Activity: Proof of Concept
Vocal Separation via Spectral Fingerprinting

Cameron Brooks

October 28, 2025

Note to Grader: Please visit `https://brookcs3.github.io/461-Design-and-Development-/` for accompanying interactive demonstrations, audio samples, spectrograms, and complete technical documentation.

Scope & Assumptions

Context

This proof-of-concept validates spectral fingerprinting for vocal separation before committing to neural network training. Built upon a forked U-Net repository (`https://github.com/brookcs3/Pytorch-UNet`) from the 2017 Kaggle car segmentation challenge, I added manual vocal separation experiments to validate technical feasibility.

Riskiest Assumptions

Three critical assumptions required validation: whether spectral patterns can distinguish vocals from instruments, whether librosa can handle spectral manipulation reliably, and whether 70-80% manual quality justifies U-Net training investment (expected 95%+). These matter because training requires 1000+ audio pairs and days of GPU compute—unknown feasibility carries significant risk.

Prototype & Setup

Method

The prototype implements manual spectral fingerprinting using 18-slice multi-scale analysis, extracting approximately 765,000 measurements per clip through 400-point frequency profiles, band energies, harmonics, and formant measurements. Perfect time-alignment in Audacity is a critical requirement for successful processing.

Setup Steps

Setup involves placing aligned audio files in `process/100-window/`, running `prepare_audio_files.py` to standardize format, then executing `sanity_check_complete.py` for 3-4 minutes of processing time.

Results & Evidence

Quality Metrics

Testing on "Intergalactic" by Beastie Boys achieved 70-80% overall separation quality, including 85-90% vocal intelligibility, 90-95% pitch accuracy, with 20-30% instrumental bleed remaining. All three assumptions were validated—spectral patterns proved unique enough, librosa's STFT/ISTFT pipeline functioned reliably, and manual success justified the future U-Net training investment.

Visual Evidence

In lieu of traditional screenshots, complete interactive demonstrations with playable audio samples, spectrograms, optimization curves, and full workflow documentation are available at `https://brookcs3.github.io/461-Design-and-Development-/`. This website serves as visual proof, containing original mixture, target acapella, and extracted vocal audio players alongside technical visualizations.

Evaluation

Challenges

Time alignment proved critical—millisecond offsets cause complete failure. The key insight emerged that separation works in the frequency domain through spectral masking but remains impossible in the time domain where waveforms are literally added together.

Next Steps

The manual POC successfully proves the core concept. Future work involves gathering 1000+ training pairs and training a U-Net to automate mask generation, targeting 95%+ quality in 100ms versus the current 3-4 minutes. This three-phase approach successfully validated feasibility before committing significant resources.

References

1. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Springer, LNCS, Vol.9351: 234-241. Available: `https://arxiv.org/abs/1505.04597`

2. milesial. (2017). Pytorch-UNet: PyTorch implementation of U-Net for Kaggle Carvana Image Masking Challenge. GitHub repository. `https://github.com/milesial/Pytorch-UNet`

3. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python. *Proceedings of the 14th Python in Science Conference*, 18-24. `https://librosa.org/`

4. Brooks, C. (2025). Vocal Separation Proof-of-Concept: Interactive Documentation. `https://brookcs3.github.io/461-Design-and-Development-/`

5. Brooks, C. (2025). Pytorch-UNet Fork with Vocal Separation Experiments. GitHub repository. `https://github.com/brookcs3/Pytorch-UNet`