MUSIC SOURCE SEPARATION USING DEEP U-NET ARCHITECTURES

Stefano Polimeno, Riccardo Corà, Riccardo Moschen

ABSTRACT

Music source separation — the task of isolating individual sound sources (such as vocals, drums, bass, and other instruments) from a mixture audio signal — has become an essential problem in the field of music information retrieval (MIR), with applications spanning music production, remixing, karaoke generation, and computational music analysis. Despite its usefulness, the task remains fundamentally challenging due to overlapping frequency content, intricate source interactions, and the diversity of musical genres and recording techniques.

Over the past decade, significant progress in source separation has been achieved through deep learning. In particular, the U-Net architecture, originally proposed for biomedical image segmentation by Ronneberger et al. [1], has proven highly effective for spectrogram-based audio separation tasks. U-Net models are characterized by a symmetric encoder-decoder structure with skip connections, which allow the network to retain fine-grained temporal and spectral information across layers. This architectural design has been successfully adapted to music source separation by Jansson et al. [2], who demonstrated that deep U-Nets trained on spectrogram representations can learn robust time-frequency masks for vocal extraction, yielding competitive results against traditional signal processing methods.

In response to the growing interest in supervised learning approaches, standardized datasets such as MUSDB18 [3] have emerged as benchmarks for the task. The MUSDB18 corpus consists of 150 professionally mixed music tracks with ground truth isolated stems for vocals, drums, bass, and other accompaniment. This dataset has become the de facto standard for training and evaluating deep source separation models, enabling fair comparison across studies and driving reproducible research in the field.

Recent advancements in source separation have focused on improving model robustness and separation quality, as measured by objective metrics like Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifacts Ratio (SAR) [4]. Methods such as Demucs [5], which operate directly in the time domain using convolutional encoder-decoder structures with BiLSTMs and gated activations, have achieved state-of-the-art performance. Similarly, models incorporating attention mechanisms and complex ratio masking have shown promise in enhancing the interpretability and fidelity of separated sources.

The primary goal of this project is to design, implement, and evaluate an enhanced U-Net-based system for music source separation using the MUSDB18 dataset. The system aims to isolate one or more target sources from a stereo mixture input, the model incorporates architectural enhancements such as attention blocks and normalization strategies (e.g., InstanceNorm) to improve convergence and generalization. Furthermore, the project evaluates the impact of loss functions, input representations, and data augmentation techniques on separation performance.

By replicating and extending U-Net-based approaches within a reproducible experimental framework, this project contributes to the ongoing development of robust and interpretable deep learning models for source separation. The evaluation will be conducted using both quantitative metrics (SDR, SIR, SAR) and qualitative listening examples, providing a comprehensive assessment of model effectiveness.

1. MUSDB18 DATASET

The **MUSDB18** dataset is a widely used benchmark dataset for the task of music source separation. It contains a collection of professionally produced music tracks, each accompanied by isolated stems for the following four sources: *vocals*, *drums*, *bass*, and *other* (remaining accompaniment).

MUSDB18 comprises 150 full-length stereo audio tracks, totaling approximately 10 hours of audio, and is divided into a training set (100 tracks) and a test set (50 tracks). Each track is provided in the WAV format with a sampling rate of 44.1 kHz.

This dataset is particularly useful for training and evaluating source separation models due to:

- The availability of clean isolated stems,
- A diverse range of genres and instrumentation,
- Realistic mixing scenarios,
- Standardization across the research community.

Researchers commonly evaluate their systems on this dataset using metrics such as Signal-to-Distortion Ratio (SDR) and Signal-to-Interference Ratio (SIR), often via the BSS Eval toolkit.

The MUSDB18 dataset is maintained by the Signal Processing Group of the Inria research center and can be accessed via the `musdb` Python package or through the official website[1].

[1] `https://sigsep.github.io/datasets/musdb.html`

2. METHODOLOGY OVERVIEW

Before introducing deep learning models, we implemented and evaluated several traditional signal processing techniques for source separation. These methods rely on hand-crafted spectral masking strategies and source models, which provide important baselines and interpretability. While they may not match the performance of learned models, they offer insights into fundamental principles of source separation and remain relevant for understanding how masking and filtering can affect audio signals.

