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)

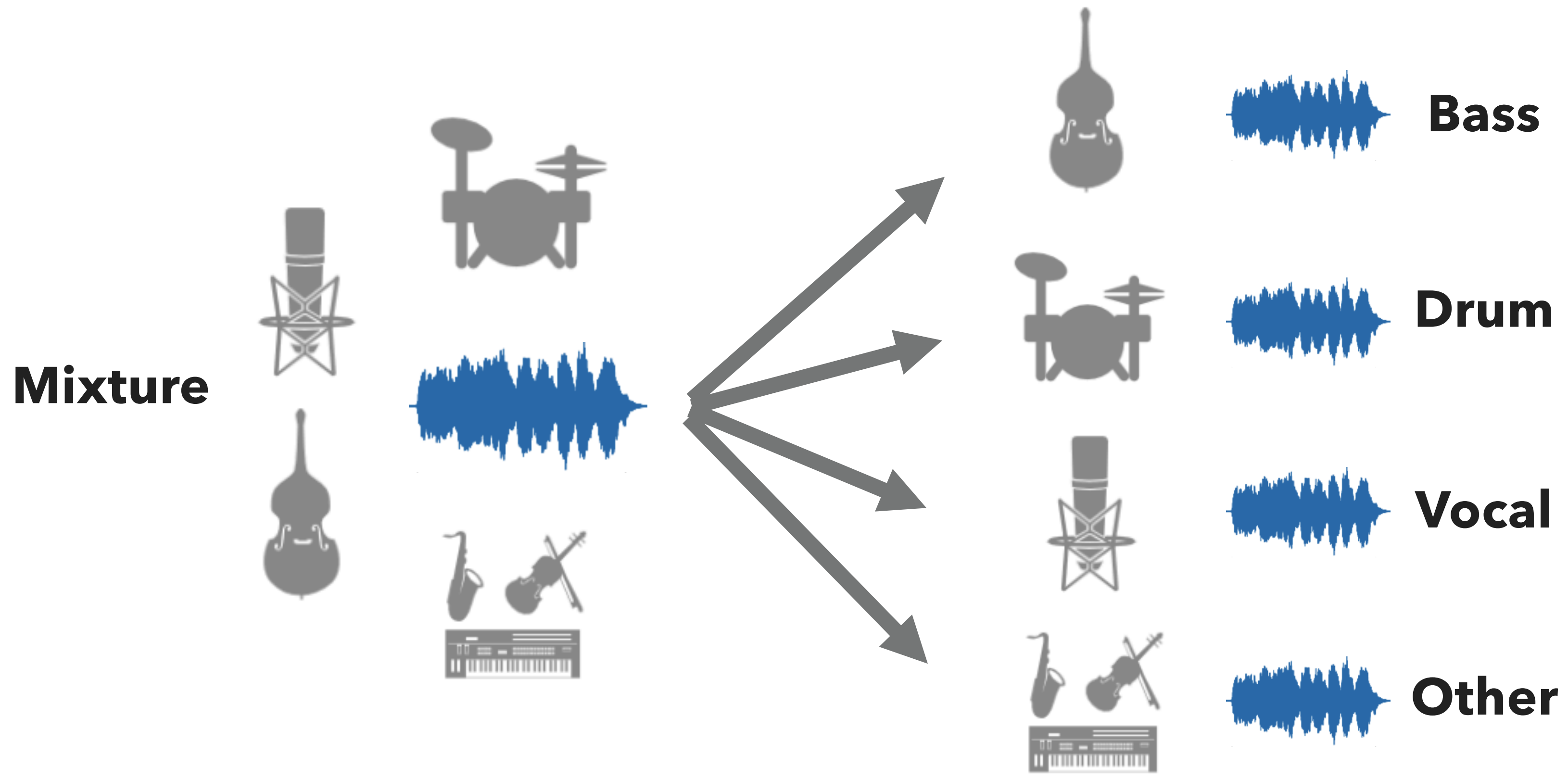
[2]Center for Music Technology, Georgia Institute of Technology

IEEE ICASSP - June 2021

MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

Music source separation

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



MITSUBISHI ELECTRIC
Changes for the Better

paper id: 2698

Existing systems

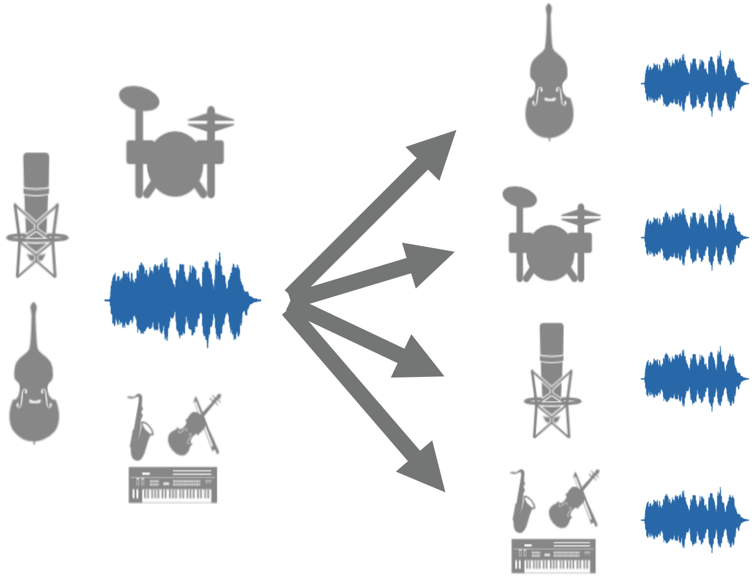
- Open-Unmix [1]
- Demucs [2]
- Conv-Tasnet [3]
- MMDenseLSTM [4]
- Spleeter [5]
- Dilated GRU [6]

[1] F-R. Stöter et al. "Open-unmix-a reference implementation for music source separation," 2019.
[2] A. Défossez et al. "Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed," arXiv:1909.01174, 2019.
[3] Y. Luo et al. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM TASLP 27.8, 2019.
[4] N. Takahashi et al. "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," IEEE IWAENC, 2018.
[5] R. Hennequin et al. "Spleeter: A fast and state-of-the art music source separation tool with pre-trained models," ISMIR, 2019.
[6] J-Y. Liu et al. "Dilated convolution with dilated GRU for music source separation," IJCAI, 2019.

