Music demixing with the sliCQ transform

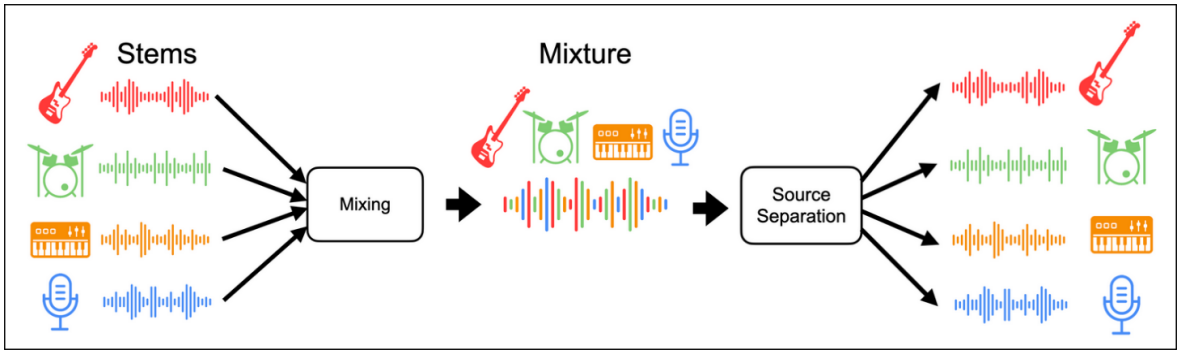
Sevag Hanssian

December 15, 2021

Music source separation, demixing, and unmixing

Music source separation: extract an estimate of an isolated source (or target) from mixed musical audio (e.g., harmonic/percussive, vocals, drums, bass, piano)

Music demixing (or unmixing): estimate multiple sources (vocals, drums, bass, other[1]) that can be summed back to the original mix. Multiple MSS subproblems, reversing the linear mixing of stems in the recording studio (stem datasets can be used for mixing and demixing)

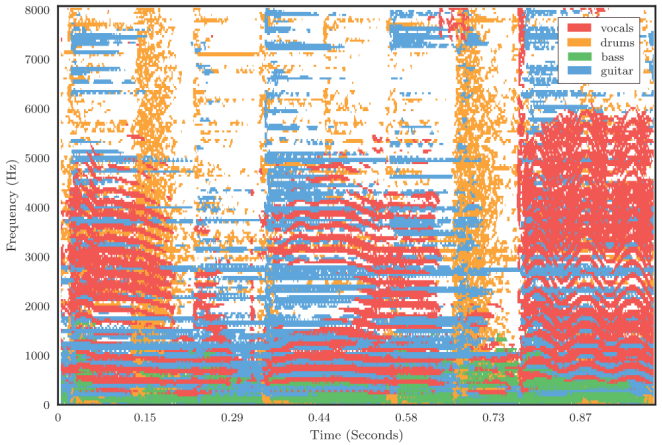


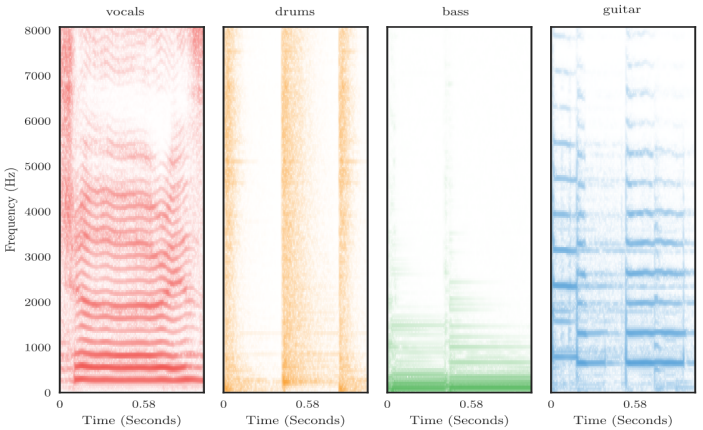
Popular approach: musical source models, which are "model-based approaches that attempt to capture the spectral characteristics of the target source"[2] with **time-frequency masks**

