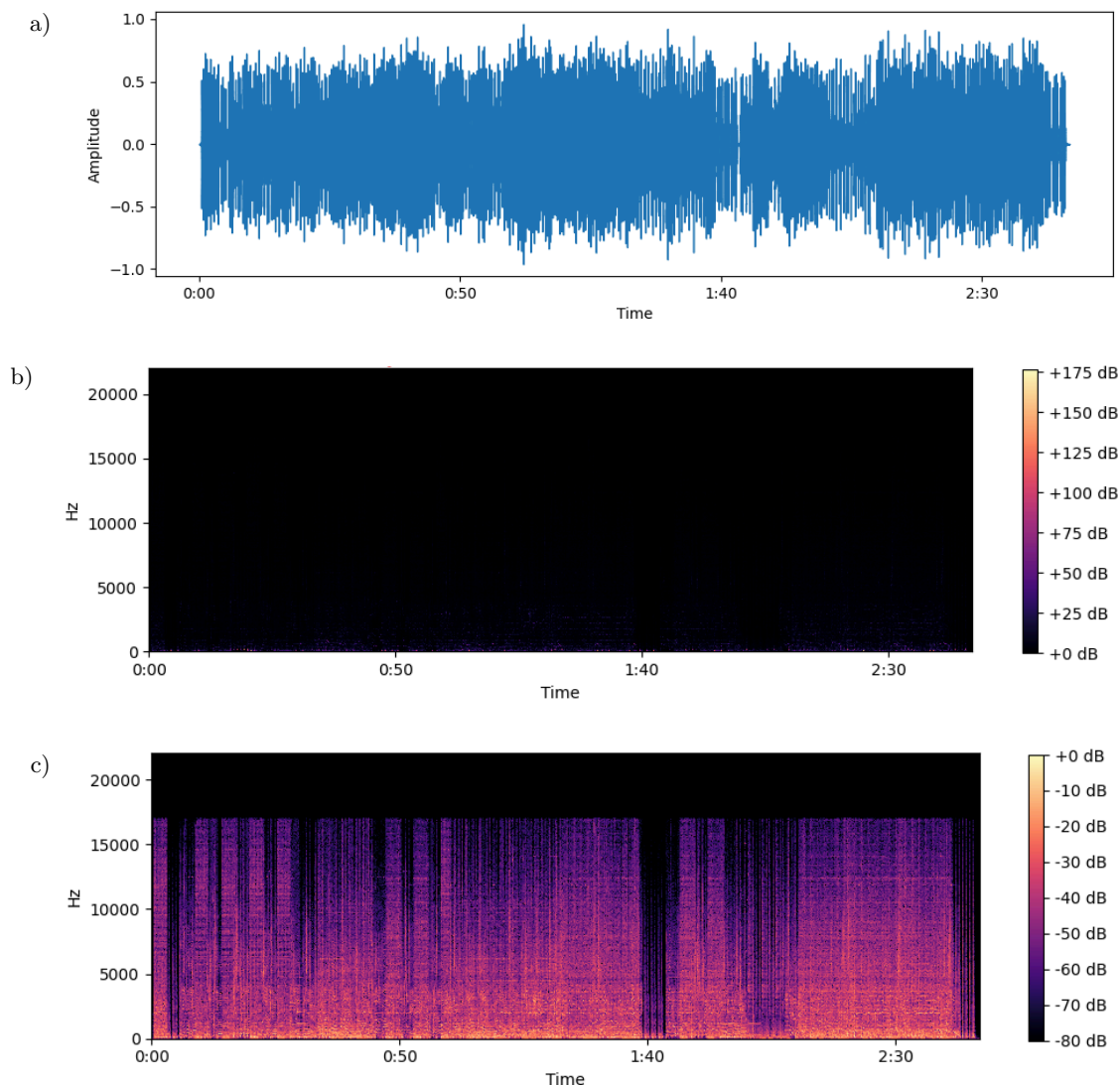


FIG. 3. Visual representations of the audio signal during preprocessing stages:
 a) time-domain waveform, b) time-frequency magnitude spectrogram, c) time-frequency decibel spectrogram.

3.3. A2F

A specialized feature extraction pipeline designed for audio source separation tasks is introduced and named A2F. Figure 4 illustrates the proposed A2F framework. This preprocessing system operates on mixed audio inputs to derive discriminative feature representations that enhance the downstream DL model's ability to separate and reconstruct individual sound sources. The A2F framework's hierarchical feature amalgamation process is constructed to capture both local spectro-temporal patterns and global acoustic characteristics, providing the separation network with optimized input features for improved source disentanglement performance.

The input to the framework is a mixed audio signal that contains a blend of multiple sources. This signal is processed in two parallel branches to extract distinct types of features.

