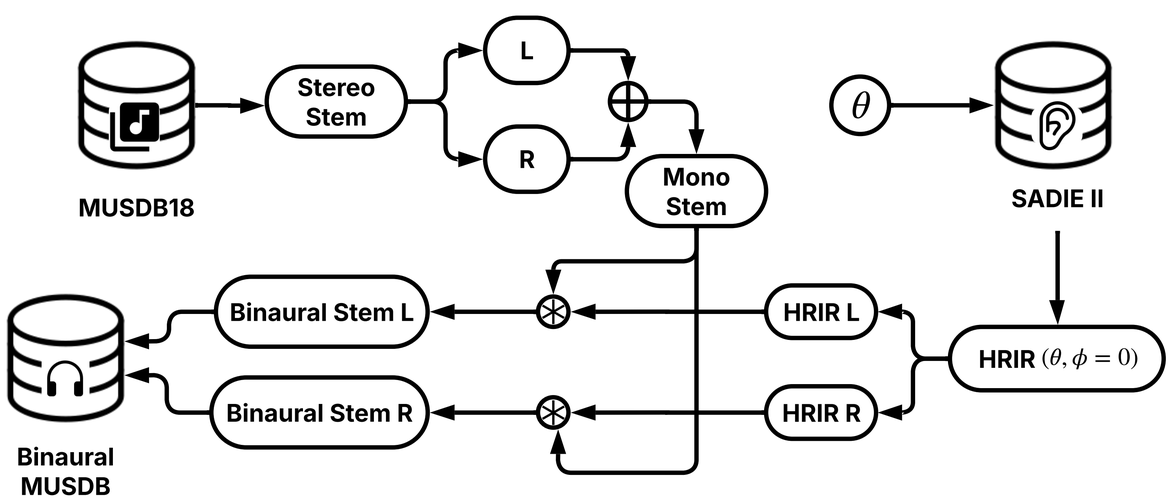
Do Music Source Separation Models Preserve Spatial Information in Binaural Audio?

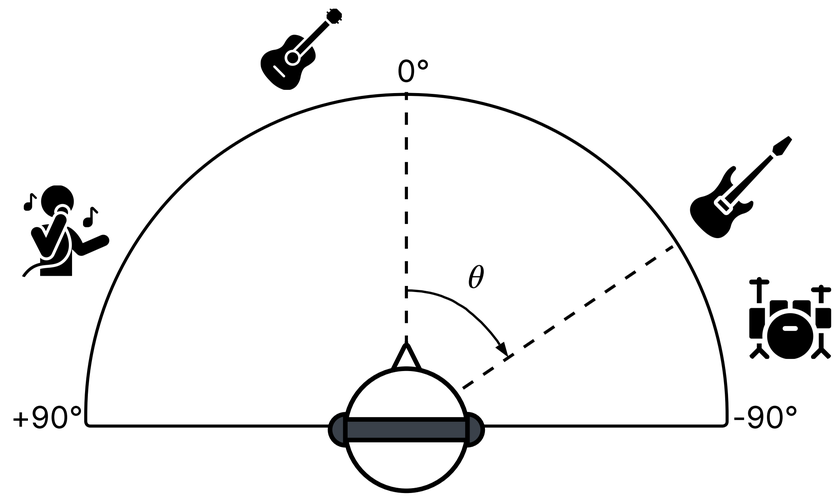
Richa Namballa [1] Dr. Agnieszka Roginska [1] Dr. Magdalena Fuentes[1, 2]

[1]Music and Audio Research Laboratory, New York University, New York, USA

[2]Integrated Design & Media, New York University, New York, USA



(a) An overview of the binaural synthesis process for the Binaural-MUSDB dataset.



(b) Binaural-MUSDB: each binaural source signal $\mathbf{s}_i$ is placed randomly along the front horizontal plane at an angle $\theta_i \in [-90°, +90°]$.

Figure 1

Background

- Immersive experiences (VR / AR) have gained popularity, requiring realistic audio stimuli.
- **Binaural audio** goes beyond standard gain-based stereo panning – filters two-channel audio to create interaural cues differing in level, time, and spectral content to simulate the location of a source in space [1, 2].
- Typically reproduced on headphones and has real-time applications in accessibility [3].
- Binaural music has received minimal attention in the music information retrieval (MIR) research community, especially in the task of **musical source separation (MSS)**.

