Better music demixing with sliCQT

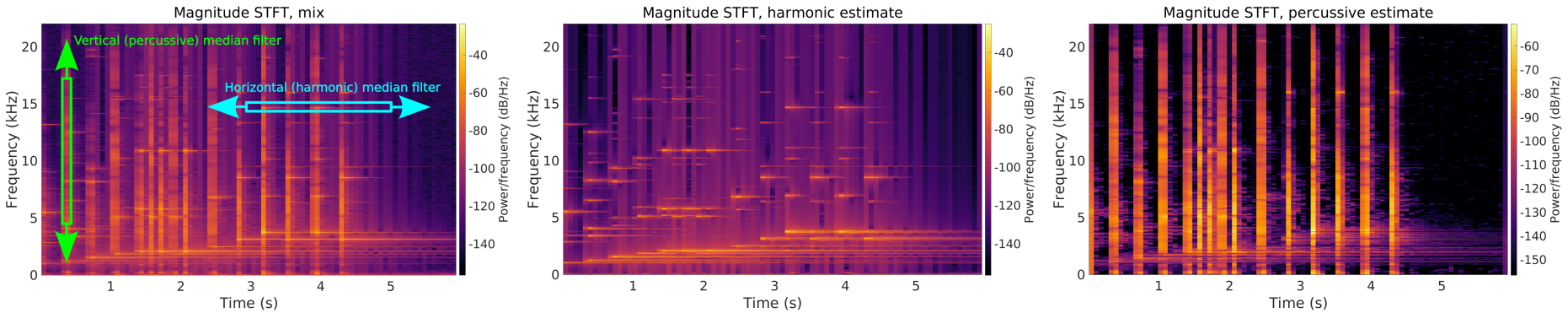
Submission to Cadenza Challenge CAD1

Sevag Hanssian

December 08, 2023

Time-frequency tradeoffs

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

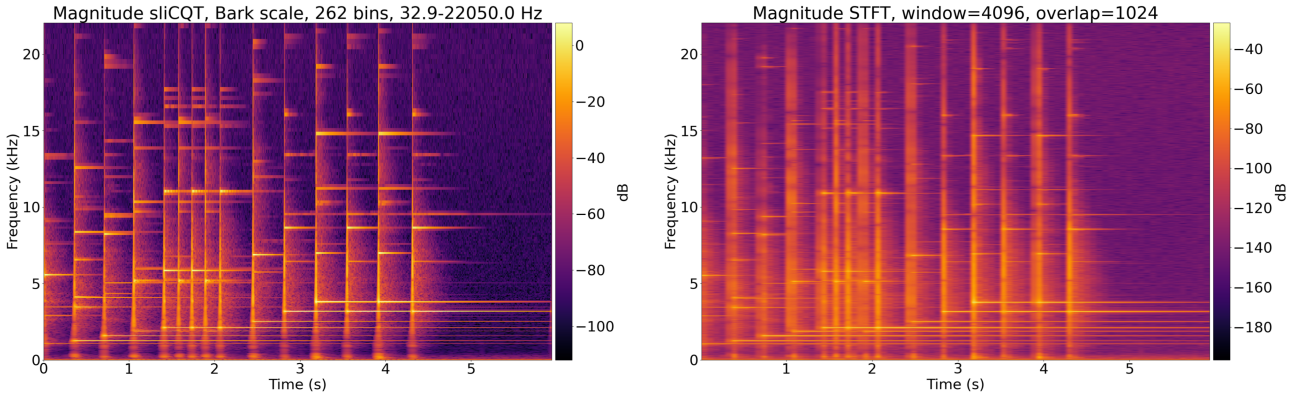


1. Short window (256) for percussion, long window (4096) for harmonic
2. Short-time Fourier Transform (STFT) window size matters per-target[2] in VDBO problems
3. In musical and auditory contexts, frequency resolution should increase from high to low frequencies (vice-versa for time resolution)[3]
4.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

