

DO MUSIC SOURCE SEPARATION MODELS PRESERVE SPATIAL INFORMATION IN BINAURAL AUDIO?

Richa Namballa
New York University
rn2214@nyu.edu

Agnieszka Roginska
New York University
ar137@nyu.edu

Magdalena Fuentes
New York University
mf3734@nyu.edu

ABSTRACT

Binaural audio remains underexplored within the music information retrieval community. Motivated by the rising popularity of virtual and augmented reality experiences as well as potential applications to accessibility, we investigate how well existing music source separation (MSS) models perform on binaural audio. Although these models process two-channel inputs, it is unclear how effectively they retain spatial information. In this work, we evaluate how several popular MSS models preserve spatial information on both standard stereo and novel binaural datasets. Our binaural data is synthesized using stems from MUSDB18-HQ and open-source head-related transfer functions by positioning instrument sources randomly along the horizontal plane. We then assess the spatial quality of the separated stems using signal processing and interaural cue-based metrics. Our results show that stereo MSS models fail to preserve the spatial information critical for maintaining the immersive quality of binaural audio, and that the degradation depends on model architecture as well as the target instrument. Finally, we highlight valuable opportunities for future work at the intersection of MSS and immersive audio.

1. INTRODUCTION

In recent years, immersive experiences have gained popularity in various forms of media such as video games, concerts, and movies. The shift to virtual and augmented reality (VR/AR) requires not only realistic visual stimuli, but authentic auditory cues as well. One common form of spatial audio used to provide the listener with directionality of sound is binaural audio. Binaural audio goes beyond traditional gain-based stereo panning by filtering two-channel audio to create interaural cues differing in level, time, and spectral content to simulate the location of a source in space [1]. Furthermore, binaural audio requires reproduction through headphones or loudspeakers equipped with crosstalk cancellation to maintain spatial imaging integrity. Level differences resulting from the “head-shadow effect”

and the Time Difference of Arrival (TDOA) of a sound at each ear provide directional cues. Frequency-dependent filtering, determined by the form of the listener’s head and specific ear (pinna) shape, causes two identical sound sources positioned differently to exhibit slightly different spectral content at each ear, further assisting localization. The two common methods of producing binaural audio are recording with a binaural dummy head and signal processing with a Head-Related Transfer Function (HRTF).

Beyond the increasing demand for immersive VR/AR experiences, binaural audio has significant potential applications in accessibility. For instance, individuals who identify as neuro-divergent or hard of hearing often benefit from enhanced auditory clarity, enabling them to isolate and focus on specific sound sources in complex acoustic environments, facilitating independent navigation and interaction in social and public settings. Binaural source separation has been shown to significantly enhance auditory accessibility by reducing background noise and emphasizing relevant auditory signals in real-time with the use of microphone-enabled headphones [2]. In this context, music source separation (MSS) in binaural audio could substantially improve how individuals engage with and enjoy musical environments such as concerts, festivals, and other live performances, enabling users to isolate specific musical elements or instruments and thus enhance their listening experience and overall participation in music events. These tools can further be utilized for recorded binaural content such as spatial audio captures of live performances or binaural field recordings.

Despite these potential benefits and growing interest, binaural audio processing has received limited attention within the music information retrieval (MIR) community, particularly concerning MSS. In this work, we investigate whether existing MSS models are able to separate binaural mixtures into their respective stems while preserving the spatial characteristics, which are crucial for the immersive experience provided by binaural audio. We create a binaural MSS dataset based off of the well-established MUSDB18-HQ dataset [3], and leverage several metrics that quantify separation quality, spatial distortion, and immersiveness to evaluate these models. Our results show that there is a considerable gap in binaural MSS performance compared with MSS in simpler stereo settings, and that this gap depends on model architecture and target source. Lastly, we discuss the shortcomings of current metrics and identify opportunities for future research.



© R. Namballa, A. Roginska, and M. Fuentes. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** R. Namballa, A. Roginska, and M. Fuentes, “Do Music Source Separation Models Preserve Spatial Information in Binaural Audio?”, in *Proc. of the 26th Int. Society for Music Information Retrieval Conf.*, Daejeon, South Korea, 2025.

2. RELATED WORK

Until now, most work on binaural source separation has been completed in the speech domain, often overlapping with the similar task of target source extraction (TSE). In particular, the speech research community describes the task as two-fold: source separation and localization [4]. We focus on prior studies concerning the former.

Early two-channel source separation models were primarily signal processing-based, with a focus on mathematical and theoretical techniques [5]. As the focus moved towards capturing directionality, models began using psychoacoustic spatial cues to improve the performance of the signal processing-based source estimation methods [6–11]. With the technological progress made in computational resources, binaural source separation models shifted to using deep learning approaches to perform source extraction in more complex environments and in real-time [2, 12–15]. Recent deep learning systems have proposed novel loss functions aimed at preserving the level, phase, and time differences between binaural channels, cues which are critical to the immersive nature of binaural audio [16, 17].