Spectral features are derived from the STFT-based spectrogram, which provides information about the frequency distribution, harmonic structure, and timbre characteristics of each source. The spectral centroid indicates the brightness of the signal, helping the model distinguish high-frequency sources such as vocals from low-frequency sources such as bass and drums. The spectral roll off reflects how energy is distributed across the spectrum, enabling the separation of harmonic instruments that exhibit smoother energy decay from percussive instruments whose energy extends sharply into higher frequencies. MFCCs contribute by capturing perceptually

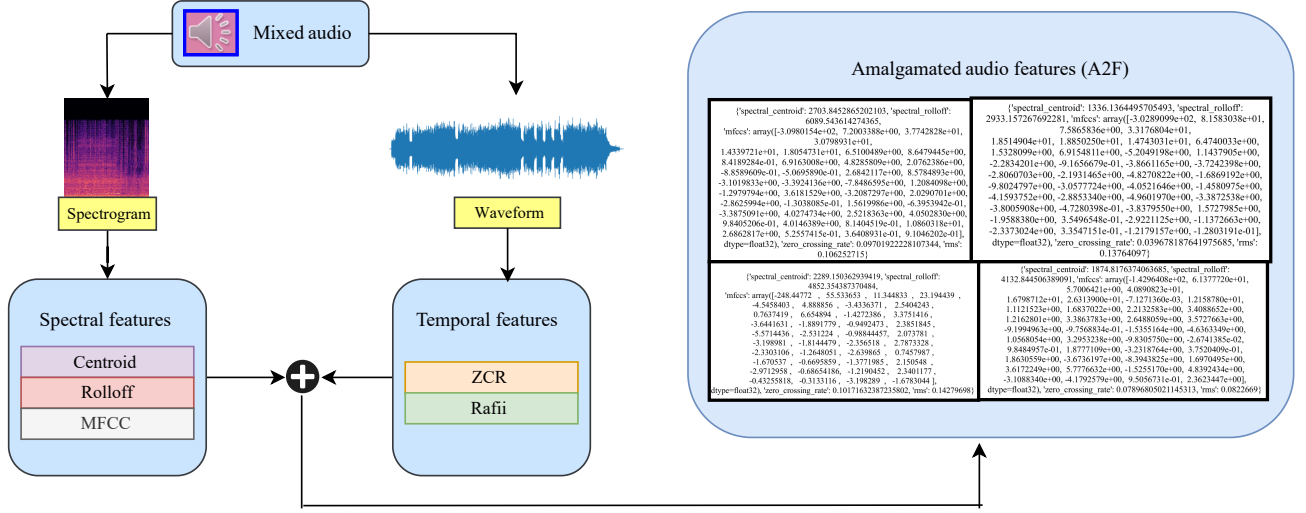


FIG. 4. Design of the proposed A2F framework.

relevant timbral information, allowing the network to identify and cluster tones from different instruments even when their frequency components overlap.

In parallel, temporal features extracted directly from the waveform, which provides insight into the signal's time-domain behavior. The zero-crossing rate (ZCR) helps differentiating between transient, noisy, or percussive sounds, which have high ZCR, and sustained or tonal sounds such as vocals or bass, which show a lower ZCR. Root-mean-square (RMS) energy captures the loudness contour of the signal, helping to segment active and inactive regions and maintain a balanced energy during source reconstruction. Together, these spectral and temporal features offer a holistic representation of the mixed audio, enabling the model to learn clearer boundaries between overlapping sources and ultimately improving the accuracy and robustness of the separation task.

3.3.1. FEATURE FUSION

A rich, multi-domain feature set is created by combining the temporal and spectral features. The resulting A2F representation, shown in the right panel of Fig. 4, contains the numerical values of the extracted spectral and temporal features. This matrix provides a complete set of input features for the source separation model, enabling a comprehensive analysis of the audio signal that specifically makes use of time-domain features to find rhythmic patterns and transients, alongside frequency-domain features to identify timbre and tonal patterns. Thanks to this robust feature representation, the model is better equipped to differentiate between overlapping sources in the mixed signal.

The A2F framework addresses some major limitations in current audio feature extraction techniques by combining noise-aware preprocessing, dual-domain feature extraction, multi-feature fusion, and integration of deep learning. With respect to noise, the proposed model demonstrates improved robustness to noise, better generalization capabilities, less reliance on handcrafted features alone. In the end, this enhances separation performance when used to separate sound events present in realistic audio environments.

3.4. ASNET

The proposed ASNet is composed of a downsampling module, a bottleneck layer, a bidirectional long short-term memory (Bi-LSTM), an upsampling module and a fully connected layer (FCL), as shown in Fig. 5.

The downsampling blocks are designed to gather the most significant musical elements from the incoming spectrogram data. When viewing the first downsampling block (downsample 1) the use of a 2D convolutional (Conv2D) layer expands the input data from 1 to 64 channels. Performing convolution using a 3×3 kernel allows the network to detect both temporal variations in harmony intensities and the harmonic relationship of notes