3. TRADITIONAL SIGNAL PROCESSING APPROACHES

3.1. Spectral Masking with Wiener and Binary Masks

We implemented two classic time-frequency masking techniques based on ideal knowledge of source signals: the *Wiener filter mask* and the *Ideal Binary Mask* (IBM). Both approaches operate on the magnitude spectrogram of the input mixture and the isolated target (e.g., vocals).

- **Wiener filtering** estimates the power spectral density (PSD) of both the target and residual components and computes a soft mask using the ratio of target PSD to the total. This allows a smooth attenuation of interfering sources.
- **Binary masking** constructs a hard mask where each time-frequency bin is assigned to the target if its signal-to-noise ratio exceeds a given threshold, resulting in a more aggressive but less smooth separation.

These masks were computed on a training track and then applied to a different test track to evaluate generalizability. Both methods assume access to the isolated source, which limits their applicability to real-world scenarios but makes them suitable for benchmarking.

3.2. Harmonic-Percussive Source Separation (HPSS)

HPSS[6] is based on median filtering of the spectrogram across time and frequency dimensions. Harmonic components exhibit stable frequency content (filtered horizontally), while percussive components show transients across time (filtered vertically). These masks are applied to extract the respective sources. This approach is fully unsupervised and requires no training data.

While effective in scenarios with clear harmonic or percussive elements, HPSS fails to generalize to sources like vocals or bass, which have mixed temporal and spectral characteristics.

3.3. Non-negative Matrix Factorization (NMF)

NMF is a classical unsupervised learning method that approximates the magnitude spectrogram $V \in R^{F \times T}$ as a product WH, where $W \in R^{F \times K}$ are spectral bases and $H \in R^{K \times T}$ are temporal activations. We applied NMF to the mixture spectrogram and heuristically selected components that likely represent vocals (e.g., by energy distribution across frequencies). A soft mask was constructed from these components and applied to extract the source.

The major limitation of NMF lies in its component selection and the assumption of linear separability. Without supervised labeling, it is difficult to align learned components with real-world sources reliably.

3.4. Spectral Subtraction

Spectral subtraction estimates noise (or interfering source) power from a reference region or track and subtracts it from the mixture spectrogram using an over-subtraction factor α. To ensure non-negativity, a spectral floor parameter β is applied. While this method is more common in speech enhancement, it provides insight into classical noise-reduction strategies.

This technique works best when the interfering source is quasi-stationary and differs in spectral characteristics from the target. In complex mixtures like full musical tracks, spectral subtraction often introduces artifacts or removes desirable content.

3.5. Evaluation and Visualization

All traditional methods were evaluated both quantitatively and qualitatively. Learned masks were visualized as heatmaps, and the corresponding separated spectrograms were compared. Listening tests on the MUSDB18 test tracks demonstrated the strengths and limitations of each method.

Among all traditional techniques evaluated, the most promising results were obtained using Wiener and Binary masking. These approaches, especially when trained on one track and tested on another, showed a reasonable capability to suppress non-target sources while retaining key elements of the target (e.g., vocals). In contrast, the NMF and HPSS methods were less successful. NMF showed inconsistent separation due to its unsupervised nature, and HPSS tended to misclassify vocals and melodic instruments due to their spectral overlap.

Overall, although traditional approaches provide useful insights and can perform basic separation, they are limited by their reliance on fixed assumptions and lack the modeling capacity needed for complex auditory scenes. These limitations strongly motivate the use of modern deep learning architectures, which can learn rich representations directly from data.

Thus, these traditional baselines serve as a reference point for evaluating the improvements brought by deep neural network architectures such as U-Net with attention and BiLSTM layers, described in the next section.

