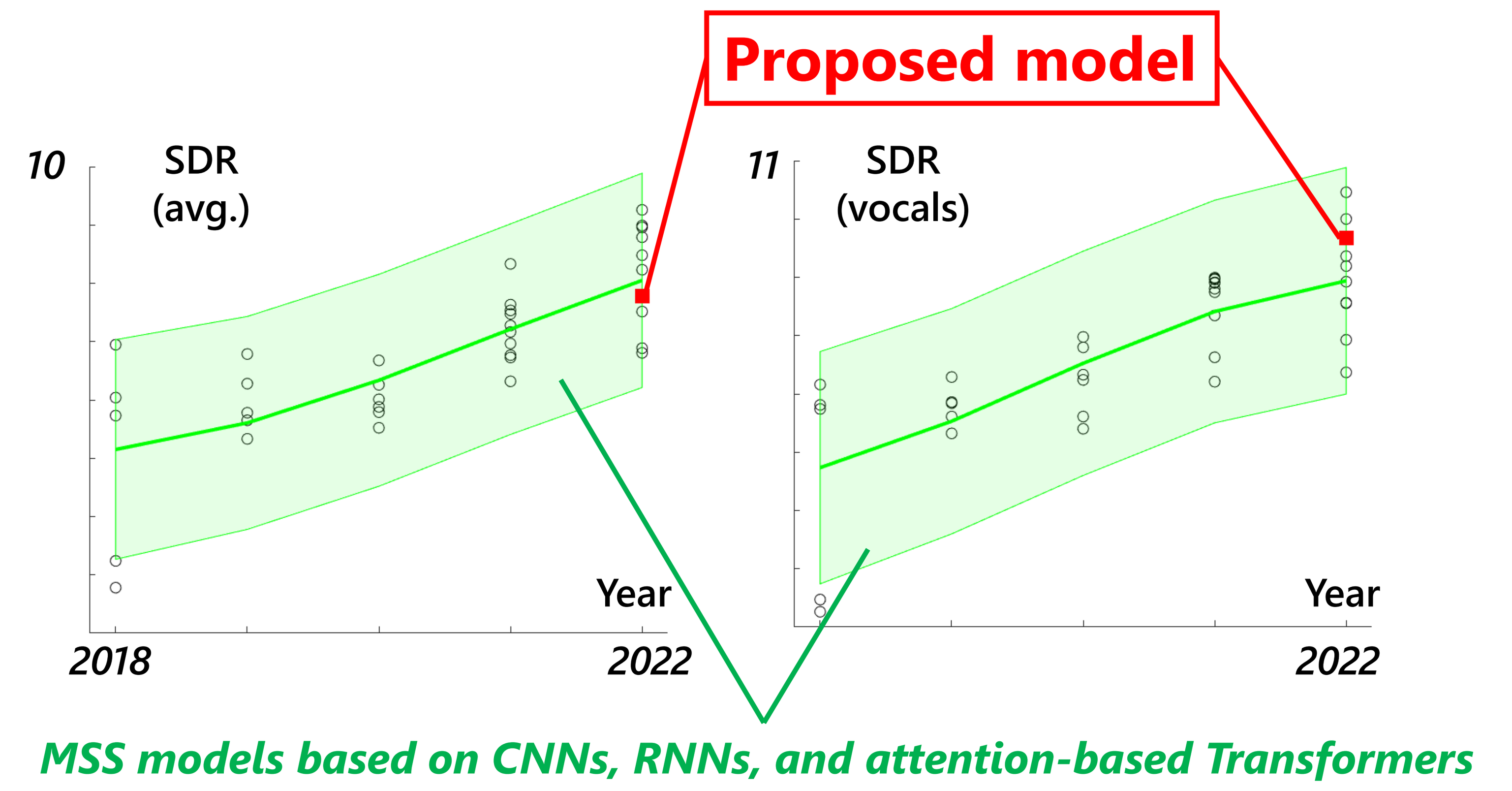
Music Source Separation (MSS) with MLP Mixing of Time, Frequency, and Channel