paper id: 2698

Existing systems

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

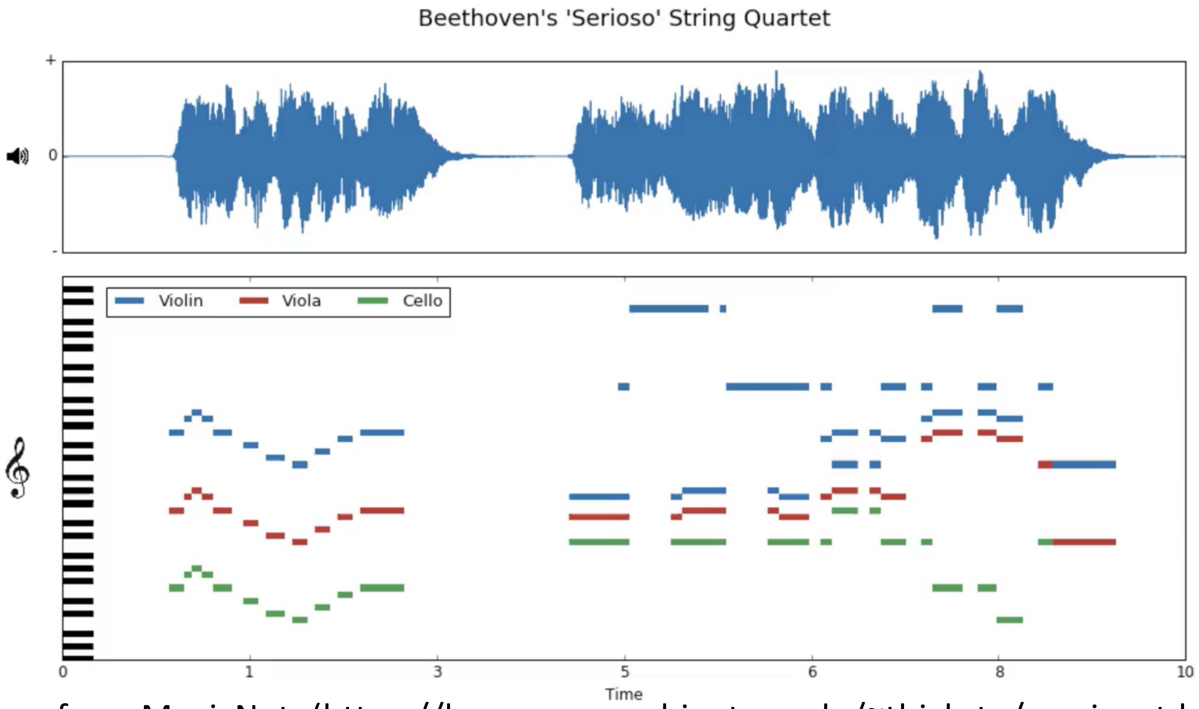


→ Supervised learning: need a dataset containing individual instrument tracks for training.
This greatly limits the data that can be used for training.

[1] F-R. Stöter et al. "Open-unmix-a reference implementation for music source separation," 2019.
[2] A. Défossez et al. "Demucs: Deep Extractor for Music Sources with extra unlabeled data remixed," arXiv:1909.01174, 2019.
[3] Y. Luo et al. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," IEEE/ACM TASLP 27.8, 2019.
[4] N. Takahashi et al. "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," IEEE IWAENC, 2018.
[5] R. Hennequin et al. "Spleeter: A fast and state-of-the art music source separation tool with pre-trained models," ISMIR, 2019.
[6] J-Y. Liu et al. "Dilated convolution with dilated GRU for music source separation," IJCAI, 2019.

What we propose

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

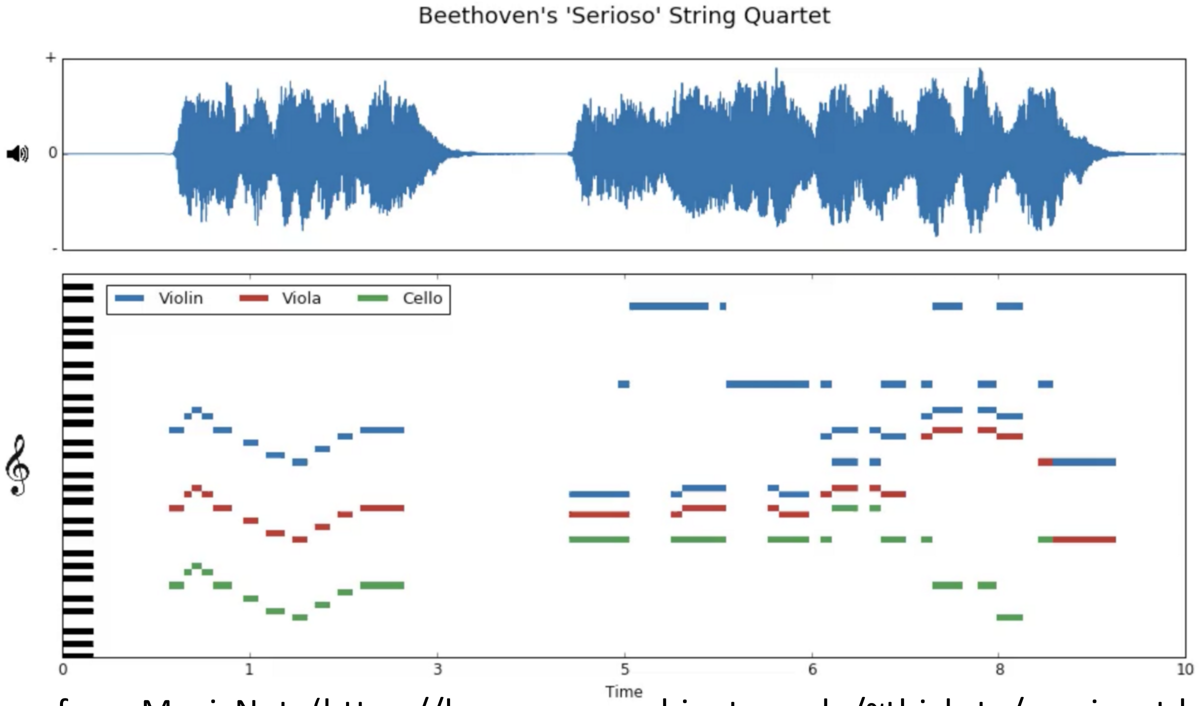


Picture from MusicNet: (https://homes.cs.washington.edu/~thickstn/musicnet.html)

[7] E. Manilow et al. "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," IEEE *WASPAA*, 2019.
[8] https://musescore.com/dashboard