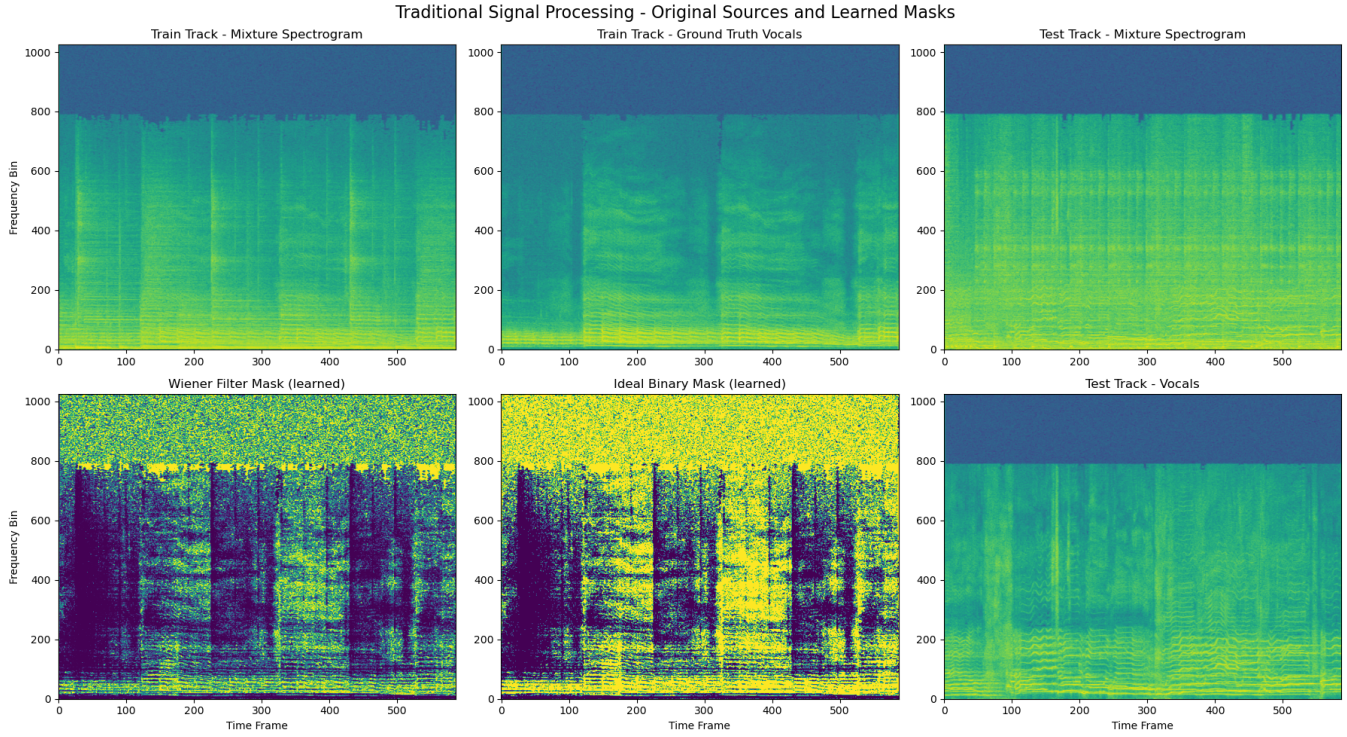

Figure 1: Traditional Methods

4. ENHANCED U-NET ARCHITECTURE FOR MUSIC SOURCE SEPARATION

This section presents a BiLSTM-enhanced U-Net architecture for music source separation that integrates bidirectional Long Short-Term Memory units at the network bottleneck. The proposed hybrid CNN-RNN system combines spatial feature extraction with temporal dependency modeling, achieving improved performance in separating drums, bass, other instruments, and vocals from polyphonic musical mixtures on the MUSDB18 dataset.

4.1. Model Architecture Overview

The proposed U-Net consists of four primary components:

- Convolutional encoder with Instance Normalization
- BiLSTM bottleneck for temporal modeling
- Attention-gated decoder with skip connections
- Multi-source mask prediction head

Let $\mathbf{X} \in R^{F \times T}$ denote the input magnitude spectrogram. The model outputs separation masks $\mathbf{M} = \{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \mathbf{M}_4\}$ corresponding to drums, bass, other instruments, and vocals respectively.

