Data-Scarce Audio Source Separation: A Comparison of Classical Methods and Deep Learning

A Genre-Wise SI-SDR Analysis of Speech and Music Separation with Limited Training Data

Namo Samuthrsindh
Massachusetts Institute of Technology
paramuth@mit.edu

Angel Cervantes
Massachusetts Institute of Technology
cerangel@mit.edu

Abstract

We explore the challenge of separating speech from instrumental music using a constrained dataset, where each source contains 100 seconds of audio. For each speech–music pair, we jointly split the data: the first 80 seconds from both the speech and the music are used for training, and the remaining 20 seconds from each source are reserved for testing. These test segments are then split and combined to create four distinct test mixtures. The dataset spans five source pairs per genre across eight distinct music genres. In this paper, we compare four approaches to speech–music separation under identical data splits: non-negative matrix factorization (NMF) with Kullback–Leibler multiplicative updates, principal component analysis (PCA), independent component analysis (ICA), a U-Net operating in the short-time Fourier transform (STFT) domain trained with a scale-invariant signal-to-distortion ratio (SI-SDR) loss, and an augmented U-Net that incorporates a global self-attention bottleneck.

Across all 40 source pairs and 160 test mixtures, the attention-augmented U-Net achieves the highest mean SI-SDR for both sources, reaching +5.17 dB for speech and +11.52 dB for music. This model consistently outperforms the baseline U-Net (+3.81 dB for speech and +10.48 dB for music), while the baseline U-Net only slightly surpasses NMF (+2.61 dB for speech and +9.66 dB for music). A closer examination of the U-Net results shows that its gains over NMF are most pronounced in genres with less repetitive structure and richer timbral complexity, such as Classical and Jazz. This suggests that methods relying on fixed bases are more limited in these settings. The addition of global self-attention appears particularly beneficial, as it helps capture long-range spectral dependencies that are important for effective separation. In contrast to U-Net and NMF, both PCA and ICA perform substantially worse, particularly in speech reconstruction, likely due to the way components are allocated between speech and music in our separation algorithms.

While the limited dataset size and controlled train/test setup restrict broad generalization, these results still provide a useful comparison of model behavior under low-data conditions.

1 Introduction

Separating speech from background music is a classic problem in audio processing, with real-world applications ranging from enhancing dialogue in videos to restoring old broadcasts. The traditional approach is to decompose source magnitude spectrograms onto pre-fit non-negative or linear bases, but in today's world, approaches using deep learning are becoming more common. These deep learning solutions learn to generate a time-frequency mask end-to-end. Each method has its own tradeoffs, especially when the data is scarce.

This paper evaluates four algorithms on a dataset designed to mimic a challenging practical scenario where the model needs to isolate a known speaker from a known piece of music with little adaptation data. For each speech-music pair, we limit the training audio to just 80 seconds per source. We then test the models on four short mixtures created from held out chunks. Then we break down our performance metrics by music genre to see exactly where each algorithm's underlying assumptions help or hurt the separation process.

2 Related Works

Audio source separation has evolved greatly over time, shifting from statistical signal processing to deep neural networks. Early approaches relied heavily on matrix factorization and subspace methods. For example, Non-negative Matrix Factorization (NMF) [6], Independent Component Analysis (ICA) [1], and Principal Component Analysis (PCA) [3] are the foundational techniques that separate sources by identifying distinct spectral bases or statistically independent components. While computationally lightweight and easy to deploy, these methods struggle when sources share overlapping frequency content or complex timbres.

Recently, deep learning has become more and more popular in audio source separation. Architectures like the U-Net [7], initially developed for biomedical image segmentation, have been successfully adapted for audio by predicting time-frequency masks directly over spectrograms. Modern systems like Spleeter [4] use this approach with massive datasets to train generalized models. However, our work diverges from these large-scale paradigms by focusing on a low-resource regime instead of aiming for general-purpose separation. We evaluate how both classical techniques and modern neural architectures adapt when only a fraction of training data is available, measuring their success using standardized metrics like the scale-invariant signal-to-distortion ratio (SI-SDR) [8].

3 Dataset

For the music source, we use the `Reubencf/fma-labeled` dataset [2], restricting the selection to samples without lyrics to ensure that the music stems do not contain overlapping speech components. Speech data is obtained from the `AIDC-AI/Marco_Longspeech` dataset [10]. The specific choice of datasets is not critical to the approach, as comparable results should be achievable using alternative datasets with similar characteristics.