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

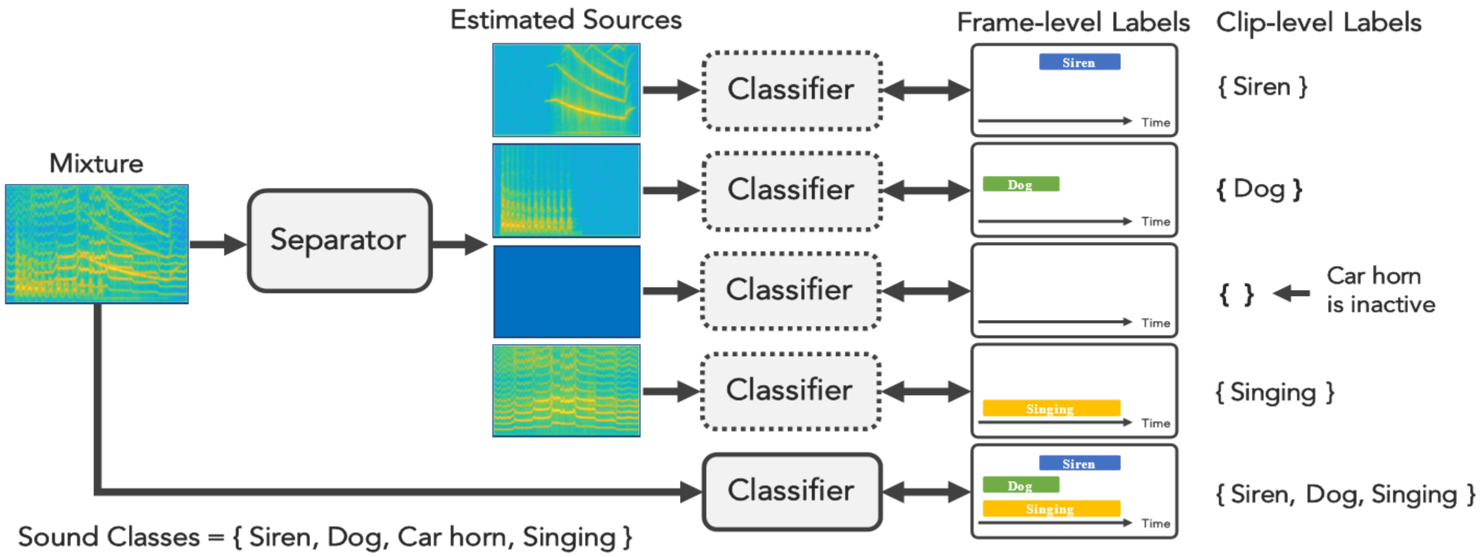


Picture from MusicNet: (https://homes.cs.washington.edu/~thickstn/musicnet.html)

[7] Manilow, Ethan, et al. "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity." IEEE *WASPAA*, 2019.
[8] https://musescore.com/dashboard

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



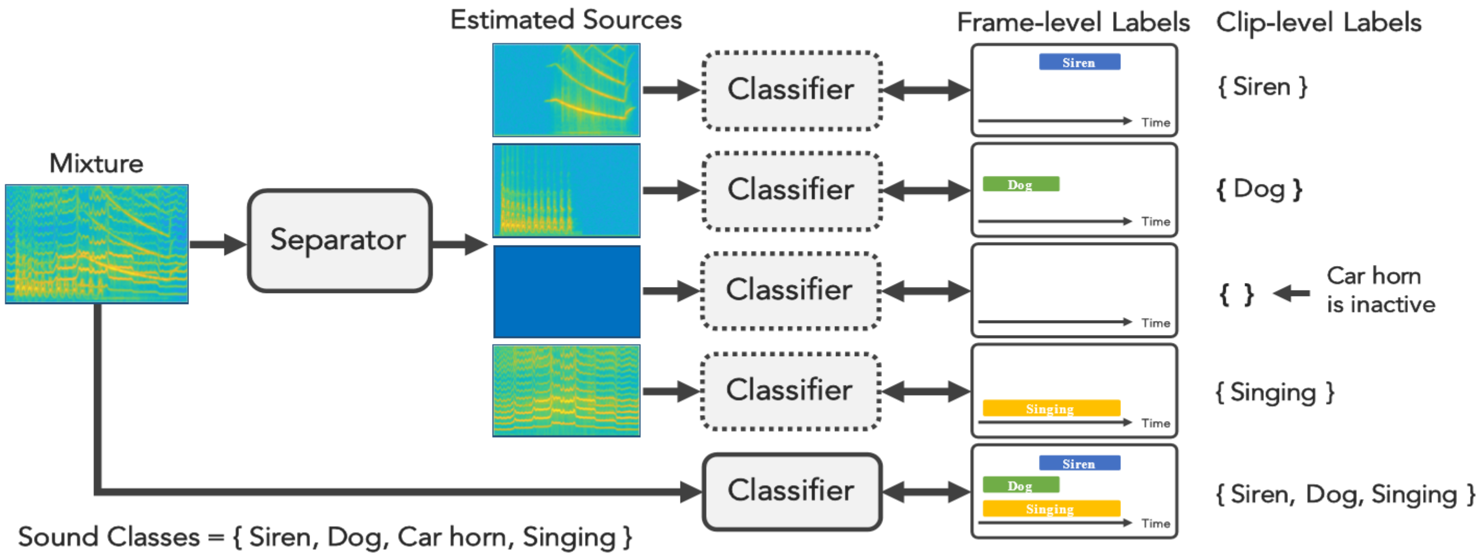
[9] F. Pishdadian, G. Wichern, J. Le Roux. "Finding strength in weakness: Learning to separate sounds with weak supervision." IEEE/ACM Trans. Audio, Speech, and Language Processing (2020).

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

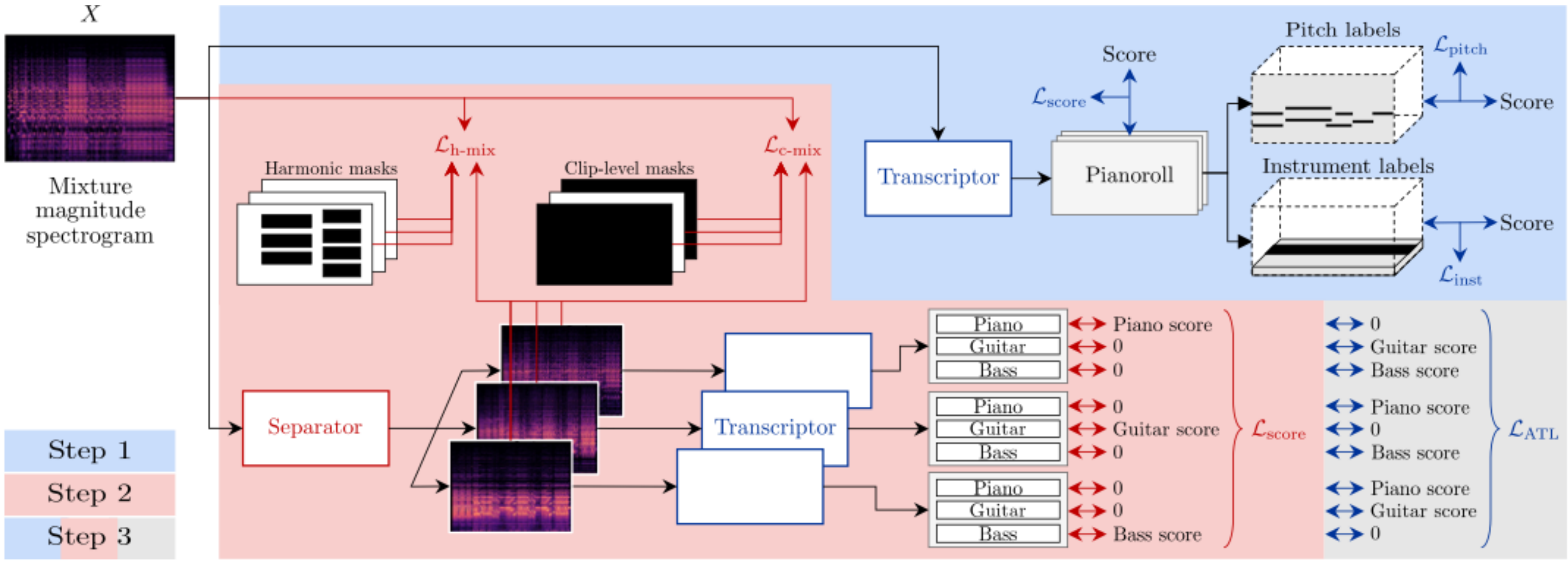
*Performance degrades on sound classes with complex and/or varied spectral structures

