Bespoke Neural Networks for Score-Informed Source Separation

Ethan Manilow
Interactive Audio Lab
Northwestern University, Evanston, IL

Bryan Pardo
Interactive Audio Lab
Northwestern University, Evanston, IL

ABSTRACT

In this paper, we introduce a simple method that can separate arbitrary musical instruments from an audio mixture. Given an unaligned MIDI transcription for a target instrument from an input mixture, we synthesize new mixtures from the transcription that sound similar to the mixture to be separated. This lets us create a labeled training set to train a network on the specific bespoke task. When this model applied to the original mixture, we demonstrate that this method can: 1) successfully separate out the desired instrument with access to only unaligned MIDI, 2) separate arbitrary instruments, and 3) get results in a fraction of the time of existing methods. We encourage readers to listen to the demos posted here: `https://git.io/JUu5q`.

1. INTRODUCTION

Source separation in the musical context is the task of extracting the audio of one or more target instruments from an audio mixture containing multiple concurrent instruments. Today's top musical source separation methods all rely on deep learning models trained on large sets of labeled training data. All machine learning methods rely on the assumption that the distribution of the training data matches the distribution of data out in the real world. When this assumption is broken, the performance of a model suffers. It is not possible to guarantee the system was trained on data relevant to the current separation task using this paradigm.

In this paper, we introduce a method that explicitly avoids the problem of mismatch between the training and real world data, but creating a bespoke training set specifically for the mixture. Inspired by Deep Image Prior [1], our method aims to separate out exactly one source from exactly one mixture by training a separation model on a dataset created specifically to be similar to the mixture that requires separation. The resulting bespoke network is trained specifically to solve the task at hand.

2. BESPOKE NEURAL NETWORKS

This paper introduces *Bespoke Neural Networks* for musical source separation, where a network is trained to only separate exactly one target source from exactly one mixture. We assume the user can provide an unaligned MIDI transcription for the target source in the mixture. We create synthesized audio from the MIDI, and synthesize the source in one or more ways [2, 3]. We then apply many types of augmentations to the synthesized source, and mix it together with background audio. We describe the specifics of augmentation, the background audio, and mixing in the next section. Using the known synthesized source as the ground truth source label, we train a small neural network to "overfit" to these local augmented examples as surrogate to the actual mixture. After small number of training iterations, we then use the trained network on the original mixture to retrieve an estimate of the desired source.

Because we assume that some transcription of the desired source is available *a priori*, we say that this method is score-informed. We note that unlike traditional methods for score-informed separation [4], our method does not need perfectly aligned score data. Other deep net methods have been proposed for score-informed separation [5] or for separating and transcribing sources [6], however they are limited in types of source types they can separate and require large amounts of time and data to train.

There are many advantages to this method over existing methods. First, our method only requires an unaligned MIDI transcription of a target source, which we assert is a much less stringent requirement than the large corpus of isolated source data needed by other methods.

Second, our method is not expected to model a large distribution of data in the hopes that it will generalize well to every unseen musical mixture. It is only expected to generalize to exactly one source within exactly one mixture. This gives our method the flexibility of separating a myriad of diverse source types–any source type that can be synthesized. Compare this to many existing methods, which are limited by whatever coarse set of source types might be available in a large enough numbers to train a model with (*e.g.,* "Vocals", "Bass", "Drums", & "Other").

Finally, because the network is designed to "overfit", we drastically reduce the its size and number of training steps, making the proposed method much faster. Our method is able to produce suitable results in *minutes*, instead of hours or days for typical source separation networks. This makes one-off bespoke training practical to use.

3. DEMOS

In this section we describe the process of creating two demo excerpts using bespoke networks. Both demo clips were created by training mask inference network with a 2 BLSTM layers with 300 units each and a dropout at 30% zeroing probability applied to the second layer. After the BLSTM layers, a fully connected layer with a sigmoid activation function estimated a mask. The whole network was trained to minimize the truncated Phase Sensitive Approximation (tPSA) [7] objective function.

Each training step, mixes were created randomly such that a gain for the background and synthesized target were chosen from a uniform distribution in the interval $[-12, 6]$ dB. The background and target were then mixed together. The details of the background audio differ for each demo, below. Dynamic range compression was applied to the mix with a reduction ratio chosen randomly from the set $\{2, 4, 8, 12, 16, 20\}$ and finally the training mixes were peak normalized. All models were trained for 2,000 steps and training times are reported using a Nvidia V100 GPU.

For these demos, we specifically chose to isolate sources that would be difficult or impossible to separate using existing deep learning source separation systems. Listen to the mixes, estimated sources, the synthesized source training data, and output from competing methods at our demo website: `https://git.io/JUu5q`

3.1 Take On Me – A-ha

The first example that we demonstrate is an excerpt from the 1984 hit "Take On Me" by A-ha. We chose to isolate the famous synthesizer melody at the beginning of the song. The rest of the mixture contains a synthesizer countermelody, a synthesizer "pad", a bass, and a drum track.

This example was created by synthesizing just one example of the target synth melody with a synthesizer that sounded loosely similar to the target source. We then created training mixes by summing the synthesized target together with the input mixture. Training took less than 5 minutes.

3.2 Reelin' in the Years – Steely Dan

The second example is an excerpt from an instrumental section of the 1972 song "Reelin' in the Years" by Steely Dan. During this excerpt, a piano, bass, drum set, a rhythm guitar, and two lead guitars two guitars are playing simultaneously. The two guitars are highly correlated, playing a harmonized melody line in diatonic thirds. We choose to isolate just *one* of the two guitars that was playing the lower part.

For this example, we found a MIDI transcription of the song online and synthesized all of the parts with 2 instrument patches for each instrument. We noticed that the transcription was incorrect in a few small but musical ways (i.e., one would not guess the MIDI was incorrect without studying the audio excerpt). To create the mixtures, we selected one patch for each instrument, and created mixes with the procedure above. Training took about 10 minutes.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced Bespoke neural networks and showed how they can be used for score-informed source separation. We demonstrated our method by showcasing two songs where we were able to quickly and effectively isolate sources that would have been neglected by other source separation systems.

In preparing these demos, we discovered that the source separation quality for a particular song is highly dependent on the augmentation strategies. We hope that future work on bespoke networks will address this problem to find a "one-size-fits-all" augmentation strategy and provide a scientific evaluation of the proposed method.

5. REFERENCES

[1] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[2] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some slakh: a dataset to study the impact of training data quality and quantity," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 45–49.

[3] M. Miron, J. Janer Mestres, and E. Gómez Gutiérrez, "Generating data to train convolutional neural networks for classical music source separation," in *Proceedings of the 14th Sound and Music Computing Conference*. Aalto University, 2017, pp. 227–33.

[4] S. Ewert, B. Pardo, M. Muller, and M. D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, 2014.

[5] M. Gover, "Score-informed source separation of choral music," Ph.D. dissertation, McGill University, 2019.

[6] E. Manilow, P. Seetharaman, and B. Pardo, "Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 771–775.

[7] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.