To the best of our knowledge, the only published work on binaural MSS thus far concerns vocal separation of binaural audio recorded with a dummy head [18]. Their approach uses various hybrid combinations of single- and multichannel-source separation algorithms to extract the vocal stems, with a focus on signal-processing methods [19–23]. The results are evaluated with standard source separation metrics [24] and subjective listener ratings. Based on the limited existing research in binaural MSS, we believe that there is a significant opportunity to explore this task using deep learning methods, inspired by recent progress in the speech community.

Regarding performance, the most common metric reported for evaluating source separation models is the Signal to Distortion Ratio (SDR), measured in decibels (dB) [24]. Specifically, for MSS, researchers often benchmark their models on the test set of MUSDB18-HQ and report the SDR both overall and by instrument type [25]. SDR (and its scale-invariant version, SI-SDR [26]) aim to reflect what portion of the estimated stem corresponds to the reference stem versus any error introduced by interference from other instruments, noise, and artifacts. While SDR is well-established for evaluating mono and stereo tracks [27], it does not specify the amount of spatial error introduced between channels in the model’s estimated output, which is essential for evaluating the quality of binaural source separation. Therefore, we leverage other metrics from the literature which reflect spatial quality.

In the immersive audio research community, there are several models used to quantify the quality of a binaural signal, such as BAM-Q [28] and MoBi-Q [29], trained on a combination of extracted binaural features and subjective quality ratings. We save the use of these models for future work in binaural MSS and choose to focus on more accessible and interpretable metrics, further explained in Section 4.1, which originate from the duplex theory of sound localization [30]. This theory states that, along horizon-

tal plane (0° elevation), humans use two auditory cues to localize the direction of a sound: the interaural time difference (ITD) and the interaural level difference (ILD). ITD refers to the difference in time of arrival, at each ear, of a sound emitted from a source. Generally, a sound will reach the ipsilateral (closest to the source) ear faster than the contralateral (farthest from the source) ear. Likewise, the ILD is the difference in a sound’s intensity as it arrives at the ipsilateral and contralateral ears. Originally, it was believed that ILD was the primary cue used for high frequency signals while ITD was for low frequencies [1]. However, recent studies have shown that broadband signals require a complex interaction of the ITD and ILD to effectively identify a sound’s location [31].

The work in [32] leverages this duplex theory of localization to propose two energy-ratio metrics for spatial evaluation: Signal to Spatial Distortion Ratio (SSR) and Signal to Residual Distortion (SRR). These measures are interpreted similarly to SDR, with SSR intended as a substitute for the Image to Spatial Distortion Ratio (ISR), proposed by [27]. The spatial error is computed by projecting the reference signal to the estimated signal and optimizing for relative changes in gain and delay. From these projections, we can separate the distortion in spatial information (spatial error) from errors such as interference in the estimated signal (residual error). The ratios of SSR and SRR are defined in Section 4.1.

3. DATASET

To directly compare the performance of various MSS models on both stereo and binaural audio, we created a binaural version of MUSDB18-HQ [3]. MUSDB18-HQ is the uncompressed, 22kHz-bandwidth version of MUSDB18 [33] containing full-length, mixed music tracks from primarily Western pop and rock genres as well as their respective stems separated into vocals, drums, bass, and “other”. The training and test sets consist of 100 and 50 songs, respectively. All audio files are stereophonic in WAV format, sampled at 44.1kHz/16b. We call our binaural dataset Binaural-MUSDB and we refer to the original MUSDB18-HQ as Stereo-MUSDB.

To construct Binaural-MUSDB, we utilized binaural synthesis to create the illusion of the source signal emitting from a specific location around the listener [1]. We use the publicly available SADIE II¹ database of HRTFs [34]. Each two-channel HRTF measurement contains the auditory spatial cues which can be superimposed onto a signal such that the listener will perceive the sound as originating from a location along the azimuth (θ) and at a given elevation (ϕ). For our synthesis, we apply the HRTF measurements for subject D1 from SADIE II, which correspond to the head and pinnae of the Neumann KU100 binaural dummy head microphone, which is the size of the average human head.

We limited the horizontal plane to $\theta \in [-90^\circ, +90^\circ]$ along the azimuth, fixed at $\phi = 0^\circ$ elevation. In spatial au-

¹ <https://www.york.ac.uk/sadie-project/database.html>

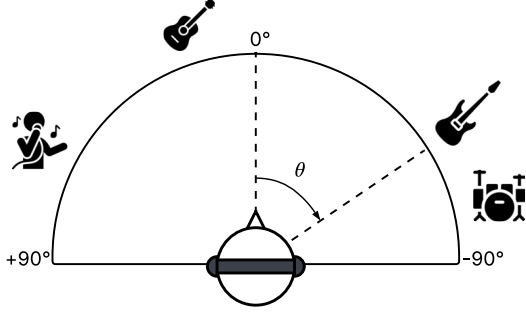


Figure 1. Binaural-MUSDB: each binaural source signal s_i is placed randomly along the horizontal plane at an angle $\theta_i \in [-90^\circ, +90^\circ]$ with the origin located directly in front of the listener. Every source has a minimum of 10° separation from the others, ensuring that there is no direct spatial overlap between stems.

