Better music demixing with sliCQT

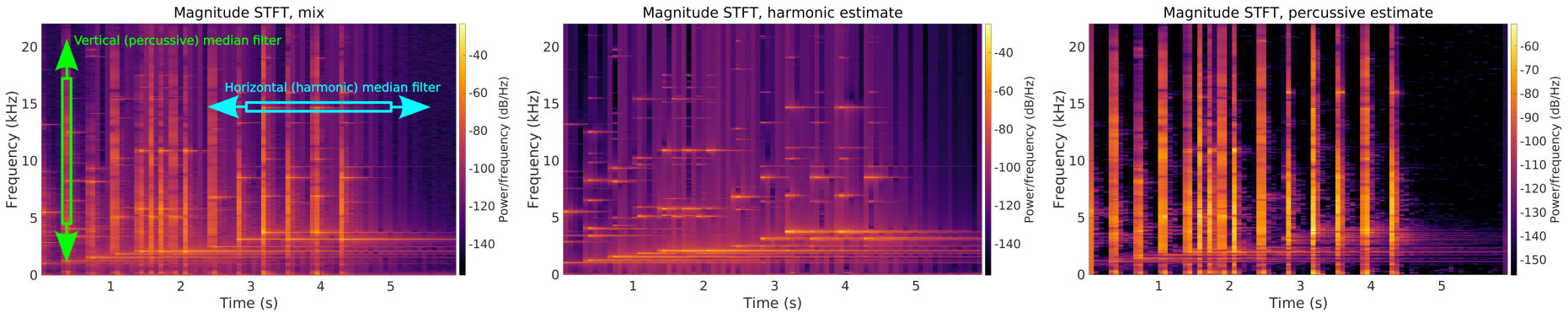
Submission to Cadenza Challenge CAD1

Sevag Hanssian

December 08, 2023

Time-frequency tradeoffs

Median-filtering harmonic/percussive source separation (HPSS)[1]



1. Short window (256) for percussion, long window (4096) for harmonic
2. Short-time Fourier Transform (STFT) window size matters per-target[2] in VDBO problems
3. In musical and auditory contexts, frequency resolution should increase from high to low frequencies (vice-versa for time resolution)[3]
4. CQT[4] uses long windows in low frequencies and short windows in high frequencies for the 12-tone Western pitch scale

1 Derry Fitzgerald. 2010; Jonathan Driedger et al. 2014.
2 Ilya Kavalerov et al. IEEE, 2019.
3 Christian Schörkhuber et al. 2012; Monika Dörfler. PhD thesis. 2002.
4 Judith Brown. (1991).

sliCQT vs. STFT

1. Nonstationary Gabor Transform (NSGT)[5], realtime sliCQ Transform[6]
2. STFT-like transforms with windows that vary with time
3. CQT motivates the NSGT/sliCQ, but can use any monotonically increasing frequency scale (log/cq, mel, Bark, etc.)
4. Outputs the familiar Fourier coefficients with perfect inverse

