Computational Analysis of Sound and Music

Music Information Retrieval – Source Separation

Dr.-Ing. Jakob Abeßer

Fraunhofer IDMT

jakob.abesser@idmt.fraunhofer.de

Source Separation

Outline

- **Source Separation**
    - Introduction
    - Tasks
    - Traditional Method
    - DL-based Methods

Source Separation

Introduction

- Music recordings
    - Mixtures of different musical instruments (sources) playing simultaneously
- Sound Separation
    - Reverse engineering the audio mixing process
    - Output: 1 stem per instrument

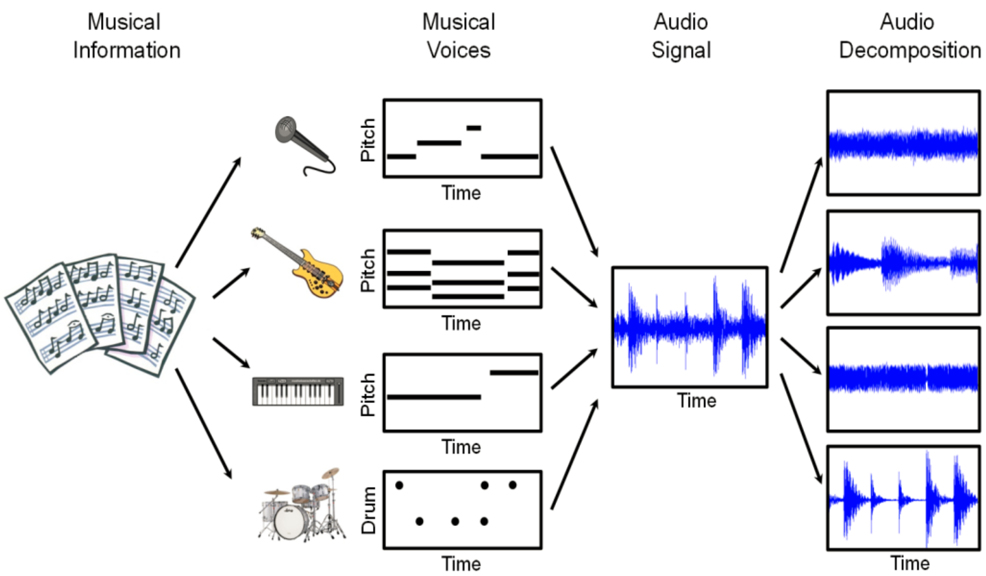


Fig-M5-1

Source Separation

Introduction

- Audio mix is influenced by
  - Instrument characteristics (timbre, note decay, …)
  - Musical performance (timing, dynamics, playing techniques, …)
  - Recording chain (microphones, room acoustics)