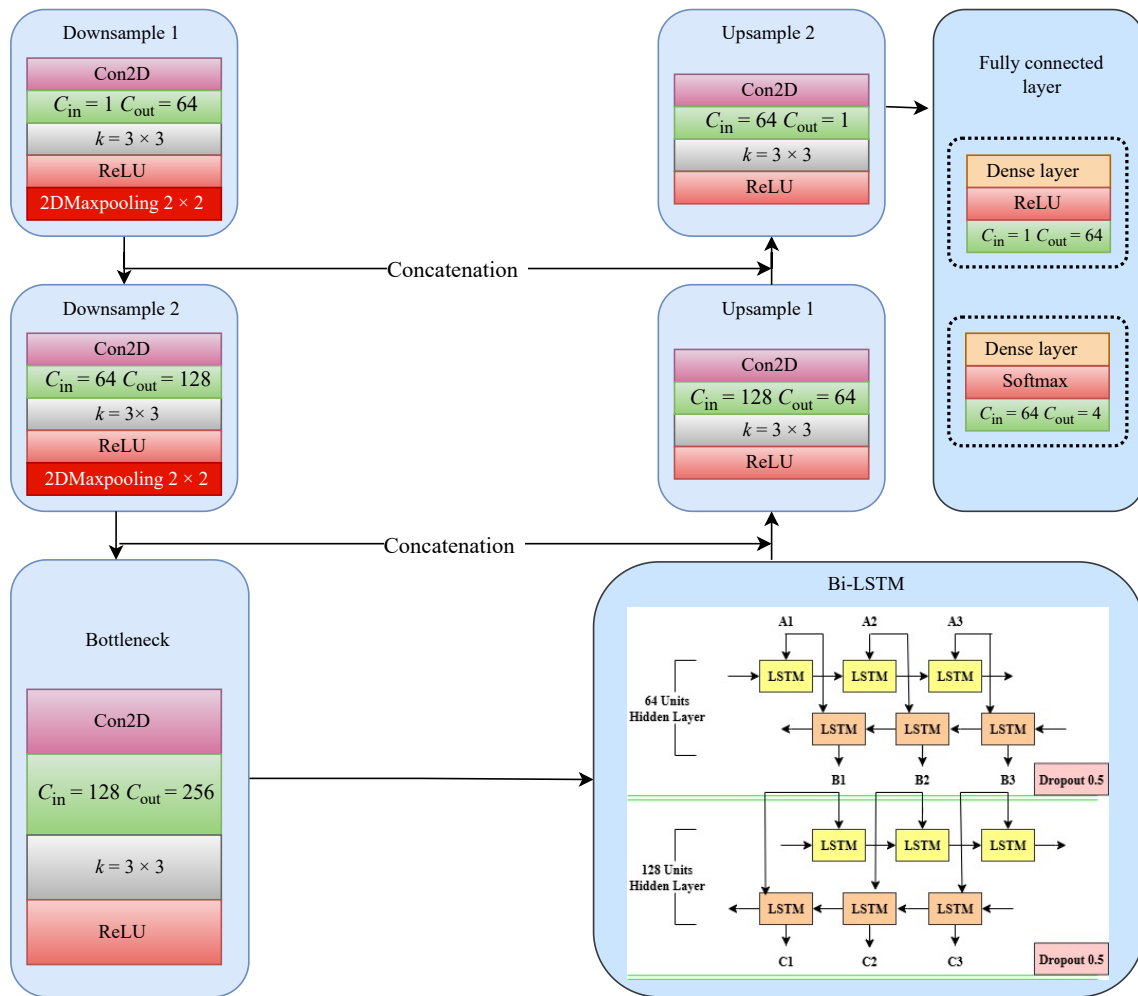


FIG. 5. Proposed ASNet architecture for audio source separation.

(timbral) to one another. Following a rectified linear unit (ReLU) activation, the maximum spectra are retained, creating a new representation with the highest harmonic (tonal) composition and instrument recognition. In the second block, downsample 2, the total number of filters is increased from 64 to 128, creating higher-level features that allow the network to learn additional patterns of timbre associated with certain instruments and notes in the music, as well as additional harmonics. Repeating this process using ReLU activation and max-pooling operations gradually reduces the data, allowing only the most significant musical features to remain, while all extraneous information such as noise is removed. Together, downsample 1 and downsample 2 allow the model to effectively establish and recognize the time-frequency relationships of music required for effective music extraction.

After the compressed features have been created, they are input into a bottleneck layer, which has the scope of expanding the depth of feature maps from 128 to 256. This layer allows the model to find highly abstract patterns in music that cannot be represented by other layers. The bottleneck layer also lets the model capture information related to more complex interactions between timbres, dense overlaps within harmonic structures, and instrument-specific signatures in spectra, leading to more precise separation of sounds. The use of a Conv2D-layer at this point lets the network apply wide-ranging time-frequency dependencies during encoding operations. Furthermore, the increased number of filters gives the network the flexibility to encode finer, smaller details such as vocal-harmonic characteristics, resonances in drums, overtones in bass, and textures from other sounds in the background. Additionally, the ReLU activation functions support the largest and sparsest high-energy musical features, therefore enhancing the detection of key harmonic points and tempo events or vice versa. In conclusion, the bottleneck layer acts as a high-level encoder within the overall structure of an end-to-end

network; it transforms the lower-level features of spectrograms into robust, rich and discriminative representations ready for subsequent use in recurrent multi-task layers, producing improved separation capabilities for each song's music and achieving improved music extraction capabilities.

Temporal modelling occurs at the Bi-LSTM block, which creates two bidirectional layers, containing 64 and 128 units, respectively. These two bidirectional layers within one block allow for capturing short-term transitions, such as note onsets and inflections that occur with voice, as well as longer strands of time, such as sustained harmonic notes, rhythmic patterns, and instrument continuities. This design is aimed at increasing the capability of the model to accurately represent complex, overlapping audio source patterns found in music where many different instruments coexist and interact throughout a time sequence. The use of a Bi-LSTM is critical because musical events rely not only on previous events in the past but also on future events, as this future context plays an important role in separating elements such as chord progressions, vocal phrasing or the sustain of notes. With the inclusion of a 0.5 dropout rate, it is also critical to provide strong regularization to prevent overfitting. The dropout rate will encourage the LSTM to identify effective time dependencies rather than memorizing actual patterns. Ultimately, this layer of the neural network block provides a strong temporal understanding that enables the model to separate, differentiate, and then extract audio sources.

