Music demixing with the sliCQ transform

Sevag Hanssian
Distributed Digital Music Archives & Libraries Lab
Schulich School of Music, McGill University, Montréal, Canada
sevag.hanssian@mail.mcgill.ca

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) [1] and CrossNet-Open-Unmix (X-UMX) [2] are models for music demixing that use the Short-Time Fourier Transform (STFT) to represent musical signals, but the time-frequency uncertainty principle states that the STFT of a signal cannot have maximal resolution in both time and frequency [3]. The tradeoff in time-frequency resolution can significantly affect music demixing results [4]. The STFT is computed by applying the Discrete Fourier Transform on fixed-size windows of the input signal, but for auditory and musical considerations, variable-sized windows are preferred to vary the time-frequency resolution by frequency region [5]. Our proposed adaptation of UMX and X-UMX, called xumx-sliCQ,[1] replaces the STFT with the sliCQT [6], an invertible transform with varying time-frequency resolution. It uses a convolutional network architecture [7] trained on the MUSDB18-HQ [8] dataset. 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, unfortunately performing worse than the original STFT-based models.

References

[1] Fabian-Robert Stöter et al. "Open-Unmix: A reference implementation for music source separation". In: *Journal of Open Source Software* 4.41 (2019), p. 1667. DOI: 10.21105/joss.01667.

[2] Ryosuke Sawata et al. "All for one and one for all: improving music separation by bridging networks". In: *arXiv preprint arXiv:2010.04228* (2021). URL: https://www.ismir2020.net/assets/img/virtual-booth-sonycsl/cUMX_paper.pdf.

[3] Dennis Gabor. "Theory of communication". In: *Journal of Institution of Electrical Engineers* 93.3 (1946), pp. 429–457. URL: http://www.granularsynthesis.com/pdf/gabor.pdf.

[4] Andrew Simpson. "Time-frequency trade-offs for audio source separation with binary masks". In: *arXiv preprint arXiv:1504.07372* (2015). URL: https://arxiv.org/abs/1504.07372.

[5] Monika Dörfler. "Gabor analysis for a class of signals called music". PhD thesis. Numerical Harmonic Analysis Group, University of Vienna, 2002. URL: http://www.mathe.tu-freiberg.de/files/thesis/gamu_1.pdf.

[6] Nicki Holighaus et al. "A framework for invertible, real-time constant-Q transforms". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.4 (2013), pp. 775–785. DOI: 10.1109/TASL.2012.2234114.

[7] Emad M. Grais, Fei Zhao, and Mark D. Plumbley. "Multi-band multi-resolution fully convolutional neural networks for singing voice separation". In: *28th European Signal Processing Conference*. 2021, pp. 261–265. DOI: 10.23919/Eusipco47968.2020.9287367.

[8] Zafar Rafii et al. *MUSDB18-HQ: an uncompressed version of MUSDB18*. 2019. DOI: 10.5281/zenodo.3338373.

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