Tomoyasu Nakano Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST), Japan

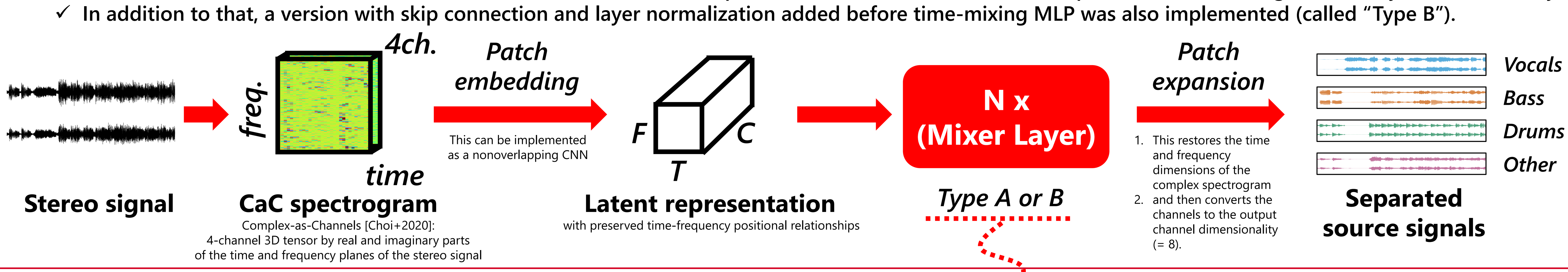
Introduction

- Music source separation (MSS) is the task of obtaining individual source signals (e.g., vocals and drums) from real music acoustic signals.
- This is an essential technique for various applications, including MIR.
- Currently, the mainstream approaches for MSS use deep neural networks, and their performance is improving year by year.
- Such deep MSS models can be classified in terms of the type of input and output used for separation and the type of architecture.
  - ✓ The input and output are selected from waveforms, amplitude spectrograms, complex spectrograms, phase spectrograms, etc.
  - ✓ The architecture is mainly selected from ResNet, DenseNet, U-Net, and Transformer and is used with layers of Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs).
  - ✓ Simpler architectures based on multilayer perceptrons (MLPs) have not been used in state-of-the-art MSS models.
- In the field of computer vision, high performance architectures based on MLPs have recently been proposed and reported to perform as well as or better than architectures using CNNs or Transformers [Tolstikhin+2021, Mansour+2022].
- Since we believe that new perspectives are important for the advancement of the research field, this paper investigates how MLP-based architectures can be effectively leveraged for MSS.



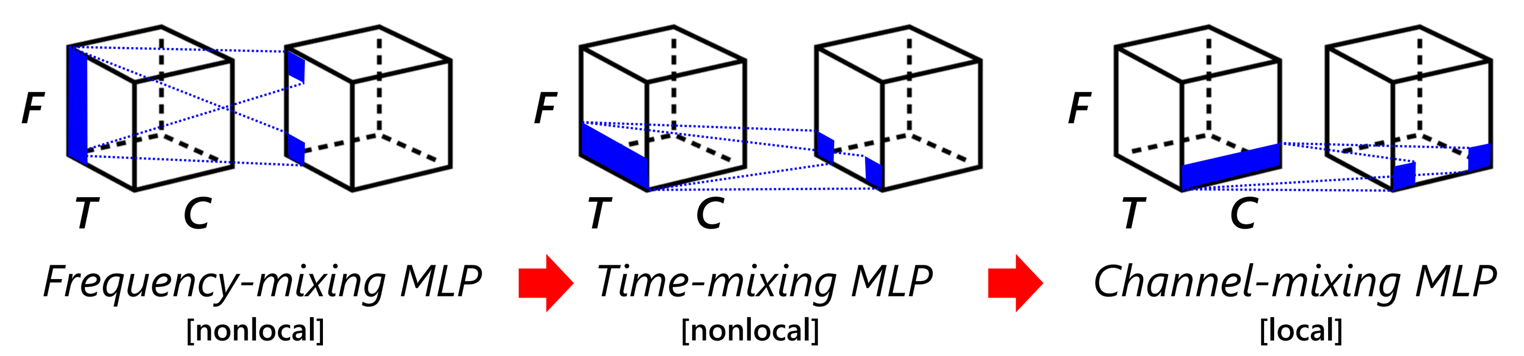
Proposed model: Time-Frequency-Channel-MLP (TFC-MLP)

- This is a model that leverages the Image-to-Image Mixer architecture [Mansour+2022] to separate music sources using a complex spectrogram as input.
  - ✓ In order to be able to consider local features suitable for MSS, we implemented a function that allows the patch size to be changed vertically and horizontally.
  - ✓ In addition to that, a version with skip connection and layer normalization added before time-mixing MLP was also implemented (called "Type B").