In our experiment, we consider eight musical genres: Hip-Hop, Pop, Rock, Electronic, Folk, Classical, Jazz, and Metal. For each

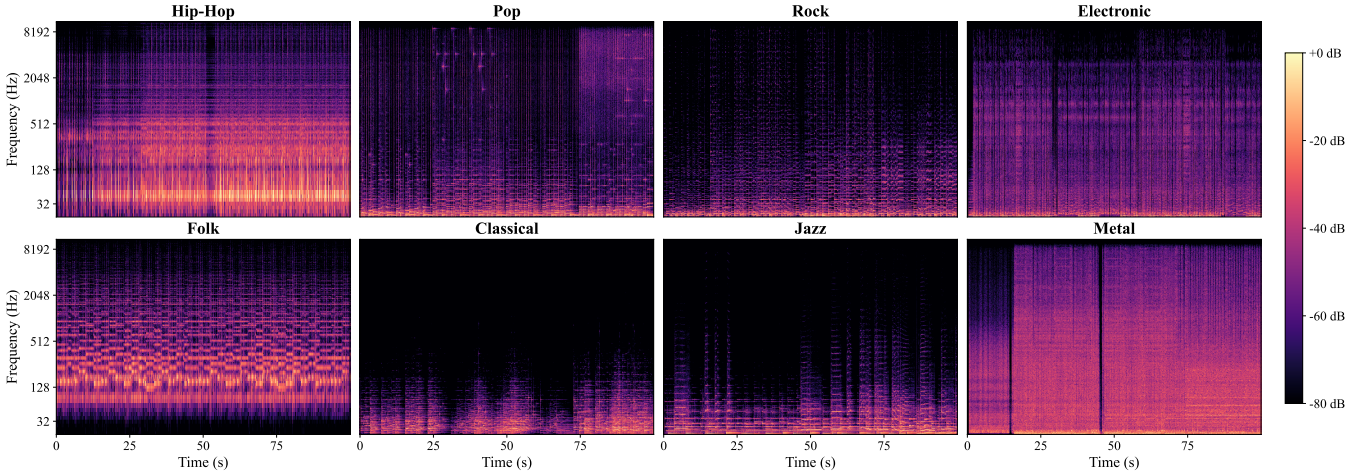

Figure 1: Representative log-magnitude spectrograms for each genre. The frequency axis is shown on a logarithmic scale, and amplitudes (in dB) are normalized relative to the maximum within each genre.

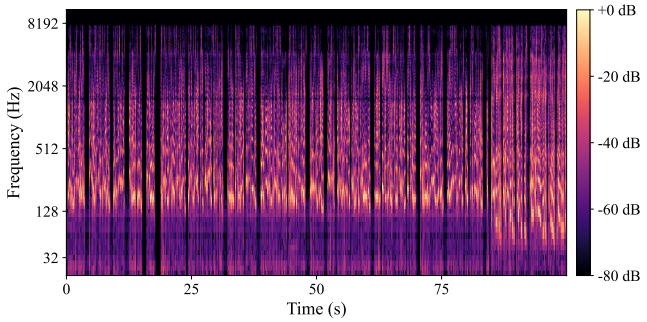

Figure 2: Representative log-magnitude spectrogram of a speech signal. The frequency axis is shown on a logarithmic scale, and amplitudes (in dB) are normalized relative to the maximum of the speech signal.

genre, five instrumental tracks are sampled and paired with distinct speech recordings, resulting in 40 speech-music pairs. All source recordings are resampled to 22,050 Hz and segmented into ten consecutive 10-second chunks. Figure 1 shows the spectrogram of a sampled 100-second music recording per genre, while Figure 2 presents a spectrogram for a sampled 100-second speech recording.

For each pair, the first eight segments are used for training, while the remaining two are reserved for testing. Let y_{sp} and y_{mu} denote the time-domain speech and music segments, respectively. Test mixtures are constructed by averaging the two sources,

$$y_{\text{mix}} = \frac{1}{2}\left(y_{\text{sp}} + y_{\text{mu}}\right),$$

where y_{mix} is the resulting mixture segment. All four possible combinations of the test segments are used, yielding a total of 160 test mixtures.