*Difficulty handling different sources that consistently appear together



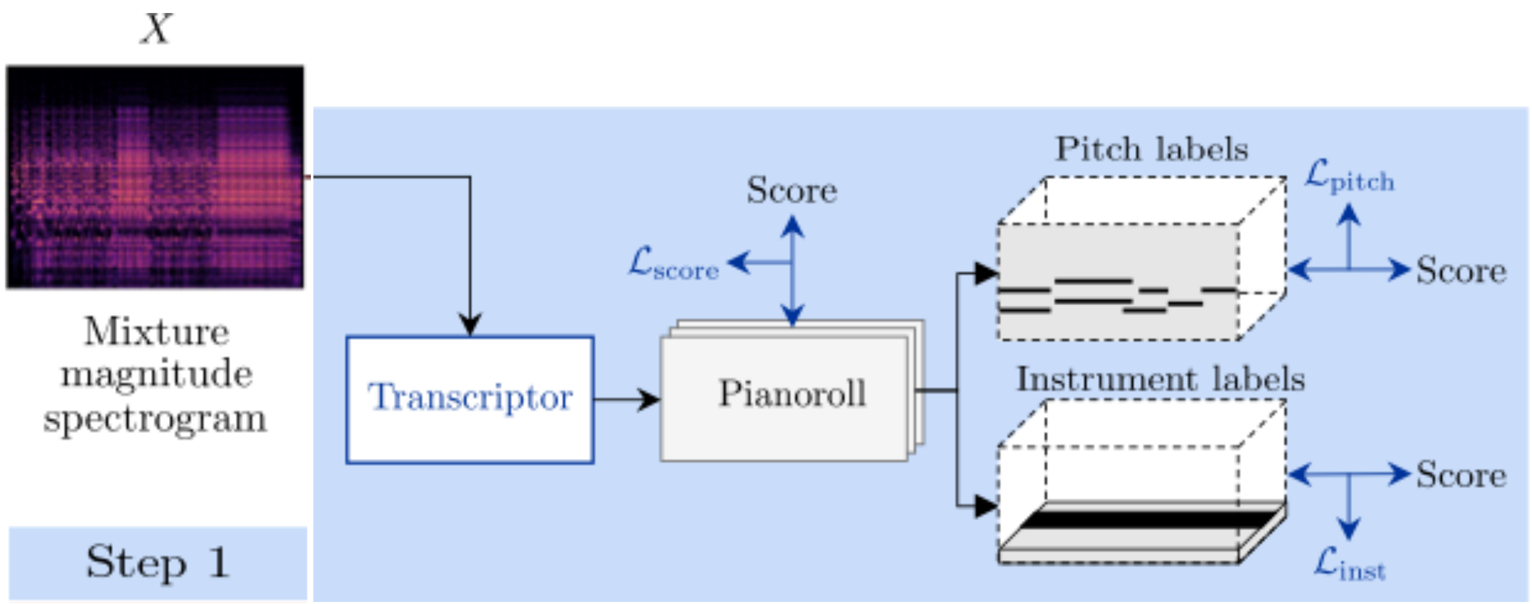
Proposed system

- We propose a three-step training strategy to further improve weakly labeled music source separation



Step 1 – Transcriptor training

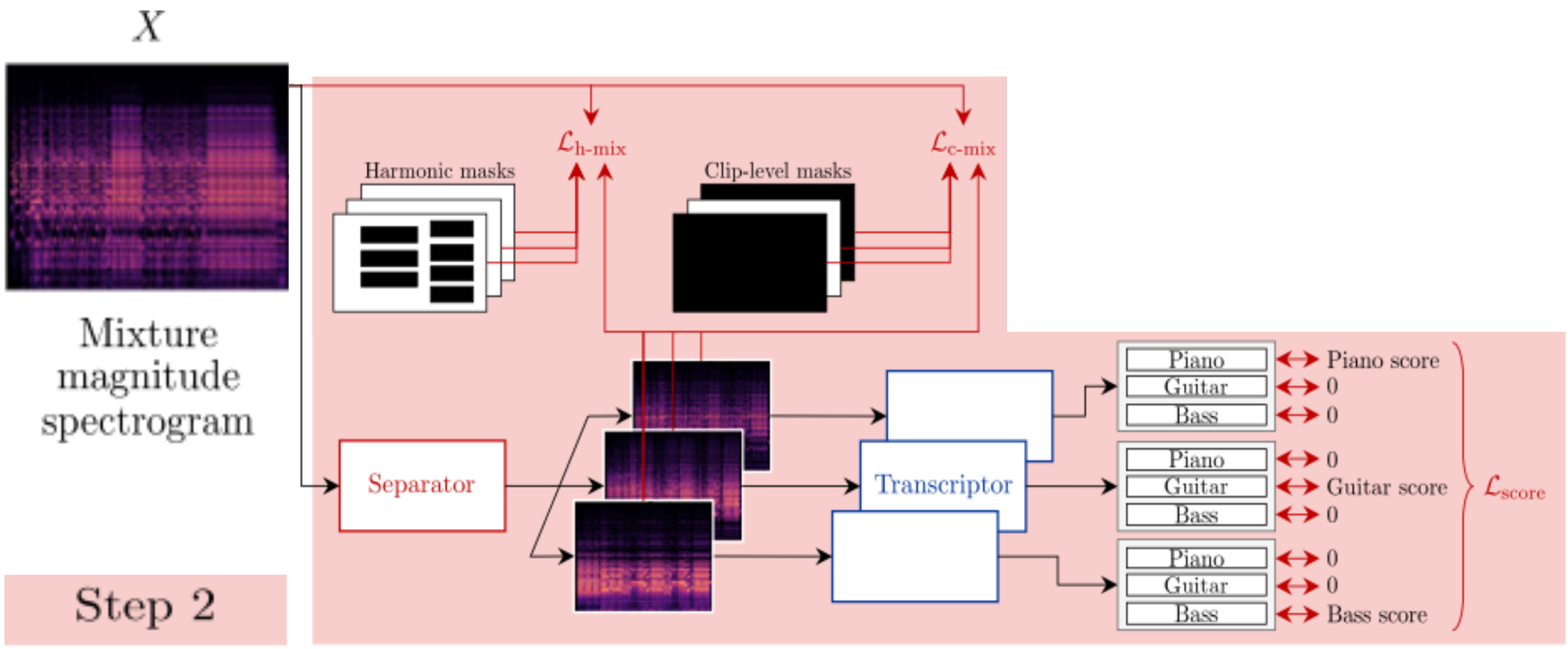
- Replace classifier with transcriptor
- Provides information in both time and frequency dimensions
- Transcriptor learns to transcribe the score of individual instruments from the music mixture
- We use the training strategy proposed in [10] to train the transcriptor



