AUTOMATIC DRUM TRANSCRIPTION USING RNN SOURCE SEPARATION AND ENERGY-BASED ONSET DETECTION

William Bellows

wbellows@ece.rocehster.edu

ABSTRACT

The aim of Automatic Drum Transcription (ADT) is to convert a recorded or synthesized percussive mixture into onset information for each constituent part within it. Onset information can be output in the form of transient tracks of impulses which can then be used to create musical scores or for drum sample replacement in digital music production. This algorithm uses RNN based source separation and energy based onset detection to isolate individual drum tracks. Drum mixtures each contain three varying parts of kick, snare and hi hat (cymbal) information. The algorithm is similarly tested with mono drum samples of the same composition. Results will show how effective RNN source separation is at separating and detecting onsets in percussive mixtures.

1. INTRODUCTION

Automatic music transcription, in general, has been a constant subject of study in the field of computer audition. The ability to simply listen to a song and immediately transcribe it into sheet music or onset information has huge applications in education and creative industries [1]. One common application is automatic drum transcription, or ADT. ADT is especially useful in music creation and production due to the common practice of full drum sample replacement. This is when a producer uses the time-based onsets, or transients, of a drum performance to insert ideal samples in place of the recorded hits. ADT makes this task even simpler by removing the original performance entirely, leaving only the transients for an artist to fill with whatever samples they choose. Learning models such as recurrent neural networks (RNN) are useful for automatic transcription applications; their superior source separation results allow for cleaner onset detection and performance [1].

2. METHODS

The proposed method for this algorithm is to intake input features (drum mixtures), convert those features into spectrograms through Short Time Fourier Transform (STFT), run RNN separation on the spectrograms [1], isolate the audio waveforms and detect musical onsets based on their energy.

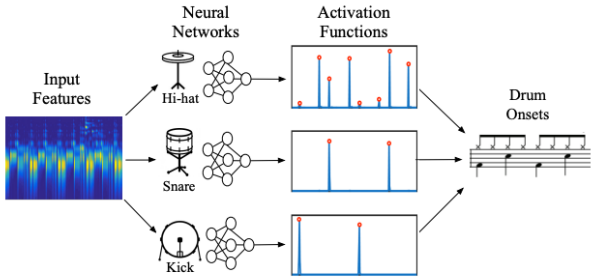

Fig. 1: Proposed method for ADT in [1]. Note the neural networks individually separating each drum track. In this paper, onsets will be detected from waveforms rather than activation functions as described in [1].

RNN source separation is a generally accepted way for accurate isolation in music based mainly on timbre rather than pitch, and because it includes temporal information. Since drums mixtures will rely solely on timbre for source separation, the use of RNN or similar networks such as bidirectional recurrent neural networks (BRNN) is advised [1]. The model used in this algorithm is based on the RNN model described in [2]. The architecture of the model is shown in figure 2 below.

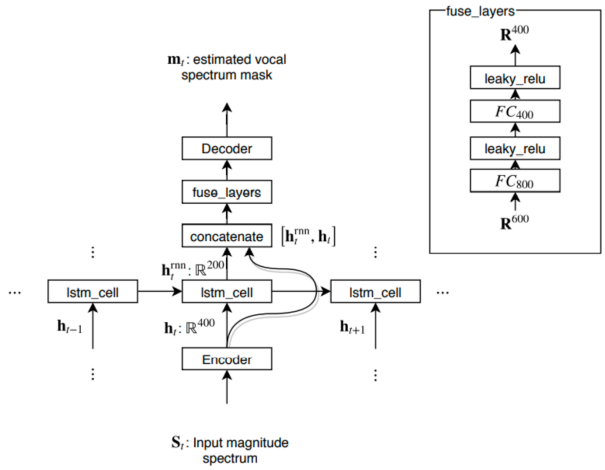

Fig. 2: Architecture of the RNN Source Separation model as shown in [2]. The methods used in this paper are directly based off of this model.

The architecture runs data through an encoder comprised of several fully-connected layers (FC) and Leaky-Relu layers which reduce data down to a smaller bin size. The encoded data is both sent to an LSTM layer, and also concatenated with the output of said layer. The LSTM

layer saves past values in order to preserve temporal information. This data is then put through a "fuse" layer which converts the data to the proper bin size before it is decoded.[2] The model outputs a spectral "mask" which, when multiplied with the input mixture spectrogram, produces the spectrogram of the separated source through cancellation [2].

Instead of just one vocal track separation [2], this model is now used to separate drum mixtures containing kick drum, snare drum, and hi hat (cymbals). Aside from being easier to train for, these parts are generally considered the most important components of a drum beat. Like typical drum separation algorithms, the network is trained three times to separate each individual part from the mixture [1]. The network is trained with a data set of sixteen different drum mixtures of the same three parts, each with a varied timbre and groove. These mixtures are read into the RNN as spectrograms, and the spectrograms of the individual parts are set as outputs for training. The library contains WAVE files of each mixture and their separated parts sampled at 44.1 kHz with 16 bit resolution. The samples are obtained from Apple Logic Pro X's drummer libraries.

After the network is trained with the given datasets, output spectrograms can be converted once again into raw audio files to be put through a simple energy-based onset detection function. The input audio file is sampled with a Hamming window with frame size of 512 samples and hop size of 256 samples. Onsets are detected through peak-picking on each of these frames by setting an energy threshold at an appropriate level for each track. A WAVE file is then compiled and written with quick bursts of white noise corresponding to each onset frame containing an energy above the given threshold.

