The Future of Music “Demixing”

Using a Band-Split Recurrent Neural Network

Christopher Landschoot

Audio Data Scientist

Background

- Music "Demixing" (audio source separation) – separating individual instruments' audio from a single music track.
- Applications:
  - Music Remixing
  - Music information retrieval (lyric recognition, automatic scoring, etc.)
  - Music education
  - And of course...
  - Karaoke!

Background

- Audio source separation has increased interest in recent years due to an increase in computing power and capabilities of neural networks.
- Previous methods utilized DSP filtering techniques for source separation, but they did not produce the desirable high-fidelity audio required for music.
- AIcrowd has created a competition, MDX-23, focused on pushing forward music source separation technology with 3 “paths”:
  - General audio source separation
  - Bleeding: Some stems have “bleed” of other audio (e.g. audio from vocalist headphones bleeds into their microphone)
  - Mislabeling: Some stems have been labeled incorrectly (e.g. “Bass” labeled as “Drums”)

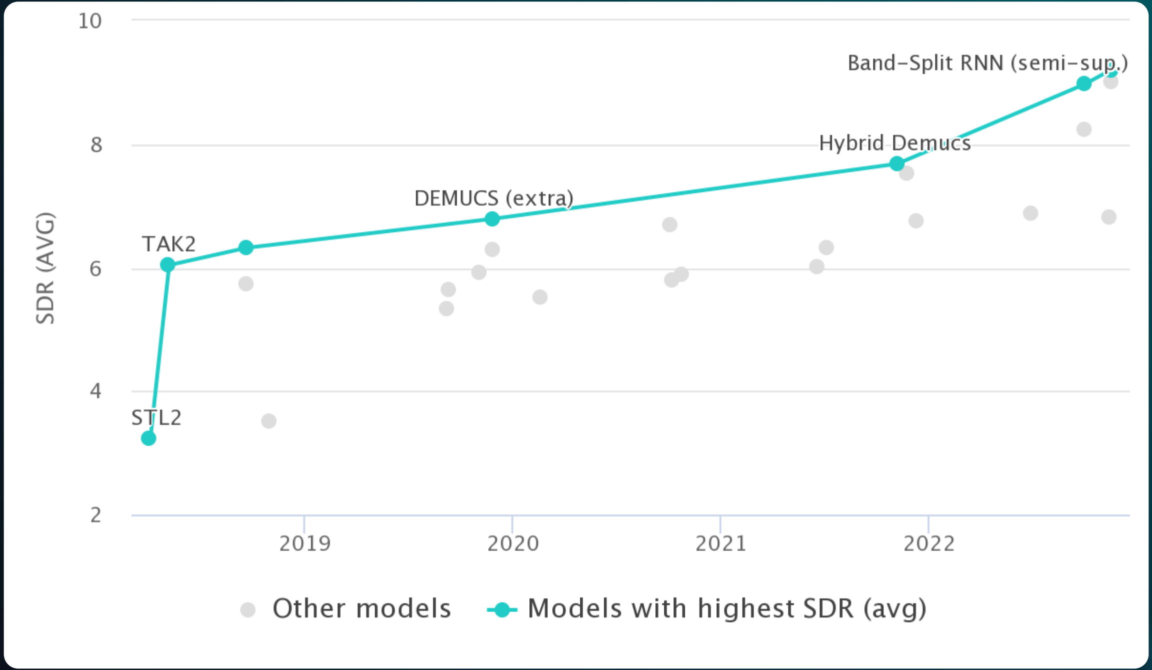
Background

- Audio source separation has increased interest in recent years due to an increase in computing power and capabilities of neural networks.
- Previous methods utilized DSP filtering techniques for source separation, but they did not produce the desirable high-fidelity audio required for music.
- AIcrowd has created a competition, MDX-23, focused on pushing forward music source separation technology with 3 “paths”:
  - **General audio source separation**

Problem Statement

- Participation in MDX-23 seeks to:
    - Explore and review current methods of audio source separation.
    - Replicate state-of-the-art modelling techniques.
    - Compare the Band-Split RNN against other methods.
    - Determine shortcomings and methods for improvement.