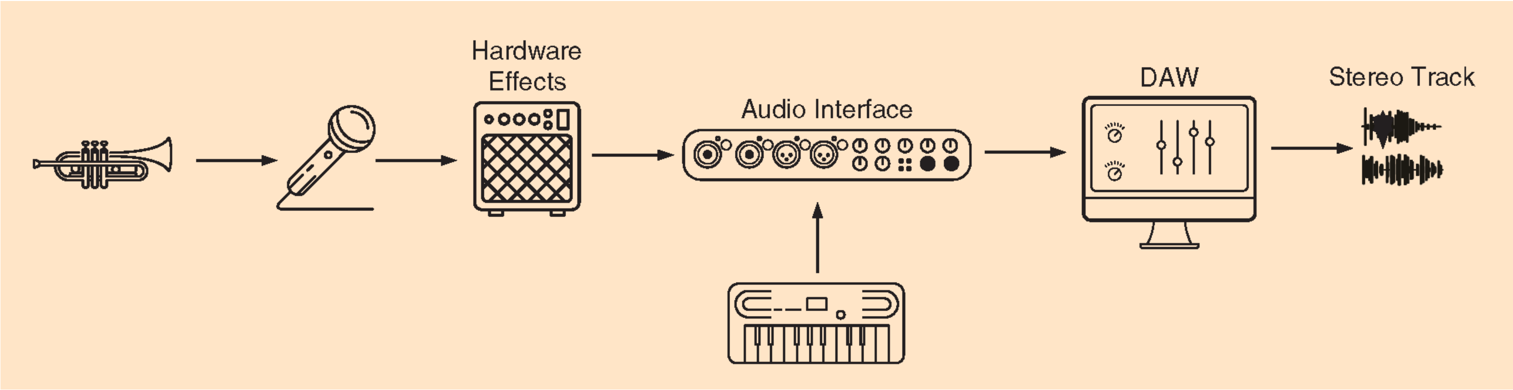


Fig-M5-2

Source Separation

Tasks

- Audio remixing
- Audio upmixing
  - Mono → stereo
  - Stereo → 5.1
- Music Analysis
  - Transcription, beat tracking, harmony analysis etc.
- Music Education
  - Solo / Backing track generation

Source Separation

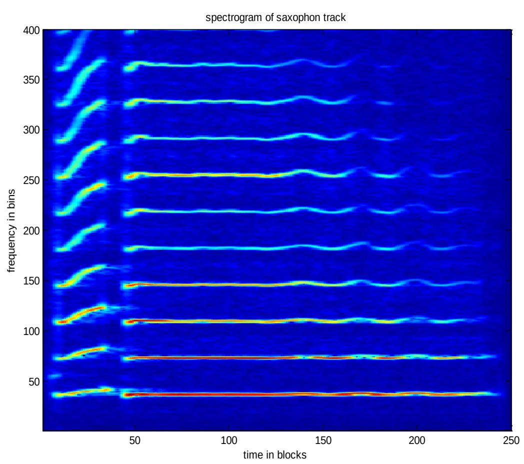
Tasks

- Harmonic/percussive separation
  - H → stable harmonic components (fundamental frequency, overtones)
  - P → transient components (drum sounds, note attacks)
- Solo/accompaniment separation
  - S → predominant melody instrument
  - A → accompanying instruments
- Singing voice separation
  - S → singing voice (male / female)
  - A → band
- Separation of all sources

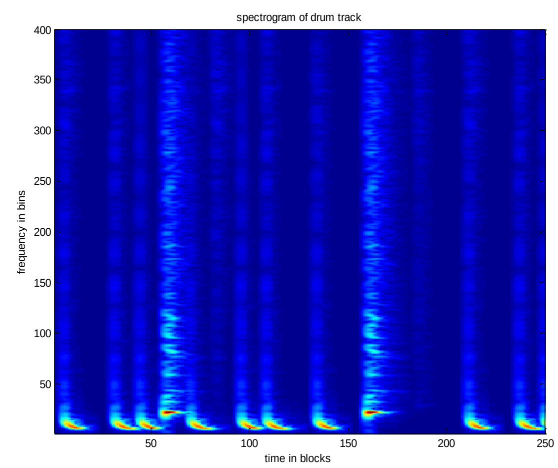
Source Separation

Traditional Approaches

- Harmonic/percussive (H/P) separation
    - Different spectral characteristics of harmonic and percussive signals