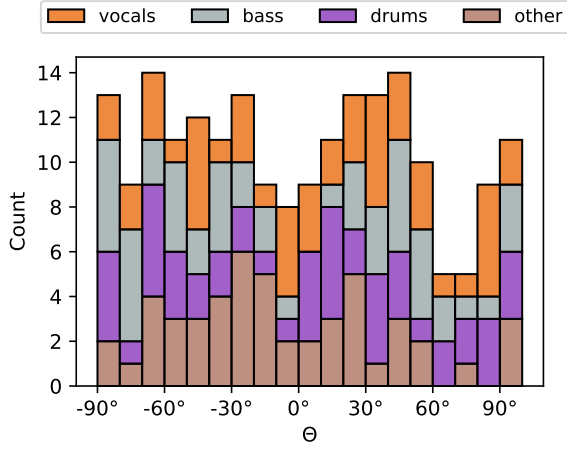


Figure 2. Distribution of instrument positions in the test set of Binaural-MUSDB. θ corresponds to the source’s location along the horizontal plane where 0° corresponds to the position directly in front of the listener.

dio, $\theta = 0^\circ$ corresponds to the location directly in front of the listener, equidistant from the left and right ears, as seen in Figure 1. While the duplex theory states that humans primarily rely on ITD and ILD for binaural localization on the horizontal plane [30], they require spectral information for disambiguating front-back locations [35]. Since we limit source locations to the front half of the sound field, we do not anticipate any significant differences in results using HRTFs other than the KU100’s.

For every song in both the training and test sets, we assigned each source i to a static location θ_i in increments of 10° . Angles for each stem in a single song were sampled randomly without replacement in the order of vocals, bass, drums, and other. Furthermore, in a given mixture, no two sources were allowed to be located at the same angle ensuring that there was a minimum of 10° separation (no direct overlap) between each stem. Each song was assigned only one set of source locations. The distribution of locations across the test set can be seen in Figure 2.

We converted the original stereo stem to mono by averaging the two channels. Next, we loaded the Head-

Related Impulse Response (HRIR), the time-domain version of an HRTF, corresponding to θ_i and convolved each HRIR channel with the mono stem signal to produce a binaural signal, with the two channels corresponding to the left and right ears. This process is visualized in Figure 3.

Finally, we summed the binaural versions of the vocals, drums, bass and other stems together and normalized the resulting signal to create the binaural mixtures which were used as the input to the MSS models described in Section 4.2. The binaural synthesis was completed for all 150 tracks with the same train-test split as Stereo-MUSDB.

4. EXPERIMENTAL SETUP

4.1 Metrics

We utilize four metrics to describe the amount of distortion introduced by the MSS models, three of which quantify the level of spatial error in the estimated stems (SSR, Δ ITD, Δ ILD) and one that measures the remaining signal distortion due to interference and artifacts introduced by the separation (SRR).

In binaural audio, it is crucial that the ITD and ILD of a sound remain unchanged after separation to allow a listener to localize the source and maintain their sense of immersion. Therefore, we quantify how well the interaural cues are preserved by measuring the change (Δ) in ITD and ILD between the estimated stem (\hat{s}) and the reference stem (s), as in [2]. To compute Δ ITD, we calculate the magnitude of the difference in ITD(\hat{s}) and ITD(s) [36].

$$\Delta\text{ITD} = |\text{ITD}(s) - \text{ITD}(\hat{s})| \quad (1)$$

We measure the ITD of each signal as the TDOA of the source in the left and right channels using the frame-wise Generalized Cross Correlation with Phase Transform (GCC-PHAT) algorithm [37], implemented by [2]. First, we segment the signal x into frames of 0.5s in length (with no overlap) and apply a Tukey window to each frame. Next, we calculate the GCC-PHAT $C(t, \tau)$ at frame t , for lags τ (in samples) corresponding to the range $[-1, 1]$ ms, and find τ^* , the value of τ which maximizes C [2, 38]. The frame-wise TDOA is computed in seconds by dividing τ^* by the sample rate f_s .

$$\text{TDOA}(x, t) = \frac{1}{f_s} \cdot \arg \max_{\tau} C(t, \tau) \quad (2)$$

The ITD of the full signal is then calculated as the weighted mode of the frame-wise TDOA. Each weight w_t is based on the Root Mean Square (RMS) energy, where x_{tc} is the signal at frame t and channel c , n is the length of the frame, and k is the sample index of the frame.

$$w_t = \max_c \left(\sqrt{\frac{1}{n} \sum_{k=0}^{n-1} x_{tc}[k]^2} \right) \quad (3)$$

Frames with a w_t less than a threshold of 5×10^{-4} are considered silent and excluded from the signal’s ITD calculation. Δ ITD is presented in microseconds (μ s) [2].

