The Andrew and Erna Viterbi Faculty of
ELECTRICAL & COMPUTER ENGINEERING

SIPL
SIGNAL AND IMAGE PROCESSING LAB

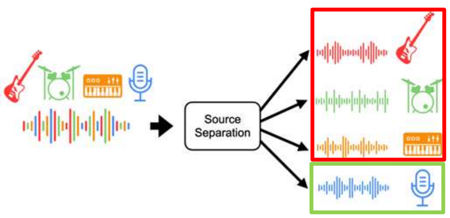
TECHNION
Israel Institute of Technology

Karaoke Maker Using Deep Learning

Izar Hasson, Supervised by Hadas Ofir

Introduction

- Audio source separation involves isolating individual components (e.g., vocals, drums, bass) from a mixed audio signal.
- This technology has applications in music production, audio restoration, and more.
- Deep learning-based approaches, have proven effective, but challenges remain in improving accuracy and handling diverse audio conditions.



Source Separation

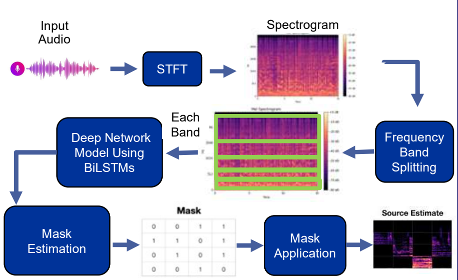
Goals

- Develop a deep learning model for audio source separation.
- Leverage spectral information and advanced neural network architectures.
- Achieve high separation accuracy across varied audio tracks.

Challenges

- Overlapping Frequencies: Sources share similar spectral components.
- Background Noise: Environmental noise complicates separation.
- Variability: Diverse genres, recording conditions, instruments, and sound quality.
- Model Complexity: Balancing accuracy and computational efficiency.

Vocal Separation

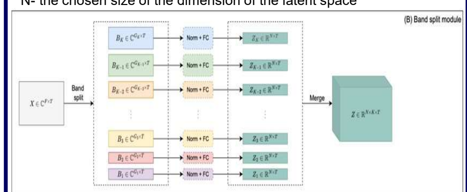


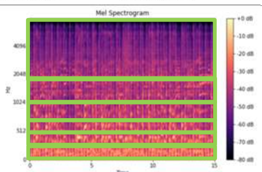
- The input is an audio mix
- After band splitting, each band is processed through our deep network
- The Mask get multiplied with the input spectrogram to get the estimated spectrogram
- The estimated spectrogram goes through ISTFT to get the audio back

Frequency Band Splitting

- Allows specialized processing for different frequency ranges
- The chosen frequency bands are designed to align with the distribution of musical components
- Band splitting improves separation accuracy

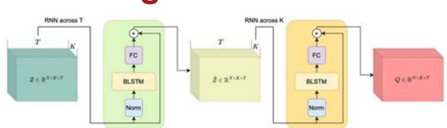
F , T - the frequency and time dimensions
K- number of bands
N- the chosen size of the dimension of the latent space





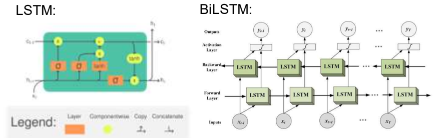
Example for Band Splitting, lower frequencies get finer resolution because that where the vocals and instruments usually at

Band and Sequence Modeling Module



F , T - the frequency and time dimensions
K- number of bands
N- the chosen size of the dimension of the latent space

- Band and Sequence Modeling Module is an efficient feature extraction Module
    - Suggested in [Yi Luo et al., 2022]
    - Processes the frequency bands sequentially to capture temporal dependencies and relationships across sub-bands.
    - The model applies bidirectional LSTM layers (BiLSTMs) in two dimensions: across time and across frequency sub-bands.
    - This dual-path processing enhances the model's ability to track temporal structures while maintaining spectral consistency, leading to improved audio source separation.

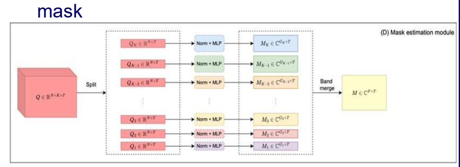


BiLSTMs (Bidirectional Long Short-Term Memory networks) are a type of recurrent neural network (RNN) that process data in both forward and backward directions.

- This allows the model to capture both past and future context, making them highly effective for sequential tasks like audio processing.

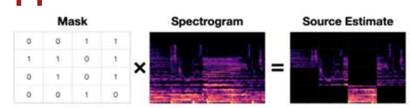
Mask Estimation

- Generating a spectral mask that highlights the target source while suppressing unwanted components.
- The refined feature maps provide a better representation of the audio for our task.
- The mask estimation module then applies a learned transformation to predict the optimal mask



The module reconnects every band to the final mask

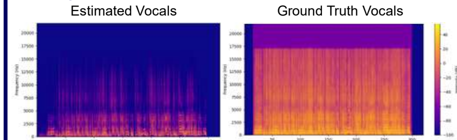
Application of the Mask



- Element-wise multiplication of the predicted spectral masks with the original spectrogram.
- The masked spectrogram is then used to produce a clean extraction of individual instruments or vocals from the mixture.

Results

- Songs were processed through our module
- For comparison to ground truth, we used 50 songs from the test set in out dataset. Meaning we know the ground truth vocals, drums etc.
- Results show a Mean Squared Error of 0.875. For comparison, the mean squared error of the original open Unmix module is 0.913.



Estimated Vocals vs. Ground Truth Vocals

- main differences are in the high frequencies, which the human ear can't notice
- The results are mostly pleasing to the human ear

Conclusions

- Successfully separated audio sources with pleasing accuracy using deep learning techniques
    - Effective band-splitting and sequential modeling frequency representation
    - BiLSTM-based sequence modeling
- Feasible for practical applications in music processing and source separation tasks

February 2025