History

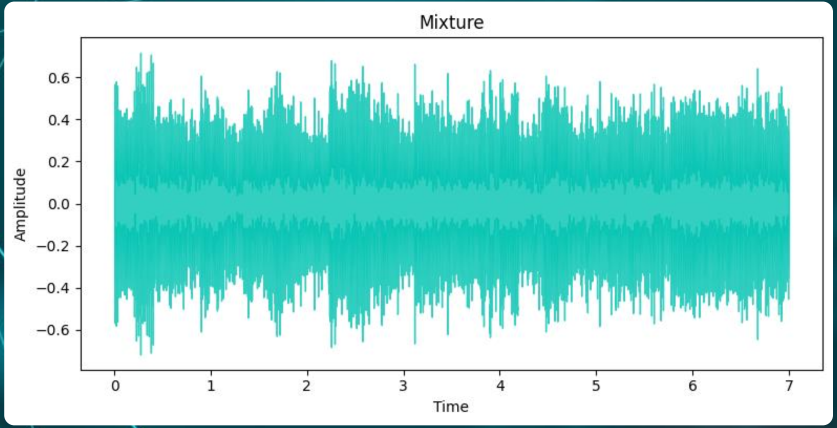


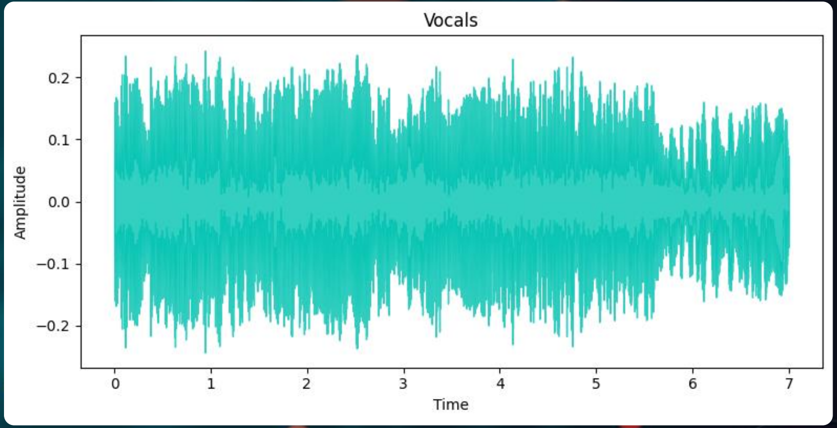
Audio Source Separation Methods (source: paperswithcode)

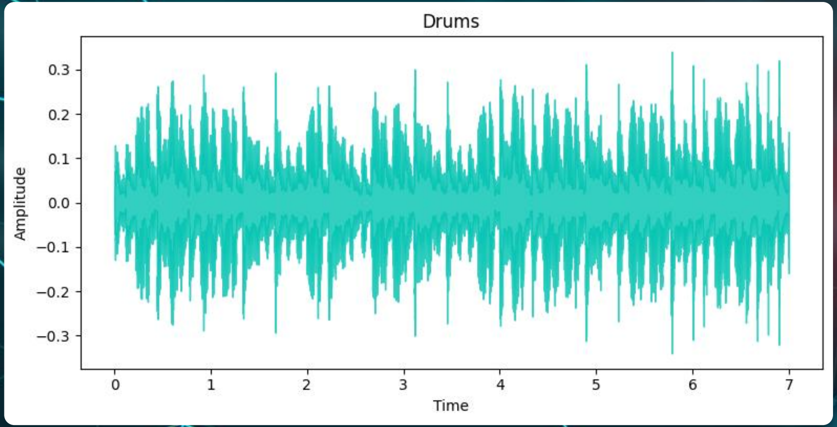
Audio Processing

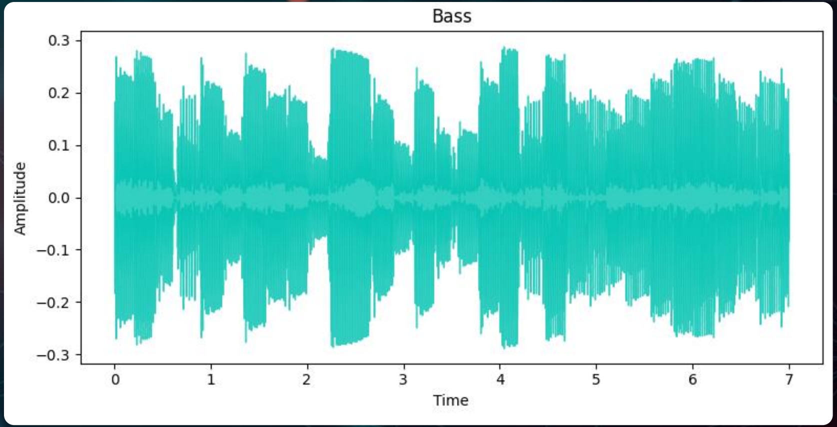
- Audio data is mainly analyzed in two domains
  - Time domain - Waveforms
  - Frequency domain – Spectrums
- The “Time-Frequency” domain, combines both concepts in one

Waveform (Time Domain)