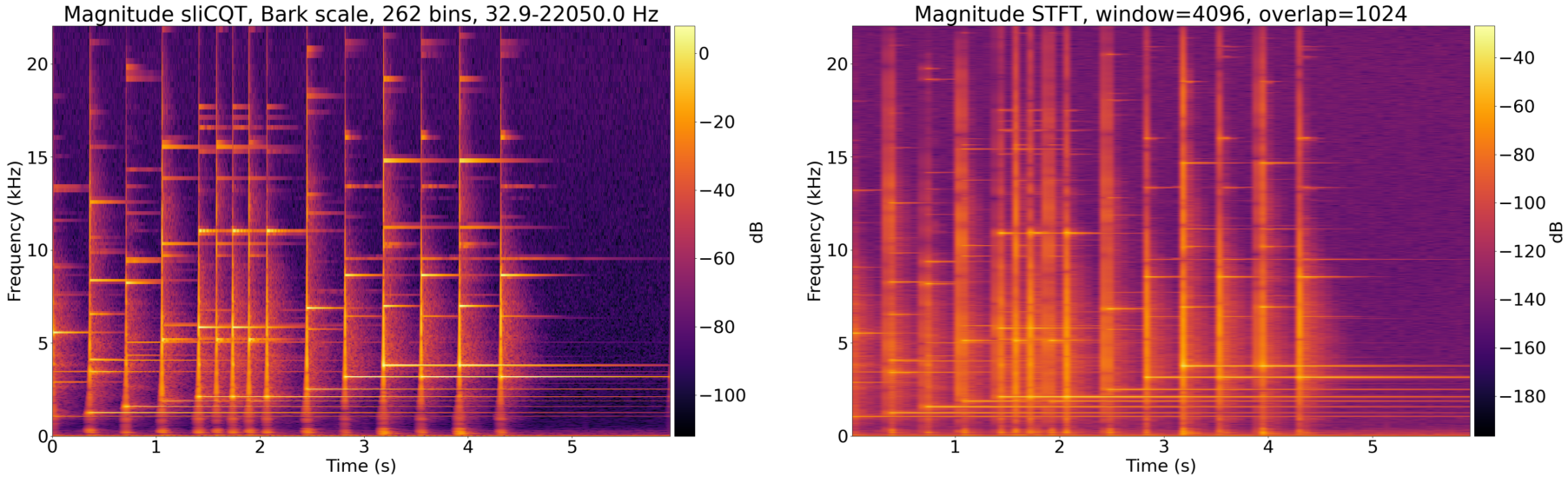


Figure: sliCQ: 262-bin Bark scale, 32.9–22050 Hz

[5] Peter Balazs et al. (2011).

[6] Gino Angelo Velasco et al. 2011; Nicki Holighaus et al. (2013); Christian Schörkhuber et al. 2014.

xumx-sliCQ v1 @ MDX 2021

Use sliCQT (Bark scale, 262 bins, 32.9–22050 Hz) and convolutional autoencoder architecture[7] to achieve 3.6 dB signal-to-distortion ratio (SDR): `https://github.com/sevagh/xumx-sliCQ/tree/v1`

