Better Music Demixing with the sliCQ Transform

Sevag Hanssian[1]

[1]Independent researcher (sevag.xyz), Montréal, Canada
sevagh@protonmail.com

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). Open-Unmix (UMX), and the improved variant CrossNet-Open-Unmix (X-UMX), are high-performing models that use Short-Time Fourier Transform (STFT) as the representation of music signals, and apply masks to the magnitude STFT to separate mixed music into four sources: vocals, drums, bass, and other.

The time-frequency uncertainty principle states that the STFT of a signal cannot be maximally precise in both time and frequency. The tradeoff in time-frequency resolution can significantly affect music demixing results. For the Cadenza Challenge in 2023, we submitted a model, xumx-sliCQ-V2,[1] which replaces the STFT with the sliCQT, a time-frequency transform with varying time-frequency resolution. Our system achieved an SDR score of 4.4 dB on the MUSDB18-HQ test set.
Index Terms: music source separation, music demixing, deep neural networks, time-frequency resolution, MUSDB18-HQ

1. Introduction

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 [1]. The sliCQ Transform (sliCQT) [2] is a time-frequency transform with complex Fourier coefficients and perfect inverse that uses varying windows to achieve nonlinear time or frequency resolution. An example application of the sliCQT is an invertible Constant-Q Transform (CQT) [3].

2. Methodology

From the guidelines of the Cadenza challenge and to ensure reproducibility, we only relied on the standard and widely-available MUSDB18-HQ dataset [4] for training and evaluation of xumx-sliCQ-V2.

In xumx-sliCQ-V2, we kept the same sliCQT parameters from the older variant, xumx-sliCQ [5]. The sliCQT parameters are 262 bins on the Bark scale between 32.9–22050 Hz, chosen in a random parameter search to maximize the mix-phase or noisy-phase oracle [5]. STFT and sliCQT spectrograms of a glockenspiel signal are shown in Figure 1.

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-V2, convolutional layers were applied separately to each time-frequency matrix, shown in Figure 2.

We made three significant changes to the older system, xumx-sliCQ, which account for the improved performance of xumx-sliCQ-V2.

2.1. Improved overlap-add

The sliCQT subdivides the input signal into "slices" of length N that are "symmetrically zero-padded to length $2N$" [2, 10]. To create a spectrogram, adjacent slices need to be 50% overlap-added with each other, with no inverse operation. In xumx-sliCQ-V2, we incorporated the slice size in the kernel and stride of the first convolution layer and last transpose convolution layer to avoid the non-invertible overlap-add procedure, shown in Figure 3.

2.2. Differentiable Wiener filtering and complex MSE

In xumx-sliCQ, the mean-squared error (MSE) loss function is applied to the magnitude spectrogram estimates of the neural network. The post-processing Wiener filtering step is then used to further improve the separation results and create the estimated complex spectrograms. Danna-Sep [6], a winning system from MDX 21, incorporated the Wiener filtering step into the neural network to output complex spectrograms, and used the complex mean-squared error as the loss function. We did the same in xumx-sliCQ-V2.

2.3. Mask-sum loss

The final activation layer of xumx-sliCQ-V2 is the sigmoid function ($\in [0.0, 1.0]$), to apply as a soft mask (or ratio mask) to the magnitude spectrogram of the mix. An underlying simplifying assumption in music demixing is that the mix is a linear sum of the sources. Therefore, the sum of the four target masks should be exactly equal to one, shown in Equation (1). In xumx-sliCQ-V2, we introduce an additional loss term called the "mask-sum loss" which computes the MSE of the sum of the esti-