Overview of frequency/ time/ channel-mixing MLPs

- TFC-MLP has a structure that alternates mixing in the frequency, time, and channel dimensions.

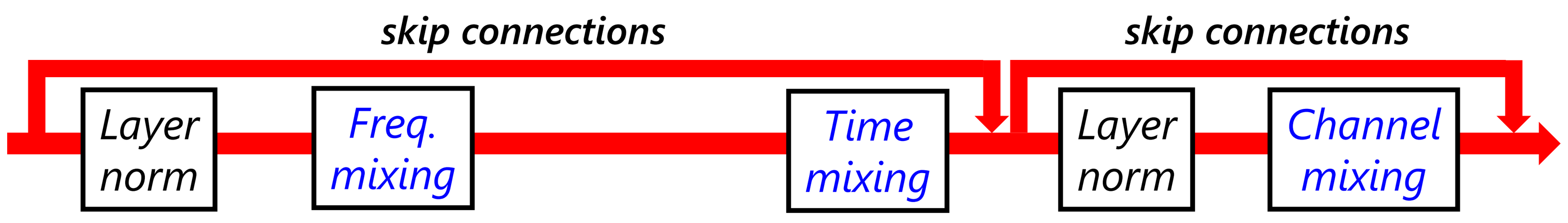


- We expect to be able to take into account the nonlocal structure.
  - ✓ e.g., to extract nonlocal relationships along the frequency axis, such as harmonic structures, by connecting the entire frequency range
- For MSS, to the best of our knowledge, there are no studies that mix the channel dimension as in TFC-MLP.
  - ✓ Such a mixer layer used in the TFC-MLP architecture has the advantage of reducing the overall memory usage compared to applying the original MLP-mixer architecture, just as the Image-to-Image Mixer reduced the memory usage.

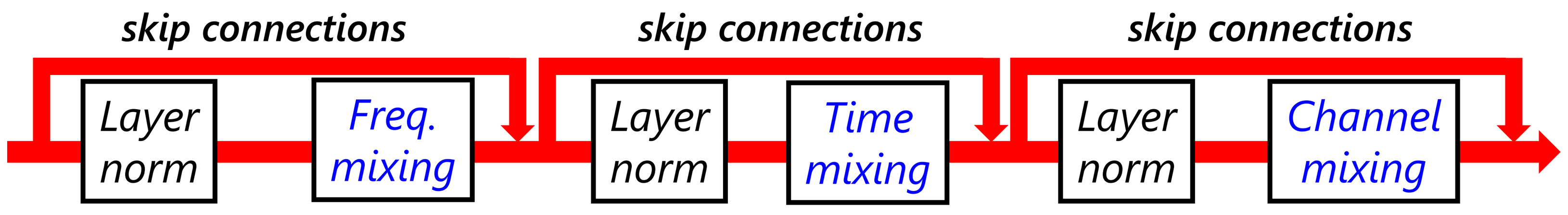
Two different implementations of the mixer layer

- The Mixer layer contains one frequency-mixing MLP, one time-mixing MLP, and one channel-mixing MLP.

Type A (Original): Same as the implementation in [Mansour+2022]