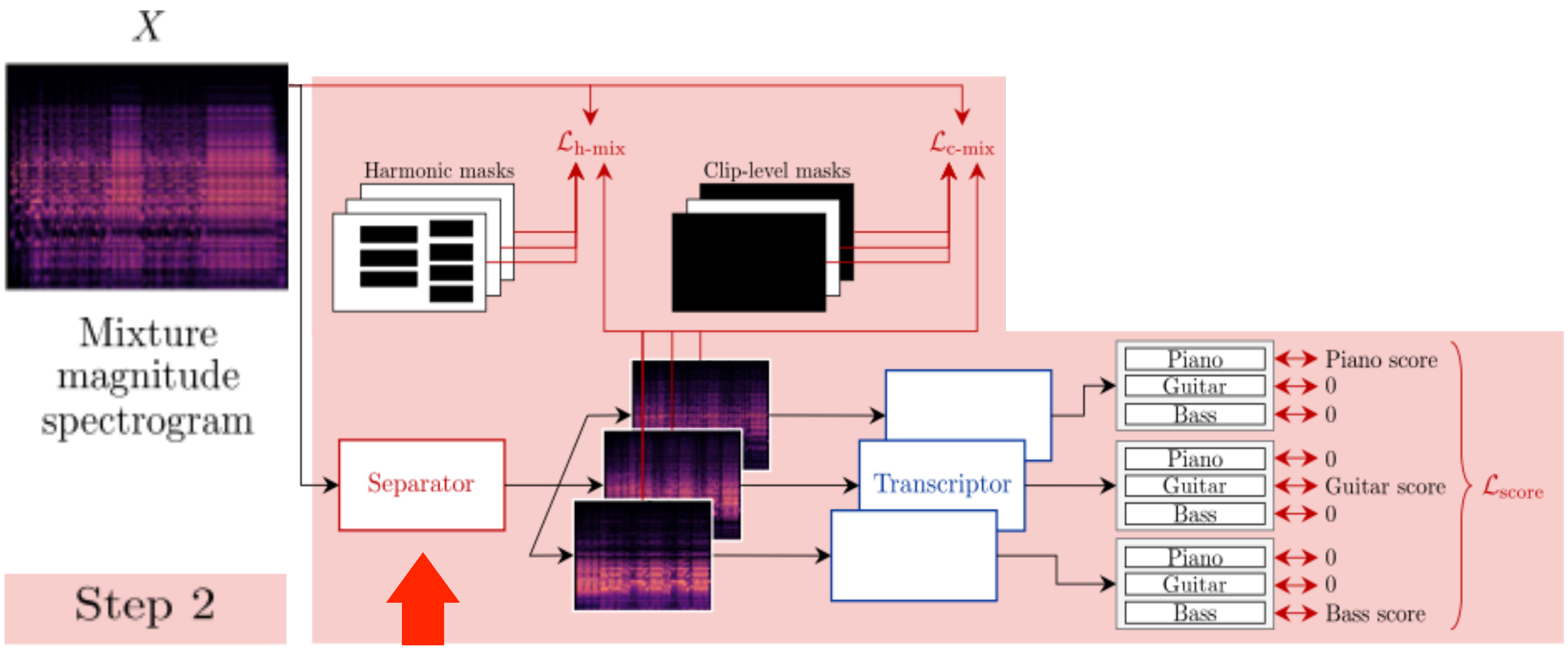
[10] Y-N. Hung et al. "Multitask learning for frame-level instrument recognition," IEEE *ICASSP*, 2019.