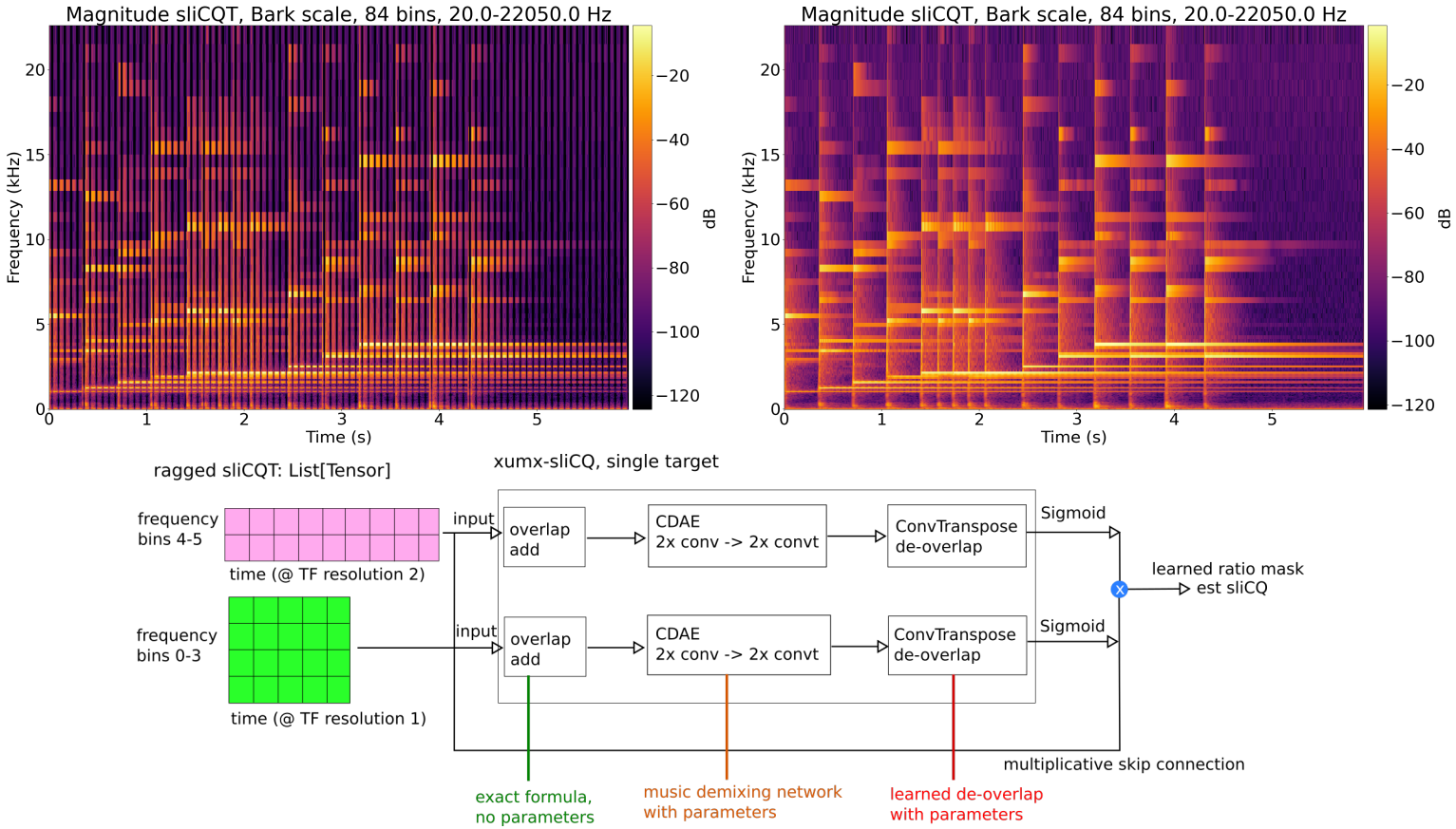
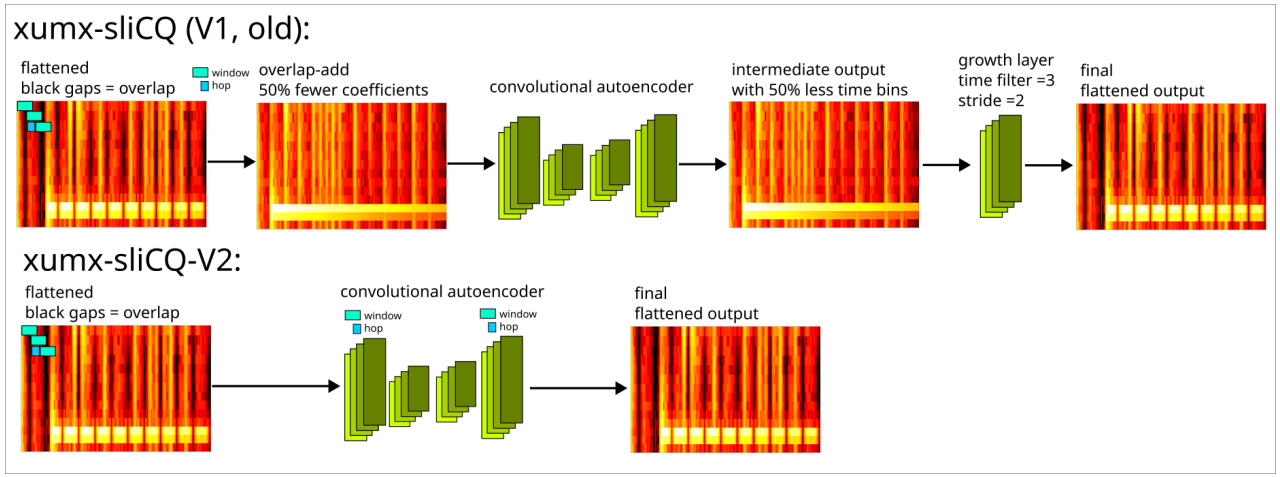


Figure: xumx-sliCQ: block diagram

[7] Emad M. Grais et al. 2017.

xumx-sliCQ v2 @ CAD1 2023

1. Bark scale may have some benefits for human listeners
2. Focused solely on VDBO demixing problem
3. **4.4 dB SDR** up from 3.6, including better handling of overlap-add, mask sum loss, differentiable Wiener filtering, and complex MSE[8]: `https://github.com/sevagh/xumx-sliCQ/tree/v2`



$$x_{\text{mix}} = x_{\text{v}} + x_{\text{d}} + x_{\text{b}} + x_{\text{o}}, |X|_{\text{mix}} = M_{\text{v}}|X|_{\text{mix}} + M_{\text{d}}|X|_{\text{mix}} + M_{\text{b}}|X|_{\text{mix}} + M_{\text{o}}|X|_{\text{mix}}$$

$$\rightarrow 1 = M_{\text{v}} + M_{\text{d}} + M_{\text{b}} + M_{\text{o}}$$

4. HAAQI score: mean of 0.094 vs. 0.255 of Baseline 1 (demucs)
5. BAQ score: mean of 41.84 vs. 41.40 of Baseline 1 (demucs)