Own



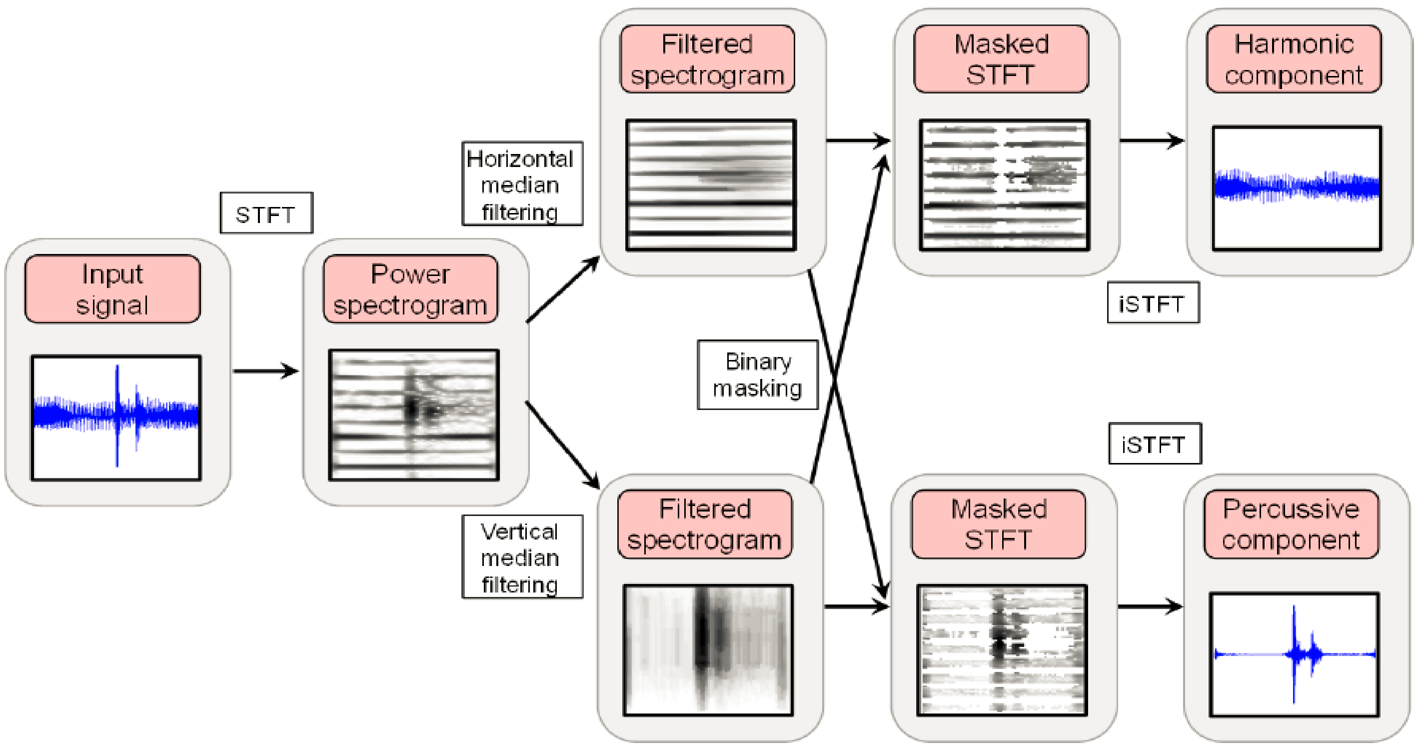
Own

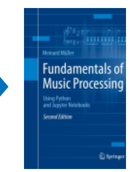
- Time-continuous (horizontal)
- Localized in frequency

- Wide-band (vertical)
- Localized in time

Source Separation

Traditional Approaches





FMP Notebooks

Fig-M5-3

Source Separation

Tasks

- Phase-based H/P separation
    - Harmonic sources → phase change values are predictable
    - Percussive sources → unpredictable phase (noise-like characteristics)
    - Instantaneous Frequency Distribution (IFD)
        - How does phase change over time?

$$\Phi(k,n) = \frac{1}{2\pi}\frac{d\phi(k,n)}{dn}$$

Instantaneous Frequency

Gradient of (unwrapped) phase over time

Source Separation

Tasks

- Phase-based H/P separation
    - Harmonic mask → phase change within range / predictable?

$$H(k,n) = \begin{cases} 1 & \text{if } \Delta_{k_{Low}} < \Phi(k,n) < \Delta_{k_{High}} \\ 0 & \text{otherwise} \end{cases}$$

    - Percussive mask

$$P(k,n) = 1 - H(k,n)$$

Source Separation

DL-based Approaches

- U-Net based [Jannson et al., 2017]
    - Input → magnitude spectrogram (mix)
    - Output → 2 soft masks (voice / other
- Issue
    - Only magnitude of STFT is modeled
    - Still phase from the mixture is used

