Transcription Is All You Need: Learning To Separate Musical Mixtures With Score As Supervision

Yun-Ning (Amy) Hung\(^1,2\), Gordon Wichern\(^1\), Jonathan Le Roux\(^1\)
\(^1\)Mitsubishi Electric Research Laboratories (MERL)
\(^2\)Center for Music Technology, Georgia Institute of Technology

IEEE ICASSP - June 2021
Music source separation

• Goal: isolate individual sources (e.g., instruments) from a music mixture
Existing systems

- Open-Unmix [1]
- Demucs [2]
- Conv-Tasnet [3]
- MMDenseLSTM [4]
- Spleeter [5]
- Dilated GRU [6]

Existing systems

• Open-Unmix [1]
• Demucs [2]
• Conv-Tasnet [3]
• MMDenseLSTM [4]
• Spleeter [5]
• Dilated GRU [6]

→ Supervised learning: need a dataset containing individual instrument tracks for training. This greatly limits the data that can be used for training.

What we propose

• Musical score is easier to obtain than separated tracks (e.g., Musescore [8] and Lakh MIDI dataset [7])

What we propose

• Musical score is easier to obtain than separated tracks (e.g., Lakh MIDI dataset [7], Musescore [8])

• Weakly supervised training: only a song and its (aligned) score needed for training

Previous work [9]

- Separate sounds based on sound activation labels
- Step 1: train a classifier to recognize sound events from a sound mixture
- Step 2: Fix the classifier, and use the classifier to guide the learning of the separator

Previous work [9]

- Separate sounds based on sound activation labels
- Step 1: train a classifier to recognize sound events from a sound mixture
- Step 2: Fix the classifier, and use the classifier to guide the learning of the separator

*Performance degrades on sound classes with complex and/or varied spectral structures

*Difficulty handling different sources that consistently appear together
Proposed system

• We propose a three-step training strategy to further improve weakly labeled music source separation
Step 1 – Transcriptor training

• Replace classifier with transcriptor
• Provides information in both time and frequency dimensions
• Transcriptor learns to transcribe the score of individual instruments from the music mixture
• We use the training strategy proposed in [10] to train the transcriptor

Step 2 – Separator training
Overview

- Separator should generate separated spectrogram for each instrument
Step 2 – Separator training

Overview

• Separator should generate separated spectrogram for each instrument
• Transcription loss: a pre-trained transcriptor acts as a critic that assesses whether the score transcribed from the separated spectrogram is close to the correct score
Step 2 – Separator training

Overview

- Separator should generate separated spectrogram for each instrument
- Transcription loss: a pre-trained transcriptor acts as a critics that assesses whether the score transcribed from the separated spectrogram is close to the correct score
- Mixture loss: separated spectrograms should sum to the mixture spectrogram
Step 2 – Separator training

Additional constraint on mixture loss

• Clip-level mask -> only activated instruments should count in mixture loss
Additional constraint on mixture loss

- Clip-level mask -> only activated instruments should count in mixture loss
- Harmonic mask -> only activated harmonic position should count in mixture loss. We assume most of the energy is in the harmonic frequencies
Harmonic mask

- Use score (fundamental frequency) to calculate harmonic mask
- Multiply with magnitude spectrogram
- Make the harmonics salient and suppress other frequencies
Step 3 – Fine-tuning

- **Overview**
- Load the pre-trained model in step 2 and fine-tune both transcriptor and separator together
Step 3 – Fine-tuning

• Overview
• Load the pre-trained model in step 2 and fine-tune both transcriptor and separator together
• Adversarial transcription loss (ATL): transcriptor attempts to detect notes from competing instruments in separated sources
Step 3 – Fine-tuning

- **Overview**
  - Load the pre-trained model in step 2 and fine-tune both transcriptor and separator together
  - Adversarial transcription loss (ATL): transcriptor attempts to detect notes from competing instruments in separated sources
  - Adversarial mixture loss (AML): transcriptor attempts to detect errors in synthetic mixtures composed of separated tracks

\[
\mathcal{L}_{AML}(\hat{X}, \hat{Y}) = -\mathcal{L}_{score}(\hat{X}, \hat{Y})
\]

\[X_a \xrightarrow{\text{Separator}} S_{a,1}, S_{a,2}, S_{b,3} \]
\[X_b \xrightarrow{\text{Separator}} S_{b,1}, S_{b,2}, S_{b,3} \]
\[X_c \xrightarrow{\text{Separator}} S_{c,1}, S_{c,2}, S_{c,3} \]