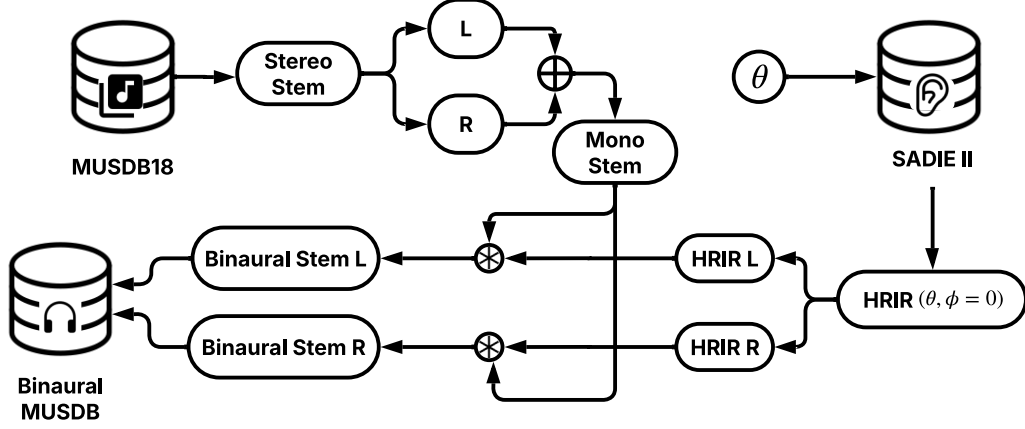


Figure 3. An overview of the binaural synthesis process for the Binaural-MUSDB dataset. For every song in MUSDB18-HQ, each source is assigned a location θ along the azimuth in the frontal portion of the horizontal plane ($\pm 90^\circ$). The corresponding HRIR (θ , elevation $\phi = 0$) is retrieved from the SADIE II database and each channel is convolved (\otimes) with the monophonic version of the source stem. The resulting signals are the left and right channels of the binaural version of the stem which are included in the dataset.

The ILD is computed as the decibel ratio of the sum of squares for each channel across the entire signal. Here, x_c represents channel c of the full signal \mathbf{x} , k is the corresponding sample index, and N is the length of the entire signal in samples. As with ITD, we report ΔILD .

$$\text{ILD}(\mathbf{x}) = 10 \cdot \log_{10} \left(\frac{\sum_{k=0}^{N-1} x_L[k]^2}{\sum_{k=0}^{N-1} x_R[k]^2} \right) \quad (4)$$

$$\Delta\text{ILD} = |\text{ILD}(\mathbf{s}) - \text{ILD}(\hat{\mathbf{s}})| \quad (5)$$

For both ΔITD and ΔILD , a lower value indicates a higher-quality spatial preservation of the interaural cue in the estimated stem.

In addition to ΔITD and ΔILD , we compute the SSR and SRR as proposed in [32] using their provided open-source implementation with its default parameters. Both metrics are computed frame-wise, reporting the median value, with a window of 1s and a hop length of 0.5s.

$$\text{SSR}(\hat{\mathbf{s}}; \mathbf{s}) = 10 \cdot \log_{10} \left(\frac{\|\mathbf{s}\|^2}{\|\mathbf{e}_{\text{spat}}\|^2} \right) \quad (6)$$

$$\text{SRR}(\hat{\mathbf{s}}; \mathbf{s}) = 10 \cdot \log_{10} \left(\frac{\|\tilde{\mathbf{s}}\|^2}{\|\mathbf{e}_{\text{resid}}\|^2} \right) \quad (7)$$

The SSR is intended to capture the spatial distortion introduced by the separation (\mathbf{e}_{spat}) into the estimated stem ($\hat{\mathbf{s}}$) while the SRR reflects only non-spatial distortion and errors such as interference and artifacts ($\mathbf{e}_{\text{resid}}$). Note that $\tilde{\mathbf{s}}$ is the projection² of \mathbf{s} into $\hat{\mathbf{s}}$, as mentioned in Section 2. Both SSR and SRR are measured in dB and a higher value indicates less distortion in the estimated signal.

4.2 Models

We evaluate the performance of three well-known pre-trained MSS models on both stereo and binaural

conditions: Hybrid Transformer Demucs Fine-Tuned (htdemucs_ft) [39], OpenUnmix (umxhq) [40], and Spleeter (spleeter:4stems) [41]. We chose these models over newer MSS models to validate our results with [32] and because all three models have official open-source implementations available for use. Both Demucs and OpenUnmix are trained on the Stereo-MUSDB training set, while Spleeter is trained on a proprietary dataset. Additionally, the version of Demucs we use is trained on an extra 800 songs not publicly identified. Each model accepts a stereophonic mixture input and returns an estimated two-channel stem.