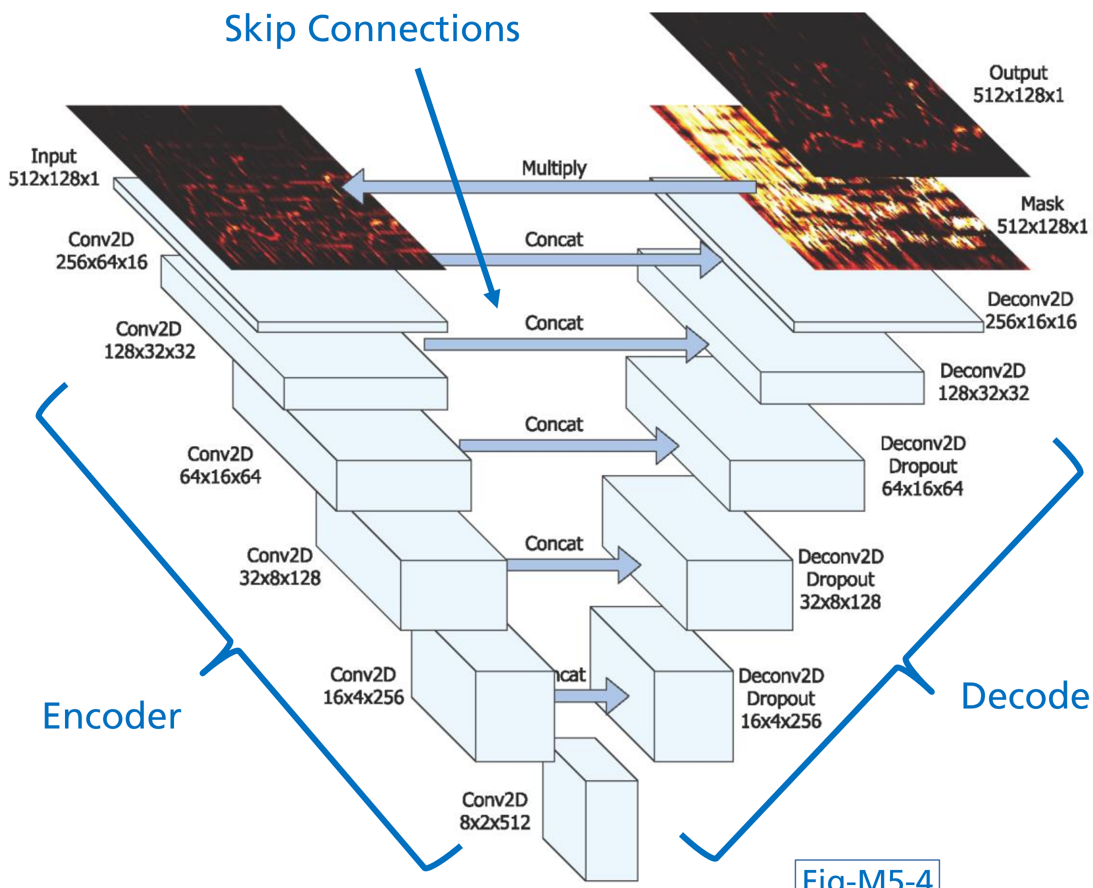


Fig-M5-4

Source Separation

DL-based Approaches

- Spleeter [Hennequin et al., 2020]
    - Open-source version for MIR research
    - 3 pre-trained models
        - 2 stems (vocals and accompaniments)
        - 4 stems (vocals, drums, bass, and other)
        - 5 stems (vocals, drums, bass, piano and other)

Spleeter Demo

Source Separation

DL-based Approaches

- Conv-TasNet [Luo & Mesgarani, 2019]
  - Time-domain speech separation network (end-to-end)
  - Encoder → optimized representation for speaker separation
  - Seperation → masks (weighting functions)
  - Decoder → invert to waveforms
  - Temporal convolutional networks (TCN)
    - Stack of 1-D dilated convolutional blocks
    - Large receptive field → model long-term dependencies