New mixture \( \hat{X} = S_{a,1} + S_{b,2} + S_{c,3} \)
Experiment

Training/Evaluation dataset

- Slakh dataset: synthetic dataset created from MIDI using professional-grade instruments
- Avoids mis-alignment between score and audio
- Choose most common three instruments: piano, distorted guitar and electric bass, for separation

Baseline system

- Proposed by Pishdadian et al. [9]

Evaluation metric

- Scale invariant signal to distortion ratio (SI-SDR)
## Separation Results

### Table 1. Separation performance (SI-SDR [dB])

<table>
<thead>
<tr>
<th>Training</th>
<th>$\mathcal{L}_{c\text{-mix}}$</th>
<th>$\mathcal{L}_{h\text{-mix}}$</th>
<th>$\mathcal{L}_{\text{AML}}$</th>
<th>$\mathcal{L}_{\text{ATL}}$</th>
<th>Bass</th>
<th>Guitar</th>
<th>Piano</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>11.1</td>
<td>5.7</td>
<td>7.7</td>
<td>8.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>isolated</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>7.5</td>
<td>1.2</td>
<td>4.2</td>
<td>4.3</td>
</tr>
<tr>
<td>isolated</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>7.8</td>
<td>0.4</td>
<td>4.1</td>
<td>4.1</td>
</tr>
<tr>
<td>isolated</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>8.4</td>
<td>1.6</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>9.0</td>
<td>2.7</td>
<td>5.3</td>
<td>5.6</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>9.1</td>
<td>2.8</td>
<td>5.4</td>
<td>5.8</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>9.0</td>
<td>2.5</td>
<td>5.7</td>
<td>5.7</td>
</tr>
<tr>
<td>Input mixture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.2</td>
<td>-5.8</td>
<td>-2.3</td>
<td>-2.3</td>
</tr>
<tr>
<td>Baseline [16]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.3</td>
<td>0.5</td>
<td>3.5</td>
<td>3.8</td>
</tr>
</tbody>
</table>
Separation Results

<table>
<thead>
<tr>
<th>Training</th>
<th>$\mathcal{L}_{c\text{-}mix}$</th>
<th>$\mathcal{L}_{h\text{-}mix}$</th>
<th>$\mathcal{L}_{AML}$</th>
<th>$\mathcal{L}_{ATL}$</th>
<th>Bass</th>
<th>Guitar</th>
<th>Piano</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.1</td>
<td>5.7</td>
<td>7.7</td>
<td>8.2</td>
</tr>
<tr>
<td>isolated</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>7.5</td>
<td>1.2</td>
<td>4.2</td>
<td>4.3</td>
</tr>
<tr>
<td>isolated</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>7.8</td>
<td>0.4</td>
<td>4.1</td>
<td>4.1</td>
</tr>
<tr>
<td>isolated</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>8.4</td>
<td>1.6</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>9.0</td>
<td>2.7</td>
<td>5.3</td>
<td>5.6</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>9.1</td>
<td>2.8</td>
<td>5.4</td>
<td>5.8</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>9.0</td>
<td>2.5</td>
<td>5.7</td>
<td>5.7</td>
</tr>
<tr>
<td>Input mixture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.2</td>
<td>-5.8</td>
<td>-2.3</td>
<td>-2.3</td>
</tr>
<tr>
<td>Baseline [16]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.3</td>
<td>0.5</td>
<td>3.5</td>
<td>3.8</td>
</tr>
</tbody>
</table>

- Our proposed system, using transcriptor, out-performs baseline system, using classifier
## Separation Results

### Table 1. Separation performance (SI-SDR[dB])

<table>
<thead>
<tr>
<th>Training</th>
<th>$\mathcal{L}_{c\text{-mix}}$</th>
<th>$\mathcal{L}_{h\text{-mix}}$</th>
<th>$\mathcal{L}_{\text{AML}}$</th>
<th>$\mathcal{L}_{\text{ATL}}$</th>
<th>Bass</th>
<th>Guitar</th>
<th>Piano</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.1</td>
<td>5.7</td>
<td>7.7</td>
<td>8.2</td>
</tr>
<tr>
<td>isolated</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>7.5</td>
<td>1.2</td>
<td>4.2</td>
<td>4.3</td>
</tr>
<tr>
<td>isolated</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>7.8</td>
<td>0.4</td>
<td>4.1</td>
<td>4.1</td>
</tr>
<tr>
<td>isolated</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>8.4</td>
<td>1.6</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>fine-tune</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>9.0</td>
<td>2.7</td>
<td>5.3</td>
<td>5.6</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>9.1</td>
<td>2.8</td>
<td>5.4</td>
<td>5.8</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>9.0</td>
<td>2.5</td>
<td>5.7</td>
<td>5.7</td>
</tr>
<tr>
<td>Input mixture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.2</td>
<td>-5.8</td>
<td>-2.3</td>
<td>-2.3</td>
</tr>
<tr>
<td>Baseline [16]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.3</td>
<td>0.5</td>
<td>3.5</td>
<td>3.8</td>
</tr>
</tbody>
</table>