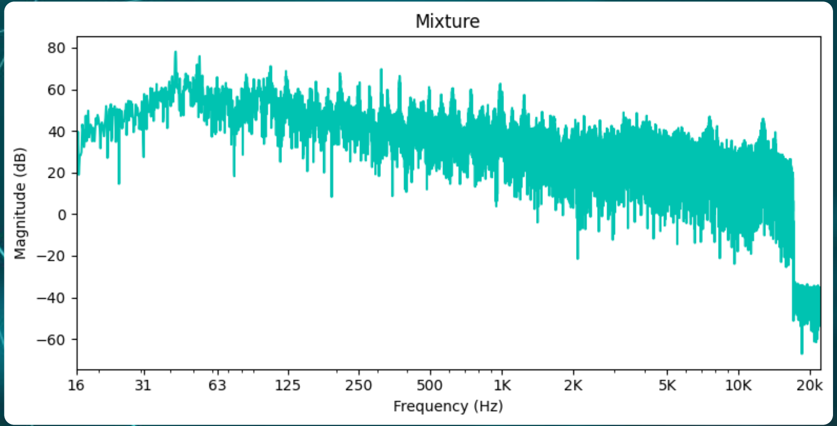


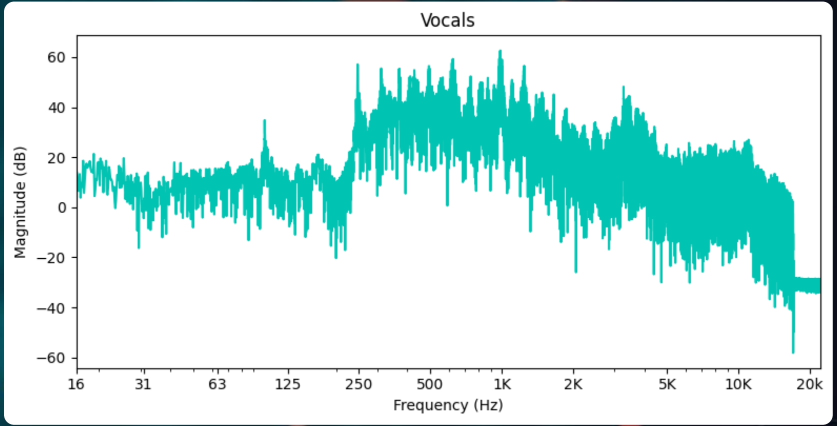


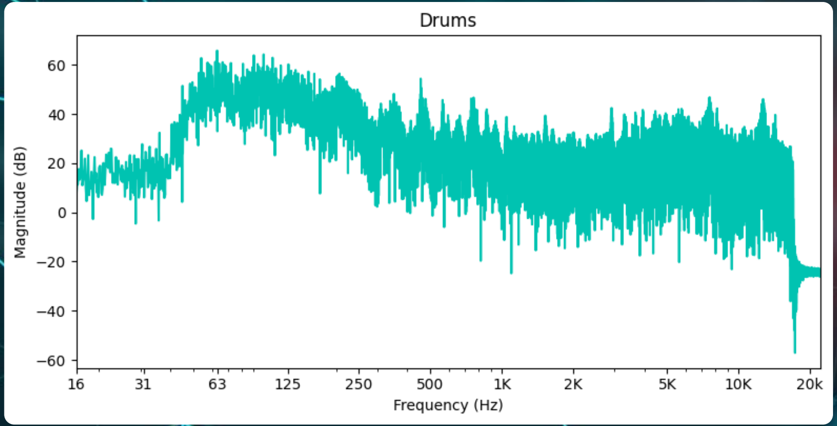


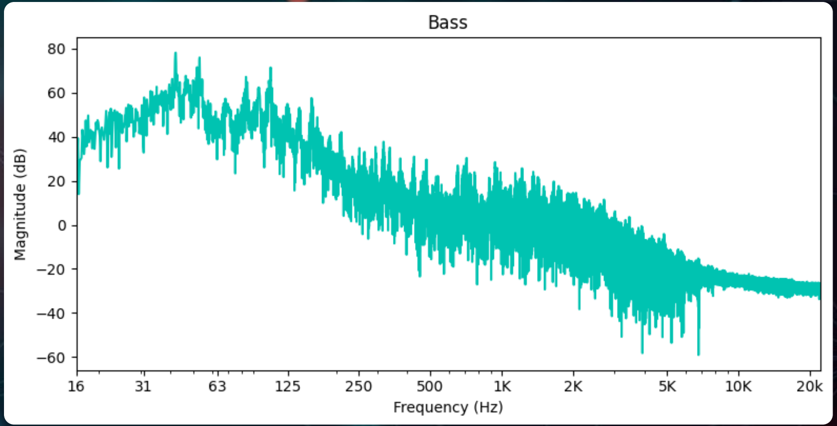


Spectrum (Frequency Domain)