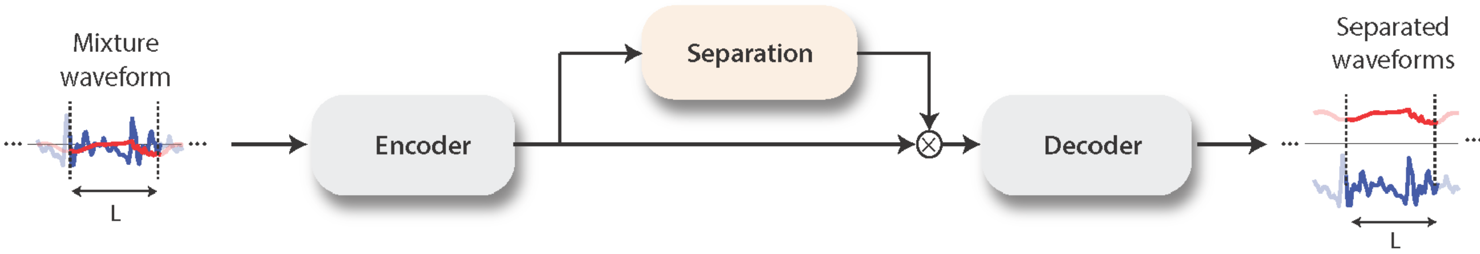


Fig-M5-5

Source Separation

DL-based Approaches

- Conv-TasNet [Luo & Mesgarani, 2019]

