# Audio Source Separation For Automatic Music Transcription

Bradford Derby, Lucas Dunker, Samarthkumar Galchar, Shashank Jarmale, Akash Setti

## 1 Project Overview

Source separation is the process of isolating individual sounds in an auditory mixture of multiple sounds [1], and has a variety of applications ranging from speech enhancement and lyric transcription [2] to digital audio production for music and film. Once a complete piece of sound is recorded, it is infamously difficult to recover the individual parts; the manual process of source separation is expensive and time-consuming, but recent developments in deep learning have brought promising new approaches for constructing distortion-free stems out of audio signals [3].

A similarly active area of music information retrieval (MIR) research is the field of automatic music transcription, or AMT. Sheet music is widely known as an effective medium for musicians to share their work; as a result, several attempts have been attempted to convert audio data into sheet music without human input [4]. Until recently, accurate AMT was nearly impossible; with recent advancements in deep learning, however, AMT accuracy for areas like classical piano has skyrocketed to rates as high as 95% [5].

In this project, we aim to create two deep learning models:

- **Audio Source Separation:** Translate digitized audio signals into their individual sources using neural networks with back propagation. [1]
- **Automatic Music Transcription:** Transcribe audio to sheet music using long short-term memory (LSTM), a type of recurrent neural network (RNN). [5]

Our goal is to utilize these deep learning models for an end-to-end automation process of converting audio to sheet music. A user can input any song they choose, and our process would output both the individual audio sources as well as sheet music for each part (vocals, piano, guitar, drums, etc).

## 2 Related Work

Both audio source separation and automatic transcription have been done before, but to our knowledge, there is no product or service published that can do both.

In [3], the authors illustrate that sound and time can indeed be separated within their own respective domains. Spectrograms (Mel Spectrograms) are utilized in order to divide the audio into smaller frames so time and frequency can be extracted. Convolutional Neural Networks (CNNs) are used to process the spectrograms . The work of [6] utilizes the MCRAE network, consisting of an encoder and a decoder. The encoder extracts features using convolutional layers, while the decoder reconstructs sources using transpose convolutional layers. This method processes audio in raw data without any pre- or post-processing of the input and output signals.

When working with automatic transcription, the authors of [5] used recurrent CNNs with bidirectional Long Short-Term Memory cells to classify audio representations of classical piano music symbolically. The CNN layers were used to find patterns in the spectrograms and LSTMs were used for handling relationships between notes.

# 3 Proposed Methods

We propose a two-stage deep learning approach to achieve our goal of end-to-end automation for converting raw song audio into sheet music categorized by instrument.

For the first stage, we will implement a convolutional neural network (CNN)-based model for audio source separation. We plan to use a U-Net architecture, which has shown promising results in recent audio separation tasks discussed in [2] and [7]. The U-Net will be trained on datasets such as MUSDB18 [8] & DSD100 [9], which consist of full-length music song tracks accompanied by their isolated stems for vocals, bass, drums, and other relevant instruments. We will apply the Short-Time Fourier Transform (STFT) to convert the raw audio into spectrograms which serve as input for our deep learning model. The network will be trained to mask the spectrogram, isolating each instrument onto a different channel. Each resulting spectrogram will be reconstructed back into a waveform for the sake of metric scoring & qualitative testing.

In the process of improving stem separation performance, we will tweak neural network hyperparameters & architecture, experiment with data augmentation techniques such as pitch shifting and time stretching, investigate the impact of dataset size, and more. We will use standard metrics to objectively measure the performance of our stem separation model, including Source-to-Distortion Ratio (SDR), Source-Image-to-Spatial-Distortion Ratio (ISR), Source-to-Interference Ratio (SIR), and Source-to-Artifacts Ratio (SAR) [6]. We will also conduct subjective evaluations based on our own listening to the separate audio tracks generated.

For the second stage, we aim to implement a method for chord estimation from audio signals using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. The method leverages Constant-Q Transform (CQT) for preprocessing, followed by specmurt analysis, to improve chord recognition accuracy by addressing overtone components in musical instrument recordings [10]. Our method is based on prior work that has shown success in using such preprocessing techniques in combination with LSTMs to capture the temporal dynamics of audio signals [4].

Initially, the system receives an audio signal in WAV format. This signal is converted into a logarithmically scaled spectrogram using the Constant-Q Transform (CQT), which maintains a constant ratio between musical intervals, allowing the mapping of musical pitch on a perceptually relevant scale [10].

We apply spectral analysis to suppress overtone frequencies, treating the audio as a convolution of the clean spectrogram and an overtone spectrum. This process isolates fundamental frequencies, yielding a clearer representation of true pitches for chord recognition. Chroma vectors are computed from the processed spectrogram which captures the combined magnitude of each pitch class across all octaves. These would serve as features for our input in neural networks.

For classification, we use an LSTM model to learn temporal relationships in the chroma vectors. The LSTM's ability to store information over time helps capture chord progressions. A Bidirectional LSTM enhances the model by considering both past and future context, improving chord transition predictions [10].

We train the LSTM using categorical cross-entropy and an Adam optimizer [4], expecting it to handle long-term dependencies better than simpler models. Finally, we will visualize the predicted chords using music21, providing a graphical representation of the model's output for interpretation and validation.

# 4 Timeline

We've broken our timeline into 2-week milestones:

- By October 1st, we aim to have established our project repository, explored similar projects, and broken down the necessary action items, with clear task allocation among team members.

- By mid-October, specifically October 15th, we anticipate having a working but rudimentary stem-splitting model using the MUSDB18 dataset. We aim to have our AMT system accurately output notes and pitches by this stage as well.
- By October 29th, we expect our model to adequately split audio on waveforms different from our training data, and the sheet music data should be in a form clear enough to integrate into a simple graphical user interface.
- As we approach November 12th, the project should be tentatively finished, allowing us to begin drafting the final proposal.
- By November 26th, we will have written the final report, polished the presentation and source code, and prepared our project for submission.

# References


[1] E. Manilow, P. Seetharman, and J. Salamon, *Open Source Tools & Data for Music Source Separation*. https://source-separation.github.io/tutorial, Oct. 2020.

[2] J. Oh, D. Kim, and S. Yun, "Spectrogram-channels u-net: a source separation model viewing each channel as the spectrogram of each source," *CoRR*, vol. abs/1810.11520, 2018.

[3] D. Parekh, D. Kharah, K. Suthar, and V. Shirsath, "Audio stems separation using deep learning," *International Journal of Engineering Research  Technology (IJERT)*, vol. 10, 2021.

[4] M. Dua, R. Yadav, D. Mamgai, and S. Brodiya, "An improved rnn-lstm based novel approach for sheet music generation," *Procedia Computer Science*, vol. 171, pp. 465–474, 2020. Third International Conference on Computing and Network Communications (CoCoNet'19).

[5] A. Grossman and J. Grossman, "Automatic music transcription: Generating midi from audio," *Stanford CS230*, Jun 2020.

[6] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," *CoRR*, vol. abs/1803.00702, 2018.

[7] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *CoRR*, vol. abs/1806.03185, 2018.

[8] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017.

[9] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation - 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings* (P. Tichavský, M. Babaie-Zadeh, O. J. Michel, and N. Thirion-Moreau, eds.), (Cham), pp. 323–332, Springer International Publishing, 2017.

[10] T. Hori, K. Nakamura, and S. Sagayama, "Music chord recognition from audio data using bidirectional encoder-decoder lstms," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1312–1315, 2017.