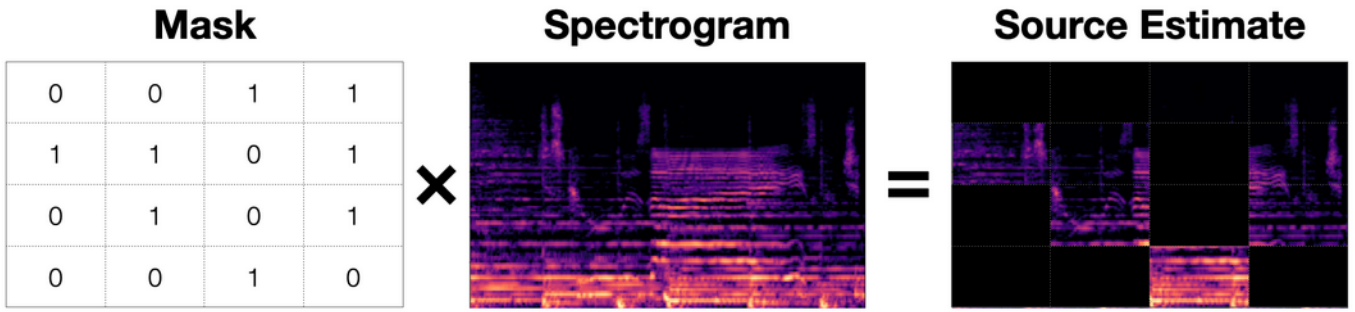
1 Zafar Rafii et al. 2019.

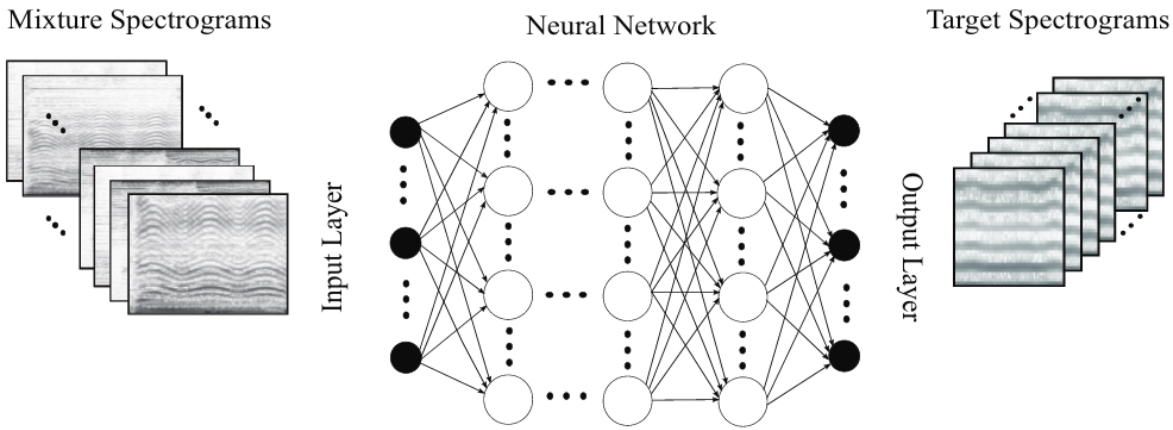
2 Estefanía Cano et al. (2018), p. 36.