paper id: 2698

Step 2 – Separator training



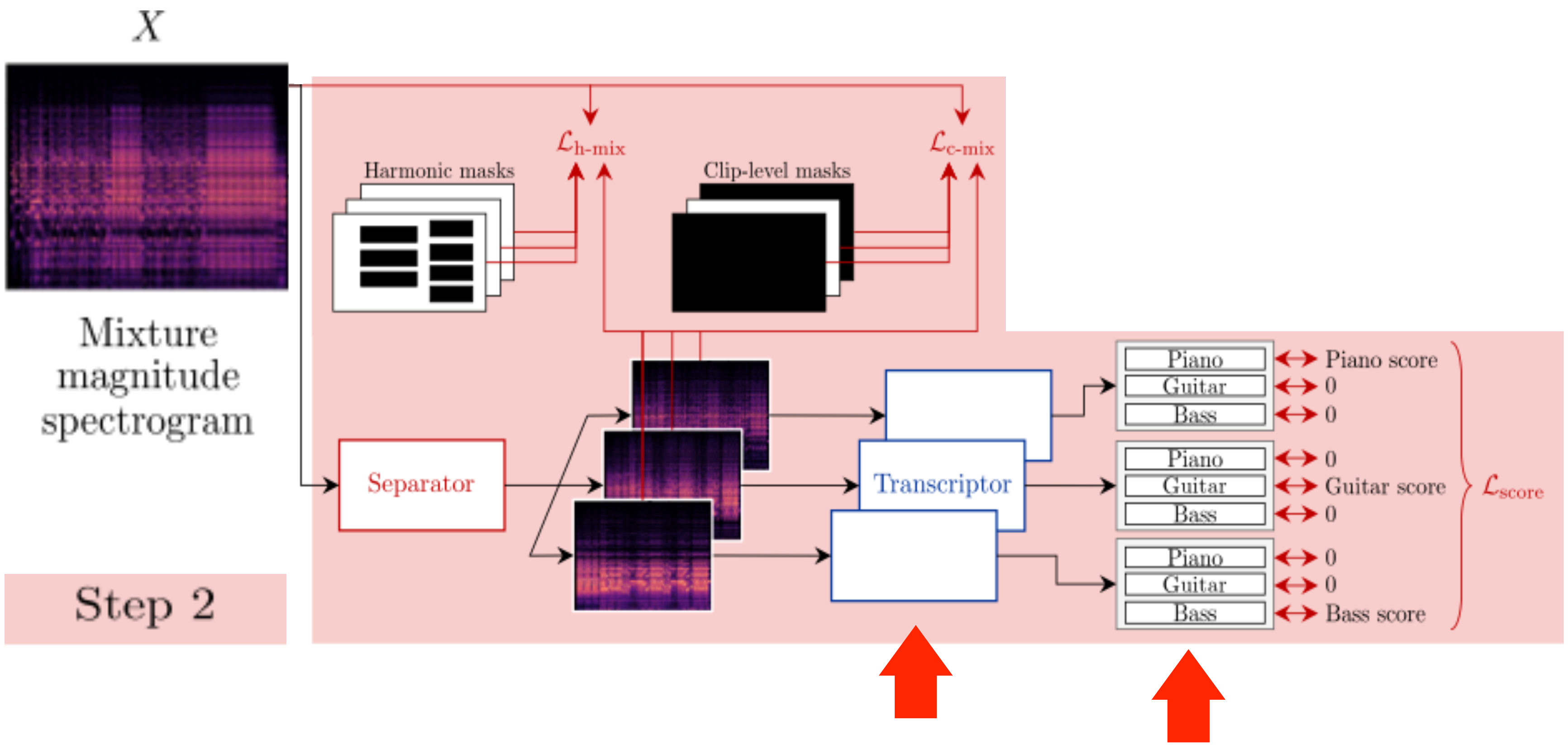
Step 2 – Separator training



Overview

- Separator should generate separated spectrogram for each instrument

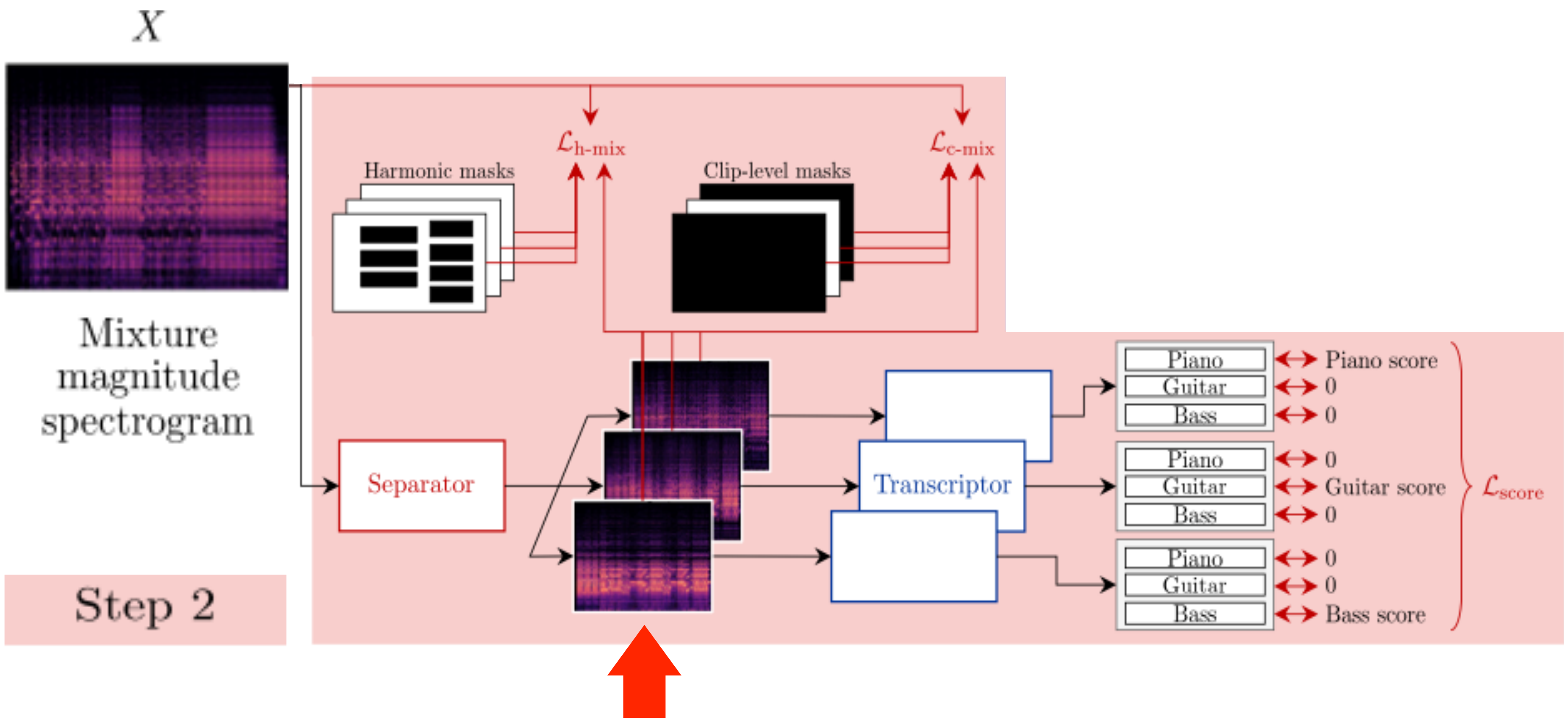
Step 2 – Separator training



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

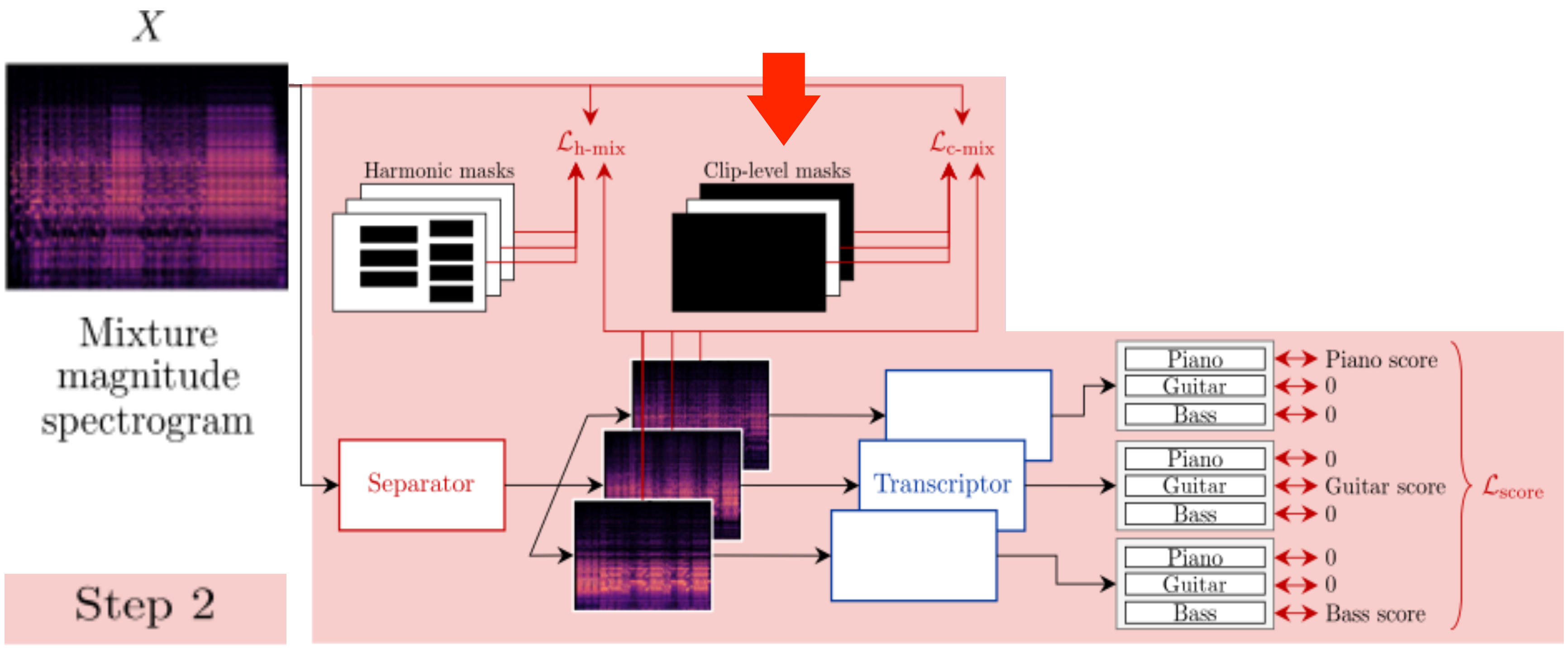
Step 2 – Separator training



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

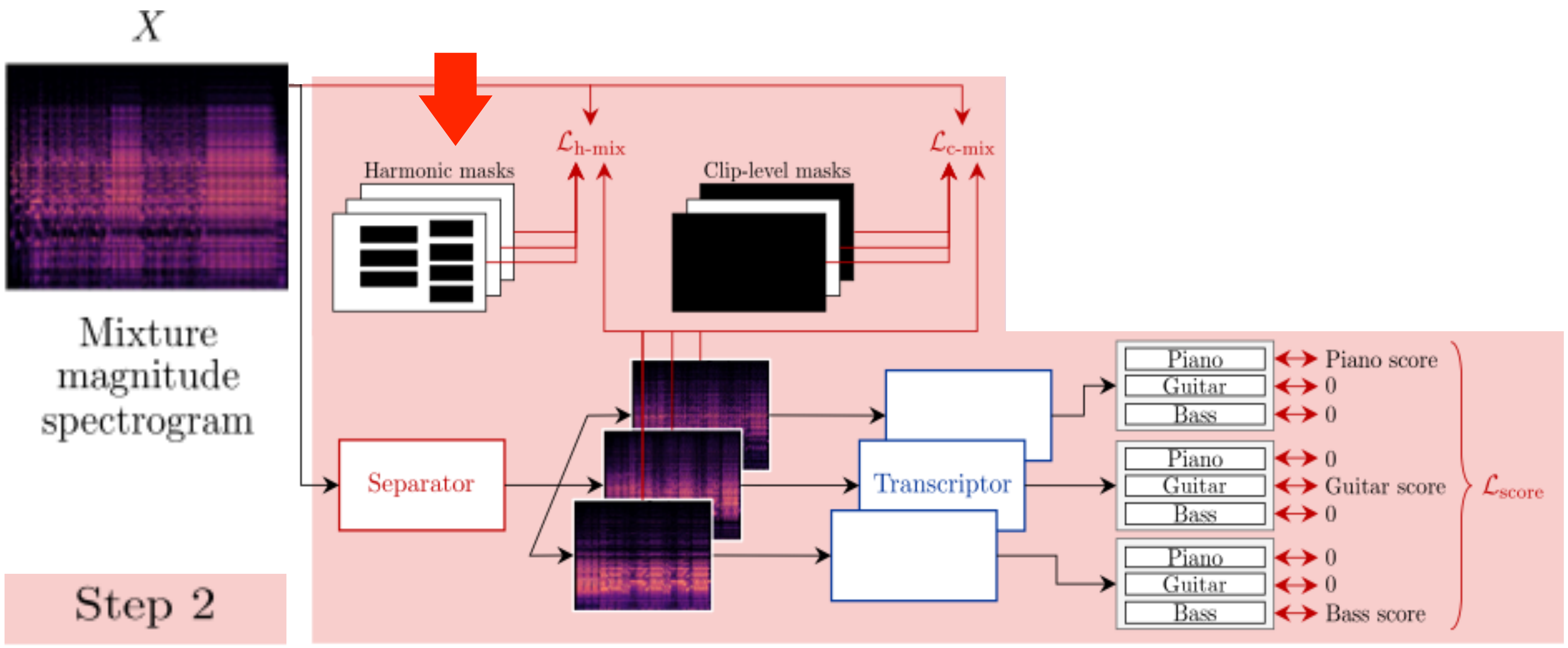