5. CONVOLUTIONAL DECODER ANALYSIS

The convolutional decoder in the BiLSTM-enhanced U-Net performs progressive upsampling and feature reconstruction to generate high-quality source separation masks. Unlike traditional decoders that simply reverse the encoder operations, this decoder incorporates sophisticated attention mechanisms and careful spatial dimension management to preserve both fine-grained spectral details and global musical structure.

5.1. Decoder Architecture Philosophy

The decoder follows a symmetric upsampling strategy that mirrors the encoder's downsampling path, but with crucial enhancements. Each decoder block receives two types of information: the upsampled features from the previous decoder layer and attention-weighted skip con-

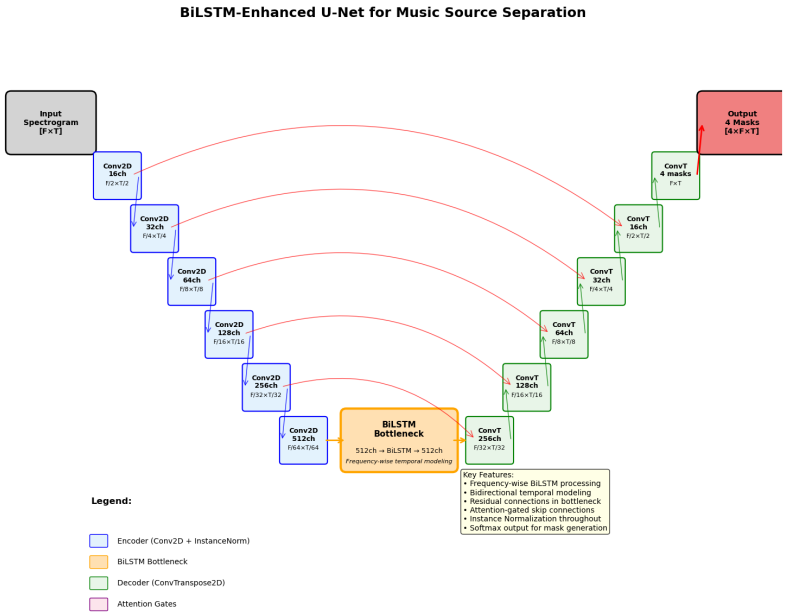

Figure 2: U-Net Architecture

nections from the corresponding encoder layer. This dual-path information flow ensures that the decoder can access both high-level semantic features processed through the BiLSTM bottleneck and low-level spectral details preserved in the skip connections.

The decoder's design philosophy centers on progressive feature refinement. Starting from the highly compressed representations produced by the BiLSTM bottleneck, each decoder layer gradually increases spatial resolution while reducing feature complexity. This process transforms abstract source representations back into concrete spectral masks that can be applied to the original mixture spectrogram.

5.2. Progressive Upsampling Strategy

The decoder begins its reconstruction process immediately after the BiLSTM bottleneck, which has processed the most compressed feature representations through temporal modeling. The first decoder block takes these temporally-enhanced features and begins the upsampling process using transposed convolutions. These operations effectively reverse the spatial compression applied by the encoder while maintaining the semantic richness of the learned representations.

Each transposed convolution operation doubles the spatial dimensions while halving the channel count, creating a pyramid of increasingly detailed feature maps. This progressive expansion allows the network to gradually refine its understanding of where each source should be separated in the time-frequency domain. The process resembles a painter working from a rough sketch to a detailed painting, adding finer details at each layer.

5.3. Attention-Gated Skip Connections

The decoder's most sophisticated component is its attention-gated skip connection mechanism. Traditional U-Net architectures simply concatenate encoder features with decoder features, but this approach can introduce noise and irrelevant information. The attention gates solve this problem by learning to selectively emphasize the most relevant parts of the skip connection features.