Music source separation with DNN and spectrograms



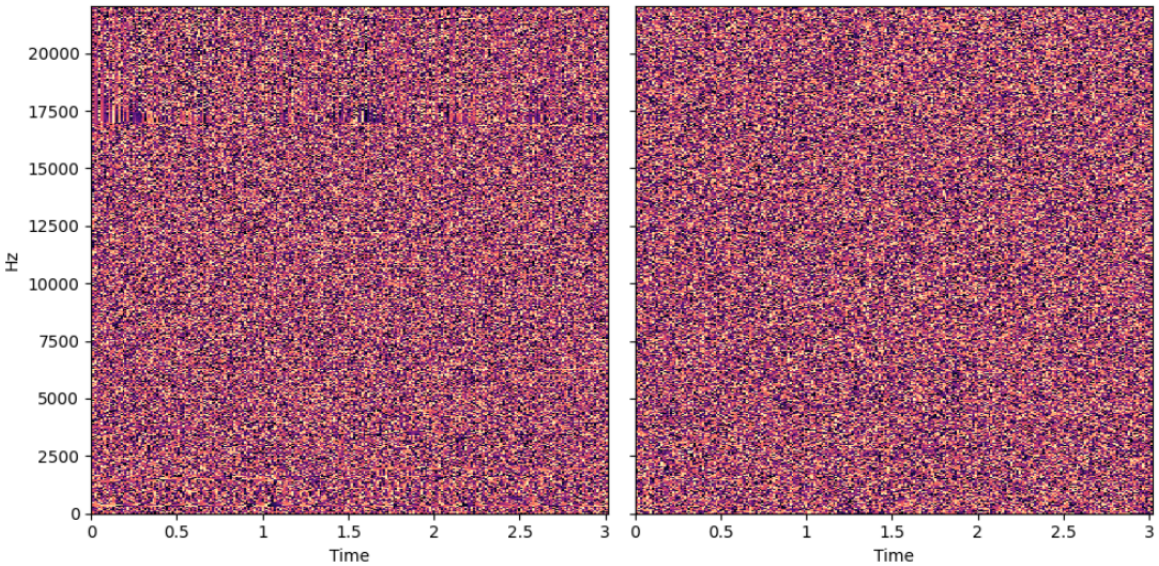






Phase performance ceiling

1. Simplifying assumption: estimate magnitude spectrograms, use the phase of the original mixed audio. Called "noisy phase"[3]. Done by Open-Unmix (UMX), CrossNet-Open-Unmix (X-UMX)[4], and many other popular & near-SOTA models
2. Why? Phase is hard to model![5]



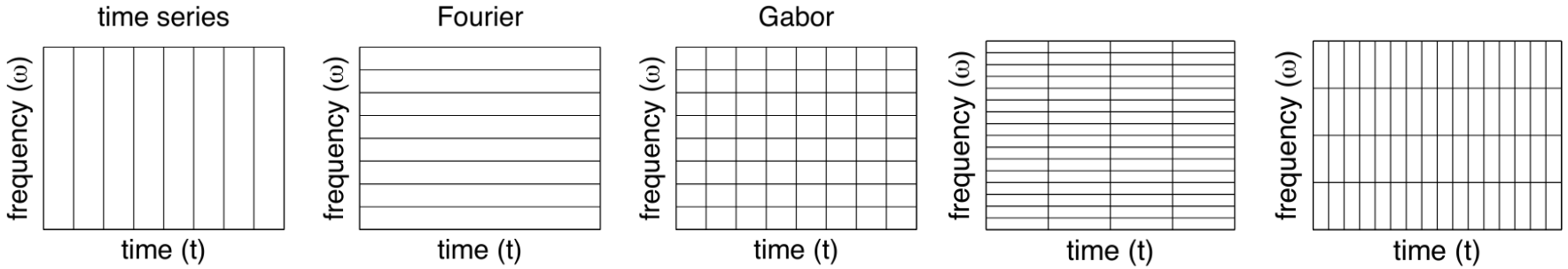
[3] Gordon Wichern et al. *arXiv preprint arXiv:1907.01160* (2019).

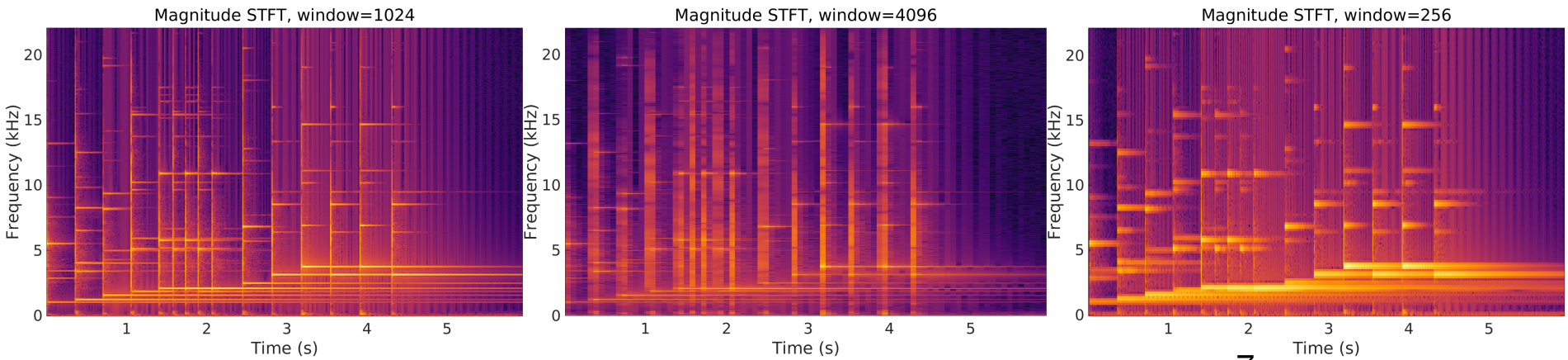
[4] Fabian-Robert Stöter et al. *Journal of Open Source Software* (2019); Ryosuke Sawata et al. *arXiv preprint arXiv:2010.04228* (2021).