Goal: investigate whether existing (stereo) MSS models are able to separate binaural mixtures into their respective stems while preserving spatial characteristics.

Binaural-MUSDB

- Synthesized a binaural version of MUSDB18-HQ [4] called **Binaural-MUSDB** to compare the performances models in both stereo and binaural settings.
- Figure 1a shows the synthesis process using **head-related transfer functions (HRTFs)** from the SADIE II database [5].
- Limited source locations on the horizontal plane to $\theta \in [-90°, +90°]$ along the **azimuth**, with the elevation fixed at $\phi = 0°$ (Figure 1b).
- For every song, 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.
- In a mixture, no two sources were allowed to be located at the same angle; there was a minimum of 10°separation (no direct overlap) between each stem.
- 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.

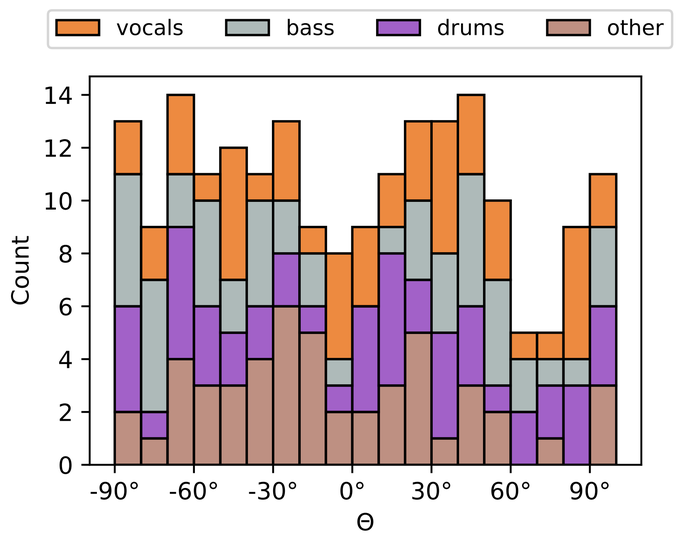


Figure 2. Distribution of instrument positions in the test set of Binaural-MUSDB.

Metrics

Aimed to describe the distortion introduced by the models.

Interaural Cues:

Change (Δ) in interaural time difference (ITD) and interaural level difference (ILD) between the estimated stem ($\hat{\mathbf{s}}$) and the reference stem ($\mathbf{s}$) [3].

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

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

$$\Delta\text{ILD} = |\text{ILD}(\mathbf{s}) - \text{ILD}(\hat{\mathbf{s}})| \tag{3}$$

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

Energy Ratios:

Signal to Spatial Distortion Ratio (SSR) captures the spatial distortion introduced by the separation ($\mathbf{e}_{\text{spat}}$) into the estimated stem ($\hat{\mathbf{s}}$).

Signal to Residual Distortion (SRR) quantifies the non-spatial distortion and errors such as interference and artifacts ($\mathbf{e}_{\text{resid}}$) [6].

$$\text{SSR}(\hat{\mathbf{s}}; \mathbf{s}) = 10 \cdot \log_{10}\left(\frac{||\mathbf{s}||^2}{||\mathbf{e}_{\text{spat}}||^2}\right) \tag{5}$$

$$\text{SRR}(\hat{\mathbf{s}}; \mathbf{s}) = 10 \cdot \log_{10}\left(\frac{||\tilde{\mathbf{s}}||^2}{||\mathbf{e}_{\text{resid}}||^2}\right) \tag{6}$$

$\tilde{\mathbf{s}}$ is the projection of $\mathbf{s}$ into $\hat{\mathbf{s}}$.

