MDX
Music Demixing Workshop

Monotimbral Ensemble Separation using Permutation Invariant Training

Saurjya Sarkar[1], **Emmanouil Benetos**[1], **and Mark Sandler**[1]

1 Centre for Digital Music, Queen Mary University of London, London, UK

In partnership with

ISMIR2021

Abstract

The majority of research in Music Source Separation is focussed around the "Music De-mixing" problem of separating vocals, drums and bass from mastered songs (Cano et al., 2018). We have seen that the quality of separation achievable is very directly related to the relative loudness of the target source in the input mixture. This makes separation of vocals, drums and bass feasible from pop songs as they are typically amongst the loudest components in a mastered pop song. Additionally they also have very distinctive spectro-temporal features which allow effective separation using spectrogram masking. Separating other sources from similar tracks can be challenging due to the lack of data, the low input SNR for the target instruments and the high spectral overlap with other sources. In our work we focus on the specific task of separating similar sounding sources from unmastered mixtures. We focus on unmastered tracks so that the input SNR of our target monotimbral sources is reasonable. While models trained on such mixtures may be ineffective in the typical music demixing paradigm, they have ample opportunities to be used as music production tools for recording ensembles in non-ideal conditions.

We first present our work on Vocal Harmony Separation, which is the task of separating multiple vocal tracks performed in a harmonised fashion from an a capella mixture. We utilise a time-domain neural network architecture re-purposed from speech separation (Luo & Mesgarani, 2019) research and modify it to separate a capella mixtures at a high sampling rate (Sarkar et al., 2021). Polyphonic vocal recordings are an inherently challenging source separation task due to the melodic structure of the vocal parts and unique timbre of its constituents. We use four-part (soprano, alto, tenor and bass) a capella recordings of Bach Chorales and Barbershop Quartets for our experiments. Unlike current deep learning based choral separation models where the training objective is to separate constituent sources based on their class (Gover & Depalle, 2019; Petermann et al., 2020), we train our model using a permutation invariant objective (Yu et al., 2017). Using this we achieve state-of-the-art results for choral music separation.

We also present our ongoing work on separating other monotimbral ensembles like string sections. To study this problem, we introduce a novel multitrack dataset generated using the Spitfire BBC Symphony Orchestra Professional sample library and the RWC classical music dataset (Goto et al., 2002). Our dataset utilizes a more realistic data generation method than other synthesized multi-track datasets due to the ability of this plugin to incorporate various articulation methods dynamically based on the input symbolic music data and a round-robin sampling technique introducing uniqueness to each note articulation. The sample library also enables us to render the dataset with various microphone configurations on which the library was recorded in, allowing us to study various recording scenarios for the same performance in the same acoustic space. We explore the monotimbral separation task of separating any 2 string instruments (i.e. Violin, Viola, Cello, Bass) by training a DPTNet (Chen et al., 2020) based model in a permutation invariant fashion.

References

Cano, E., FitzGerald, D., Liutkus, A., Plumbley, M. D., & Stöter, F.-R. (2018). Musical source separation: An introduction. *IEEE Signal Processing Magazine*, *36*(1), 31–40.

Chen, J., Mao, Q., & Liu, D. (2020). Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. *arXiv Preprint arXiv:2007.13975*.

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2002). RWC music database: Popular, classical and jazz music databases. *Ismir*, *2*, 287–288.

Gover, M., & Depalle, P. (2019). *Score-informed source separation of choral music* [Master's thesis]. McGill University.

Luo, Y., & Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(8), 1256–1266.

Petermann, D., Chandna, P., Cuesta, H., Bonada, J., & Gómez Gutiérrez, E. (2020). Deep learning based source separation applied to choir ensembles. *Proc. Of the 21st International Society for Music Information Retrieval Conference (ISMIR)*.

Sarkar, S., Benetos, E., & Sandler, M. (2021). Vocal Harmony Separation Using Time-Domain Neural Networks. *Proc. Interspeech 2021*, 3515–3519. https://doi.org/10.21437/Interspeech.2021-1531

Yu, D., Kolbæk, M., Tan, Z.-H., & Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 241–245.