[5] `https://source-separation.github.io/tutorial/basics/phase.html#why-we-don-t-model-phase`

Time-frequency tradeoff in the STFT and MDX

Joint time-frequency analysis is important for signals whose frequencies change with time.[6] Take the Fourier transform of local windows of the signal, i.e., the STFT. Change window size to trade off time and frequency:





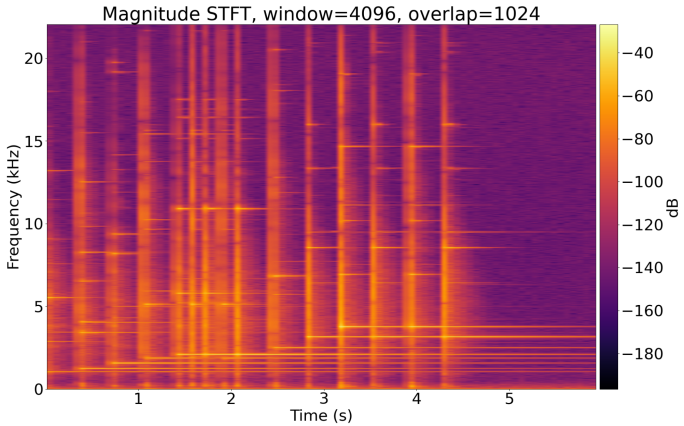
In music source separation, window size matters per-target.[7] Short-window for percussion, long-window for harmonic

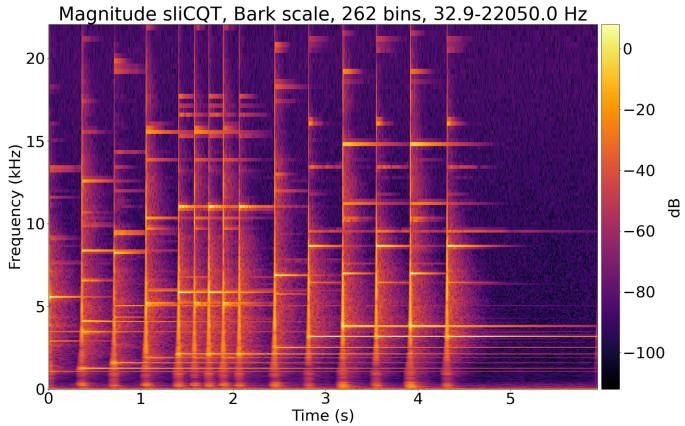
6 Dennis Gabor. *Journal of Institution of Electrical Engineers* (1946).

7 Andrew Simpson. *arXiv preprint arXiv:1504.07372* (2015).

CQT, NSGT, sliCQT

1. For musical and auditory reasons, we want high frequency resolution at low frequencies and high time resolution at high frequencies[8]
2. CQT[9] uses long windows in low frequencies and short windows in high frequencies for the 12-tone Western pitch scale
3. Nonstationary Gabor Transform (NSGT) and sliCQT[10] are TF transforms with Fourier coefficients, perfect inverse, and varying windows to create a varying time-frequency resolution
4. sliCQT params chosen for max quality of the noisy-phase waveform