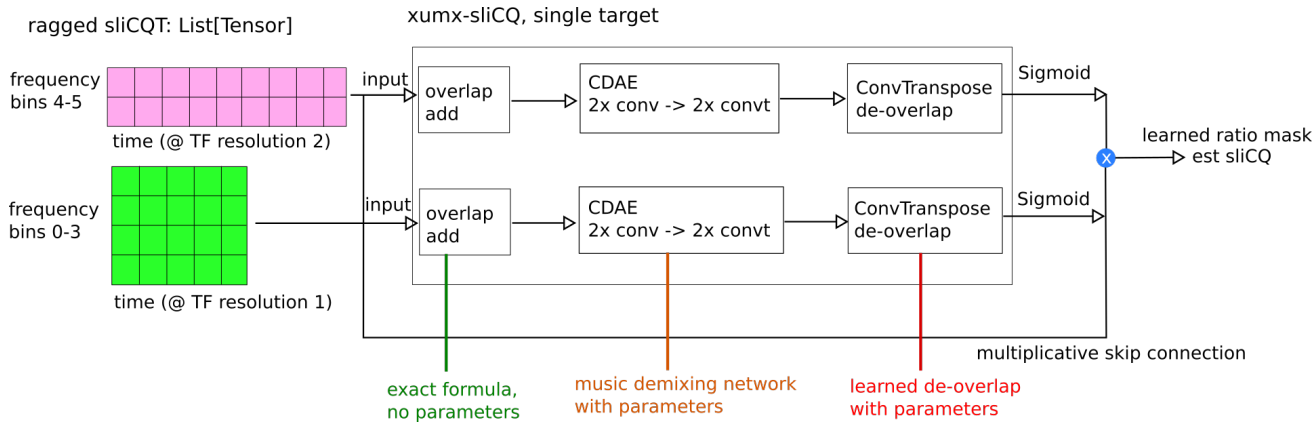
1 Derry Fitzgerald. 2010; Jonathan Driedger et al. 2014.
2 Ilya Kavalerov et al. IEEE, 2019.
3 Christian Schörkhuber et al. 2012; Monika Dörfler. PhD thesis. 2002.
4 Judith Brown. (1991).

xumx-sliCQ v1 @ MDX 2021

1. sliCQ Transform[5] is an STFT-like transform with **perfect inverse** that uses time-varying windows for nonuniform spectral analysis



2. Bark scale (262 bins 32.9–22050 Hz) sliCQT + convolutional denoising autoencoder (CDAE) architecture[6] to achieve 3.6 dB SDR