Both OpenUnmix and Spleeter have inputs in the frequency domain, while Demucs is a hybrid model, operating in both the waveform and spectrogram domains. Spleeter uses a U-net architecture (CNN-based) [42] to estimate a time-frequency mask for each source and applies it to the input mixture’s magnitude spectrogram to generate the spectrogram of the estimated stem [43]. OpenUnmix operates similarly, however, it uses a bi-directional LSTM model (RNN-based) to estimate the mask [44]. All three models use a L1 loss function to minimize the error between the estimated and reference signals.

To preserve the temporal structure of the input audio, both OpenUnmix and Spleeter apply the original input mixture’s phase to the estimated magnitude spectrogram before inversion to the time domain to construct the final predicted stem. On the other hand, since Demucs functions in two domains, the model has to combine the estimated time and frequency representations to provide the final synthesized waveform. In the original hybrid version of Demucs [45], the model required careful hyperparameter tuning to align the temporal and spectral representations of the estimated signal so they could be summed in the waveform domain. However, in the newest version of the model [39], the authors claim that the transformer addresses this bottleneck through its flexible architecture.

To compare the separation performance in stereo and

² Due to space constraints, we encourage readers to reference the original publication [32] for the precise mathematical definition of $\tilde{\mathbf{s}}$.

binaural settings, we apply these models to the test sets of Stereo-MUSDB and Binaural-MUSDB.

5. RESULTS AND DISCUSSION

In this section, we analyze and discuss the performance of the three MSS models by looking at the different metrics in the binaural and stereo datasets, considering the effect on individual instruments, and identifying the effect of spatial distortion in the different locations along the azimuth.

Table 1. SRR results from the MSS models across the two datasets using median values. The best results are highlighted in **bold** and the second best are underlined.

Dataset	Model	SRR (dB) ↑				
		Bass	Drums	Other	Vocals	Overall
Binaural	Demucs	8.90	10.58	<u>4.10</u>	<u>4.37</u>	<u>6.91</u>
	OpenUnmix	3.37	6.75	1.19	2.37	3.51
	Spleeter	1.53	4.71	0.11	0.00	2.01
Stereo	Demucs	<u>8.36</u>	<u>9.86</u>	6.36	6.08	7.39
	OpenUnmix	1.72	4.82	2.90	2.40	3.14
	Spleeter	1.25	4.51	3.31	2.76	3.21

5.1 Stereo vs. Binaural Performance

Based on the median SRR values shown in Table 1, we observe a relatively consistent separation quality across the stereo and binaural datasets, suggesting that introducing spatial cues does not dramatically impact the ability of models to isolate instruments from one another. The SRR serves as a proxy of separation quality in spatial audio settings as it considers all residual distortions that are not spatial. Demucs appears to outperform the other two models in SRR for both datasets, which aligns with its original SDR-based ranking reported on the test set of Stereo-MUSDB [40, 41, 45].

The median spatial metrics in Table 2 show that the MSS models introduce substantial spatial distortion when applied to binaural audio. For reference, SSR values around 10dB relate to noticeable spatial distortion, while values below that indicate severe spatial distortion, based on trends seen in other energy-ratio metrics [24, 32, 46]. Note that spatialization in stereo tracks traditionally uses gain-based panning, so a median ΔITD of 0 μs is not unexpected. Upon closer inspection, a few ΔITD values were nonzero, indicating that some interchannel temporal distortion is introduced by the models, even in the stereo stems.

Demucs shows a considerable performance drop from stereo to binaural conditions, especially in SSR, compared to the other models. A plausible explanation is that, by operating directly on waveforms, Demucs implicitly learned stereo spatial cues based on amplitude differences and struggled to effectively interpret the subtler spectral information characteristic of binaural audio. In turn, Open-Unmix occasionally achieves superior results in binaural settings compared to stereo, likely due to its frequency-domain masking approach that preserves the original mixture’s phase, inadvertently maintaining the spatial integrity. Similarly, Spleeter, also employing frequency-domain masking, demonstrates stable and

sometimes improved performance on binaural audio, reinforcing that preserving the original phase of a mixture can be beneficial for spatial cue accuracy. Nevertheless, none of the models’ binaural metrics match Demucs’s stereo performance level, demonstrating considerable room for improvement in retaining binaural spatial cues.