Then, the purpose of the upsampling blocks is to gradually recover the source signal from its compressed representation by progressively increasing the size of the representation while improving the quality of the reconstructed musical information. The first upsampling block reduces the number of channels from 128 to 64 during using a Conv2D layer, so the model still has the basic harmonic and timbral information necessary for the reconstruction of the original audio while eliminating the redundant features created by the encoder. Furthermore, the addition of the ReLU activation layer helps highlight spectral elements such as dominant harmonics and transient signals. In the second upsampling block, reducing the number of channels from 64 to 1 helps the network focus on generating a clean, single-channel spectrogram while preserving the frequency characteristics of the instrument to correctly reconstruct the extracted source. The concatenation operation that occurs at this point allows the model to combine high-level semantic details learned in the deeper layers of the network with low-level spatial information from the early encoder layers. This integration helps the model recover fine frequency details, harmonic continuity, and timbral subtleties that were impacted by the downsampling process. The combination of filters and concatenation operations allows the network to increase the size of the audio representation, increase its quality, and build the target source audio in both time and frequency with sufficient temporal and spectral resolution.

The architecture ends with an FCL that provides the final classification by processing the high-level feature extracted throughout the network. The dense layer with 64 units generates a consolidated learned representation of the temporal and spectral feature sets. It achieves this by utilizing a ReLU activation to discover stronger features (dominant harmonics, transient peaks, and instrument-specific signatures), allowing for more accurate identification of instrument categories. The final softmax layer takes the output from the dense layer and converts it to probabilities, which enables the model to identify what portions of the time-frequency spectrogram belong to each instrument (vocals, drums, bass, and other). The classification-based output from the network is vital to providing targeted instrument assignments for audio extraction, thus facilitating a higher accuracy rate with respect to each separated audio component.

3.5. POST-PROCESSING

Post-processing serves as the critical final stage in audio source separation, transforming the model's spectrogram predictions into usable audio signals. As illustrated in Fig. 6, this involves two key conversions: (a) shows the inverse short-time Fourier transform (ISTFT) spectrogram (magnitude), visualizing frequency content over time after ISTFT reconstruction, while (b) displays the reconstructed audio waveform in the time domain, illustrating amplitude variations across time. This stage effectively closes the loop between neural network outputs and practical audio applications, enabling immediate playback or downstream analysis.