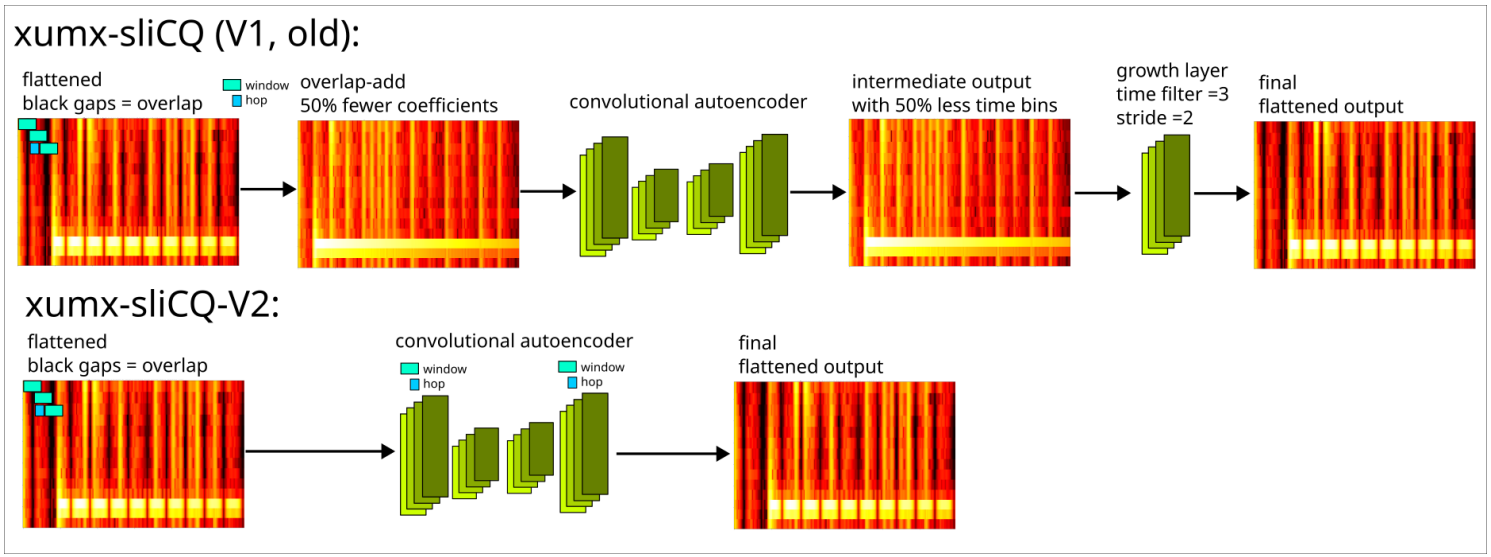


5 Gino Angelo Velasco et al. 2011; Nicki Holighaus et al. (2013).

6 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. Better handling of overlap-add, mask sum loss, differentiable Wiener filtering, and complex MSE[7]: github.com/sevagh/xumx-sliCQ



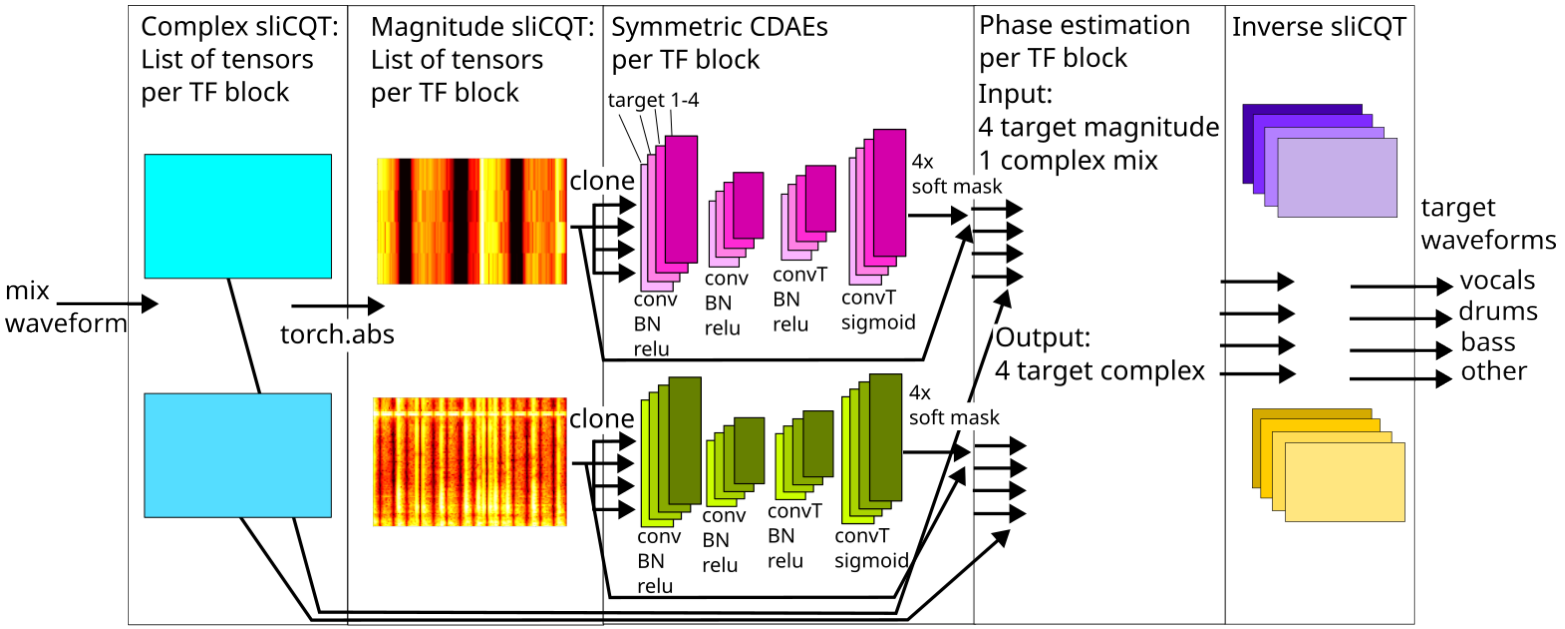
$$x_{\text{mix}} = x_{\text{v}} + x_{\text{d}} + x_{\text{b}} + x_{\text{o}}$$