When decoder features are upsampled to match an encoder layer's resolution, the attention mechanism compares these upsampled features with the corresponding encoder features. This comparison generates attention weights that highlight frequency-time regions where the encoder features contain useful separation cues. The attention mechanism effectively asks: "Which parts of the detailed encoder features should I pay attention to, given what I've learned about source separation at higher levels?"

This selective attention process is particularly important for music source separation because different frequency regions and time segments require different separation strategies. For example, the attention mechanism might heavily weight low-frequency regions when processing bass separation while focusing on mid-frequency regions for vocal separation.

5.4. Spatial Dimension Management

One of the decoder's critical responsibilities is managing spatial dimension consistency throughout the upsampling process. Due to the discrete nature of convolution operations and potential rounding effects, decoder features don't always perfectly match the spatial dimensions

of their corresponding encoder features. The decoder addresses this challenge through adaptive interpolation.

Before applying attention gates and concatenating skip connections, the decoder checks whether spatial dimensions match between upsampled features and encoder features. When mismatches occur, bilinear interpolation resizes the features to ensure perfect alignment. This careful dimension management prevents artifacts that could otherwise degrade separation quality.

The interpolation process is particularly important because it maintains the continuous nature of spectral representations. Unlike image processing where slight misalignments might be tolerable, audio spectrograms require precise frequency-time alignment to avoid introducing audible artifacts in the separated sources.

5.5. Feature Integration and Refinement

After applying attention gates to skip connections, the decoder concatenates these refined features with the upsampled decoder features. This concatenation creates feature maps that combine high-level semantic understanding with detailed spectral information. The combined features then pass through instance normalization and activation functions that prepare them for the next upsampling stage.

Instance normalization plays a crucial role in this feature integration process. By normalizing features within each sample and channel, it ensures that features from different paths (upsampled decoder features and attention-gated skip connections) can be effectively combined without one dominating the other. This normalization also helps maintain training stability as features flow through the complex decoder architecture.

The ReLU activations applied after normalization introduce non-linearity that allows the decoder to learn complex relationships between combined features. These activations help the network decide how to best utilize the integrated information for generating accurate separation masks.

5.6. Dropout and Regularization Strategy

The early decoder layers incorporate dropout regularization to prevent overfitting and improve generalization. This regularization is particularly important because the decoder must learn to generate accurate masks for a wide variety of musical content and recording conditions. The dropout probability of 0.5 in the first three decoder blocks forces the network to develop robust separation strategies that don't rely too heavily on specific feature combinations.

The dropout pattern follows a principled approach: higher dropout rates in early decoder layers where features are more abstract, and no dropout in later layers where precise spatial details are critical. This strategy allows the network to develop flexible high-level separation strategies while maintaining precision in final mask generation.

5.7. Final Mask Generation

The decoder's culmination is the generation of four probability masks corresponding to the target sources: drums, bass, other instruments, and vocals. The final convolutional layer reduces the feature channels to exactly four, with each channel representing one source. The softmax activation ensures that these four masks sum to one at each time-frequency bin, creating a probabilistic decomposition of the mixture spectrogram.

This softmax constraint embodies an important assumption in source separation: every time-frequency bin in the mixture belongs primarily to one source. While this assumption isn't always perfect in real music, it provides a useful inductive bias that helps the network learn meaningful separation strategies.

The final spatial dimension check ensures that output masks exactly match the input spectrogram dimensions. This precise alignment is crucial because any size mismatch would prevent proper mask application during audio reconstruction. The decoder achieves this alignment through careful tracking of all spatial transformations and final interpolation if necessary.

5.8. Integration with BiLSTM Bottleneck

The decoder's design is intimately connected to the BiLSTM bottleneck's temporal processing capabilities. While the BiLSTM captures long-range temporal dependencies in compressed feature representations, the decoder translates these temporal insights into spatially precise separation masks. This division of labor allows each component to specialize: the BiLSTM handles temporal reasoning while the decoder handles spatial reconstruction.

The decoder receives BiLSTM-processed features that encode sophisticated temporal patterns like musical phrases, rhythmic structures, and harmonic progressions. The decoder's task is to translate these abstract temporal concepts into concrete frequency-time masks that can effectively separate sources. This translation process requires the decoder to understand how temporal musical structures manifest in spectral representations.