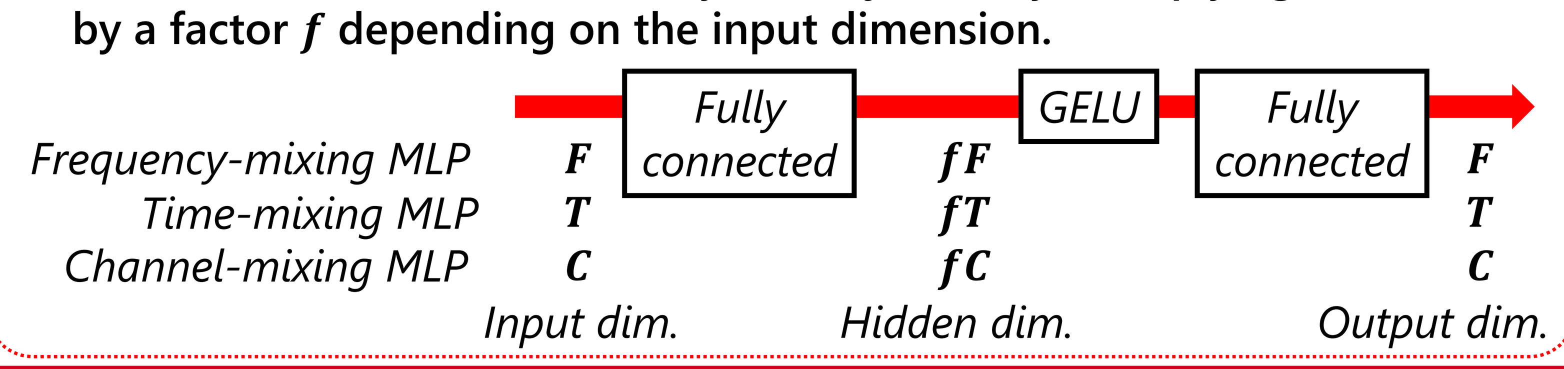


Type B (Variant): We expected that the additional skip connections and layer norm would help optimize and yield higher performance.



Mixing MLPs

- The dimension at the hidden layer is adjusted by multiplying it by a factor f depending on the input dimension.



Evaluation settings & Results

- ✓ STFT frame size: 4096
- ✓ STFT hop size: 1024
- ✓ Time frames: 512
- ✓ C: 256
- ✓ f: 4

- Using the MUSDB18-HQ dataset (44.1kHz) [Rafii+2019]
  - ✓ Training: 86 songs / Valid: 14 songs / Test: 50 songs
- TFC-MLP provides competitive results to the SoTA MSS models

TFC-MLP (Type A) outperformed SoTA models

SDRs in MUSDB18-HQ

Model	Avg.	Vocals	Drums	Bass	Other
TFC-MLP: Type B (ours)	7.17	8.92	6.95	6.83	5.96
KUIELab-MDX-Net (w/o Demucs)	7.28	8.91	7.07	7.33	5.81
TFC-MLP: Type A (ours)	**7.3**	**8.91**	**7.18**	**6.96**	**6.14**
KUIELab-MDX-Net	7.48	8.97	7.2	7.83	5.9
Hybrid Transformer Demucs	7.52	7.93	7.94	8.48	5.72
Hybrid Demucs	7.64	8.35	8.12	8.43	5.65
Band-Split RNN	8.24	10.01	9.01	7.22	6.7

SDRs in MUSDB18-HQ (+extra training data)

Model	Avg.	Vocals	Drums	Bass	Other
TFC-MLP: Type A (ours) 120	**7.78**	**9.68**	**7.75**	**7.23**	**6.46**
Hybrid Demucs 800	8.34	8.75	9.31	9.13	6.18
Hybrid Transformer Demucs 150	8.49	8.56	9.51	9.76	6.13
Hybrid Transformer Demucs 800	8.8	8.93	10.05	9.78	6.42
Band-Split RNN 1750	8.97	10.47	10.15	8.16	7.08
Hybrid Transformer Demucs 800	9	9.2	10.08	10.39	6.32
Sparse HT Demucs 800	9.27	9.37	10.83	10.47	6.41

Comparison with the state-of-the-art models

- TFC-MLP has some similarities to the SoTA MSS models, which potentially have led to the competitive performance achieved
  - ✓ Extract nonlocal relationships
    - ***The frequency-mixing MLP*** is similar to the full connection of frequency dimensions in **TDF [Choi+2020]** and the band-level RNN applied across band dimensions in **Band-Split RNN [Luo+2022]**
    - ***The time-mixing MLP*** is similar to the sequence-level RNNs applied across time dimensions in **Band-Split RNN [Luo+2022]**
  - ✓ Extract local relationships
    - ***The patch embedding*** is related to the increase in channel dimensionality in the encoder part, such as **Hybrid Transformer Demucs [Rouard+2022]**
    - ***The channel-mixing MLP*** is similar to 1x1 convolution used in **KUIELab-MDX-Net [Kim+2021]** to enhance the independently estimated sources

Contributions

1. We proposed a simpler MLP-centric MSS architecture that achieves competitive performance compared to state-of-the-art models
2. We discussed the similarities and differences between the state-of-the-art models and TFC-MLP, and suggested directions for future research