4 Methods

All four algorithms operate on magnitude short-time Fourier transform (STFT) representations computed with FFT size $N_{\text{FFT}} = 2048$, hop length $H = 512$, and a Hann window. The STFT produces a complex spectrogram $D \in \mathbb{C}^{F\times T}$, where $F = N_{\text{FFT}}/2 + 1$ is the number of frequency bins and T is the number of time frames. Also, let $V \in \mathbb{R}^{F\times T}$ denote the magnitude spectrogram, i.e., V is obtained as the element-wise magnitude of D.

Performance is evaluated using SI-SDR. Let s and $\hat{s}$ denote the reference and estimated time-domain signals, respectively, both mean-normalized prior to evaluation. SI-SDR is defined as

$$\text{SI-SDR}(\hat{s}, s) = 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|\hat{s} - s_{\text{target}}\|^2},$$

where the projected target component is

$$s_{\text{target}} = \frac{\langle \hat{s}, s\rangle}{\|s\|^2 + \epsilon}\, s,$$

and ϵ is a small constant for numerical stability. Higher SI-SDR indicates better reconstruction quality.

4.1 NMF

To perform source separation using NMF [6], we learn per-source basis matrices $W_{\text{sp}}, W_{\text{mu}} \in \mathbb{R}^{F\times K}$, where $K = 32$ is the number of components per source. NMF minimizes the Kullback–Leibler divergence [5] using multiplicative updates for 1500 iterations.

For separation, V is factorized using the concatenated matrix $W = [W_{\text{sp}} \,|\, W_{\text{mu}}] \in \mathbb{R}^{F\times 2K}$ to get an activation matrix $H \in \mathbb{R}^{2K\times T}$. After that, H is partitioned into H_{sp} and H_{mu}, yielding source spectrogram estimates $V_{\text{sp}} = W_{\text{sp}}H_{\text{sp}}$ and $V_{\text{mu}} = W_{\text{mu}}H_{\text{mu}}$. Separation masks are then computed as

$$M_{\text{sp}} = \frac{V_{\text{sp}}}{V_{\text{sp}} + V_{\text{mu}} + \epsilon}, \qquad M_{\text{mu}} = \frac{V_{\text{mu}}}{V_{\text{sp}} + V_{\text{mu}} + \epsilon},$$

and applied to the mixture spectrogram prior to inverse STFT to obtain time-domain estimates.

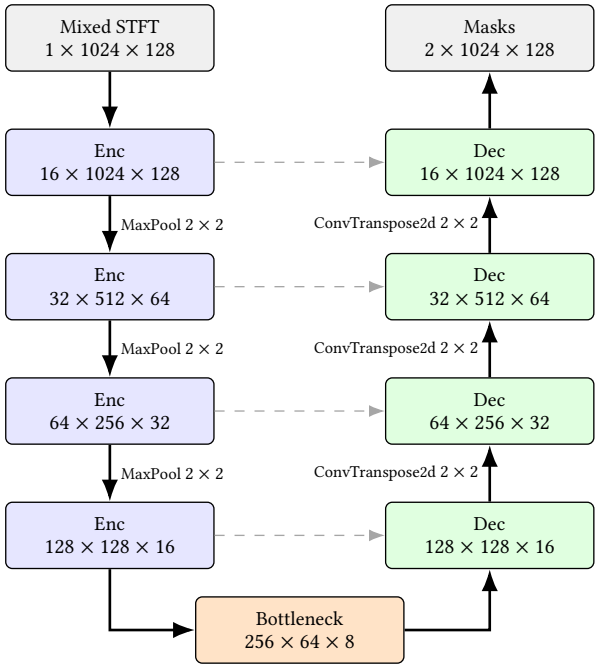

Figure 3: Four-level U-Net for STFT-domain source separation. The encoder uses 2×2 max-pooling for downsampling, while the decoder uses 2×2 transposed convolutions for upsampling. Skip connections fuse encoder and decoder features at matching resolutions. The model outputs two soft masks via a 1×1 convolution with sigmoid activation.

4.2 PCA and ICA

PCA and ICA follow a unified pipeline. We first construct a joint magnitude matrix $V_{\text{joint}} = [V_{\text{sp}} \mid V_{\text{mu}}] \in \mathbb{R}^{F \times 2T}$, formed by concatenating speech and music training spectrograms along the time axis. Each frequency bin is then standardized using the joint mean and standard deviation. A subspace model with rank $K = 32$ is learned using either PCA or ICA, yielding a set of components and corresponding activation scores. Components are assigned to either speech or music by comparing their average activation energy on the speech and music segments of the training data. In particular, components with higher energy on speech are assigned to W_{sp}, and the remainder to W_{mu}. If all components are assigned to a single source, a balanced split is used as a fallback.