- Our proposed system, using transcriptor, out-performs baseline system, using classifier
- Using masking constraint can further improve the separation
Separation Results

Table 1. Separation performance (SI-SDR[dB])

<table>
<thead>
<tr>
<th>Training</th>
<th>$\mathcal{L}_{c\text{-mix}}$</th>
<th>$\mathcal{L}_{h\text{-mix}}$</th>
<th>$\mathcal{L}_{\text{AML}}$</th>
<th>$\mathcal{L}_{\text{ATL}}$</th>
<th>Bass</th>
<th>Guitar</th>
<th>Piano</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.1</td>
<td>5.7</td>
<td>7.7</td>
<td>8.2</td>
</tr>
<tr>
<td>isolated</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td>7.5</td>
<td>1.2</td>
<td>4.2</td>
<td>4.3</td>
</tr>
<tr>
<td>isolated</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>7.8</td>
<td>0.4</td>
<td>4.1</td>
<td>4.1</td>
</tr>
<tr>
<td>isolated</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
<td>8.4</td>
<td>1.6</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>9.0</td>
<td>2.7</td>
<td>5.3</td>
<td>5.6</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>9.1</td>
<td>2.8</td>
<td>5.4</td>
<td>5.8</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>9.0</td>
<td>2.5</td>
<td>5.7</td>
<td>5.7</td>
</tr>
<tr>
<td>Input mixture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.2</td>
<td>-5.8</td>
<td>-2.3</td>
<td>-2.3</td>
</tr>
<tr>
<td>Baseline [16]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.3</td>
<td>0.5</td>
<td>3.5</td>
<td>3.8</td>
</tr>
</tbody>
</table>

- Our proposed system, using transcriptor, out-performs baseline system, using classifier
- Using masking constraint can further improve the separation
- Fine-tuning transcriptor and separator can further improve separation result
Separation Results

Table 1. Separation performance (SI-SDR[dB])

<table>
<thead>
<tr>
<th>Training</th>
<th>$\mathcal{L}_{c\text{-mix}}$</th>
<th>$\mathcal{L}_{h\text{-mix}}$</th>
<th>$\mathcal{L}_{\text{AML}}$</th>
<th>$\mathcal{L}_{\text{ATL}}$</th>
<th>Bass</th>
<th>Guitar</th>
<th>Piano</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.1</td>
<td>5.7</td>
<td>7.7</td>
<td>8.2</td>
</tr>
<tr>
<td>isolated</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>7.5</td>
<td>1.2</td>
<td>4.2</td>
<td>4.3</td>
</tr>
<tr>
<td>isolated</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>7.8</td>
<td>0.4</td>
<td>4.1</td>
<td>4.1</td>
</tr>
<tr>
<td>isolated</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>8.4</td>
<td>1.6</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>9.0</td>
<td>2.7</td>
<td>5.3</td>
<td>5.6</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>9.1</td>
<td>2.8</td>
<td>5.4</td>
<td>5.8</td>
</tr>
<tr>
<td>fine-tune</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>9.0</td>
<td>2.5</td>
<td>5.7</td>
<td>5.7</td>
</tr>
<tr>
<td>Input mixture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.2</td>
<td>-5.8</td>
<td>-2.3</td>
<td>-2.3</td>
</tr>
<tr>
<td>Baseline [16]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.3</td>
<td>0.5</td>
<td>3.5</td>
<td>3.8</td>
</tr>
</tbody>
</table>

- Our proposed system, using transcriptor, out-performs baseline system, using classifier
- Using masking constraint can further improve the separation
- Fine-tuning transcriptor and separator can further improve separation result
- Compared to baseline system, we close a significant gap from the mixture SI-SDR to the supervised setting
Conclusion / takeaway

- We proposed a method to train a music source separation system based on musical score only, without any supervision from isolated tracks.

- We proposed a masking strategy and an adversarial fine-tuning strategy to further improve the system.
Future work

• Semi-supervised learning: combine our proposed training strategy with supervised learning
• Expand to vocals and drums
• Integrate with audio to score alignment algorithms
• Experiments on real-world data
Listening demo!
Thank you!

Paper id 2698