TimbreNet: A Convolutional Network for Blind Audio Source Separation

Scott Reid
shrcor@stanford.edu

Nathaniel Okun
nokun@stanford.edu

Abstract

We aim to separate the instruments in multi-instrument audio tracks, a process commonly known as audio source separation. Solving this problem would be exceptionally useful for musicians. Mixed musical sources are difficult to analyze, transcribe and sample. Background noise, distortion and the tendency for harmonics to mix together complicate the task for even professional musicians. Improvements made in this domain can be applied more generally to dozens of important problems from dialogue transcription to seismic monitoring [3].

Data

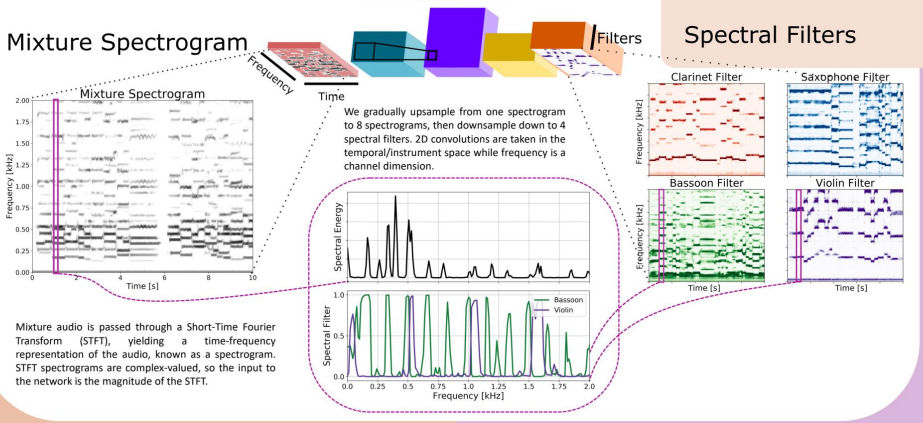
We performed most of our experiments on the Bach10 dataset [1]. The dataset consists of 10 J.S Bach chorales performed by four different instruments - a saxophone, bassoon, clarinet and violin. Each chorale is roughly 30 seconds - the dataset has a cumulative length of 5 minutes and 34 seconds. Each instrument was recorded in isolation while the musician listened to the recordings of others through a headphone.

We found that 5 minutes of audio was insufficient to perform well on the task. We split the dataset into slices of roughly one second and permuted them to create the equivalent 255 years of distinct audio. We train on 3 hours of permuted audio due to memory and time constraints.

References

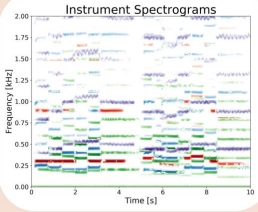
[1] Zhiyao Duan and Bryan Pardo, "Soundprism: an online system for score-informed source separation of music audio," IEEE Journal of Selected Topics in Signal Process., vol. 5, no. 6, pp. 1205-1215, 2011.
[2] F.R Stöter, A. Liutkus and N. Ito, "The 2018 signal separation evaluation campaign," International Conference on Latent Variable Analysis and Signal Separation., 2018
[3] J.J Sallas, D. Corrigan and K.P Allen (1998). United States Patent no. 5721710
[4] Emmanuel Vincent, Rémi Gribonval, Cédric Févotte. Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech and Language Processing, Institute of Electrical and Electronics Engineers, 2006, 14 (4), pp.1462–1469.

TimbreNet



Spectral Filters

It is very difficult for the network to output Short-Time Fourier Transform (STFT) spectrograms for each instrument because STFT spectrograms are complex-valued. Instead, the network outputs spectral filters. Each spectral filter is the same shape as the input mixture spectrogram, and elements are bounded between zero and one via a sigmoid output layer. We multiply the instrument spectral filter by the input spectrogram, yielding a complex STFT instrument spectrogram. Then we invert the instrument spectrograms, yielding the separated instrument audio.

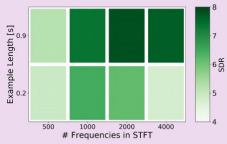


Evaluation

We evaluated our model by calculating the extracted sources' similarity to ground truth as represented by the Source to Distortion Ratio (SDR). The evaluation estimates the predicted source as the sum of the target source, noise, interference and other artifacts using orthogonal projections [3,4].

$$s_{\text{predicted}} \approx s_{\text{target}} + e_{\text{interference}} + e_{\text{noise}} + e_{\text{artifact}}$$

$$\text{SDR} = 10\log_{10}\left(\frac{\|s_{\text{predicted}}\|^2}{\|e_{\text{interference}} + e_{\text{noise}} + e_{\text{artifact}}\|^2}\right)$$



After training for 60 epochs with 6000 examples, we reached an SDR of 7.8 for optimal hyperparameters.

Unordered Loss

When the network is trained, the ordering of instrument filters at the network output may not match the ordering of instruments in training examples. To solve this problem, we developed an "unordered" loss function that returns the squared error from the best pairing of output and training example instrument spectrograms.

We first define the vector of each possible pairwise squared error between test and training instrument spectrograms:

$$\mathbf{p} = [\ \|\hat{y}_1 - y_1\|^2 \quad \|\hat{y}_1 - y_2\|^2 \quad \cdots \quad \|\hat{y}_N - y_N\|^2\ ]^\top$$

Here, N is the number of instruments and $\mathbf{p}$ is an N^2 dimensional vector. Next, we define $\mathbf{A}$, an $N! \times N^2$ matrix. Each of $\mathbf{A}$'s rows is a distinct permutation of the standard basis vectors. For $N = 2$,

$$\mathbf{A} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \\ \mathbf{e}_2 & \mathbf{e}_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

The product $\mathbf{Ap}$ yields a vector of possible losses for each instrument pairing. Our loss is the minimal value of $\mathbf{Ap}$,

$$J = -\text{LogSumExp}(-\mathbf{Ap})$$

Future: More Instruments

TimbreNet's performance on the Bach10 dataset proves that our architecture, combined with permuted data augmentation, is a promising avenue for blind monaural audio separation. With the Unsorted Loss function, we hope to train TimbreNet on a much larger dataset with many more instruments. Such a dataset can be obtained by scraping the internet for solo instrument audio.

More work must to be done to gain intuition about how TimbreNet works. We suspect that the upsampled dimension contains information about the timbre, or sound quality, of the instruments present in the mixture. It would be very interesting to interpret the network and test this hypothesis.

TimbreNet's could be used for improved audio transcription. The task of transcribing audio from multiple instruments requires the ability to separate audio sources. Thus, multi-instrument audio transcription can be viewed as a transfer learning task from audio transcription.