Similarity Metrics in the Source Separation of Percussive Sounds

Christopher Grabow, Acoustics – Penn State University
Advisor: Dr. Tyler Dare

Introduction

The process of source separation involves isolating individual components from an all-encompassing track or file. In the field of acoustics, this process can be applied to various audio recordings to extract key features for additional analysis or processing. In the past decade, many music source separation (MSS) models have been developed to extract key components of a song through the use of spectrograms. Figure 1 shows an example of these extracted components in spectrogram form for a short song snippet. While the separation of general features is quite good, many current MSS models lack a higher level of specificity for individual instruments and sounds. In addition, the "Drums" in a given song are transient and broadband in nature. This indicates that spectrograms are not ideal for the separation of individual drum sounds since there are no defining frequency features for a drum hit. Instead, techniques utilizing the time domain will be explored.

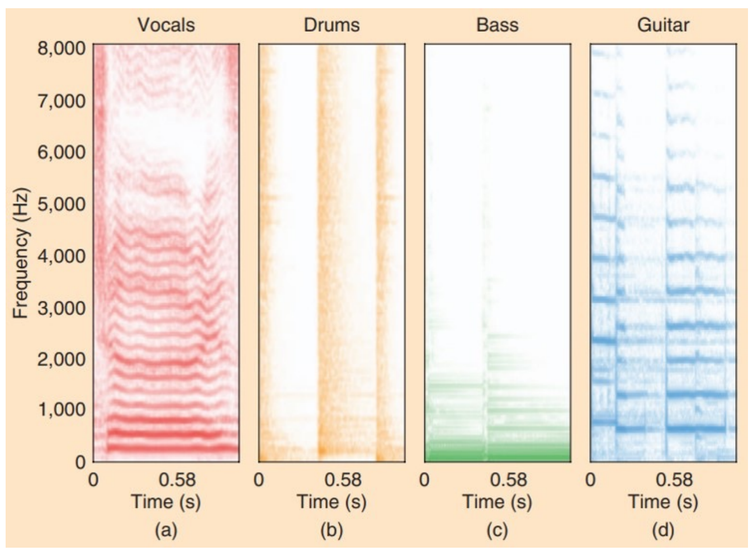

Figure 1. Spectrogram of individual stems for a song snippet. [1]

Objectives

The main objective of this research is to develop a new source separation algorithm that functions solely in the time domain and separates a drum track into its individual percussive sounds. The techniques used to perform the source separation are two similarity metrics known as dynamic time warping (DTW) [2] and normalized cross correlation (NCC) [3]. The new algorithm is trained with a database of digital drum sounds and tested with constructed drum tracks containing individual and polyphonic drum hits.

Methodology

The source separation process involves several steps to isolate the individual drum sounds within a given track. A summarized process of the new algorithm is shown in the flowchart in Fig. 2. The similarity metrics (DTW and NCC) play a key role in matching drum sounds from the database to those in a given track. They both function by assigning a single numerical value to the level of similarity between two time series. For DTW, this is known as the DTW distance, where two identical time series will have a distance of zero. For NCC, it is the correlation value which has a range from 0 to 1, where a value of 1 indicates identical time series. The DTW metric is unique in that it will stretch or compress the time series to find the optimal level of similarity. This feature provides additional benefits beyond the capabilities of the NCC metric. Each similarity metric is also capable of time-aligning a pair of time series as described in the following step. Finally, gradient descent signal subtraction is a machine learning method that finds the ideal amount of the identified drum signal to subtract out of the track section. The following equation mathematically illustrates this process,

$$y_f(t) = y_i(t) - \alpha * x(t)$$

where $y_i(t)$ is the track section signal, x(t) is the selected drum signal, α is a vector of values between 0 and 1 with a small, uniform step size, and $y_f(t)$ is the resulting track section after subtraction.

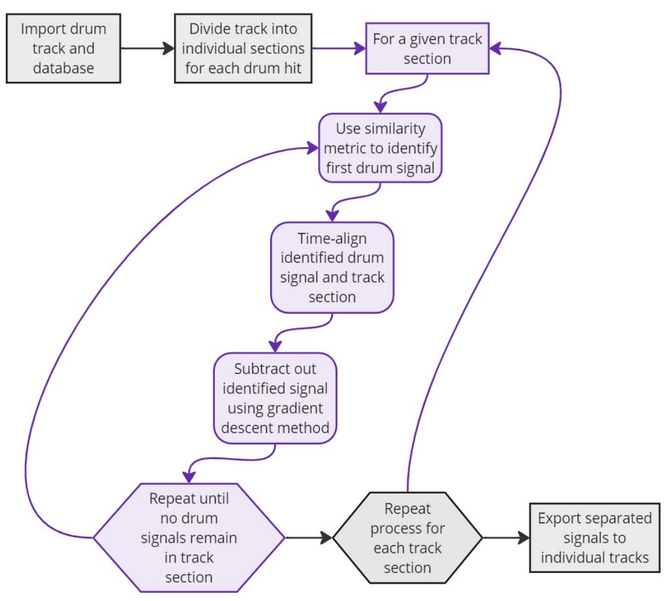

Figure 2. Flowchart of proposed source separation algorithm

Results / Analysis

Prior to testing the new algorithm against full-scale drum tracks, a series of testing was conducted with 50 polyphonic drum signals. The purpose of this testing was to determine the accuracy of both similarity metrics and find specific values for the correlation limits and noise threshold in the algorithm. The results of this testing are shown below in Table 1.