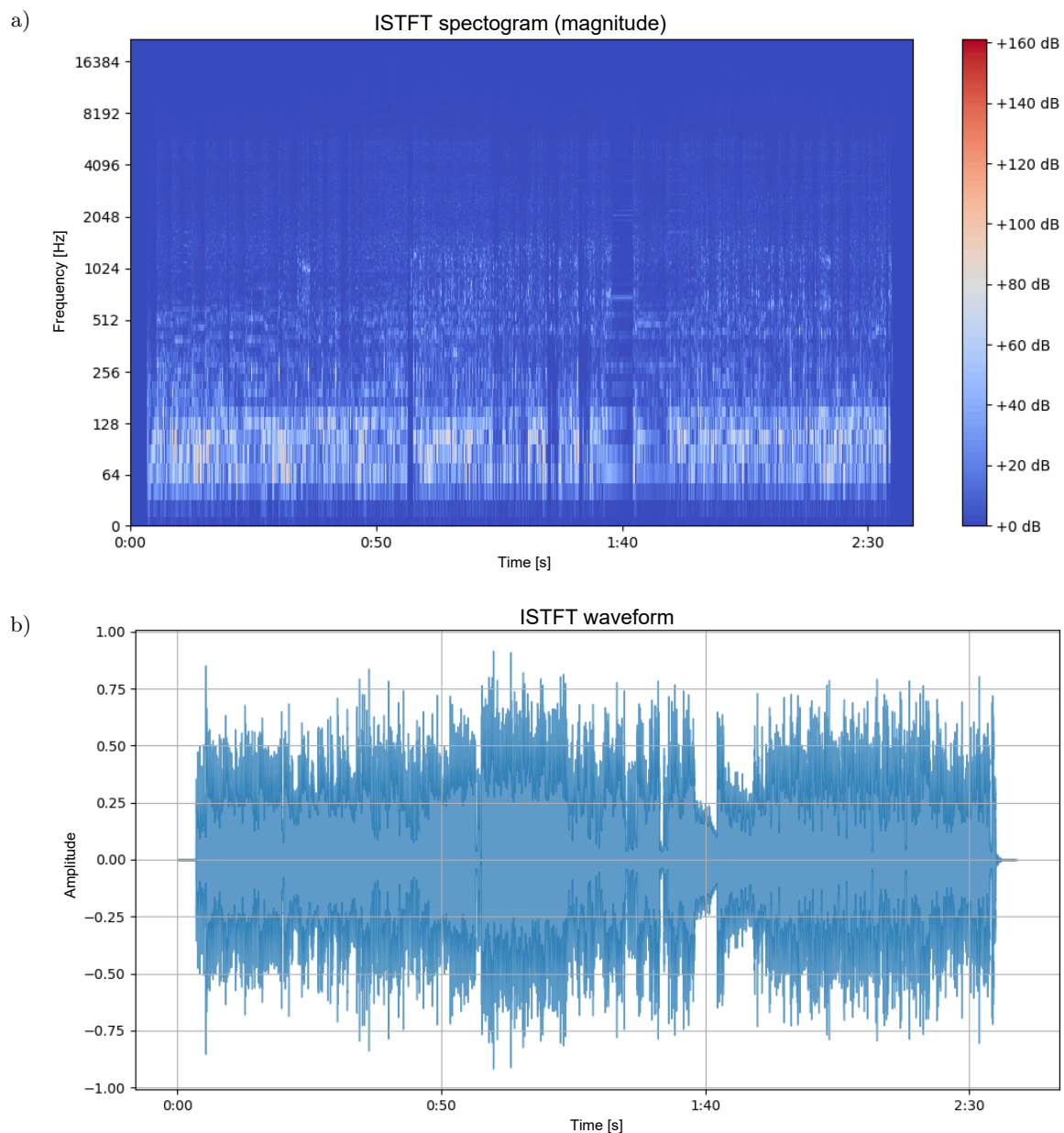


FIG. 6. Inverse spectrogram: a) time-frequency inverse magnitude spectrogram, b) time-domain inverse waveform.

4. EXPERIMENT

This section presents the experimental configuration designed to evaluate the proposed ASNet framework. It describes the primary aspects of model validation, the datasets used for training and validating, the evaluation metrics used to evaluate performance, and a comparison with existing methods, to show the improvements achieved by the proposed model. The objective of this experimental study is to demonstrate that the model not only performs well in source separation but also generalizes well across a variety of audio mixtures.

4.1. POST-PROCESSING

To objectively evaluate the effectiveness of the proposed audio source separation method, two widely accepted evaluation metrics are employed: signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR). When compared to the original actual sources, these measurements evaluate how well the separated audio sources perform.

SDR describes the fidelity of a separated audio signal by measuring the amount of distortion, artifacts, or errors that remain after separation. An increased SDR score indicates that there are very little distortion or ‘artifacts’ in the signal, and that the reconstructed audio is a clean and accurate representation of the original source. Conversely, a lower SDR score indicates poorer audio quality and implies that the separated signal contains some distortion, missing components, and/or visible artifacts that diminish clarity.

SIR quantifies how well a model separates the target source signal from any unwanted sounds associated with other instruments or vocal sounds in the original mixture. A higher SIR score indicates that the separated source has been effectively isolated from any competing audio sources and has clearly defined separation boundaries. A lower SIR score indicates that interference remains, with other instruments still audible in the extracted signal; therefore, this indicates that the separation was less efficient. These two metrics evaluate complementary aspects of performance: SDR evaluates the quality of the reconstruction, and SIR evaluates the strength of the separation of the source.

4.2. PERFORMANCE COMPARISON ANALYSIS

This section evaluates and compares the effectiveness of the proposed ASNet models in achieving the desired source separation outcomes. The comparison provides insights into optimal approaches and potential areas for improvement. The metrics SDR and SIR are both measured in decibels.

Table 1 displays the SDR and SIR scores achieved by the proposed ASNet model on three benchmark datasets: MUSDB, DSD100, and MUSDB18-HQ. The evaluation covers four audio source separation tasks: drums (D), vocals (V), bass (B), and other (O). The model demonstrates its best overall performance on the MUSDB18-HQ dataset, achieving the highest SDR and SIR scores among all evaluated datasets. Specifically, the SDR improved by approximately 5.4 % over MUSDB and 12.9 % over DSD100, while the SIR showed an improvement of about 2.7 % over MUSDB and 1.8 % over DSD100. The Conv2D block (256 filters) compresses and enriches features before passing them to the Bi-LSTM block, preserving fine details yielding a higher SDR. Finally, dropout and FCL layers prevent overfitting, boosting model’s generalization capacities across datasets.

TABLE 1. Evaluation of the ASNet model across various benchmark datasets.

Metric	MUSDB				DSD100				MUSDB18-HQ			
	D	V	B	O	D	V	B	O	D	V	B	O
SDR [dB]	11.98	10.73	11.21	10.81	11.19	10.64	10.33	10.68	12.63	11.42	12.01	11.14
SIR [dB]	9.27	9.52	9.51	9.51	9.55	9.46	9.59	9.11	9.57	9.61	9.66	9.67

Descriptive statistics such as the mean, median, and standard deviation (SD) reveal different but complementary dimensions of model behavior when evaluating audio separation, as these metrics provide a snapshot of the model’s overall separation capability, a representation of typical separation performance when outliers or other unusual audio samples exist, and an opportunity to assess how consistently well the model separates multiple sample types (i.e., either as a result of the presence of several high-performing samples or as a result of consistently high separation accuracy for many samples). Together, the mean and median provide insight into model strength. While the mean provides an average value of separation success across all samples tested, the median provides a stable assessment of typical audio separation when outliers or other unusual audio samples are present. A third metric, the SD serves as another indicator of separation consistency by providing a measure of variability in the model’s performance among multiple recordings. A low SD value implies reliable separation performance, while a high SD value signifies sporadic performance or inconsistent fidelity or separation across samples. Table 2 illustrates these results across the MUSDB, DSD100, and MUSDB18-HQ datasets: the MUSDB dataset has closely aligned mean and median SDR values, and moderate, low variability in the mean values; the DSD100 dataset has a small range of SDR values with no outliers and a much lower SD value; finally, the MUSDB18-HQ dataset has a very high average SDR value of 11.80 dB and a larger SDR SD value (0.570), indicating high level of separation quality alongside consistent performance, as demonstrated by its very low SIR SD

TABLE 2. Descriptive statistics of SDR and SIR metrics across the MUSDB, DSD100, and MUSDB18-HQ datasets.

