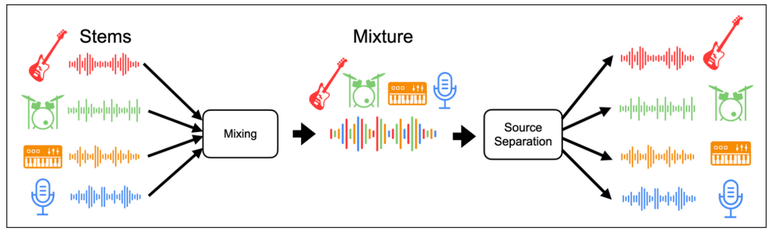
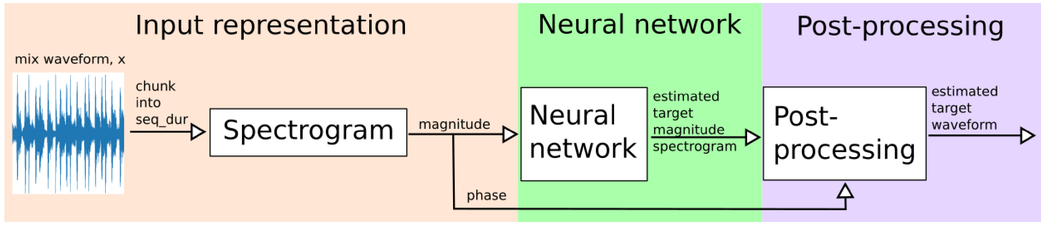
Music source separation and music demixing

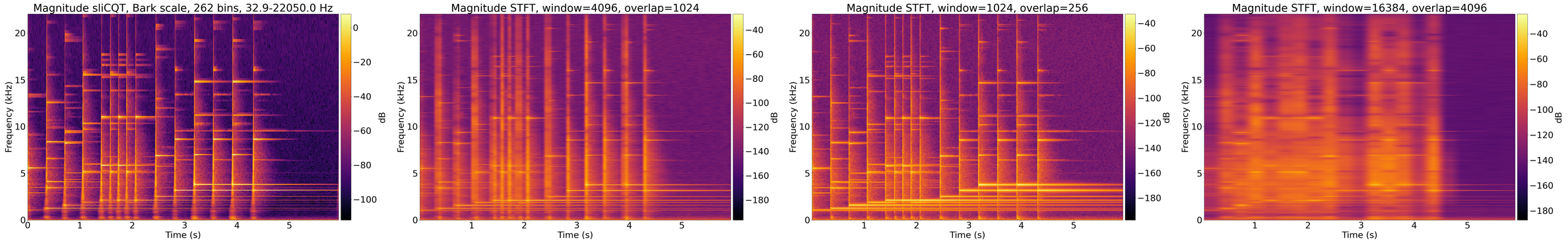
1. Music source separation is the task of extracting an estimate of one or more isolated sources or instruments (e.g., drums or vocals) from musical audio
2. Music demixing or unmixing separates the music into an estimate of **all** of its stems (that can be summed back to the original mixture)
3. Many music source separation models use magnitude spectrograms and discard phase (they use phase of the mix a.k.a the "noisy phase" to create a waveform)





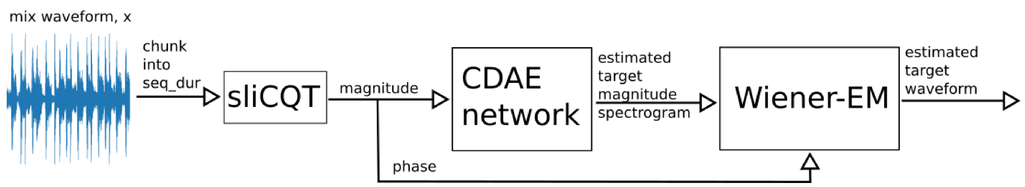
Time-frequency resolution: STFT, sliCQT, and human auditory system

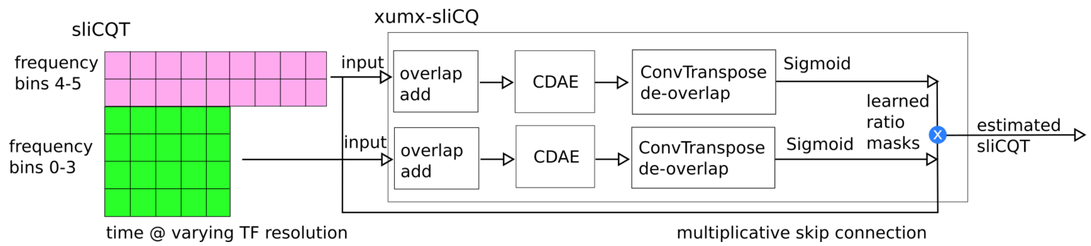
1. The Fourier transform represents the spectrum of audio as a sum of infinite sinusoids; interesting sounds (e.g., music, speech) have spectra that change with time
2. Short-Time Fourier Transform (STFT) or Gabor transform,[1] take the spectrum of consecutive, overlapping, finite-duration, fixed-size windowed frames of the audio signal
3. Signal contains only time information; spectrum contains only frequency information; STFT has a fixed time-frequency resolution determined by the window duration
4. "We have conducted the first direct psychoacoustical test of the Fourier uncertainty principle in human hearing, by measuring simultaneous temporal and frequency discrimination. Our data indicate that human subjects often beat the bound prescribed by the uncertainty theorem, by factors in excess of 10"[2]
5. Nonstationary Gabor Transform (NSGT) and sliCQT (realtime NSGT):[3] time-frequency transform with Fourier coefficients, varying time-frequency resolution, and perfect inverse. Can demonstrate good tonal/transient representation without a tradeoff, and captures more musical information than the STFT



Result: xumx-sliCQ

1. sliCQT parameters chosen by maximizing SDR of "noisy phase" oracle: **7.42 dB** SDR (median of sources and tracks) vs. 6.23 on MUSDB18-HQ validation set
2. Overall system adapted from UMX, XUMX, and CDAE:[4] convolutional layers applied to ragged sliCQT[5]
3. xumx-sliCQ:[6] **3.67 dB** SDR (median of sources and tracks) vs. 5.91 (xumx with STFT) on MUSDB18-HQ test set (trained only on MUSDB18-HQ)





[1]Allen et al., 1977; Gabor, 1946, [2]Oppenheim et al., 2012, p. 4, [3]Balazs et al., 2011; Holighaus et al., 2013, [4]Stöter et al., 2019; Sawata et al., 2021; Grais et al., 2021, [6]`https://github.com/sevagh/xumx-sliCQ`, [5]Hanssian, 2021