Music demixing with the sliCQ transform

Sevag Hanssian

November 12, 2021

1 Introduction / Motivation

Music source separation, or music demixing, is the task of decomposing musical audio into its constituent sources, which are typically isolated instruments (e.g., drums, bass, and vocals). Music demixing algorithms commonly operate on the Short-Time Fourier Transform (STFT) of the audio signal (Cano et al. 2018). Due to the time-frequency uncertainty principle (Gabor 1946), the STFT of a signal cannot be maximally precise in both time and frequency, and the tradeoff in time-frequency resolution can significantly affect music demixing results (Simpson 2015). In this thesis, the sliCQT (Holighaus et al. 2013), an STFT-like transform with varying time-frequency resolution, is explored as a replacement for the STFT in a state-of-the-art deep learning model for music demixing.

The STFT is computed by applying the discrete Fourier Transform on fixed-size windows of the input signal. Dörfler (2002) argues, based on auditory and musical considerations, that musical signals should be analyzed with long windows in the low-frequency regions to capture detailed harmonic information, and short windows in the high-frequency regions to capture transients with sharp temporal localization. The sliCQ Transform, or sliCQT, is a realtime implementation of the Nonstationary Gabor Transform (NSGT) of Balazs et al. (2011). The NSGT is an invertible time-frequency transform that applies the Fourier Transform on windows of the input signal that can be varied by frequency region. An important application of the NSGT and sliCQT is to implement the Constant-Q Transform (CQT) (Brown 1991) for music analysis, which uses a logarithmic frequency scale to better show the relationship between the fundamental frequency of a musical sound and its harmonics.

Machine learning models for music source separation have achieved recent success (Stöter, Liutkus, and Ito 2018), and the STFT is used by several of the top performers. Open-Unmix (Stöter et al. 2019) was released as a state-of-the-art baseline and reference implementation for music demixing based on the STFT and published as open-source software.[1] In this thesis, Open-Unmix will be adapted to use the sliCQT in place of the STFT to investigate the viability of using the sliCQT for music demixing.

2 Previous Work

Computational source separation has a long history (Cano et al. 2018), originating from the task of separating speech from background noise. Speech algorithms could not be generalized easily to music, and techniques more specific to music source separation were developed as a result (Rafii et al. 2018). Accordingly, musical source models arose that exploit the distinct spectral characteristics of the target sources (e.g., harmonic, percussive, or vocals) in the STFT domain, such as Kernel Additive Modeling or Nonnegative Matrix Factorization (Cano et al. 2018).

1. https://github.com/sigsep/open-unmix-pytorch

Model-based methods are "prone to large errors and poor performance" (Rafii et al. 2018, 13), and manipulating time-frequency resolution is one possible strategy to improve their results. Driedger, Müller, and Disch (2014) used multiple STFTs with different window sizes, Fitzgerald and Gainza (2010) replaced the STFT with the CQT, and Wolf, Mallat, and Shamma (2014) used a custom time-frequency transform based on wavelets. More recently, data-driven models based on deep neural networks (DNN) surpassed previous approaches (Stöter, Liutkus, and Ito 2018; Rafii et al. 2018).

The earliest approaches for DNN-based music demixing started with fully connected networks (FCN), but these needed many parameters due to the large size of input music spectrograms, which limited the networks to operate on sliding windows of under one second of temporal context (Cano et al. 2018). Recurrent neural networks (RNNs) (Uhlich et al. 2017) and convolutional neural networks (CNNs) (Grais and Plumbley 2017) have both been used to overcome this limitation, as they need fewer parameters for long temporal sequences of input data (Cano et al. 2018). A recent example of music demixing in the STFT domain with an RNN architecture is Open-Unmix (Stöter et al. 2019).

3 Proposed Research / Methodology

The adaptation of Open-Unmix to use the sliCQT will be done in two steps, using the reference implementation of Open-Unmix[2] as a starting point. First, the reference Python implementation of the sliCQT[3] needs to be modified to use PyTorch,[4] because Open-Unmix uses PyTorch as its GPU deep learning framework. Next, the code of Open-Unmix should be modified to replace the STFT with the sliCQT, using an RNN or CNN architecture, while tuning the parameters of the sliCQT to surpass the performance of the STFT.