3. EXPERIMENTATION

The training for each of the three models gave very promising results. For each of the datasets of isolated kick drums, snare drums, and hi hats, the loss between target and output spectrograms rapidly dropped below 0.1 on each run. A sample loss curve is shown in figure 3 below.

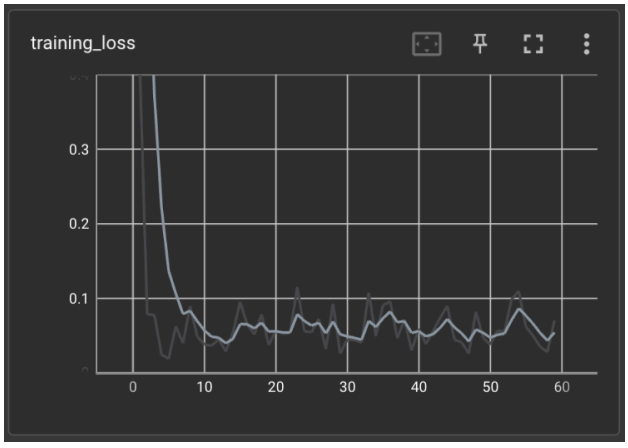

Fig. 3: Training loss curve. Shows the data loss of the output spectrogram versus the target spectrogram for the separated parts of the sample drum mixtures.

After the network had been adequately trained, experiments were conducted with a different set of drum mixtures. These mixtures were also restricted to kick, snare and hi hat tracks, and similarly be of a variety of timbres and rhythmic patterns. Other drum samples, such as ones containing tom-toms or ride cymbals, would produce inaccurate results due to their different timbres [1].

These experiments are meant to determine the robustness of the trained RNN model, suggesting whether or not it needs to be trained with a larger and more diverse data set in order to produce more accurate results. Results from testing using other drums mixtures from the same sample in Logic Pro X as used in training proved to be extremely accurate. The separated outputs, especially for the kick and snare drums, were very well-defined aside from some audible artefacts which were quiet enough to be ignored by the onset detection function. On the other hand, acoustic drum recordings such as those found in the MDB Drums Dataset [3] provided less desirable results. The separations were noisy, distorted and had a great deal of harmonic loss, although some onsets were still strong and audible. These results were expected, as the training was performed only on a very small dataset consisting of highly-produced and isolated drum samples.

4. CONCLUSIONS

Based on the results mention above, it is evident that a larger dataset is required for training. The RNN source separation algorithm performs very well on drumbeats that are "familiar" to it, e.g. those taken from the sample library in Logic Pro X. It would be useful to include a larger dataset of acoustically recorded drum mixtures along with their separated parts, if such a dataset existed. The drums tracks contained in the MDB Drums library are useful for testing [3], however they only contain drum mixtures and cannot be used to train separation networks. If given the time and resources, it would be beneficial to train networks using a large dataset of several hundred acoustically recorded drum mixtures on a variety of drums and cymbals. This would include not only the three explored in this paper (kick, snare and hi hat), but also ride cymbals, crash cymbals, and tom-toms of various tunings.

The goal of this algorithm is to use the transient WAVE files from the onset detection function in different applications of music production and education. These files can be used for full drum sample replacement, or simply to learn drum parts by analyzing each individual percussive track. The screen capture below shows transient-based drum sample replacement, a common practice in modern digital music production.

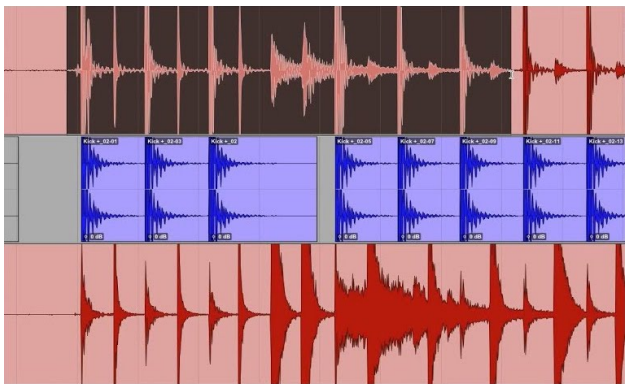

Fig. 3: Drum sample replacement in Pro Tools. One can add samples which overlap the transients of an existing drum performance or, in the case of the algorithm, onset tracks of the drum performance.

In the future, it would be very useful to develop a .pdf writer which could take the temporal onset information gathered from the input material and notate sheet music for the user. This would be very similar to automatic piano transcription, which aims to write a musical score based on the input audio. With a musical score writer, a user could easily transcribe any drumbeat without have to write notes manually.

5. REFERENCES

1. R. Stables, J. Hockman, and C. Southall, "Automatic drum transcription using bi-directional recurrent neural networks.," *Welcome to BCU Open Access Repository*, 07-Aug-2016. [Online]. Available: https://www.open-access.bcu.ac.uk/4101/. [Accessed: 06-Dec-2022].

2. A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 Signal Separation Evaluation Campaign," *SpringerLink*, 15-Feb-2017. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-53547-0_31. [Accessed: 06-Dec-2022].

3. C. Southall, C.-W. Wu, A. Lerch, and J. Hockman, "MDB drums: An annotated subset of MedleyDB for automatic drum transcription," *Welcome to BCU Open Access Repository*, 23-Oct-2017. [Online]. Available: https://www.open-access.bcu.ac.uk/6179/. [Accessed: 06-Dec-2022].