SDX
Sound Demixing Workshop 2023

Multi-Resolution and Noise Robust Methods for Audio Source Separation

Nabarun Goswami [1] **and Tatsuya Harada**[^1,2]

1 The University of Tokyo, Japan **2** RIKEN, Japan

Published: 04 November 2023

Abstract

Audio source separation comprises the separation of different source sounds from a mixed signal. The source signals can be slow or fast, varying with similar or contrasting frequency profiles. To solve this challenging problem, several methods have been proposed that utilize carefully designed frequency band-splitting (Luo & Yu (2023)) or hybrid time-frequency domain methods (Rouard et al. (2023)). In this work, we propose to use the multi-resolution analysis (MRA) capabilities of the Discrete Wavelet Transform (DWT). DWT processes the signal at several scales by successively reducing the temporal resolution and extracting the low-frequency approximation and high-frequency details at each scale.

We propose two neural network architectures to leverage the MRA: Wavelet-HTDemucs (WHTDemucs) and DWT-Transformer-UNet (DTUNet), each designed to enhance the separation of audio sources. WHTDemucs extends the HTDemucs (Rouard et al. (2023)) model by introducing a third DWT branch, with the frequency branch acting as a residual bridge between the temporal and DWT branches. Meanwhile, DTUNet adopts a more simplified architecture, with independent encoders and decoders for MRA signals, complemented by a single cross-transformer to combine with the temporal branch. A source-independent post filter is applied to further enhance the output.

We also propose a noise-robust training methodology to tackle the challenge of corrupted training data, in terms of bleeding and label noise as defined by the MDX challenge (Fabbro et al. (2023)). We combine several loss functions such as L1 loss, Mixture Consistency loss (Wisdom et al. (2019)), unsupervised MixIT Loss (Wisdom et al. (2020)), and Mean Teacher loss (Tarvainen & Valpola (2017)), which are applied in a scheduled manner throughout the training process. We also use a model trained with the noise-robust method to filter out corrupt data from the dataset and train smaller models on the cleaned subsets. Finally, we apply ensembling and blending (Uhlich et al. (2017)) to further boost the separation performance. We submitted our methods to the SDX 2023 challenge and achieved 2nd position in the label noise and 3rd in the bleeding leaderboards of the Music Demixing track. We also achieved 3rd position in the Cinematic Demixing track (Uhlich et al. (2023)), competition data-only leaderboard.

Fabbro, G., Uhlich, S., Lai, C.-H., Choi, W., Martínez-Ramírez, M., Liao, W., Gadelha, I., Ramos, G., Hsu, E., Rodrigues, H., Stöter, F.-R., Défossez, A., Luo, Y., Yu, J., Chakraborty, D., Mohanty, S., Solovyev, R., Stempkovskiy, A., Habruseva, T., … Mitsufuji, Y. (2023). *The sound demixing challenge 2023 - music demixing track.* https://arxiv.org/abs/2308.06979

Luo, Y., & Yu, J. (2023). Music source separation with band-split RNN. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.*

Rouard, S., Massa, F., & Défossez, A. (2023). Hybrid transformers for music source separation. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech*

and Signal Processing (ICASSP), 1–5.

Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, *30*.

Uhlich, S., Fabbro, G., Hirano, M., Takahashi, S., Wichern, G., Roux, J. L., Chakraborty, D., Mohanty, S., Li, K., Luo, Y., Yu, J., Gu, R., Solovyev, R., Stempkovskiy, A., Habruseva, T., Sukhovei, M., & Mitsufuji, Y. (2023). *The sound demixing challenge 2023 - cinematic demixing track*. https://arxiv.org/abs/2308.06981

Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., & Mitsufuji, Y. (2017). Improving music source separation based on deep neural networks through data augmentation and network blending. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 261–265.

Wisdom, S., Hershey, J. R., Wilson, K., Thorpe, J., Chinen, M., Patton, B., & Saurous, R. A. (2019). Differentiable consistency constraints for improved deep speech enhancement. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 900–904. https://doi.org/10.1109/ICASSP.2019.8682783

Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R., Wilson, K., & Hershey, J. (2020). Unsupervised sound separation using mixture invariant training. *Advances in Neural Information Processing Systems*, *33*, 3846–3857.