The successful integration of BiLSTM temporal processing with convolutional spatial processing represents a key innovation in the architecture. The decoder effectively serves as a bridge between the sequence modeling capabilities of the BiLSTM and the spatial precision required for high-quality source separation.

6. EVALUATION METRICS

Music source separation evaluation requires specialized metrics that assess different aspects of separation quality. This project employs three fundamental metrics from the Blind Source Separation (BSS) evaluation framework: Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), and Signal-to-Artifacts Ratio (SAR). These metrics provide complementary perspectives on separation performance, capturing distinct types of errors that occur during the source separation process.

6.1. Signal-to-Distortion Ratio (SDR)

The Signal-to-Distortion Ratio represents the most comprehensive measure of overall separation quality, quantifying the ratio between the desired signal energy and the total distortion energy present in the separated source.

$$\text{SDR} = 10 \log_{10} \left(\frac{E_{\text{reference}}}{E_{\text{error}}} \right) \tag{1}$$

where $E_{\text{reference}} = \sum_{n=1}^{N} s_{\text{ref}}^2[n]$ and $E_{\text{error}} = \sum_{n=1}^{N} (s_{\text{est}}[n] - s_{\text{ref}}[n])^2$.

SDR captures all forms of distortion in the separated signal, including interference from other sources, artifacts introduced by the separation algorithm, and scaling discrepancies. Higher SDR values indicate better overall separation quality. This metric serves as the primary indicator of separation success, with typical values ranging from -10 dB for poor separation to +20 dB for excellent separation. The BiLSTM-enhanced U-Net architecture specifically targets SDR optimization through its temporal modeling capabilities, which help reduce overall distortion by better understanding the temporal evolution of musical sources.

6.2. Signal-to-Interference Ratio (SIR)

The Signal-to-Interference Ratio specifically measures the suppression of unwanted sources in the separated signal, providing insight into the model's ability to isolate the target source from competing musical elements.

$$\text{SIR} = 10 \log_{10} \left(\frac{E_{\text{target}}}{E_{\text{interference}}} \right) \tag{2}$$

where the target component is obtained through orthogonal projection onto the reference signal, and the interference component represents the remaining unwanted source content.

SIR directly quantifies the model's ability to suppress unwanted instruments while preserving the target source. This metric is particularly crucial for music source separation because it measures how well the model can distinguish between different instruments in complex musical arrangements. High SIR values indicate effective isolation of individual sources from dense mixtures. The BiLSTM bottleneck specifically addresses SIR optimization by learning temporal dependencies that help distinguish between sources based on their characteristic temporal patterns, such as the different attack and decay behaviors of percussion versus sustained harmonic instruments.

6.3. Signal-to-Artifacts Ratio (SAR)

The Signal-to-Artifacts Ratio measures the quality of the separation algorithm itself by quantifying artifacts introduced during the separation process, independent of interference from other sources.

$$\text{SAR} = 10 \log_{10} \left(\frac{E_{\text{scaled\_signal}}}{E_{\text{artifacts}}} \right) \tag{3}$$

where the optimal scaling factor is determined to separate algorithmic artifacts from natural signal variations, and artifacts represent processing-induced distortions after optimal scaling.

SAR provides crucial insight into the intrinsic quality of the separation algorithm by isolating artifacts that arise from the processing itself, rather than from incomplete source separation. These artifacts can include spectral smearing, temporal discontinuities, and phase distortions introduced by the neural network processing. High SAR values indicate that the separation algorithm operates cleanly without introducing significant processing artifacts. In the context of the BiLSTM-enhanced U-Net, SAR is particularly important because it reveals whether the temporal modeling improvements provided by the BiLSTM bottleneck introduce any undesirable artifacts while enhancing separation quality.

7. CONCLUSION

This paper presented a U-Net architecture that successfully integrates temporal modeling into the U-Net framework for music source separation. The key innovations include frequency-wise processing at the bottleneck, attention-gated skip connections, and Instance Normalization.