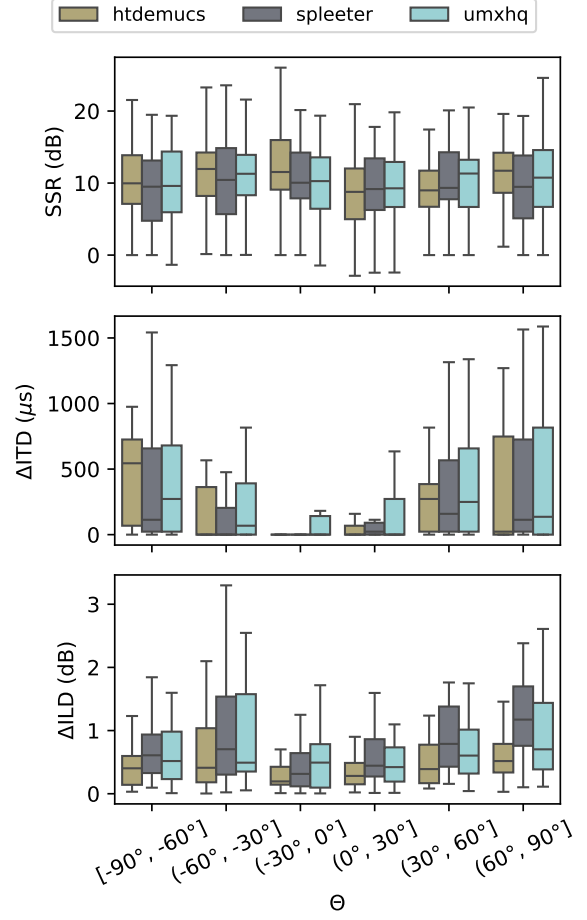


Figure 4. Distributions of spatial metrics (SSR, ΔITD , ΔILD) by model and angle, aggregated across all sources.

5.2 Performance by Angle

Figure 4 shows the overall spatial distortion across all three spatial metrics by model and angle bin along the azimuth. We observe that SSR and ΔILD remain relatively consistent across angles, whereas the ITD notably distorts more the farther the source is positioned from the origin (larger $|\theta|$), displaying a U-shaped effect. One source of this tendency could be that strongly lateralized signals have minimal overlap in time-domain amplitude between the left and right channels. Cross-correlation relies on shared, correlated energy between channels so, in these cases, even minor disturbances from separation reduce channel similarity substantially, making accurate lag estimation challenging. This pattern could also imply that current MSS models are better at preserving amplitude-based spatial information (e.g., gain-based panning) than phase-based cues, and that they are introducing temporal disturbances. Additionally, the ΔITD distribution highlights a potential limitation in the SSR metric. Although it has been designed to account

Table 2. Spatial metric results (SSR, Δ ITD, Δ ILD) from the MSS models for the two datasets using median values. The best results are highlighted in **bold** and the second best are underlined.

Dataset	Model	SSR (dB) \uparrow					Δ ITD μ s \downarrow					Δ ILD (dB) \downarrow				
		Bass	Drums	Other	Vocals	Overall	Bass	Drums	Other	Vocals	Overall	Bass	Drums	Other	Vocals	Overall
Binaural MUSDB	Demucs	9.13	10.39	<u>12.62</u>	8.70	10.59	476.19	0.00	<u>22.68</u>	0.00	68.03	0.20	0.31	0.57	0.42	0.39
	OpenUnmix	<u>10.94</u>	<u>12.22</u>	11.04	8.20	10.43	521.54	0.00	226.76	0.00	90.7	0.41	0.38	0.72	0.73	0.50
	Spleeter	10.63	11.86	9.96	5.22	9.86	544.22	<u>22.68</u>	<u>22.68</u>	<u>22.68</u>	<u>22.68</u>	0.44	0.52	0.99	0.74	0.64
Stereo MUSDB	Demucs	17.18	20.63	14.11	13.42	16.01	0.00	0.00	0.00	0.00	0.00	0.08	0.07	0.11	0.05	0.08
	OpenUnmix	9.74	12.12	10.09	<u>11.22</u>	10.73	0.00	0.00	0.00	0.00	0.00	<u>0.12</u>	0.10	0.24	<u>0.08</u>	<u>0.12</u>
	Spleeter	8.69	11.54	11.31	10.18	<u>10.78</u>	0.00	0.00	0.00	0.00	0.00	0.15	<u>0.08</u>	<u>0.23</u>	0.10	<u>0.12</u>

for all spatial distortions in accordance with the duplex theory [30, 32], it may be more sensitive to level differences rather than time of arrival changes (as it does not reflect the U-shaped behavior observed in Δ ITD). Further research with synthetic signals is needed to clarify how SSR values respond to phase distortions, whether the metric or its implementation requires revision, and how sensitive ITD and ILD calculations are to small artifacts.

5.3 Performance by Instrument

When looking at instrument-specific performance in Tables 1 and 2, we see that bass and “other” instruments exhibit higher spatial distortion (Δ ITD) compared to vocals and drums. Bass instruments predominantly occupy narrow, low frequency bands, where localization relies heavily on subtle time differences rather than level. Because these low-frequency sounds have longer wavelengths, even minor phase distortions introduced during the separation process can lead to significant perceived spatial errors. This trait is reflected in the cross-correlation calculations of ITD, which require larger sample lags (τ) to properly align the channels. Similarly, the “other” category often includes a diverse collection of complex and spectrally dense instruments with broader spatial positioning, resulting in diffused or ambiguous spatial cues.