Results

Dataset	Model	Overall (Median)			
		SSR (dB) ↑	SRR (dB) ↑	ΔITD (μs) ↓	ΔILD (dB) ↓
Binaural	Demucs	10.59	6.91	68.03	0.39
	OpenUnmix	10.43	3.51	90.7	0.50
	Spleeter	9.86	2.01	22.68	0.64
Stereo	Demucs	**16.01**	**7.39**	**0.00**	**0.08**
	OpenUnmix	10.73	3.14	**0.00**	0.12
	Spleeter	10.78	3.21	**0.00**	0.12

Dataset	Model	Bass (Median)			
		SSR (dB) ↑	SRR (dB) ↑	ΔITD (μs) ↓	ΔILD (dB) ↓
Binaural	Demucs	9.13	**8.90**	476.19	0.20
	OpenUnmix	10.94	3.37	521.54	0.41
	Spleeter	10.63	1.53	544.22	0.44
Stereo	Demucs	**17.18**	8.36	**0.00**	**0.08**
	OpenUnmix	9.74	1.72	**0.00**	0.12
	Spleeter	8.69	1.25	**0.00**	0.15

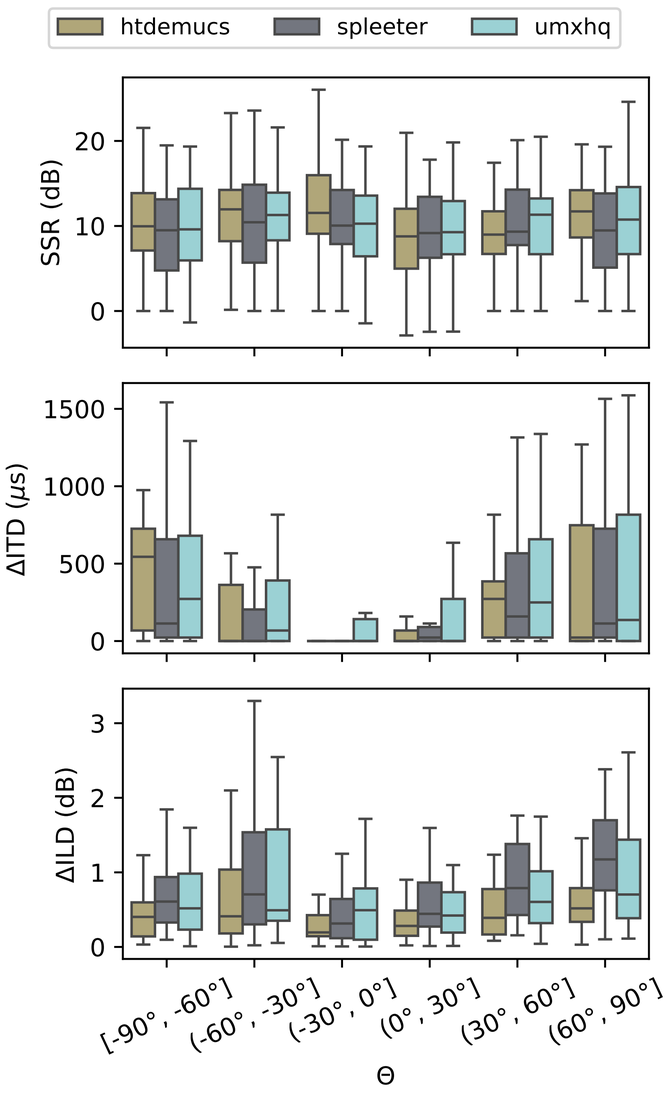


Figure 3. Distributions of spatial metrics aggregated across all sources.

Future Work

- Stability of random binaural synthesis and diverse HRIRs.
- Better understanding of metrics' sensitivity and relationship to existing immersive audio metrics.
- Robust perceptual evaluation studies.
- Evaluate more recent state-of-the-art MSS models.
- Train baseline model on binaural data and modify loss function to penalize spatial distortion.

Listen Here!

References

[1] 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.

[2] A. Roginska and P. Geluso, *Immersive Sound*. Focal Press, 2017.

[3] 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*, pp. 1–15, 2023.

[4] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ - an uncompressed version of MUSDB18," Dec. 2019.

[5] 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.

[6] K. N. Watcharasupat and A. Lerch, "Quantifying spatial audio quality impairment," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 746–750, IEEE, 2024.