For separation, the mixture magnitude spectrogram is standardized using the same statistics, and activations are estimated via a least-squares projection onto the concatenated basis $W = [W_{\text{sp}} \mid W_{\text{mu}}]$. Source spectrograms are then reconstructed from the corresponding basis–activation pairs and rescaled to the original domain. We then apply the masking trick followed by the inverse STFT, similar to the NMF case.

4.3 U-Net

A four-level encoder–decoder U-Net is used to estimate two masks corresponding to speech and music (Figure 3). The model operates on magnitude STFT representations with the Nyquist bin removed so that the frequency dimension is divisible by 16. Each input is a log-magnitude spectrogram, normalized per sample.

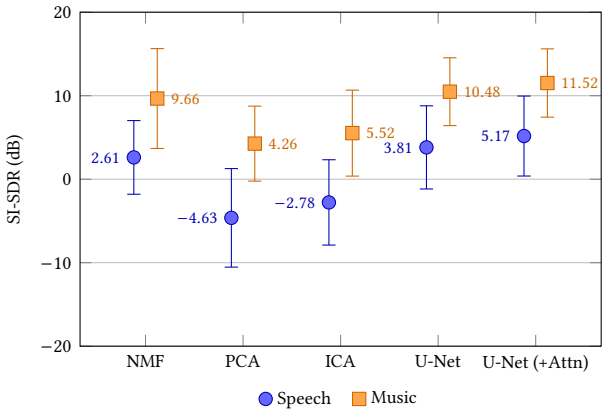

Figure 4: Comparison of source separation performance across selected methods, measured by SI-SDR (dB). Marks indicate mean performance over 160 test mixtures, with error bars representing ±1 standard deviation.

The architecture consists of four encoder blocks followed by a bottleneck and a symmetric decoder with skip connections. Each block contains two convolutional layers with Group Normalization and LeakyReLU activations. Downsampling and upsampling are performed using max-pooling and transposed convolutions, respectively.

Training uses mixture generation from random 3-second segments sampled from the training set, where each source is independently perturbed by up to 3 dB. The model is trained to predict soft masks applied in the STFT domain, and source estimates are reconstructed via inverse STFT. The optimization objective is the mean negative SI-SDR over the two sources:

$$\mathcal{L} = -\tfrac{1}{2}\big(\text{SI-SDR}(\hat{s}_{\text{sp}}, s_{\text{sp}}) + \text{SI-SDR}(\hat{s}_{\text{mu}}, s_{\text{mu}})\big).$$

Optimization is performed with a batch size of 256, for up to 50 epochs. A single held-out validation set, constructed from one fixed chunk per source, is used for model selection. Early stopping is applied if the validation loss does not improve for 8 consecutive epochs. At inference, signals are processed in 3-second overlapping windows (50% overlap) using Hann-windowed overlap-add reconstruction.

4.4 Attention-Augmented U-Net

In addition to the baseline U-Net, we evaluate an augmented U-Net where the standard bottleneck is replaced with a Global Self-Attention Bottleneck. After the initial convolutional cascade, the feature map $x \in \mathbb{R}^{B \times C \times F \times T}$ where $B = 256$ is the batch size, $C = 256$ is the number of channels, $F = 64$ is the frequency dimension, and $T = 8$ is the time-frame dimension–is flattened spatially into a sequence of tokens $x_{\text{flat}} \in \mathbb{R}^{B \times FT \times C}$. A multi-head self-attention layer with 8 heads [9] is applied to allow each time-frequency bin to attend to every other bin in the 3-second chunk. The sequence is then reshaped back to the original dimensions before being passed to the decoder.