5.4 Performance by Model

As mentioned previously, Demucs exhibits a significant performance drop from stereo to binaural conditions in terms of spatial distortion. In contrast, the frequency-domain models, Open-Unmix and Spleeter, display more consistent spatial performance across these two settings. Nevertheless, all models perform well below the level achieved by Demucs in stereo, suggesting that none are yet optimized for binaural spatial fidelity. Future research should explore training the models directly on binaural audio and adjusting the loss functions used during training to explicitly penalize distortions in ITD and ILD to improve spatial cue preservation, using systems inspired by the speech community [2, 16, 17].

5.5 Perceptual Considerations

While we primarily relied on objective metrics for our evaluation, preliminary subjective listening by the authors suggests noticeable spatial distortions, particularly affecting bass instruments. These distortions align with our

quantitative findings and indicate substantial spatial artifacts caused by inaccuracies in phase preservation. To provide a clearer illustration of these effects, selected audio examples demonstrating typical spatial distortions identified in our analysis are made available on an accompanying demonstration webpage, along with the open-source data and code repository.³

6. CONCLUSION AND FUTURE WORK

We investigated the capabilities of existing music source separation (MSS) models applied to binaural audio. Our analysis revealed a considerable gap in MSS performance between binaural and stereo settings. This performance disparity was influenced significantly by both the specific architecture of the model and the target audio source. We identify several avenues of planned future work which will address the limitations of this study and build the foundation for subsequent binaural MSS models.

Data. The binaural data was synthesized with a random placement of sources and a single set of HRTF measurements. We hope to examine the stability of the results concerning the random seed initialization in the positioning of sources and the effect of their overlap. Additionally, we can validate the the impact of using diverse HRTFs (corresponding to various pinnae) when synthesizing the data.

Metrics. We believe the current metrics require further investigation to better understand their sensitivity to changes in phase versus level. Moreover, we can explore existing binaural quality models established by the immersive audio community and perform a perceptual study to validate all metrics.

Modeling. Since MSS research has progressed significantly, we hope to evaluate newer state-of-the-art MSS models’ performance on binaural audio. We also plan to train a simple baseline MSS model on the binaural dataset with the option for data augmentations (e.g., noise, reverberation) to simulate diverse binaural conditions. Lastly, we will modify existing MSS model architectures to account for the preservation of spatial cues, such as with loss functions that minimize changes in ITD and ILD.

These paths for future research show promise in designing models specifically trained for binaural MSS with the goal of bridging immersive audio with music information retrieval for both cultural and accessibility applications.

³ <https://richa-namballa.github.io/binaural-mss-demo/>

7. REFERENCES

- [1] A. Roginska and P. Geluso, *Immersive Sound*. Focal Press, 2017.
- [2] B. Veluri, M. Itani, J. Chan, T. Yoshioka, and S. Gollakota, “Semantic hearing: Programming acoustic scenes with binaural hearables,” in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–15.
- [3] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “MUSDB18-HQ - an uncompressed version of MUSDB18,” Dec. 2019. [Online]. Available: <https://doi.org/10.5281/zenodo.3338373>
- [4] A. Deleforge and R. Horaud, “The cocktail party robot: Sound source separation and localisation with an active binaural head,” in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2012, pp. 431–438.
- [5] K. Torkkola, “Blind separation for audio signals-are we there yet?” in *First International Workshop on Independent Component Analysis and Blind Source Separation*, 1999, pp. 239–244.
- [6] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” *IEEE ASSP magazine*, vol. 5, no. 2, pp. 4–24, 1988.
- [7] A. Jourjine, S. Rickard, and O. Yilmaz, “Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures,” in *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5. IEEE, 2000, pp. 2985–2988.
- [8] H. Viste and G. Evangelista, “On the use of spatial cues to improve binaural source separation,” in *Proceedings of 6th International Conference on Digital Audio Effects (DAFx-2003)*, 2003, pp. 209–213.
- [9] S. Schulz and T. Herfet, “Binaural source separation in non-ideal reverberant environments,” in *Proceedings of 10th International Conference on Digital Audio Effects (DAFx-2007)*, Bordeaux, France, 2007.
- [10] C. Kim, K. Kumar, and R. M. Stern, “Binaural sound source separation motivated by auditory processing,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5072–5075.
- [11] R. Abdipour, A. Akbari, M. Rahmani, and B. Naser-sharif, “Binaural source separation based on spatial cues and maximum likelihood model adaptation,” *Digital Signal Processing*, vol. 36, pp. 174–183, 2015.
- [12] S. Zakeri and M. Geravanchizadeh, “Supervised binaural source separation using auditory attention detection in realistic scenarios,” *Applied Acoustics*, vol. 175, p. 107826, 2021.
- [13] Y. Yang, G. Sung, S.-F. Shih, H. Erdogan, C. Lee, and M. Grundmann, “Binaural angular separation network,” in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1201–1205.
- [14] X. Zhang and D. Wang, “Deep learning based binaural speech separation in reverberant environments,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1075–1084, 2017.
- [15] B. Veluri, J. Chan, M. Itani, T. Chen, T. Yoshioka, and S. Gollakota, “Real-time target sound extraction,” in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [16] C. Hernandez-Olivan, M. Delcroix, T. Ochiai, N. Tawara, T. Nakatani, and S. Araki, “Interaural time difference loss for binaural target sound extraction,” in *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2024, pp. 210–214.
- [17] V. Tokala, E. Grinstein, M. Brookes, S. Doclo, J. Jensen, and P. A. Naylor, “Binaural speech enhancement using deep complex convolutional transformer networks,” in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 681–685.
- [18] P. Kasak, R. Jarina, D. Ticha, and M. Jakubec, “Hybrid binaural singing voice separation,” in *2023 33rd International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2023, pp. 1–6.
- [19] S. Rickard, “The DUET blind source separation algorithm,” in *Blind Speech Separation*. Springer, 2007, pp. 217–241.
- [20] A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
- [21] J. Driedger, M. Müller, and S. Disch, “Extending harmonic-percussive separation of audio signals,” in *Proceedings of the 15th International Society for Music Information Retrieval Conference*. ISMIR, 2014, pp. 611–616.
- [22] P. Seetharaman, F. Pishdadian, and B. Pardo, “Music/voice separation using the 2D Fourier transform,” in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 36–40.
- [23] Z. Rafii and B. Pardo, “Repeating pattern extraction technique (REPET): A simple method for music/voice separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 73–84, 2012.
- [24] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