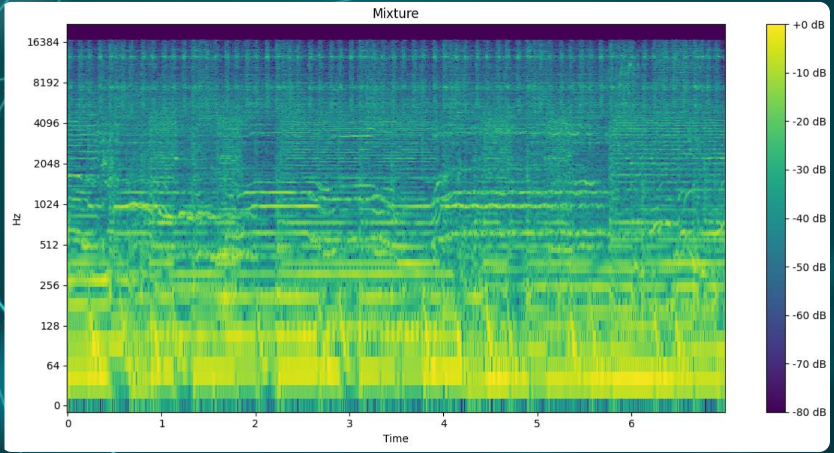


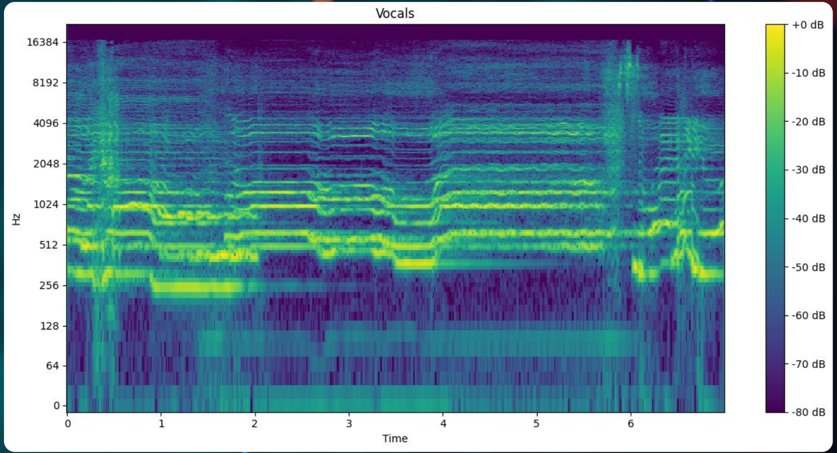


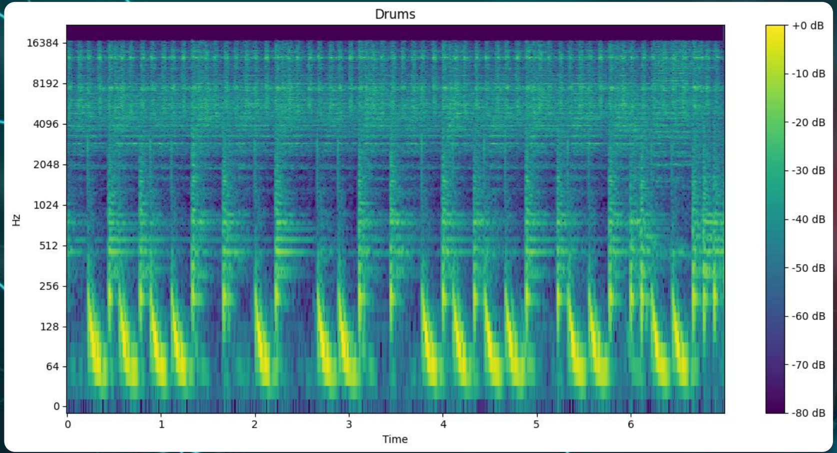


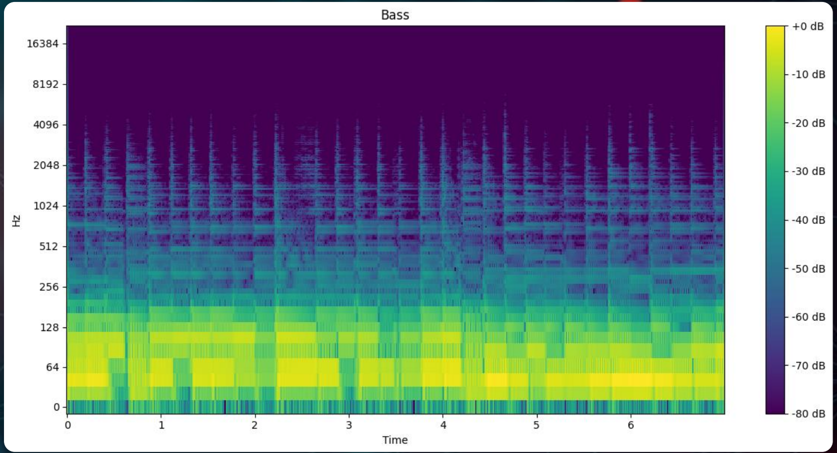


Spectrogram (Time-Frequency Domain)