	NMF		PCA		ICA		U-Net (Base)		U-Net (+Attn)	
Genre	Speech	Music	Speech	Music	Speech	Music	Speech	Music	Speech	Music
Hip-Hop	+4.16±5.66	+16.64±5.01	−9.36±5.25	+6.63±3.09	−4.07±3.52	+11.23±4.06	+2.76±4.16	+14.71±2.40	+4.64±4.63	+16.02±3.18
Pop	+2.20±1.84	+10.80±3.09	−6.26±3.14	+5.04±2.34	−4.17±2.73	+6.27±1.88	+3.29±1.86	+11.07±2.03	+4.26±1.95	+11.66±2.51
Rock	+1.45±6.34	+9.33±4.00	−4.46±5.46	+4.65±3.90	−3.36±5.93	+5.36±3.41	+1.25±7.29	+9.52±3.10	+2.59±6.92	+10.41±2.88
Electronic	+3.94±3.84	+11.11±4.77	−4.79±6.81	+4.93±3.36	−2.21±5.04	+7.03±4.13	+4.31±3.36	+10.89±5.31	+6.34±3.31	+12.42±4.41
Folk	+3.61±2.66	+5.99±6.49	−0.64±6.27	+1.99±4.03	+0.46±5.55	+3.06±4.42	+4.38±4.56	+6.22±4.16	+6.23±3.64	+7.96±4.73
Classical	+3.49±4.18	+5.08±6.82	−0.72±4.82	+0.71±6.88	+0.25±4.42	+1.63±8.13	+7.71±5.41	+9.07±3.80	+8.70±4.89	+9.99±4.19
Jazz	+1.88±4.42	+7.49±5.22	−2.80±4.22	+3.91±5.03	−1.75±4.65	+4.90±4.43	+5.89±4.57	+11.13±2.86	+6.92±4.25	+11.89±3.02
Metal	+0.12±2.59	+10.85±1.98	−7.99±3.76	+6.23±1.49	−7.37±3.24	+4.65±0.78	+0.91±2.40	+11.28±1.95	+1.70±2.55	+11.78±1.72

Table 1: Comparison of source separation performance across genres, measured by SI-SDR (dB). Values denote mean ± standard deviation over 20 mixtures per genre, where best results per genre are highlighted (speech in blue, music in teal).

5 Results

We evaluate four separation algorithms across eight musical genres using SI-SDR as the primary metric. Results are presented from two perspectives: Table 1 reports genre-level SI-SDR for speech and music sources, while Figure 4 aggregates these results into global averages per method.

The results exhibit three observations. (1) Music separation performance is higher than speech separation performance across all methods and genres. (2) The attention-augmented U-Net achieves the highest average performance, followed by the baseline U-Net and NMF. The baseline U-Net remains closely competitive with NMF, with the two methods showing comparable results across several settings. (3) PCA and ICA fail to reliably reconstruct speech, often producing negative SI-SDR, while still maintaining partial effectiveness for music.

5.1 Music vs. Speech

A first consistent trend is a performance gap between music and speech. As shown in Figure 4, music SI-SDR exceeds speech SI-SDR for all methods: NMF (+7.05 dB), PCA (+8.89 dB), ICA (+8.30 dB), U-Net (+6.67 dB), and Augmented U-Net (+6.35 dB). This gap is also observed across all genres in Table 1, suggesting that the structure of music signals is more favorable for separation than speech signals under our selected models.

A finer-grained view indicates that this difference is not uniform across genres. Genres with strong rhythmic regularity tend to yield higher SI-SDR, whereas less structured genres consistently show lower performance. For example, Hip-Hop achieves the highest and most stable performance across all methods, which is consistent with its strong repetitive structure, as illustrated in Figure 1. In contrast, Folk and Classical consistently show lower performance. The reason is that, these genres exhibit less repetition and more variable temporal structure, making them more similar in nature to speech signals, as shown in Figure 2.

5.2 Comparison of Leading Methods

Figure 4 shows that the Attention-Augmented U-Net achieves the highest overall performance among all evaluated methods, surging to +5.17 dB for speech and +11.52 dB for music. The baseline U-Net and NMF follow, reaching +3.81 / +10.48 dB and +2.61 / +9.66 dB, respectively. Beyond average performance, differences in robustness across genres are also observed. Both U-Net variants generally exhibit more stable performance, especially in structured musical settings. For instance, U-Net shows lower deviation (2.40 dB) compared to NMF (5.01 dB) on Hip-Hop, suggesting a consistent separation behavior in highly regular rhythmic conditions for U-Net.