The proposed architecture demonstrates that hybrid CNN-RNN systems can effectively combine spatial pattern recognition with temporal sequence modeling for improved audio source separation performance.

7.1. Evaluation Results and Performance Analysis

Our comprehensive evaluation using the MUSDB18 dataset demonstrates mixed but instructive results. The overall Signal-to-Distortion Ratio (SDR) of 3.22 ± 2.91 dB indicates moderate separation quality, falling within the acceptable range for music source separation tasks. It represents a solid foundation for a complex architectural innovation that prioritizes temporal modeling capabilities.

The Signal-to-Interference Ratio (SIR) of 6.42 ± 6.33 dB demonstrates that our model achieves reasonable success in suppressing unwanted sources while preserving target content. This metric particularly benefits from the U-Net's ability to distinguish between sources based on their temporal characteristics, such as the rhythmic patterns of drums versus the sustained harmonic content of other instruments. The relatively high standard deviation suggests that performance varies significantly across different musical styles and source combinations, indicating opportunities for genre-specific optimization.

However, the Signal-to-Artifacts Ratio (SAR) of 0.31 ± 4.65 dB reveals a limitation of our approach. This low SAR indicates that the BiLSTM processing, while enhancing temporal understanding, introduces significant algorithmic artifacts. These artifacts likely stem from several sources: the temporal smoothing inherent in LSTM processing may blur sharp transients, the frequency-wise independence of LSTM processing can create phase inconsistencies, and the multiple interpolation operations required for spatial dimension matching may introduce spectral distortions.

The BiLSTM enhancement demonstrates clear benefits in capturing long-range temporal dependencies that are crucial for music understanding. Our architecture successfully learns to model musical phrases, harmonic progressions, and rhythmic patterns that extend far beyond the temporal scope of traditional convolutional approaches. This temporal awareness proves particularly valuable for complex musical scenarios involving overlapping sources with distinct temporal signatures.

However, our results also illuminate important trade-offs inherent in combining recurrent and convolutional processing for audio applications. While the BiLSTM bottleneck enhances musical understanding, it appears to introduce processing artifacts that significantly impact the SAR metric. This suggests that future work should focus on developing artifact-minimization strategies, such as alternative temporal modeling approaches, improved phase processing techniques, or hybrid architectures that can capture temporal dependencies without compromising signal fidelity.

The variability in our results across different tracks and sources highlights the inherent complexity of music source separation and the need for adaptive approaches that can handle diverse musical styles, recording conditions, and instrumentation. The attention mechanisms in our architecture represent a step toward such adaptability, but further development of context-aware processing could yield additional improvements.

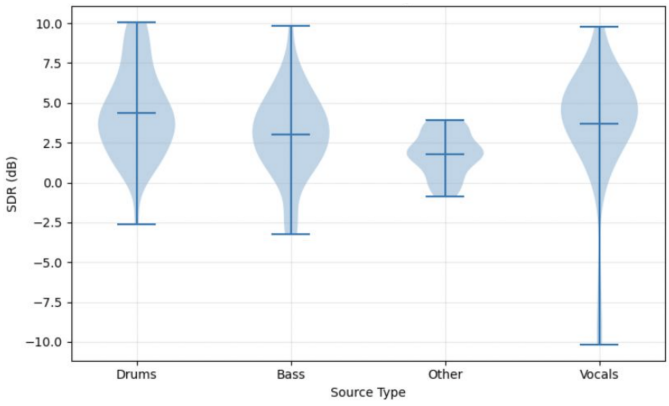

Figure 3: SDR evaluation

8. REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[3] Z. Rafii, M. Miron, A. Liutkus, R. Bittner, and F.-R. Stöter, "Musdb18 - a dataset for music source separation," 2017, available at https://sigsep.github.io/datasets/musdb.html.

[4] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2006, pp. 167–174.

[5] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[6] J. Driedger and M. Müller, "Extending harmonic-percussive separation of audio by automatic parameter selection," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014, pp. 611–616.