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

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

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

xumx-sliCQ v2: results



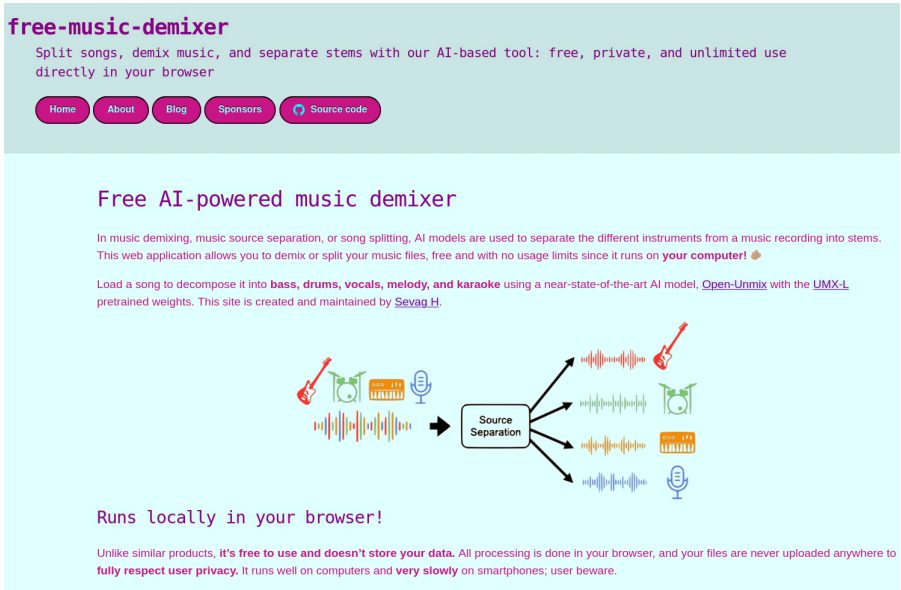
1. 4.4 dB SDR up from 3.6
2. HAAQI score: mean of 0.094 vs. 0.255 of Baseline 1 (demucs)
3. BAQ score: mean of 41.84 vs. 41.40 of Baseline 1 (demucs)
4. More efficient (using bfloat16 for faster training and inference, etc.)
5. Weights are 60 MB
6. Fast realtime variant with 4.0 dB SDR using causal convolutions

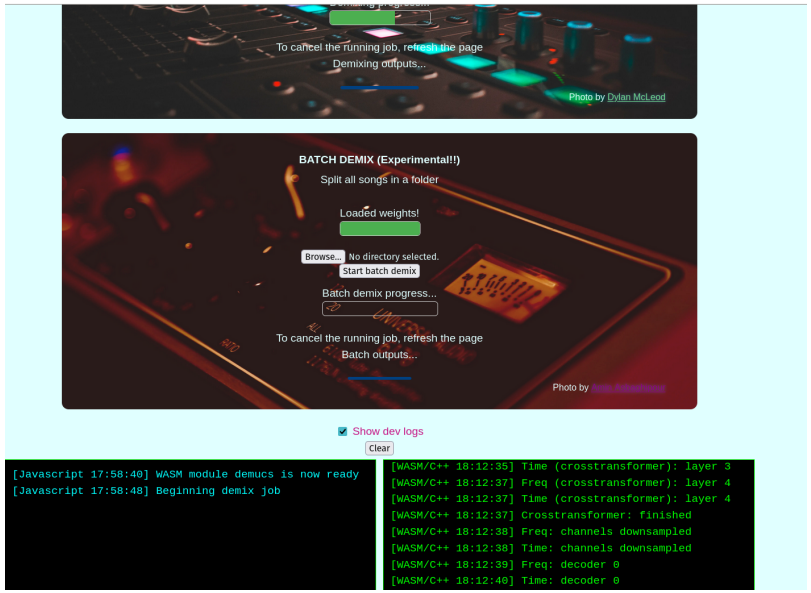
New demixing-related project

Aim of these systems are to improve listening experience for those with different hearing.[8] 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





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