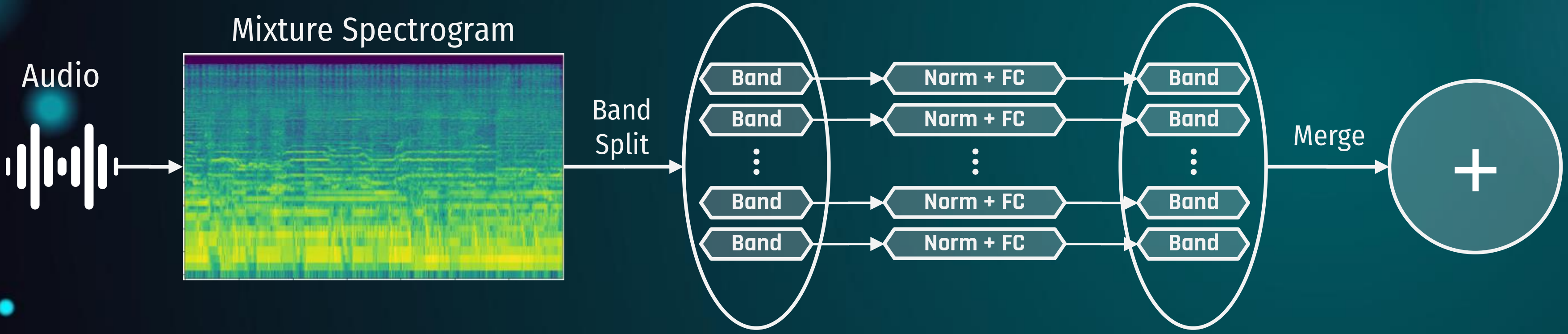


Band-Split
Recurrent Neural Network

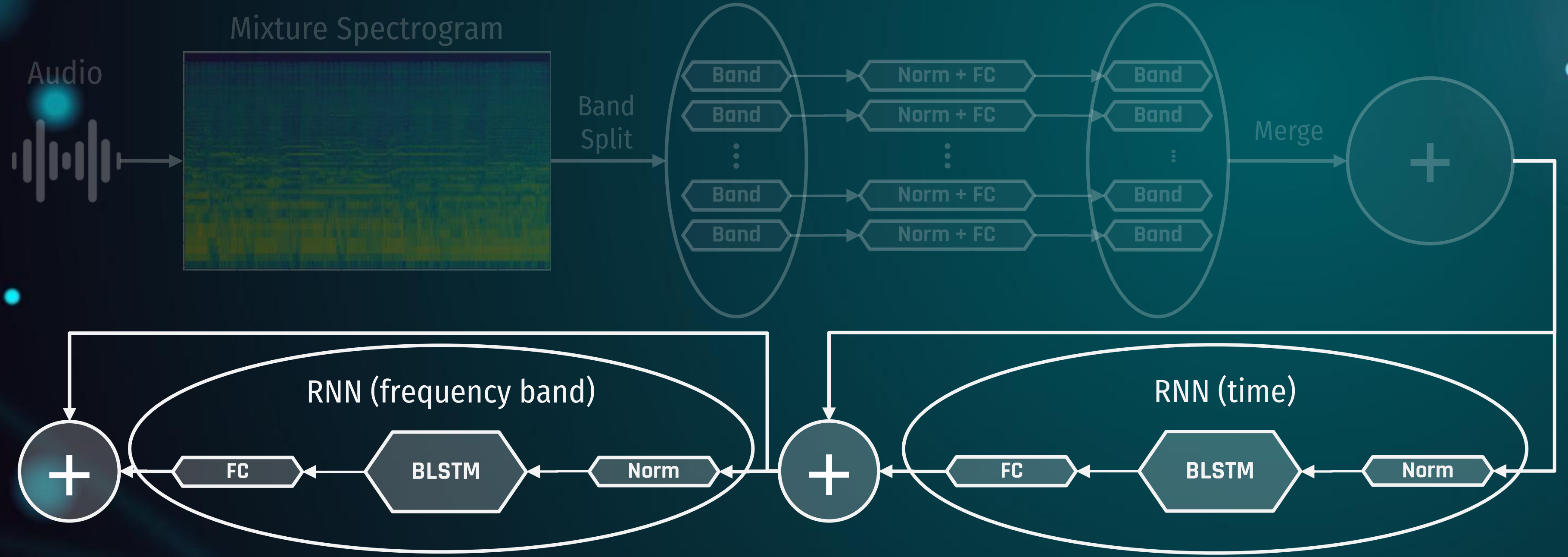
Band-Split Recurrent Neural Network

3 Modules

- Band-split module
- Band and sequence separation modeling module
- Mask estimation module