Table 1. Results of testing for 50 polyphonic drum signals.

	NCC	DTW
Correctly identifies at least one drum signal:	49 times	12 times
Correctly identifies all drum signals:	42 times	9 times
Upper Correlation Limit (relative to metric):	0.75	272
Lower Correlation Limit (relative to metric):	0.23	340
Root-Mean-Square Noise Threshold:	0.004	0.012

The results clearly show that DTW is not an effective similarity metric when it comes to source separation. The DTW calculation itself does not perform in an ideal manner when a signal contains added noise or effects. For this problem, drum signals layered on top of each other acts as noise and makes the metric less effective. However, NCC performed above expectations and was able to separate out the majority of drum signals. The use of NCC in testing the algorithm on full-length drum tracks showed promising results. An example of a source separated track is shown in Fig. 3.

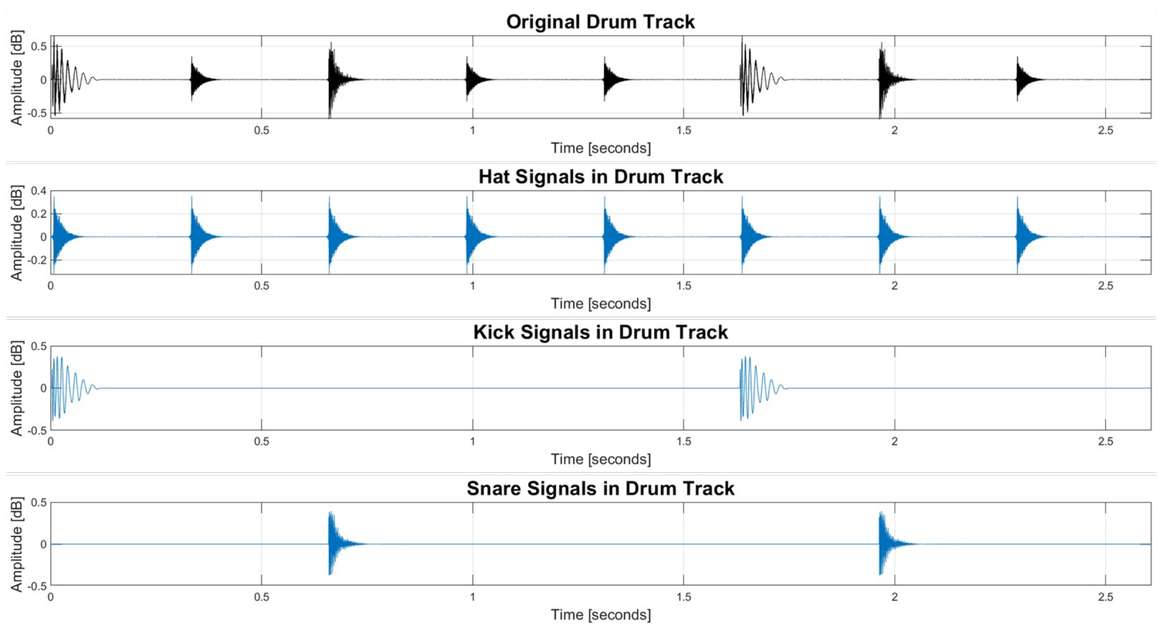

Figure 3. Drum track with separated drum signals (hi-hat, kick, and snare)

Future Objectives

The results from testing indicate that the new source separation algorithm shows promise in particular applications. Many improvements can be made to the model to make it more robust and function with a greater number of drum tracks. First, an onset detection function can be implemented to more accurately divide up the track into sections. Second, significantly more drum sounds can be added to the database to capture a wider variety of tracks. Finally, more testing can be completed with additional polyphonic drum signals to improve the accuracy of the correlation limits and noise threshold.

Acknowledgements

This work was partially supported by the **PIPELINE**: **P**enn State **I**ntern **P**ipelin**E** **LI**nks to **N**avy **E**ngineering program, ONR grant #N000142312656. The Penn State PIPELINE Program motivates and connects students and faculty to careers and research opportunities with the Navy technical workforce.

References

[1] E. Cano et al., "Musical Source Separation: An Introduction", IEEE Signal Processing Magazine 36, 31-40 (2018).
[2] M. Herrmann and G. Webb, "Amercing: an intuitive and effective constraint for dynamic time warping", Pattern Recognition 137, (2023).
[3] D. Brown, "ACS 503: Signal Analysis for Acoustics and Vibration Notes: Correlation", (2022).
[4] D. Huang, "Machine Learning for Engineering", (2024).