Step 2 – Separator training



Additional constraint on mixture loss

- Clip-level mask -> only activated instruments should count in mixture loss

Step 2 – Separator training

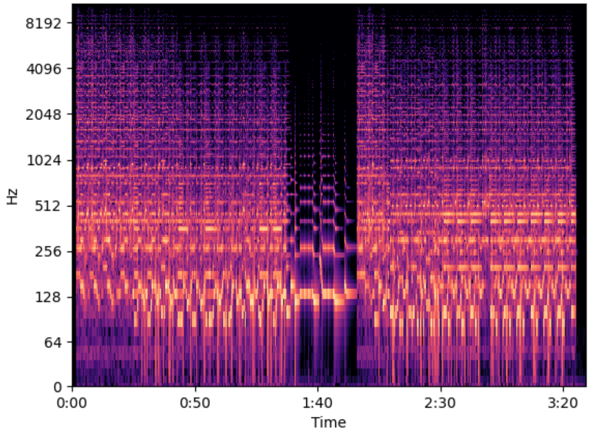


Additional constraint on mixture loss

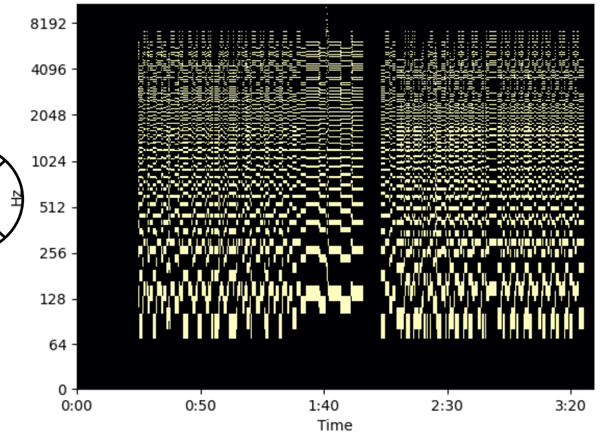
- Clip-level mask -> only activated instruments should count in mixture loss
- Harmonic mask -> only activated harmonic position should count in mixture loss. We assume most of the energy is in the harmonic frequencies