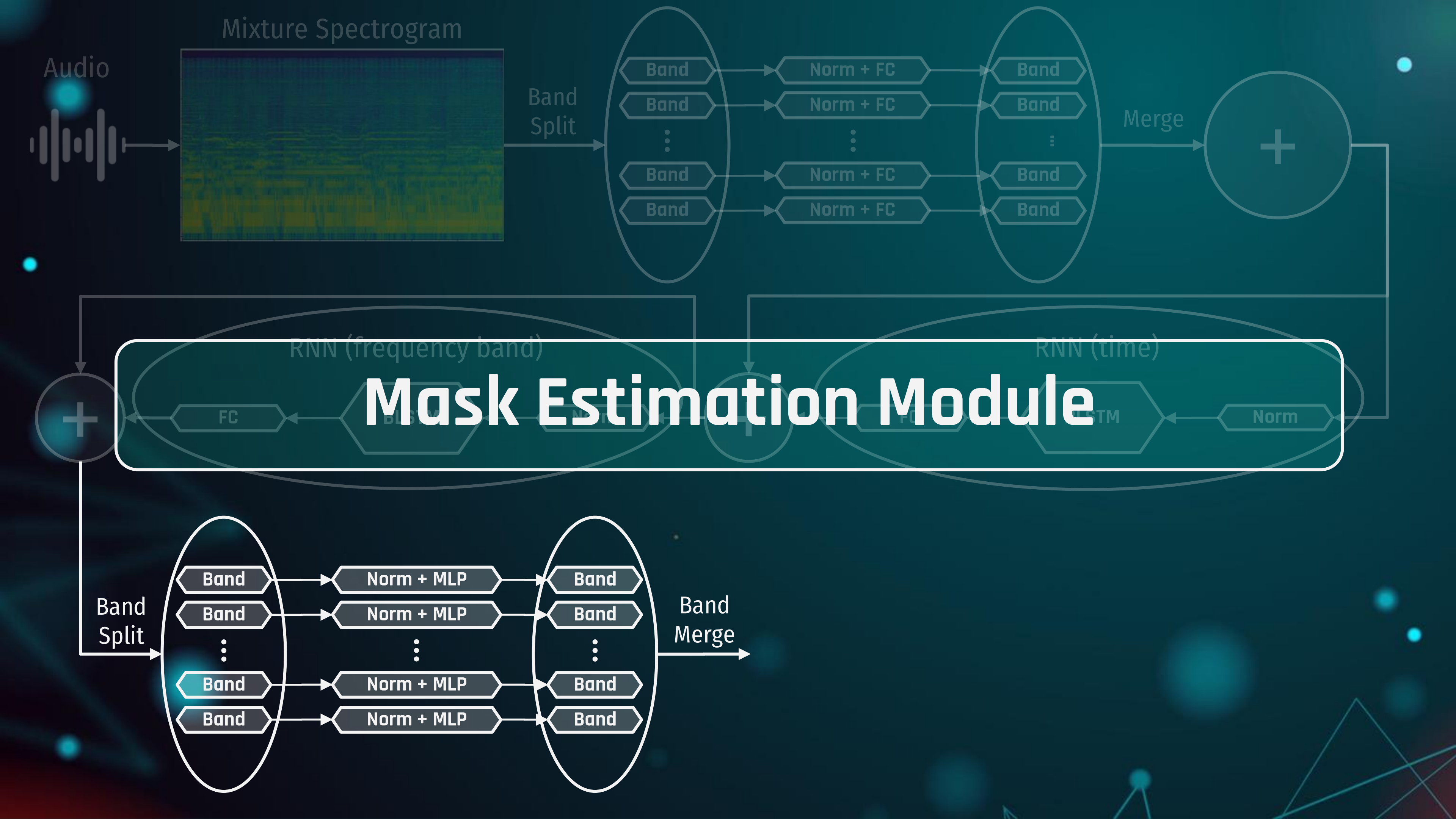


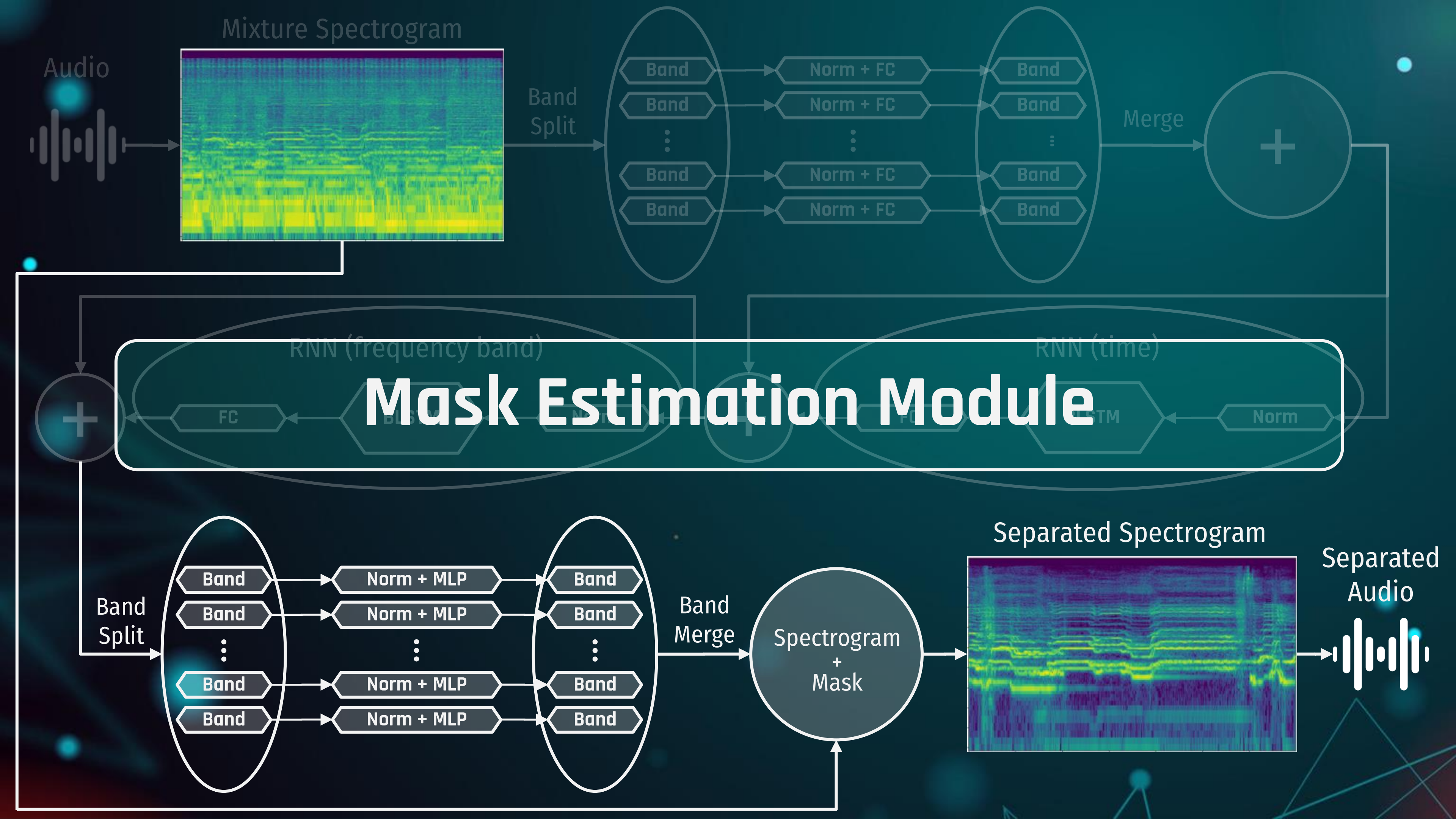
Band-Split Module

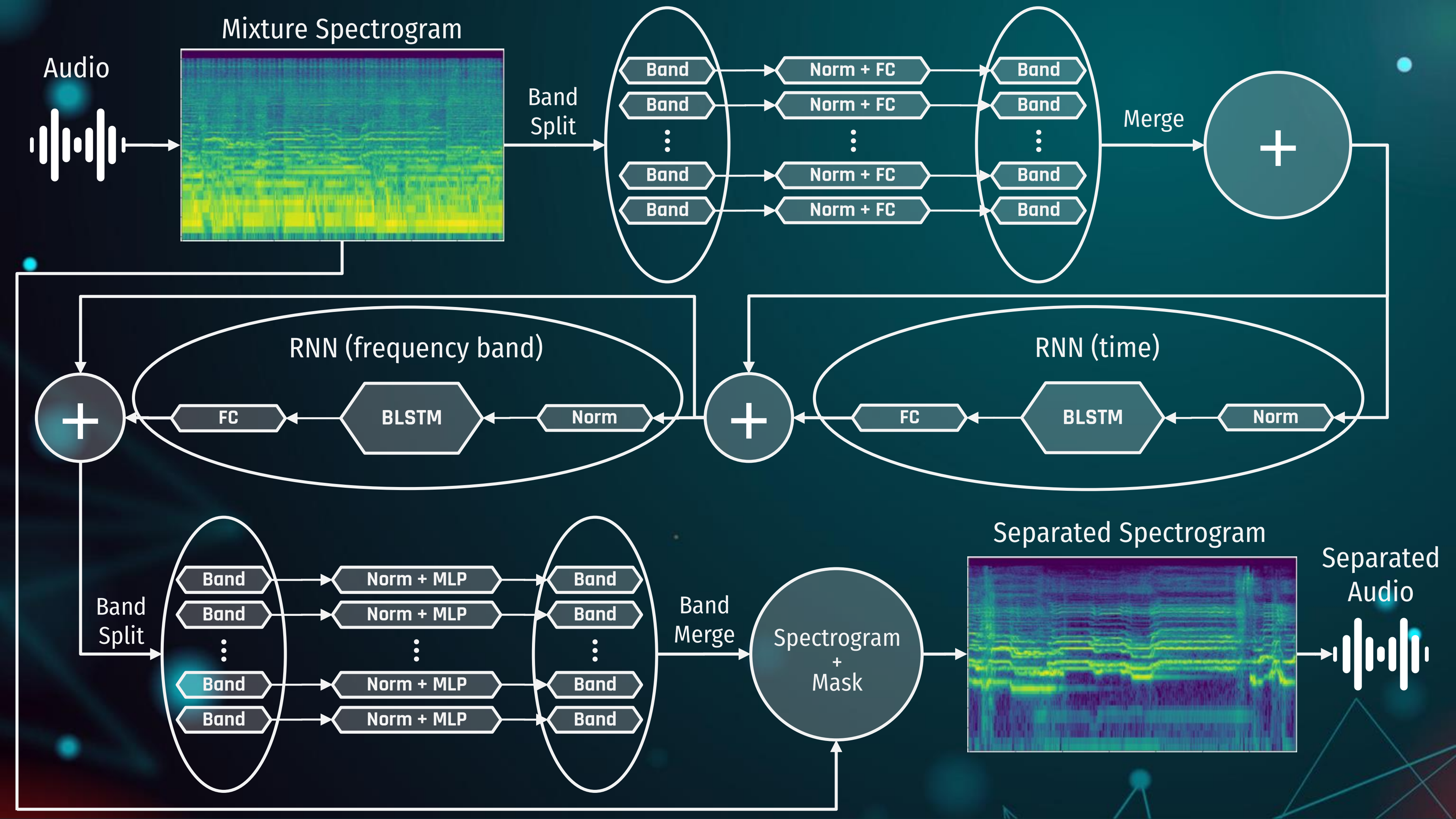


Band and Sequence Separation Modeling Module

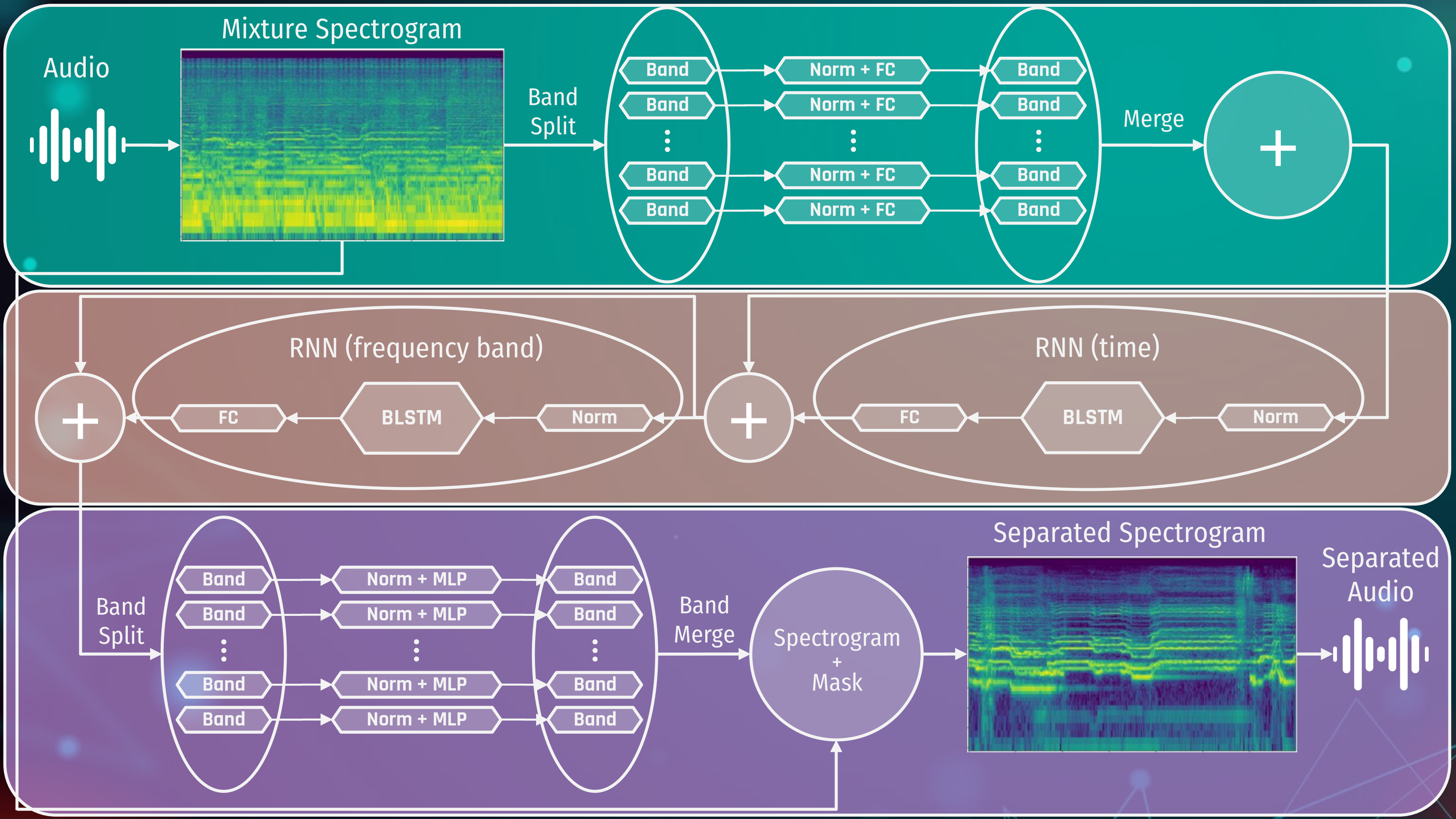
Mask Estimation Module




Mask Estimation Module
Audio
Mixture Spectrogram
Band Split
Band
Band
Band
Band
Norm + FC
Norm + FC
Norm + FC
Norm + FC
Band
Band
Band
Band
Merge
RNN (frequency band)
RNN (time)
FC
Norm
Band Split
Band
Band
Band
Band
Norm + MLP
Norm + MLP
Norm + MLP
Norm + MLP
Band
Band
Band
Band
Band Merge
Spectrogram + Mask
Separated Spectrogram
Separated Audio



Audio
Mixture Spectrogram
Band Split
Band
Band
Band
Band
Norm + FC
Norm + FC
Norm + FC
Norm + FC
Band
Band
Band
Band
Merge
+
RNN (frequency band)
FC
BLSTM
Norm
+
RNN (time)
FC
BLSTM
Norm
+
Band Split
Band
Band
Band
Band
Norm + MLP
Norm + MLP
Norm + MLP
Norm + MLP