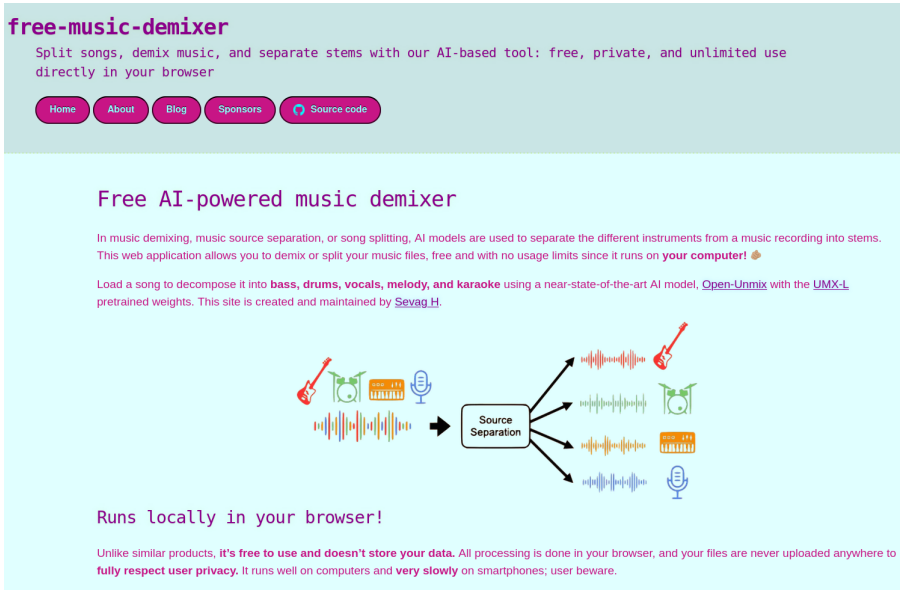
[8] Chin-Yun Yu et al. *arXiv preprint arXiv:2112.03752* (2021).

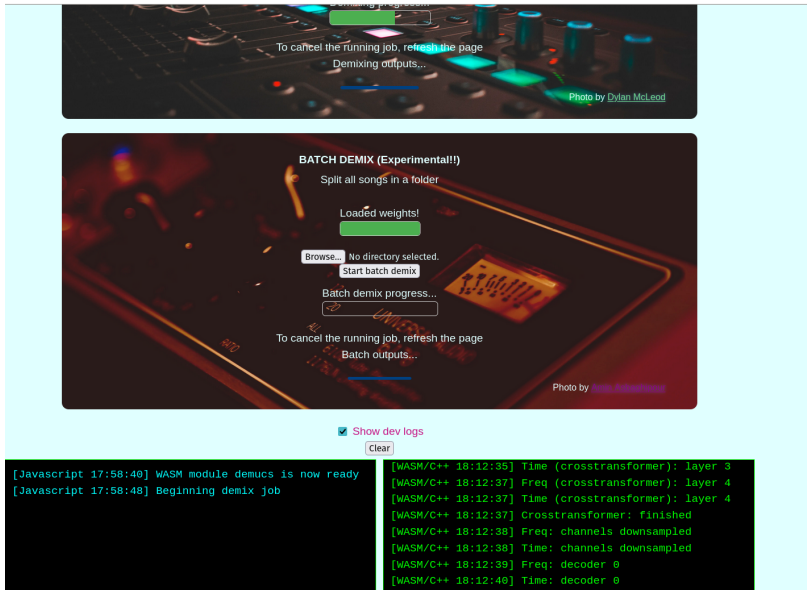
New demixing-related project

Aim of these systems are to improve listening experience for those with different hearing.[9] VDBO models are not very accessible (inscrutable Python errors, need >64GB RAM, etc.)

`https://freemusicdemixer.com`

Optimized C++ inference for UMX + Demucs, compiled to WebAssembly, running in the web, client-side on your browser, under 4 GB of memory





[9] http://cadenzachallenge.org/about