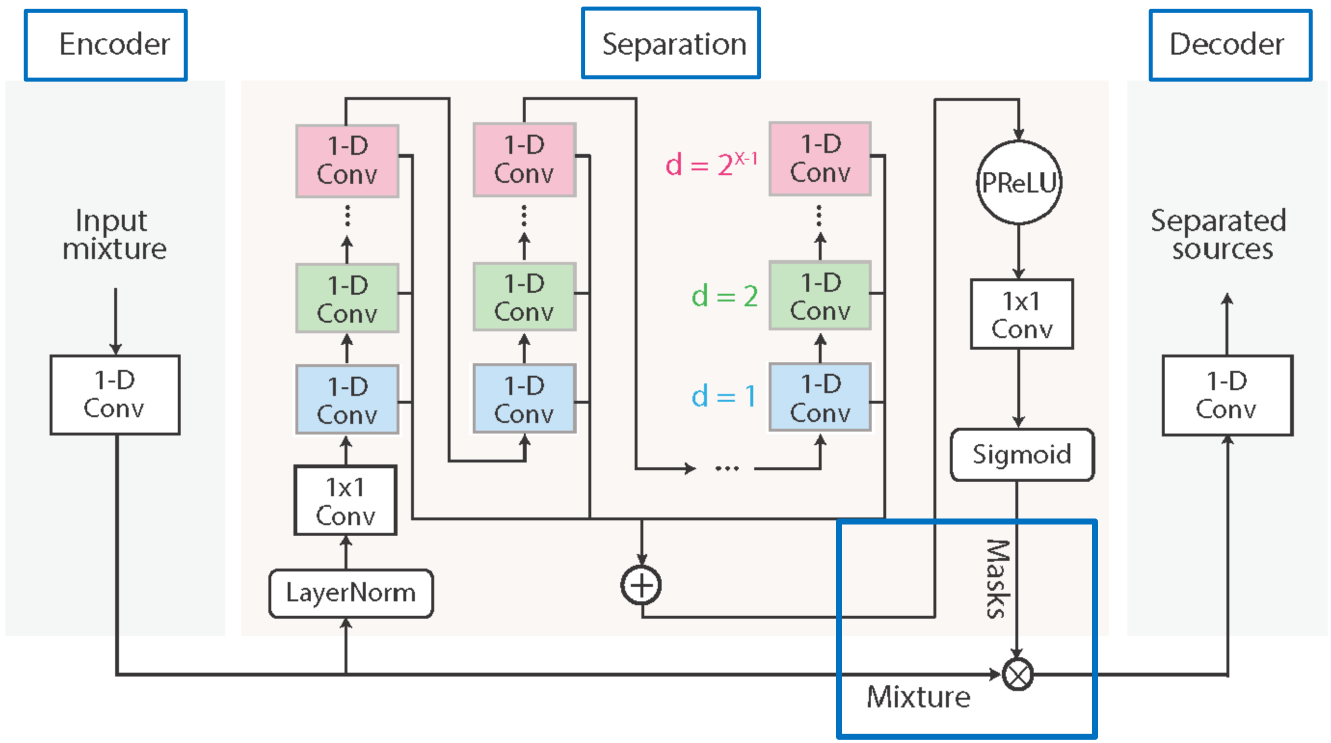


Fig-M5-6

Evaluation Metrics

Objective Metrics

- Signal-to-Distortion Ratio (SDR)

$$SDR = 10 \log_{10} \left(\frac{||s_{\text{target}}||^2}{||e_{\text{total}}||^2} \right)$$

  - Higher SDR – higher separation quality

- s_{target} – original source signal
- e_{total} – total error between separated and original signal

- Signal-to-Interference Ratio (SIR)

$$SIR = 10 \log_{10} \left(\frac{||s_{\text{target}}||^2}{||e_{\text{interf}}||^2} \right)$$

  - Higher SIR – better isolation from other sources

- e_{interf} – interference error (from other sources)

Evaluation Metrics

Objective Metrics

- Signal-to-Artifacts Ratio (SAR)

$$SAR = 10\log_{10}\left(\frac{\left\|s_{\text{target}}\right\|^2}{\left\|e_{\text{artif}}\right\|^2}\right)$$

  - Higher SAR – higher separation quality & fewer artifacts

- e_{artif} –artifact error (unwanted distortions from separation)

Evaluation Metrics

Perceptual Metrics

- PEASS (Perceptual Evaluation methods for Audio Source Separation) [Emiya et al., 2010]
    - Set of metrics to assess the perceptual quality of separated signals.
    - Overall Perceptual Score (OPS): Overall perceived quality.
    - Target-related Perceptual Score (TPS): Perceived quality of the target signal.
    - Interference-related Perceptual Score (IPS): Perceived level of interference.
    - Artifacts-related Perceptual Score (APS): Perceived level of artifacts.

Source Separation Research - Online Demos

- Time-Domain Source Separation
  - Online Demo: Separation of Vocals, Bass, Drums, Others
- Score-Informed Drum Separation
  - Online Demo: Non-Negative Matrix Factor Deconvolution (NMFD) for decomposing drum breakbeats into kick drum, snare drum, and hi-hat
- Cascaded Harmonic-Residual-Percussive Separation
  - Online Demo: Mid-level timbre feature to describe timbre changes in music recordings

References

Images

Fig-M5-1: [Müller, 2021], p. 422, Fig. 8.1

Fig-M5-2: [Cano et al., 2019], p. 3, Fig. 3

Fig-M5-3: [Müller, 2021], p. 425, Fig. 8.3

Fig-M5-4: [Jansson, 2017], p. 3, Fig. 1

Fig-M5-5: [Luo & Mesgarani, 2019], p. 3, Fig. 1(A)

Fig-M5-6: [Luo & Mesgarani, 2019], p. 3, Fig. 1(B)

References

References

Müller, M. (2021). Fundamentals of Music Processing - Using Python and Jupyter Notebooks (2nd ed.). Springer.

Cano, E., Fitzgerald, D., Liutkus, A., Plumbley, M. D., & Stoter, F. R. (2019). Musical Source Separation: An Introduction. IEEE Signal Processing Magazine, 36(1), 31–40.

Emiya, V., Vincent, E., Harlander, N., Hohmann, V. (2010): The PEASS Toolkit - Perceptual Evaluation methods for Audio Source Separation. International Conference on Latent Variable Analysis and Signal Separation, St. Malo, France.

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., & Weyde, T. (2017). Singing Voice Separation with Deep U-Net Convolutional Networks. Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 745–751. Suzhou, China

Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software, 5(50), 2154.

Luo, Y., & Mesgarani, N. (2019). Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(8), 1256–1266.