- [25] “Sound Demixing Workshop.” [Online]. Available: <https://sdx-workshop.github.io/>
- [26] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR—half-baked or well done?” in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.
- [27] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, “First stereo audio source separation evaluation campaign: data, algorithms and results,” in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 552–559.
- [28] J.-H. Fleßner, R. Huber, and S. D. Ewert, “Assessment and prediction of binaural aspects of audio quality,” *Journal of the Audio Engineering Society*, vol. 65, no. 11, pp. 929–942, 2017.
- [29] J.-H. Fleßner, S. D. Ewert, B. Kollmeier, and R. Huber, “Quality assessment of multi-channel audio processing schemes based on a binaural auditory model,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1340–1344.
- [30] L. Rayleigh, “XII. on our perception of sound direction,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 13, no. 74, pp. 214–232, 1907.
- [31] L. R. Bernstein, “Auditory processing of interaural timing information: new insights,” *Journal of Neuroscience Research*, vol. 66, no. 6, pp. 1035–1046, 2001.
- [32] K. N. Watcharasupat and A. Lerch, “Quantifying spatial audio quality impairment,” in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 746–750.
- [33] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “The MUSDB18 corpus for music separation,” Dec. 2017. [Online]. Available: <https://doi.org/10.5281/zenodo.1117372>
- [34] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, “A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database,” *Applied Sciences*, vol. 8, no. 11, p. 2029, 2018.
- [35] P. M. Hofman and A. J. Van Opstal, “Spectro-temporal factors in two-dimensional human sound localization,” *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2634–2648, 1998.
- [36] C. Han, Y. Luo, and N. Mesgarani, “Real-time binaural speech separation with preserved spatial cues,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6404–6408.
- [37] C. Knapp and G. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.
- [38] T. May, S. Van De Par, and A. Kohlrausch, “A probabilistic model for robust localization based on a binaural auditory front-end,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 1–13, 2010.
- [39] S. Rouard, F. Massa, and A. Défossez, “Hybrid transformers for music source separation,” in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [40] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, “Open-Unmix - a reference implementation for music source separation,” *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019. [Online]. Available: <https://doi.org/10.21105/joss.01667>
- [41] R. Hennequin, A. Khelif, F. Voituret, and M. Moussallam, “Spleeter: a fast and efficient music source separation tool with pre-trained models,” *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020.
- [42] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 18th International Conference, Proceedings, Part III*. Munich, Germany: Springer, 2015, pp. 234–241.
- [43] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, “Singing voice separation with deep U-net convolutional networks,” in *Proceedings of the 18th International Society for Music Information Retrieval Conference*. Suzhou, China: ISMIR, 2017, pp. 23–27.
- [44] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, “Improving music source separation based on deep neural networks through data augmentation and network blending,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 261–265.
- [45] A. Défossez, “Hybrid spectrogram and waveform source separation,” in *Proceedings of the Music Demixing (MDX) Workshop*, 2021.
- [46] “Music source separation on MUSDB18,” March 2025. [Online]. Available: <https://paperswithcode.com/sota/music-source-separation-on-musdb18>