MDX
Music Demixing Workshop

Music demixing with the sliCQ transform

Sevag Hanssian[1]

1 McGill University

In partnership with

ISMIR2021

Abstract

Music source separation, or music demixing, is the task of decomposing a song into its constituent sources, which are typically isolated instruments (e.g., drums, bass, and vocals). The Music Demixing Challenge[1] (Mitsufuji et al., 2021) was created to inspire new demixing research. Open-Unmix (UMX) (Stöter et al., 2019), and the improved variant CrossNet-Open-Unmix (X-UMX) (Sawata et al., 2021), were included in the challenge as the baselines. Both models use the Short-Time Fourier Transform (STFT) as the representation of music signals.

The time-frequency uncertainty principle states that the STFT of a signal cannot be maximally precise in both time and frequency (Gabor, 1946). The tradeoff in time-frequency resolution can significantly affect music demixing results (Simpson, 2015). Our proposed adaptation of UMX replaced the STFT with the sliCQT (Holighaus et al., 2013), a time-frequency transform with varying time-frequency resolution. Unfortunately, our model xumx-sliCQ[2] (Hanssian, 2021) achieved lower demixing scores than UMX.

Background

The STFT is computed by applying the Discrete Fourier Transform on fixed-size windows of the input signal. From both auditory and musical motivations, variable-size windows are preferred, with long windows in low-frequency regions to capture detailed harmonic information with a high frequency resolution, and short windows in high-frequency regions to capture transients with a high time resolution (Dörfler, 2002). The sliCQ Transform (sliCQT) (Holighaus et al., 2013) is a realtime variant of the Nonstationary Gabor Transform (NSGT) (Balazs et al., 2011). These are time-frequency transforms with complex Fourier coefficients and perfect inverses that use varying windows to achieve nonlinear time or frequency resolution. An example application of the NSGT/sliCQT is an invertible Constant-Q Transform (CQT) (Brown, 1991).

Method

In music demixing, the oracle estimator represents the upper limit of performance using ground truth signals. In UMX, the phase of the STFT is discarded and the estimated magnitude STFT of the target is combined with the phase of the mix for the first estimate of the waveform. This is sometimes referred to as the "noisy phase" (Wichern et al., 2019), described by Equation 1.

$$\hat{X}_{\text{target}} = |X_{\text{target}}| \cdot \angle X_{\text{mix}} \tag{1}$$

The sliCQT parameters were chosen randomly in a 60-iteration search for the largest median SDR across the four targets (vocals, drums, bass, other) from the noisy-phase waveforms of the MUSDB18-HQ (Rafii et al., 2019) validation set. The sliCQT parameters of 262 frequency bins on the Bark scale between 32.9–22050 Hz achieved 7.42 dB in the noisy phase oracle,

[1] https://www.aicrowd.com/challenges/music-demixing-challenge-ismir-2021
[2] https://github.com/sevagh/xumx-sliCQ

beating the 6.23 dB of the STFT with the UMX window and overlap of 4096 and 1024 samples respectively. STFT and sliCQT spectrograms of a glockenspiel signal[3] are shown in Figure 1.

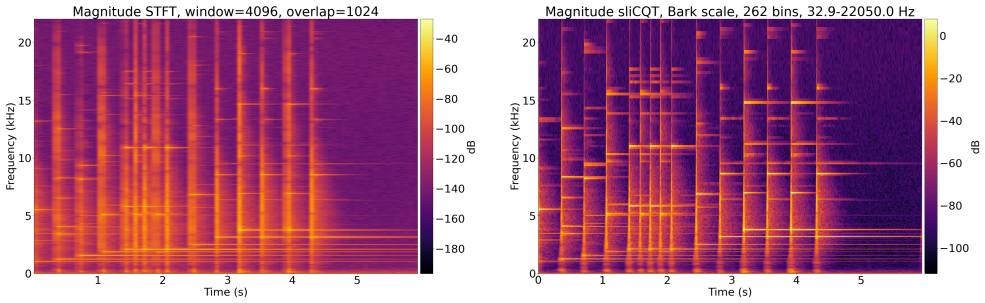

Figure 1: STFT and sliCQT spectrograms of the musical glockenspiel signal.

The STFT outputs a single time-frequency matrix where all of the frequency bins are spaced uniformly apart and have the same time resolution. The sliCQT groups frequency bins, which may be nonuniformly spaced, in a ragged list of time-frequency matrices, where each matrix contains frequency bins that share the same time resolution. In xumx-sliCQ, convolutional layers adapted from an STFT-based vocal separation model (Grais et al., 2021) were applied separately to each time-frequency matrix, shown in Figure 2.

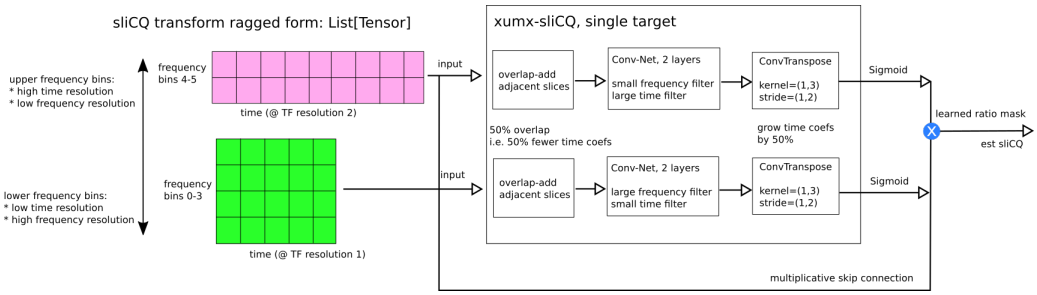

Figure 2: Example of convolutional layers applied to a ragged sliCQT.