Band
Band
Band
Band
Band Merge
Spectrogram
+
Mask
Separated Spectrogram
Separated Audio



Audio
Mixture Spectrogram
Band Split
Band
Band
Band
Band
Norm + FC
Norm + FC
Norm + FC
Norm + FC
Band
Band
Band
Band
Merge
+
RNN (frequency band)
+
FC
BLSTM
Norm
+
RNN (time)
FC
BLSTM
Norm
Band Split
Band
Band
Band
Band
Norm + MLP
Norm + MLP
Norm + MLP
Norm + MLP
Band
Band
Band
Band
Band Merge
Spectrogram + Mask
Separated Spectrogram
Separated Audio

Data Augmentation

- Reverse audio – BLSTM achieves this
- Gain scaling – randomly scales the gain (aka volume) of the signal
- Random crop – randomly selects “chunks” of audio each iteration

Performance

Audio	SDR	Sample
Mixture	--	
Vocals	10.47	
Bass	8.16	
Drums	10.15	
Other	7.08	

Conclusions

- This Band-Split RNN framework out-performs all other models in nearly every category and far outperforms vocal separation.
    - Including: Meta's Demucs, KUIELab's MDX-Net, PyTorch Open-Unmix, Deezer Spleeter
- This is a novel framework that was published in 2022 and still has a great deal of possible tuning, particularly regarding instruments other than vocals.

Conclusions

- Methods for improvement:
  - The 4 sources of Vocals, Bass, Drums, and Other are not all-encompassing.
    - Create datasets that consist of more diverse stems (e.g. acoustic guitar, strings, etc.)
  - Better choice of band-splitting (currently determined through rough grid-search)
  - Tune hyperparameters:
    - Frame size (how much audio is analysis at a time: Used 3 seconds
    - Hop size (spacing between frames, aka overlap): 2.5 seconds
    - Adjust dimensions in Band Split and Mask Estimation modules
    - Adjust dimensions and number of BLSTM layers in Band and Sequence module

Thank you!

Any Questions?