More pronounced differences are observed in genres with less regular structure. In particular, the Attention U-Net outperforms NMF by +5.31 dB on Classical speech and +4.99 dB on Classical music, which is consistent with the hypothesis that nonlinear models with global attention are better suited to capture complex and less structured time–frequency patterns. As shown in Figure 1, Classical and Jazz exhibit less repetitive temporal structure and a more dispersed time–frequency representation, reducing the effectiveness of linear additive models such as NMF.

The comparison between the baseline and attention-augmented U-Nets serves as a target ablation of the bottleneck architecture. The baseline U-Net architecture was already outperforming classical methods, but the addition of global self-attention led to a substantial gain of +1.36 dB on speech and +1.04 dB on music globally.

We observe that the attention mechanism provides the largest relative gains in Classical (+0.99 dB speech) and Jazz (+1.03 dB speech). In these genres, musical events have long-range temporal dependencies and complex harmonic overtones that the local convolutions might struggle with in the baseline U-Net architecture. By flattening the spatial dimensions into a sequence, the attention layer allows the model to see the entire 3-second context simultaneously, resolving which spectral peaks belong to the instruments vs the formants of the speaker.

5.3 Failure on Speech Separation

PCA and ICA exhibit a qualitatively different behavior from the other methods. On speech, both methods produce negative SI-SDR values: −4.63 dB for PCA and −2.78 dB for ICA. In contrast, both methods still achieve positive SI-SDR for music, suggesting that their learned representations retain partial structure. This asymmetry can be traced to the shared-component decomposition and assignment procedure used in both methods.

PCA and ICA are trained on a concatenation of speech and music, and the resulting components are assigned to either speech or music based on activation energy. Music signals typically exhibit stronger and more structured patterns, resulting in higher activation across

a larger number of components. Consequently, a disproportionate number of components are assigned to the music source, leading to an imbalance in the component allocation. As a larger fraction of components is assigned to the music subspace, relatively few components are available to represent speech. The number of components for speech representation is therefore underparameterized, which limits reconstruction quality.

6 Limitations

Several limitations should be considered when interpreting our results. Firstly, the number of experiments done per genre is relatively small. We use five music-speech pairs per genre, with only four mixtures as tests per pair. This limited sample size may not be sufficiently representative of the full diversity within each genre class, potentially affecting the generalizability of our findings. Furthermore, all music data as well as speech data is drawn from a single dataset, raising concerns about diversity and the representativeness of our benchmark.

Apart from that, no hyperparameter tuning was performed for any of the evaluated methods. We manually chose all hyperparameters. Thus, the reported performance may underestimate the true capability of each method, and comparisons across methods may not reflect optimal configurations. Moreover, we evaluated only a single architectural and feature configuration for each method. For example, mel spectrograms could be explored as an alternative input representation to the STFT, or the U-Net architecture could be extended to predict the separated signal directly rather than a mask. Exploring these design choices could yield meaningful performance improvements and provide a more complete picture of each method's potential.

7 Conclusion

Across all methods, music consistently yields higher SI-SDR performance than speech. This is likely because music tends to have a more repetitive structure and less frequency variation compared to human speech, which exhibits dynamic temporal patterns and a broader spectral range.

Among all evaluated schemes, the U-Net augmented with a global self-attention bottleneck achieved the best SI-SDR performance, decisively outperforming the baseline U-Net, NMF, ICA, and PCA. While the baseline U-Net's advantage over NMF was narrow and inconsistent across genres, the inclusion of the attention mechanism allows the model to effectively resolve overlapping spectral features, yielding substantial improvements in genres like Classical and Jazz. Given that the U-Net architectures require substantially greater computational resources to train, NMF remains a practical choice when computational resources are strictly limited, but the attention-augmented U-Net is the clear choice for maximizing separation quality.

References

[1] Pierre Comon. Independent component analysis, a new concept? *Signal Process.*, 36(3):287–314, April 1994.

[2] Reuben Fernandes. Fma labeled — multi-attribute music dataset (gemini), 2026. Labels generated with gemini-flash-latest on the Creative-Commons subset of the Free Music Archive.

[3] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. PCA beginnings.

[4] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020.

[5] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[6] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[8] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr - half-baked or well done?, 2018.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[10] Fei Yang, Xuanfan Ni, Renyi Yang, Jiahui Geng, Qing Li, Chenyang Lyu, Yichao Du, Longyue Wang, Weihua Luo, and Kaifu Zhang. Longspeech: A scalable benchmark for transcription, translation and understanding in long speech. *arXiv preprint arXiv:2601.13539*, 2026.