Dataset	Mean		Median		Standard deviation	
	SDR [dB]	SIR [dB]	SDR [dB]	SIR [dB]	SDR [dB]	SIR [dB]
MUSDB	11.18	9.45	11.01	9.51	0.495	0.105
DSD100	10.71	9.43	10.66	9.51	0.309	0.190
MUSDB18-HQ	11.80	9.63	11.72	9.64	0.570	0.040

value (0.040). The combination of mean, median, and SD provides a detailed, multi-angle view of the model's ability to separate audio accurately across diverse datasets.

Utilizing SDR and SIR across four sources, Table 3 compares the ASNet's performance under various channel configurations (128, 256, and 512) on the MUSDB18-HQ dataset. At 11.80 dB average SDR, the 256-channel configuration produced the highest average SDR value, followed by the 512-channel configuration at the average SDR value of 6.86 dB and the 128-channel configuration at the average SDR value of 5.03. This represents an approximate 134.6% improvement compared to the SDR using 128 channels, and approximately 71.9% increase over the 512-channel setup. SIR values were comparable across all channels with slight variation, indicating no difference in interference suppression regardless of channel amount. These results demonstrate that a 256-channel configuration provides the most effective balance for source separation performance in ASNet.

TABLE 3. Implementation results of the ASNet architecture based on channel-wise configurations.

Channel count	SDR				SIR			
	B [dB]	D [dB]	V [dB]	O [dB]	B [dB]	D [dB]	V [dB]	O [dB]
128 channels	5.01	4.22	6.23	4.66	9.33	9.29	9.33	9.22
256 channels	12.01	12.63	11.42	11.14	9.57	9.61	9.66	9.67
512 channels	6.38	6.97	8.01	6.09	9.48	9.52	9.47	9.42

Table 4 compares the proposed ASNet model with several existing DL approaches for audio source separation on the MUSDB18-HQ dataset, based on SDR for four sources along with total network parameter count. Despite having a significantly smaller parameter count ($0.31 \text{ M} \times 4$), ASNet outperforms all other models in SDR across all separated source types, achieving 11.42 dB (V), 12.63 dB (D), 12.01 dB (B), and 11.14 dB (O). In contrast, larger models such as Res-U-Net ($102 \text{ M} \times 4$) and Mel-Roformer (84.2M) deliver lower SDR, indicating less accurate source separation. Importantly, ASNet is an interesting case that is able to outperform heavily parameterized architectures, such as BSRNN, DTT, and Hybrid Demucs, when considering complexity and computational efficiency. This indicates that the specifications ASNet must give up for its lightweight design actually favor performance in source separation. All of this proves why ASNet is a better model, offering high performance and reduced computational complexity.

TABLE 4. Performance comparison of the proposed ASNet against existing DL approaches based on SDR and parameter count.

Model	Params	V	D	B	O
Res-U-Net (KONG <i>et al.</i> , 2021)	$102.0 \text{ M} \times 4$	8.98	6.62	6.04	5.29
Mel-Roformer (WANG <i>et al.</i> , 2023)	84.2 M	11.21	9.91	9.64	7.81
Hybrid Demucs (DEFOSSEZ <i>et al.</i> , 2021)	$83.6 \text{ M} \times 4$	8.13	8.24	8.76	5.59
Band-split recurrent neural network (BSRNN)	$37.6 \text{ M} \times 4$	10.01	9.01	7.22	6.70
DTT (CHEN <i>et al.</i> , 2024)	$5.0 \text{ M} \times 4$	10.12	7.74	7.45	6.92
ASNet (proposed)	$0.31 \text{ M} \times 4$	11.42	12.63	12.01	11.14

4.3. DISCUSSION

To quantify the computational performance regarding execution speed, latency, throughput, and memory consumption of the audio source separation model, we performed tests using an NVIDIA RTX 3090 GPU and an Intel Xeon Silver 4210 CPU in conjunction with the Pytorch library. We also utilized Python libraries such as

Librosa and soundfile for the preprocessing and inference stages of the ASNet model. This approach allowed us to maintain consistency in input formatting and minimize I/O bottlenecks by providing high-quality audio data and an efficient method for loading audio files, resampling, and generating spectrograms. On average, ASNet processes a 5-second audio sample in 42.7 ms on the RTX 3090 and 315.4 ms on the Xeon 4210, yielding a throughput rate of 23.4 samples/second (audio signals) on the RTX 3090 vs. 3.1 samples/second on the Xeon 4210. Maximum memory usage for the audio source separation model was 2.6 GB on the RTX 3090 vs. 4.2 GB on the Xeon 4210, indicating that the increased speed, efficiency, and suitability for real-time music separation tasks provided by utilizing a GPU in conjunction with Librosa and soundfile for audio processing far outweigh the additional memory resources required by using a CPU.

