Project 5 Proposal
Creating Karaoke Tracks Using CNN

Catherine Magsino

Introduction

Since one of my favorite hobbies is to sing karaoke with my friends and family, my passion project will consist of using CNN to create my own karaoke tracks. In other words, I'll be able to remove vocals from a full music mix, resulting in only background music.

MVP

For my MVP, I will be using Librosa to initially separate the audio of a song with both vocals and background music. Through this method, I will:

- Use Short-time Fourier transform (STFT) to compute the spectrogram magnitude and phase
- Create a filter by comparing frames using cosine similarity and aggregate similar frames by taking their (per-frequency) median value
- Use a margin to reduce bleed between the vocals and instrumentation masks
- Apply the vocal and instrumentation masks to the full audio
- Plot the spectrograms of the full spectrum, background, and vocals
- Use inverse Short-time Fourier transform (ISTFT) to convert the spectrogram back to the time domain
- Output the background and vocals back to audio

After MVP

I'll work on improving my MVP by training a convolutional neural network to classify human voice. To pick out the voice, it will separate the voice's unique spectrogram from the spectrograms of instruments. As a result, the input would be full audio (vocals+background music), while the output would be either just the background music or the vocals.

Data

Data will be retrieved from the DAMP (Digital Archive of Mobile Performances) Dataset Project, which includes audio from thousands of acapella performances from Smule users. Full audio performances may be retrieved from YouTube.

Lyrics

Time permitting, I'd like to use speech recognition to sync lyrics to the audio.