[1]https://github.com/sevagh/xumx-sliCQ/tree/v2

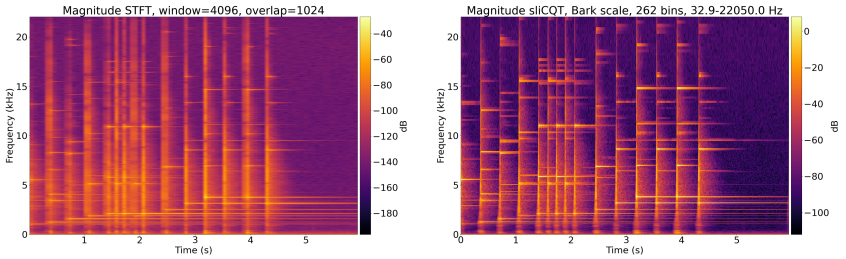

Figure 1: STFT and sliCQT spectrograms of the musical glockenspiel signal

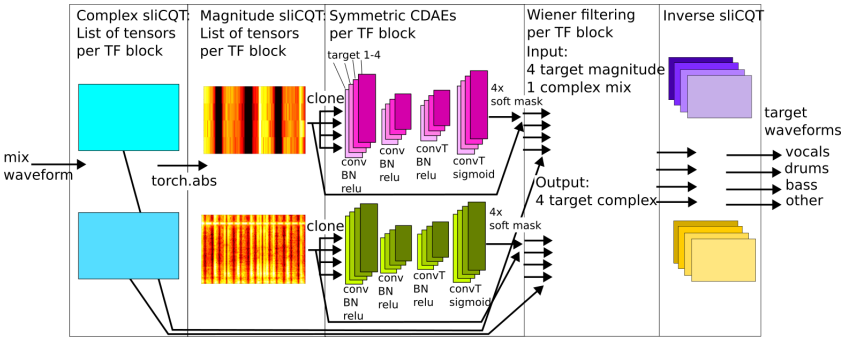

Figure 2: Convolutional denoising autoencoders (CDAE) applied to the ragged sliCQT

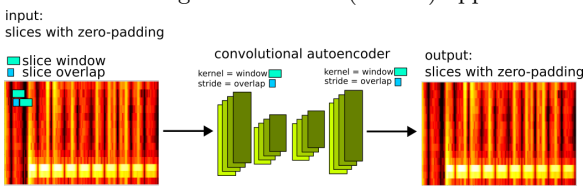

Figure 3: Incorporating slice zero-padding into the convolutional layers

mated masks with the expected total mask of ones.

$$
\begin{aligned}
x_{\text{mix}} &= x_{\text{v}} + x_{\text{d}} + x_{\text{b}} + x_{\text{o}} \\
|X|_{\text{mix}} &= M_{\text{v}}|X|_{\text{mix}} + M_{\text{d}}|X|_{\text{mix}} + M_{\text{b}}|X|_{\text{mix}} + M_{\text{o}}|X|_{\text{mix}} \\
1 &= M_{\text{v}} + M_{\text{d}} + M_{\text{b}} + M_{\text{o}} \qquad (1)
\end{aligned}
$$

3. Results

Our model, xumx-sliCQ-V2, was trained on MUSDB18-HQ. On the test set, xumx-sliCQ-V2 achieved a total SDR of 4.4 dB versus the 4.64 dB of UMX and 5.54 dB of X-UMX, performing worse than the original STFT-based models, but better than the first xumx-sliCQ which scored 3.6 dB.

4. Discussion

The Cadenza Challenge Task 1 applies further processing on top of the VDBO demixing system, and the final evaluation uses the HAAQI metric. We chose to focus only on the VDBO music demixing problem, and used BSS metrics to measure the improvement of xumx-sliCQ-V2 over xumx-sliCQ.

We believe that the mask-sum loss term of xumx-sliCQ-V2 to enforce the linear additive mixture assumption can be generally applied to any system based on spectrogram-masking and potentially improve their results.

5. Conclusion

We presented xumx-sliCQ-V2, an improved variant of xumx-sliCQ, which replaces the STFT of X-UMX with a Bark-scale sliCQT, a nonuniform time-frequency transform whose characteristics more closely map to the human auditory system. The total SDR went from 3.6 dB to 4.4 dB, demonstrating significantly improved music demixing results from the changes in xumx-sliCQ-V2.

6. References

[1] M. Dörfler, "Gabor analysis for a class of signals called music," Ph.D. dissertation, Numerical Harmonic Analysis Group, University of Vienna, 2002.

[2] N. Holighaus, M. Dörfler, G. A. Velasco, and T. Grill, "A framework for invertible, real-time constant-q transforms," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 4, pp. 775–785, 2013.

[3] J. Brown, "Calculation of a constant q spectral transform," Journal of the Acoustical Society of America, vol. 89, no. 1, pp. 425–434, 1991.

[4] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner. (2019) Musdb18-hq: an uncompressed version of musdb18.

[5] S. Hanssian, "Music demixing with the slicq transform," in MDX21 workshop at ISMIR 2021, 2021.

[6] C.-Y. Yu and K.-W. Cheuk, "Danna-sep: Unite to separate them all," 2021.