The ASNet model is trained on three commonly used benchmark datasets: MUSDB18-HQ, MUSDB, and DSD100 with fixed training and validation splits. All models were validated within the same experimental configuration for fair comparison with the baselines. Source separation performance is measured using track-wise SDR and SIR, calculated using standard formulations obtained from the original and separated audio signals. Instead of using the museval framework, the evaluation process was managed through custom metric implementations to allow for complete control of the evaluation process, consistency across all datasets, and ease of integration into the proposed ASNet pipeline. The final performance scores are the average SDR and SIR for each track across the entire test dataset. To allow for comparison and replication of previous work, the baseline results are either reproduced or cited according to the same evaluation protocol established for the ASNet model evaluation compared with other models.

The proposed A2F framework outperforms baseline feature extraction methods by integrating both spectral (STFT, spectral centroid, spectral roll-off, MFCC) and temporal (ZCR, RMS) features. Unlike single-domain methods, A2F's hierarchical fusion captures fine-grained spectral-temporal cues and global acoustic patterns simultaneously. This multi-domain representation significantly enhances source separation accuracy compared to traditional feature pipelines. ASNet via a Bi-LSTM leverages bidirectional sequential modeling to capture temporal dependencies, improving continuity and accuracy in overlapping audio sources. Without the Bi-LSTM, the model relies solely on convolutional features, limiting its ability to learn long-range temporal relationships. The integration of the Bi-LSTM thus significantly enhances source separation performance compared to a purely CNN-based ASNet. Future work will focus on extending ASNet to real-time audio source separation by optimizing inference latency and reducing computational overhead. Model compression and lightweight deployment strategies will be explored to enable integration into streaming and live music applications.

The aim of performing a subjective evaluation, in addition to calculating objective SDR and SIR metrics, was to evaluate the quality of the separated audio recordings as judged by human listeners. In this evaluation, the proposed ASNet model was compared against two standard methods frequently used in music source separation: DTT and BSRNN. The results of the subjective evaluation indicate that the proposed ASNet model achieved substantially higher perceptual quality and naturalness than the two baseline methods on average. The majority of listeners reported fewer audible musical artifacts, better preservation of timbre, and improved temporal continuity in the output from ASNet when processing mixed audio samples consisting of many overlapping musical sources compared to the other two models. Therefore, the difference between the objective and subjective evaluations provides evidence of a strong correlation between objective improvements and perceived audio quality, indicating that the proposed method is well-suited for practical applications.

4.3.1. LIMITATIONS

The learning curves show that the validation loss fluctuated during the early and late stages of training, with the MUSDB and MUSDB18-HQ datasets showing the greatest fluctuation. It appears that these validation samples may have been somewhat sensitive or biased with respect to the training data. This could suggest that there was some degree of overfitting occurring despite applying regularization techniques such as dropout. Although the overall results still provide a good level of confidence in its accuracy, further advances can likely be made using more sophisticated methods, such as adaptive early stopping, curriculum learning, and dataset-aware augmentation.

4.3.2. SCOPE FOR FUTURE WORK

The goal of the strategies just mentioned is to improve the stability of convergence during model training and decrease the oscillation of the validation loss while maintaining the complexity of the model.

Another limitation of the current framework is that it relies only on using audio spectrogram representations. Although they work well, the use of audio-only representations may limit the ability of the model to separate sources that overlap heavily in very difficult audio environments. Future improvements may come from incorporating adaptive learning techniques and trying out different types of inputs, such as audio-visual fusion, to increase both accuracy and robustness. For instance, integrating information about the movements of instruments or the actions of performers could significantly help in separating overlapping audio sources.

Future research will extend beyond symbolic music representations to convert extracted source signals into MIDI format for the automated generation of improved musical scores. This expanded system will include pitch detection, onset-offset estimation, and note-segmentation models that work together with the separated source files to create a single pipeline for audio-to-score conversions. With this type of extension, there will be a seamless transition from audio source separation to automatic music transcription. Thus, the proposed framework could be utilized in various applications, such as providing educational opportunities for learning about music, digitally preserving music through archiving, analyzing music from an academic perspective, and providing assistance to musicians in compositional and performance endeavors.

In order to improve the fidelity of transcription and robustness, other avenues of exploration will involve incorporating transformer-based temporal modeling, perceptual loss functions, and the application of cross-domain approaches and learning techniques for joint audio and symbolic musical representation. Furthermore, one avenue of continued exploration is in the field of near real-time deployments and assessments of live and/or noisy musical/audio recordings.

5. CONCLUSION

The proposed ASNet, integrated with the A2F framework, demonstrates a significant advancement in audio source separation. By effectively combining spectral and temporal features, ASNet achieves superior separation quality. The architecture adopts a U-Net-inspired encoder-decoder structure to extract hierarchical feature representations and preserves spatial resolution through skip connections. This architecture allows the network to maintain crucial audio details while learning abstract features at multiple scales. Additionally, to enhance the modeling of long-range temporal dependencies, ASNet incorporates Bi-LSTM layers. These layers enable the network to capture both past and future contextual information. Notably, the 256-channel configuration of ASNet yields the highest SDR and SIR, despite having a considerably smaller parameter count. This effectiveness highlights ASNet's potential for implementation on devices with limited resources and applications that run in real time. Future improvements may also include incorporating adaptive learning strategies, as well as evaluating different modalities of input (e.g., adding audio-visual cues), to improve separation accuracy and robustness. Also, applying quantization-aware training could allow ASNet to be deployed on edge devices without a loss in performance.

