Project Proposal: Deep Music Source Separation

Winter 2026 Final Project

Team Proposal

1 Executive Summary

Instead of selecting one of the three pre-defined options (Pose Estimation, Hallucinations, Inpainting), we propose an **"Alternative Project"** focused on **Music Source Separation (MSS)**.

The Goal: To take a mixed music track (a "mixture") and use Deep Learning to extract a single instrument stem (e.g., piano) from it.

The Research Angle: We will implement and compare distinct deep learning paradigms, specifically investigating the impact of **Time-Domain modeling** and **Attention mechanisms** on audio quality:

1. **Baseline:** Frequency Domain U-Net (Spectrogram-based).
2. **Model B1:** Time Domain Demucs with **LSTM** (Recurrent modeling).
3. **Model B2:** Time Domain Demucs with **Self-Attention** (Transformer modeling).

This project fulfills the course requirement to "train a model... working with a non-trivial amount of data" and extends learned material in a "research context".

2 Methodology & Architectures

We will implement three variations to isolate the effects of different architectural choices.

2.1 Approach A: Frequency Domain (Baseline)

- **Concept:** The audio is transformed into a **Magnitude Spectrogram** (via STFT) and treated as an image. The network predicts a soft mask to separate the source.
- **Architecture: 2D U-Net** with standard 2D convolutions.
- **Limitation:** Discards phase information, leading to imperfect reconstruction. Used here as a baseline to demonstrate the superiority of time-domain methods.

2.2 Approach B: Time Domain (Main Contribution)

- **Core Concept:** The model operates directly on the raw **1D Waveform**. This preserves phase information and allows the network to learn its own optimal frequency decomposition.
- **Backbone:** A **Simplified Demucs** architecture using strided 1D convolutions for the Encoder and transposed 1D convolutions for the Decoder.

We will compare two specific variations of the **Bottleneck** (the deepest layer of the network):

Model B1: Recurrent Bottleneck (LSTM)

- **Mechanism:** Uses a bi-directional **LSTM** (Long Short-Term Memory) in the bottleneck.
- **Hypothesis:** LSTMs excel at capturing **local continuity** and smooth transitions between notes, but may struggle with long-term dependencies (e.g., a drum beat repeating every 4 seconds).

Model B2: Attention Bottleneck (Transformer)

- **Mechanism:** Replaces the LSTM with a **Multi-Head Self-Attention** layer (Transformer Encoder) and positional embeddings.
- **Hypothesis:** Attention provides a **global receptive field**, allowing the model to attend to similar sounds across the entire input window. We expect this to improve performance on repetitive structures (rhythm) but potentially require more data to converge than the LSTM.

3 Dataset & Training Strategy

We will use **Supervised Learning**, creating perfect "Ground Truth" data efficiently.

3.1 Dataset: MUSDB18

This is the standard academic dataset for Music Source Separation, containing ∼150 tracks with isolated stems (Vocals, Drums, Bass, Other). This allows us to construct perfect ground-truth targets (Y) and inputs (X).

3.2 Data Augmentation & Curriculum Strategy

To improve generalization and robustness, we will implement a dynamic data augmentation pipeline:

- **Random Mixing (Remixing):** We will generate synthetic mixtures on-the-fly by summing stems from *different* tracks (e.g., combining "Drums from Song A" with "Piano from Song B").
- **Curriculum Learning:** We will experiment with a progressive training difficulty schedule:
  - **Stage 1 (Simple):** Train on mixtures of 2 sources (Target + 1 Distractor).
  - **Stage 2 (Intermediate):** Increase to 3 sources.
  - **Stage 3 (Full):** Train on the full mixture (Target + All Distractors).

4 Experiments & Comparative Analysis

The core of our report will focus on comparing the three models (Baseline, B1, B2).

4.1 Quantitative Metrics

We will evaluate all models on the held-out test set using the standard `museval` library:

- **SDR (Signal-to-Distortion Ratio):** The primary metric for source separation quality.

- **SIR (Signal-to-Interference Ratio):** Measures how much of the "other" instruments leaked into the prediction.
- **SAR (Signal-to-Artifacts Ratio):** Measures "robotic" or "metallic" noise introduced by the model.

4.2 Qualitative Analysis (The "Listen" Test)

We will specifically analyze:

- **Phase Coherence:** Does Model B (Time Domain) produce crisper audio than Model A (Spectrogram)?
- **Long-Range Structure:** Does Model B2 (Attention) handle repetitive drum beats better than Model B1 (LSTM)?
- **Transient Response:** Which model better preserves the sharp "attack" of a piano key press?

5 Work Plan

1. **Week 1:** Setup data loader (MUSDB18) and implement the Baseline (U-Net).
2. **Week 2:** Implement Model B1 (Demucs + LSTM) and verify waveform reconstruction.
3. **Week 3:** Implement Model B2 (Demucs + Attention) and run training comparisons.
4. **Week 4:** Evaluation, metric calculation, and report writing.

6 Feasibility Logistics

- **Submission Deadline:** 5.3.2026.
- **Libraries:** PyTorch (core), `musdb` (data loading), `torchaudio`.
- **Compute Strategy:** We will downsample audio to 16kHz or 22kHz to manage memory usage for the Attention mechanism.