Music Track Separation Techniques: From Classical Mathematical Methods to Modern Deep Learning and Transformer Approaches

Janaka V. Wijayakulasooriya

July 22, 2025

Contents

1 Introduction

Music track separation, also known as *source separation* or *music demixing*, refers to isolating individual sound sources (vocals, drums, bass, etc.) from a mixed audio recording.

Applications

- Karaoke systems
- Remixing and music production
- Music information retrieval
- Audio enhancement and restoration
- Forensic audio analysis

2 Classical Mathematical Techniques

2.1 Spectral Subtraction

Concept: Estimate the noise spectrum and subtract it from the noisy signal.
Use in separation: Removes background noise or simple accompaniment from vocals.
Limitation: Introduces musical noise artifacts.

Reference: Boll, S. F. (1979). *Suppression of acoustic noise in speech using spectral subtraction.* IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(2), 113-120. doi:10.1109/TASSP.1979.1163209

2.2 Independent Component Analysis (ICA)

Concept: Assumes observed signals are linear mixtures of statistically independent sources.
Algorithm: FastICA.
Limitations:

- Requires at least as many microphones (channels) as sources.
- Limited performance in underdetermined (single-channel) scenarios.

Reference: Hyvärinen, A., & Oja, E. (2000). *Independent component analysis: algorithms and applications.* Neural Networks, 13(4-5), 411-430. doi:10.1016/S0893-6080(00)00026-5

2.3 Non-negative Matrix Factorization (NMF)

Concept: Decomposes a non-negative spectrogram matrix V into basis W and activations H such that $V \approx WH$.
Application: Useful for monaural music separation.
Limitations:

- Requires manual selection of number of components.
- Results depend on initialization.

Reference: Lee, D. D., & Seung, H. S. (1999). *Learning the parts of objects by non-negative matrix factorization.* Nature, 401(6755), 788-791. doi:10.1038/44565

2.4 Sinusoidal Modeling

Concept: Models sound as a sum of time-varying sinusoids plus noise.
Use case: Effective for separating tonal components (e.g. vocals) from noise-like percussive components.

Reference: Serra, X., & Smith, J. O. (1990). *Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition.* Computer Music Journal, 14(4), 12-24. doi:10.2307/3680784

3 Modern Machine Learning Techniques

3.1 Deep Neural Networks (DNN)

Concept: Learn complex non-linear mappings from mixture spectrograms to source spectrograms.
Example: Deep clustering and mask-inference networks for music separation.

Reference: Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016). *Deep clustering and conventional networks for music separation: Strong together.* ICASSP 2016. doi:10.1109/ICASSP.2016.7471631

3.2 Convolutional Neural Networks (CNN)

Concept: Capture local patterns in spectrograms for mask estimation.
Example: U-Net architectures for source separation.

Reference: Jansson, A., et al. (2017). *Singing voice separation with deep U-Net convolutional networks.* ISMIR 2017. PDF Link

3.3 Recurrent Neural Networks (RNN) and LSTM

Concept: Model temporal dependencies for sequential data like audio.
Example: Bidirectional LSTM layers for better temporal context in vocal separation.

Reference: Huang, P.-S., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2015). *Joint optimization of masks and deep recurrent neural networks for monaural source separation.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(12), 2136-2147. doi:10.1109/TASLP.2015.2468583

3.4 Deep Clustering

Concept: Maps each time-frequency bin to an embedding space where sources are clustered separately.

Reference: Hershey et al. (2016). Same as above.

4 Transformer-Based State-of-the-Art Approaches

4.1 Transformers for Music Separation

Concept: Self-attention mechanisms model long-range dependencies efficiently.

Example 1: Demucs

Architecture: Encoder-decoder with bidirectional LSTM and convolutional layers.
Tool: Demucs by Facebook AI Research.

Reference: Défossez, A., Usunier, N., Bottou, L., & Bach, F. (2019). *Music source separation in the waveform domain.* arXiv preprint arXiv:1911.13254. arXiv Link

Example 2: Spleeter

Tool: Deezer's Spleeter for vocals and accompaniment separation.
Reference: Hennequin, R., et al. (2020). *Spleeter: A fast and efficient music source separation tool with pre-trained models.* Journal of Open Source Software, 5(50), 2154. doi:10.21105/joss.02154

Example 3: SpecTNT

Architecture: Transformer-based Time-domain Network leveraging global context.
Reference: Lin, J., et al. (2021). *SpecTNT: Spectro-temporal neural network for music source separation.* arXiv preprint arXiv:2107.01902. arXiv Link

5 Evaluation Metrics

- **SDR**: Signal-to-Distortion Ratio
- **SIR**: Signal-to-Interference Ratio
- **SAR**: Signal-to-Artifacts Ratio

Reference Benchmark: SiSEC MUS Dataset (SiSEC 2018)

6 Summary Table

Technique	Key Idea	Limitations	Example
Spectral Subtraction	Noise spectrum subtraction	Musical noise artifacts	Speech enhancement
ICA	Statistical independence	Needs multiple channels	Stereo source separation
NMF	Non-negative factorization	Initialization dependency	Vocal/accompaniment separation
DNN	Learn complex mappings	Requires large datasets	Deep clustering networks
CNN	Capture local features	Limited global context	U-Net vocal separation
RNN/LSTM	Temporal modeling	Vanishing gradients	Bidirectional LSTM
Transformers	Self-attention for global dependencies	High compute	SpecTNT, Demucs

7 Conclusion

Music track separation has evolved from **classical signal processing techniques** to **powerful deep learning and transformer-based models**, enabling near-professional isolation quality for applications in creative production, intelligent retrieval, and immersive user experiences.

Recommended Further Reading

1. Vincent, E., et al. (2018). *Audio source separation: A practical introduction.* Wiley.

2. Luo, Y., & Mesgarani, N. (2019). *Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation.* IEEE/ACM TASLP, 27(8), 1256-1266.

3. Défossez, A. (2021). *Hybrid spectrogram and waveform source separation.* arXiv preprint arXiv:2106.05212.