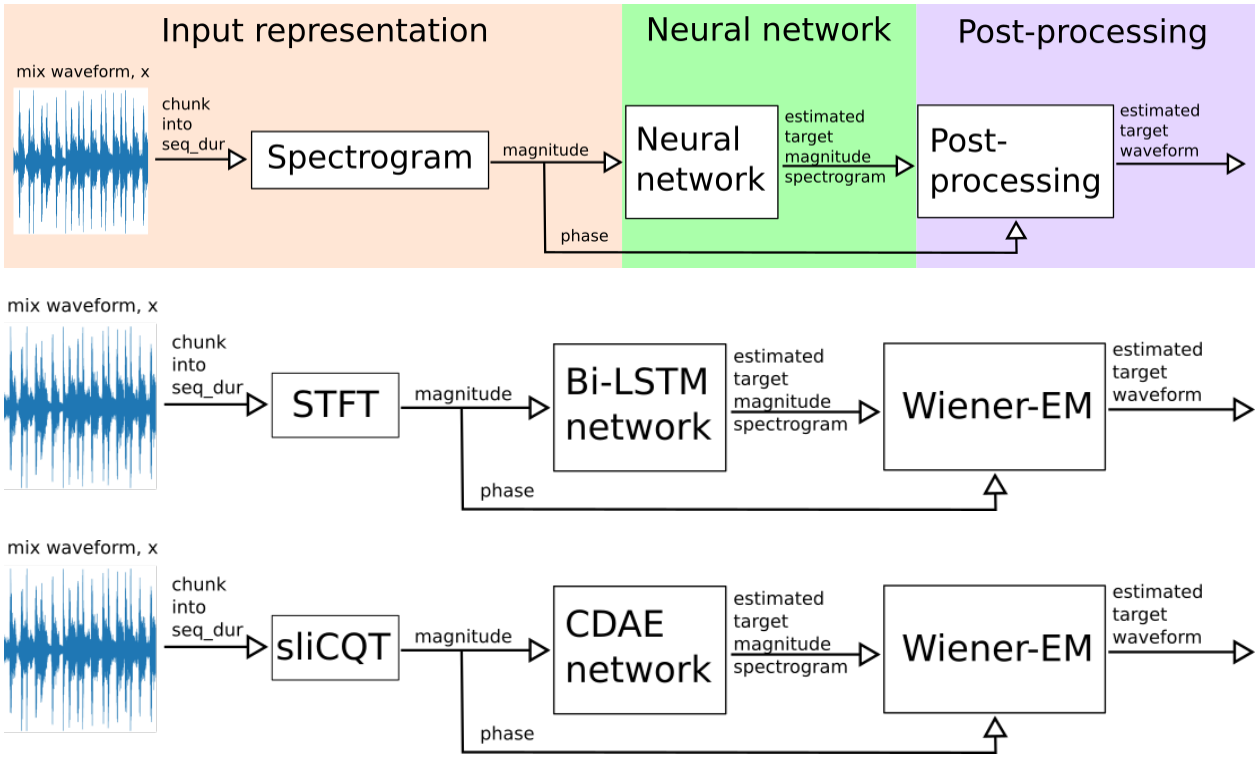
8 Monika Dörfler. PhD thesis. 2002.

9 Judith Brown. (1991).

10 Peter Balazs et al. (2011); Nicki Holighaus et al. (2013).

xumx-sliCQ

1. My goal: improve Open-Unmix by replacing STFT with sliCQT
2. My model submitted to the MDX21 challenge and workshop: `https://github.com/sevagh/xumx-sliCQ`
3. Use Convolutional Denoising Autoencoder[11] neural architecture
4. Scored 3.6 dB vs. 4.6 dB (UMX) and 5.54 dB (X-UMX); there is still room for improvement



11 Emad M. Grais et al. 2017; Emad M. Grais et al. 2021.

MDX 21 winners and current trends

1. Previously, music demixing systems were submitted to and evaluated at SiSEC (Signal Separation Evaluation Campaign). This year: MDX (Music Demixing Challenge) ISMIR 2021 @ AICrowd, follow-up MDX21 workshop, satellite @ ISMIR 2021
2. **ISMIR 2021**: Model that uses the complex spectrogram (i.e. includes phase) and uses complex masks[12]
3. **MDX21**: 1: Demucs++[13] (waveforms + complex spectrogram), 2: KUIELAB-MDX-Net[14] (waveforms + magnitude spectrogram), 3: Danna-Sep[15] (waveform + magnitude spectrogram, use complex spectrogram in loss function)

Properties in common: blending networks, waveforms (implicitly includes phase), complex spectrograms/masks, mixing spectrogram and waveform models

[12] Qiuqiang Kong et al. 2021.

[13] Alexandre Défossez. *arXiv preprint arXiv:2111.03600* (2021).

[14] Minseok Kim et al. *arXiv preprint arXiv:2111.12203* (2021).

[15] Chin-Yun Yu et al. *arXiv preprint arXiv:2112.03752* (2021).