Results

Our model, xumx-sliCQ, was trained on MUSDB18-HQ. On the test set, xumx-sliCQ achieved a median SDR of 3.6 dB versus the 4.64 dB of UMX and 5.54 dB of X-UMX, performing worse than the original STFT-based models. The overall system architecture of xumx-sliCQ is similar to X-UMX, shown in Figure 3.

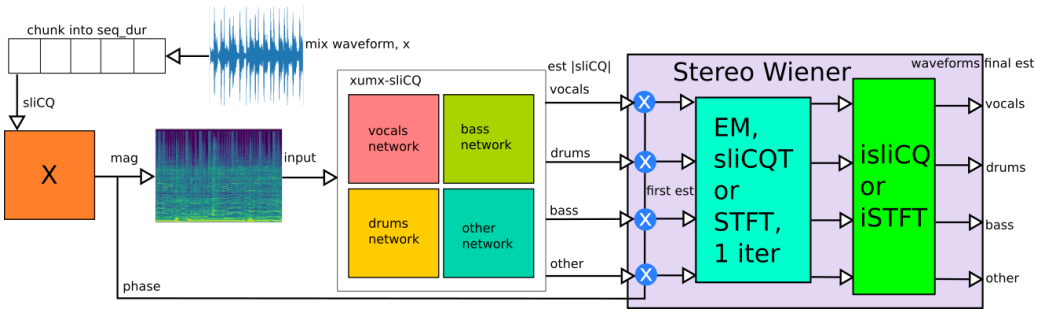

Figure 3: xumx-sliCQ overall system diagram.

[3] https://github.com/ltfat/ltfat/blob/master/signals/gspi.wav

Acknowledgements

Thanks to my colleagues Néstor Nápoles López and Timothy Raja de Reuse, and to my master's thesis supervisor Prof. Ichiro Fujinaga, for help throughout the creation of xumx-sliCQ.

References

Balazs, P., Doerfler, M., Jaillet, F., Holighaus, N., & Velasco, G. A. (2011). Theory, implementation and applications of nonstationary gabor frames. *Journal of Computational and Applied Mathematics*, *236*(6), 1481–1496. https://doi.org/10.1016/j.cam.2011.09.011

Brown, J. (1991). Calculation of a constant q spectral transform. *Journal of the Acoustical Society of America*, *89*(1), 425–434. https://www.ee.columbia.edu/%C2%A0dpwe/papers/Brown91-cqt.pdf

Dörfler, M. (2002). *Gabor analysis for a class of signals called music* [PhD thesis, NuHAG, University of Vienna]. http://www.mathe.tu-freiberg.de/files/thesis/gamu_1.pdf

Gabor, D. (1946). Theory of communication. *Journal of Institution of Electrical Engineers*, *93*(3), 429–457. http://www.granularsynthesis.com/pdf/gabor.pdf

Grais, E. M., Zhao, F., & Plumbley, M. D. (2021). Multi-band multi-resolution fully convolutional neural networks for singing voice separation. *28th European Signal Processing Conference*, 261–265. https://doi.org/10.23919/Eusipco47968.2020.9287367

Hanssian, S. (2021). Xumx-sliCQ: Music demixing with the sliCQ transform and PyTorch for the ISMIR 2021 music demixing challenge. In *GitHub repository*. GitHub. https://github.com/sevagh/xumx-sliCQ

Holighaus, N., Dörfler, M., Velasco, G. A., & Grill, T. (2013). A framework for invertible, real-time constant-q transforms. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(4), 775–785. https://doi.org/10.1109/TASL.2012.2234114

Mitsufuji, Y., Fabbro, G., Uhlich, S., & Stöter, F.-R. (2021). Music demixing challenge at ISMIR 2021. *arXiv Preprint arXiv:2108.13559*. https://arxiv.org/abs/2108.13559

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., & Bittner, R. (2019). *MUSDB18-HQ: An uncompressed version of MUSDB18*. https://doi.org/10.5281/zenodo.3338373

Sawata, R., Uhlich, S., Takahashi, S., & Mitsufuji, Y. (2021). All for one and one for all: Improving music separation by bridging networks. *arXiv Preprint arXiv:2010.04228*. https://www.ismir2020.net/assets/img/virtual-booth-sonycsl/cUMX_paper.pdf

Simpson, A. (2015). Time-frequency trade-offs for audio source separation with binary masks. *arXiv Preprint arXiv:1504.07372*. https://arxiv.org/abs/1504.07372

Stöter, F.-R., Uhlich, S., Liutkus, A., & Mitsufuji, Y. (2019). Open-unmix: A reference implementation for music source separation. *Journal of Open Source Software*, *4*(41), 1667. https://doi.org/10.21105/joss.01667

Wichern, G., Antognini, J., Flynn, M., Zhu, L. R., McQuinn, E., Crow, D., Manilow, E., & Le Roux, J. (2019). WHAM!: Extending speech separation to noisy environments. *arXiv Preprint arXiv:1907.01160*. https://arxiv.org/abs/1907.01160