The MUSDB18-HQ dataset (Rafii et al. 2019) and BSS (Blind Source Separation) eval metrics (Vincent, Gribonval, and Févotte 2006) are standard for training and evaluating music demixing systems, and they have been used in previous Signal Source Separation Evaluation Campaigns (SiSEC) (Stöter, Liutkus, and Ito 2018). The model in this thesis will use the MUSDB18-HQ dataset for both training and evaluation, using BSS eval metrics to allow fair comparisons with Open-Unmix and other published models.

4 Contributions / Summary

In music demixing approaches that use the STFT, choosing the appropriate time-frequency resolution plays an important role. In this thesis, the sliCQT with varying time-frequency resolution is explored as a replacement for the STFT in a state-of-the-art deep learning model for music demixing. It is hoped that the resulting model will demonstrate the viability of music demixing with the sliCQT.

2. https://github.com/sigsep/open-unmix-pytorch
3. https://github.com/grrrr/nsgt
4. https://pytorch.org

5 References

Balazs, Peter, Monika Doerfler, Florent Jaillet, Nicki Holighaus, and Gino Angelo Velasco. 2011. "Theory, implementation and applications of nonstationary Gabor frames." *Journal of Computational and Applied Mathematics* 236 (6): 1481–1496.

Brown, Judith. 1991. "Calculation of a constant Q spectral transform." *Journal of the Acoustical Society of America* 89 (1): 425–434.

Cano, Estefanía, Derry Fitzgerald, Antoine Liutkus, Mark Plumbley, and Fabian-Robert Stöter. 2018. "Musical source separation: An introduction." *IEEE Signal Processing Magazine* 36 (1): 31–40.

Dörfler, Monika. 2002. "Gabor analysis for a class of signals called music." PhD diss., Numerical Harmonic Analysis Group, University of Vienna.

Driedger, Jonathan, Meinard Müller, and Sascha Disch. 2014. "Extending harmonic-percussive separation of audio signals." In *Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR),* 611–616.

Fitzgerald, Derry, and Mikel Gainza. 2010. "Single channel vocal separation using median filtering and factorisation techniques." *ISAST Transactions on Electronic and Signal Processing* 4 (1): 62–73.

Gabor, Dennis. 1946. "Theory of communication." *Journal of Institution of Electrical Engineers* 93 (3): 429–457.

Grais, Emad M., and Mark D. Plumbley. 2017. "Single channel audio source separation using convolutional denoising autoencoders." In *IEEE Global Conference on Signal and Information Processing,* 1265–1269.

Holighaus, Nicki, Monika Dörfler, Gino Angelo Velasco, and Thomas Grill. 2013. "A framework for invertible, real-time constant-Q transforms." *IEEE Transactions on Audio, Speech, and Language Processing* 21 (4): 775–785.

Rafii, Zafar, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2019. "MUSDB18-HQ: an uncompressed version of MUSDB18." https://doi.org/10.5281/zenodo.3338373.

Rafii, Zafar, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, Derry FitzGerald, and Bryan Pardo. 2018. "An overview of lead and accompaniment separation in music." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26 (8): 1307–1335.

Simpson, Andrew. 2015. "Time-frequency trade-offs for audio source separation with binary masks." *arXiv preprint arXiv:1504.07372.*

Stöter, Fabian-Robert, Antoine Liutkus, and Nobutaka Ito. 2018. "The 2018 signal separation evaluation campaign." In *Latent Variable Analysis and Signal Separation. Lecture Notes in Computer Science,* edited by Yannick Deville, Sharon Gannot, Russell Mason, Mark Plumbley, and Dominic Ward, 10891: 293–305. Springer International Publishing.

Stöter, Fabian-Robert, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. 2019. "Open-Unmix: A reference implementation for music source separation." *Journal of Open Source Software* 4 (41): 1667.

Uhlich, Stefan, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, and Yuki Mitsufuji. 2017. "Improving music source separation based on deep neural networks through data augmentation and network blending." In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 261–265.

Vincent, Emmanuel, Rémi Gribonval, and Cédric Févotte. 2006. "Performance measurement in blind audio source separation." *IEEE Transactions on Audio, Speech, and Language Processing* 14 (4): 1462–1469.

Wolf, Guy, Stéphane G. Mallat, and Shihab Shamma. 2014. "Audio source separation with time-frequency velocities." In *IEEE International Workshop on Machine Learning for Signal Processing,* 1–6.