FUNDINGS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHORS' CONTRIBUTIONS

Shunmugalingam Parvathi Sakthidevi – data analysis, writing, review, and editing. Chandrasekar Divya – supervision.

ACKNOWLEDGMENTS

This work was supported by the Manonmaniam Sundaranar University, Centre for Information Technology and Engineering, Tirunelveli, Tamil Nadu, India.

REFERENCES

1. CHEN J., VEKKOT S., SHUKLA P. (2024), Music source separation based on a lightweight deep learning framework (DTTNET: DUAL-PATH TFC-TDF UNET), [in:] *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 656–660, <https://doi.org/10.1109/ICASSP48485.2024.10448020>.
2. CHUANG S.Y., WANG H.M., TSAO Y. (2022), Improved lite audio-visual speech enhancement, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **30**: 1345–1359, <https://doi.org/10.1109/TASLP.2022.3153265>.
3. DÉFOSSEZ A. (2021), Hybrid spectrogram and waveform source separation, *arXiv*, <https://doi.org/10.48550/arXiv.2111.03600>.
4. DENG B., YANG H., KIM N.Y. (2024), A denoising autoencoder based on U-Net and bidirectional long short-term memory for multi-level random telegraph signal analysis, *Engineering Applications of Artificial Intelligence*, **135**: 108685, <https://doi.org/10.1016/j.engappai.2024.108685>.
5. HAMZA A. *et al.* (2022), Deepfake audio detection via MFCC features using machine learning, *IEEE Access*, **10**: 134018–134028, <https://doi.org/10.1109/ACCESS.2022.3231480>.
6. HUBER D.M., CABALLERO E., RUNSTEIN R. (2023), *Modern Recording Techniques: A Practical Guide to Modern Music Production*, Focal Press.
7. ISSA R.J., AL-IRHAYM Y.F. (2021), Audio source separation using supervised deep neural network, *Journal of Physics: Conference Series*, **1879**(2): 022077, <https://doi.org/10.1088/1742-6596/1879/2/022077>.
8. KONG Q., CAO Y., LIU H., CHOI K., WANG Y. (2021), Decoupling magnitude and phase estimation with deep ResUNet for music source separation, *arXiv*, <https://doi.org/10.48550/arXiv.2109.05418>.
9. LUO Y., YU J. (2023), Music source separation with band-split RNN, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **31**: 1893–1901, <https://doi.org/10.1109/TASLP.2023.3271145>.
10. O'CONNELL D.C., KOWAL S. (2009), Transcription systems for spoken discourse, [in:] *The Pragmatics of Interaction*, pp. 240–254, John Benjamins Publishing Company.
11. RAFII Z., LIUTKUS A., STÖTER F.R., MIMILAKIS S.I., BITTNER R. (2019), MUSDB18-HQ-an uncompressed version of MUSDB18, *Zenodo*, <https://doi.org/10.5281/zenodo.3338372>.
12. RAJABY E., SAYEDI S.M. (2022), A structured review of sparse fast Fourier transform algorithms, *Digital Signal Processing*, **123**: 103403, <https://doi.org/10.1016/j.dsp.2022.103403>.
13. RENUKADEVI P., JOHN S., SHIVANI N. (2024), Forensic science: AI-powered image and audio analysis, [in:] *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 1519–1525, <https://doi.org/10.1109/ICOSEC61587.2024.10722068>.
14. REZAUL K.M. *et al.* (2024), Enhancing audio classification through MFCC feature extraction and data augmentation with CNN and RNN models, *International Journal of Advanced Computer Science and Applications*, **15**(7): 37–53, <https://doi.org/10.14569/IJACSA.2024.0150704>.
15. SANDERS M.E., KANT E., SMIT A.L., STEGEMAN I. (2021), The effect of hearing aids on cognitive function: a systematic review, *PLoS One*, **16**(12): e0261207, <https://doi.org/10.1371/journal.pone.0261207>.
16. SUN C. *et al.* (2021), A convolutional recurrent neural network with attention framework for speech separation in monaural recordings, *Scientific Reports*, **11**(1): 1434, <https://doi.org/10.1038/s41598-020-80713-3>.

17. TAKAHASHI N., MITSUFUJI Y. (2020), D3Net: densely connected multidilated DenseNet for music source separation, *arXiv*, <https://doi.org/10.48550/arXiv.2010.01733>.
18. TONG W. *et al.* (2024), SCNet: sparse compression network for music source separation, [in:] *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1276–1280, <https://doi.org/10.1109/ICASSP48485.2024.10446651>.
19. WANG J.-C., LU W.-T., WON M. (2023), Mel-band roformer for music source separation, *arXiv*, <https://doi.org/10.48550/arXiv.2310.01809>.
20. WU X., DANG B., ZHANG T., WU X., YANG Y. (2024), Spatiotemporal audio feature extraction with dynamic memristor-based time-surface neurons, *Science Advances*, **10**(14): eadl2767, <https://doi.org/10.1126/sciadv.adl2767>.
21. ZHANG Y. *et al.* (2025), Comparison of performance for cochlear-implant listeners using audio processing strategies based on short-time fast Fourier transform or spectral feature extraction, *Ear and Hearing*, **46**(1): 163–183, <https://doi.org/10.1097/AUD.0000000000001565>.
22. ZÖLZER U. (2022), *Digital Audio Signal Processing*, John Wiley & Sons.