Harmonic mask

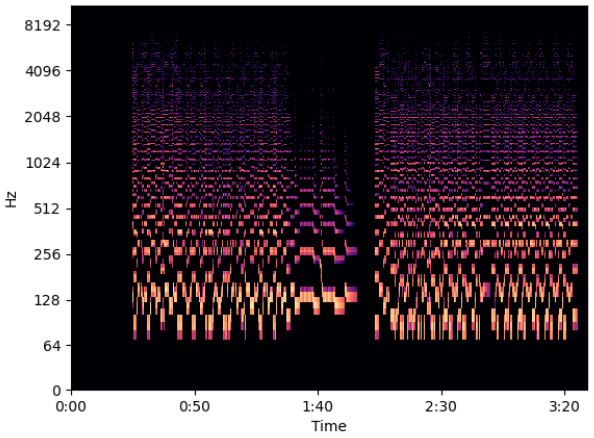
Magnitude spectrogram



Harmonic mask

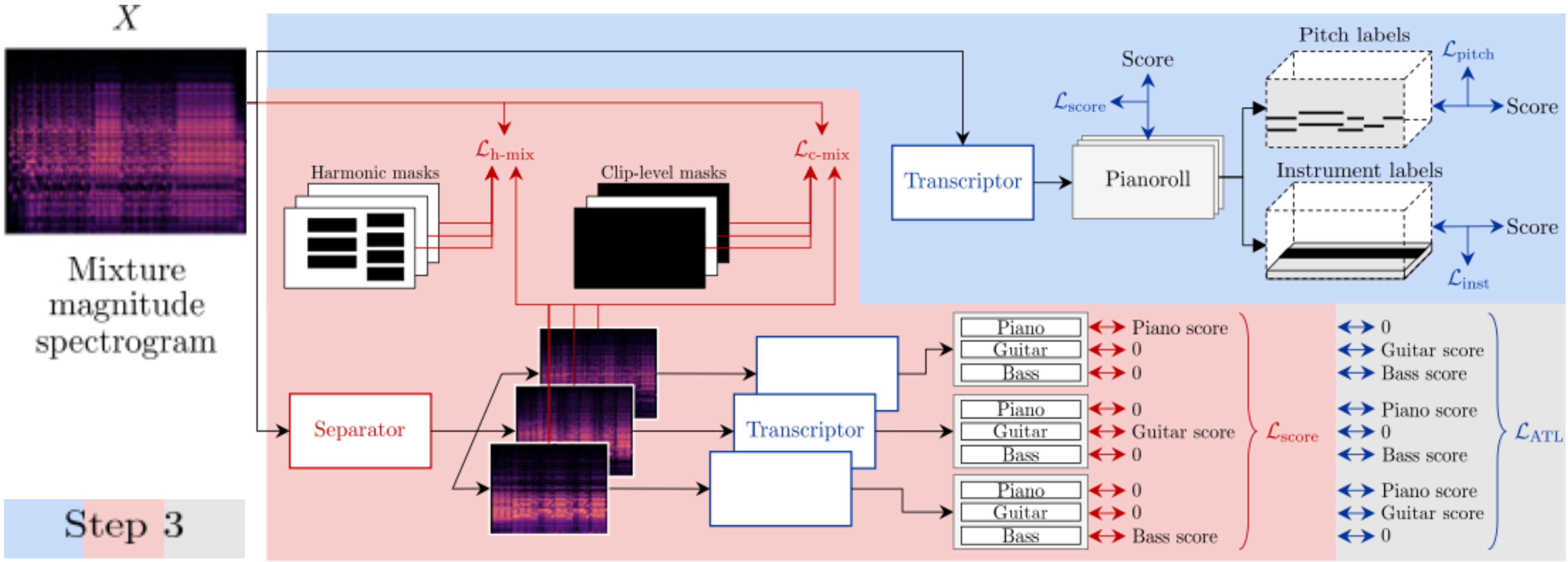


Filtered spectrogram



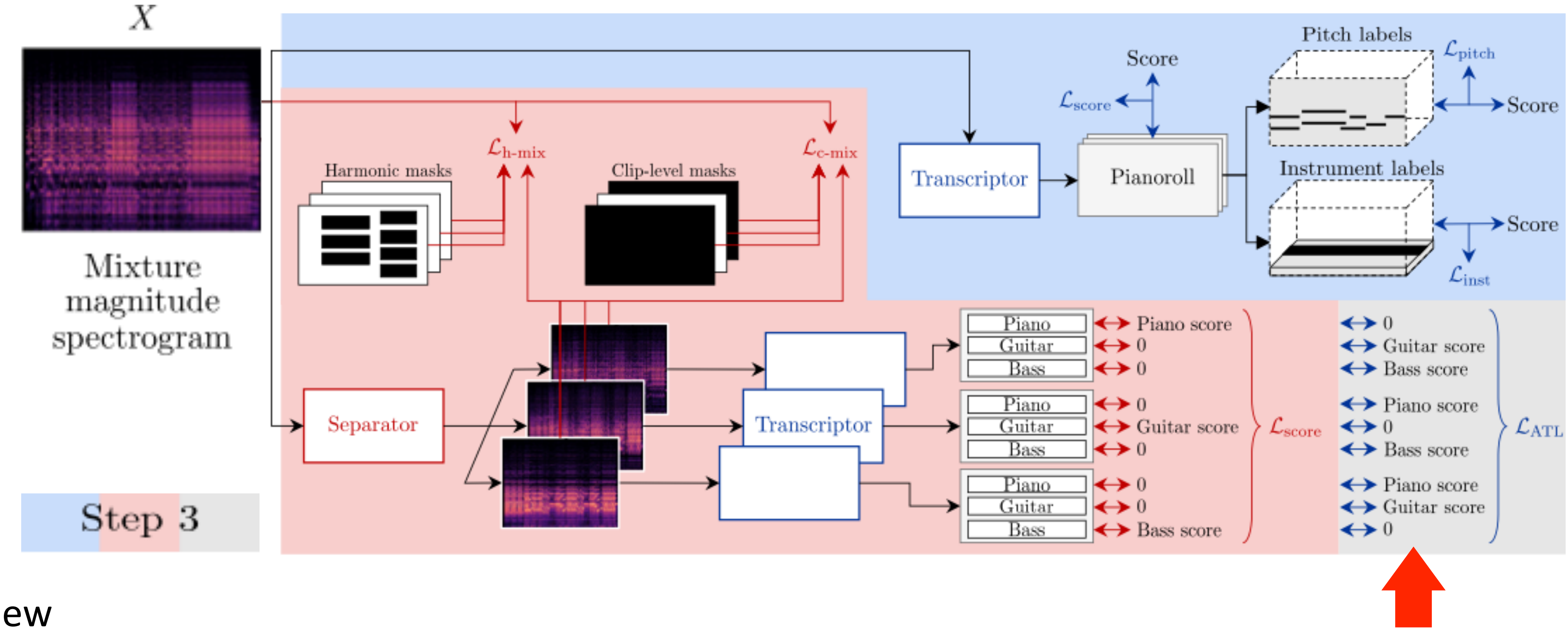
- Use score (fundamental frequency) to calculate harmonic mask
- Multiply with magnitude spectrogram
- Make the harmonics salient and suppress other frequencies

Step 3 – Fine-tuning



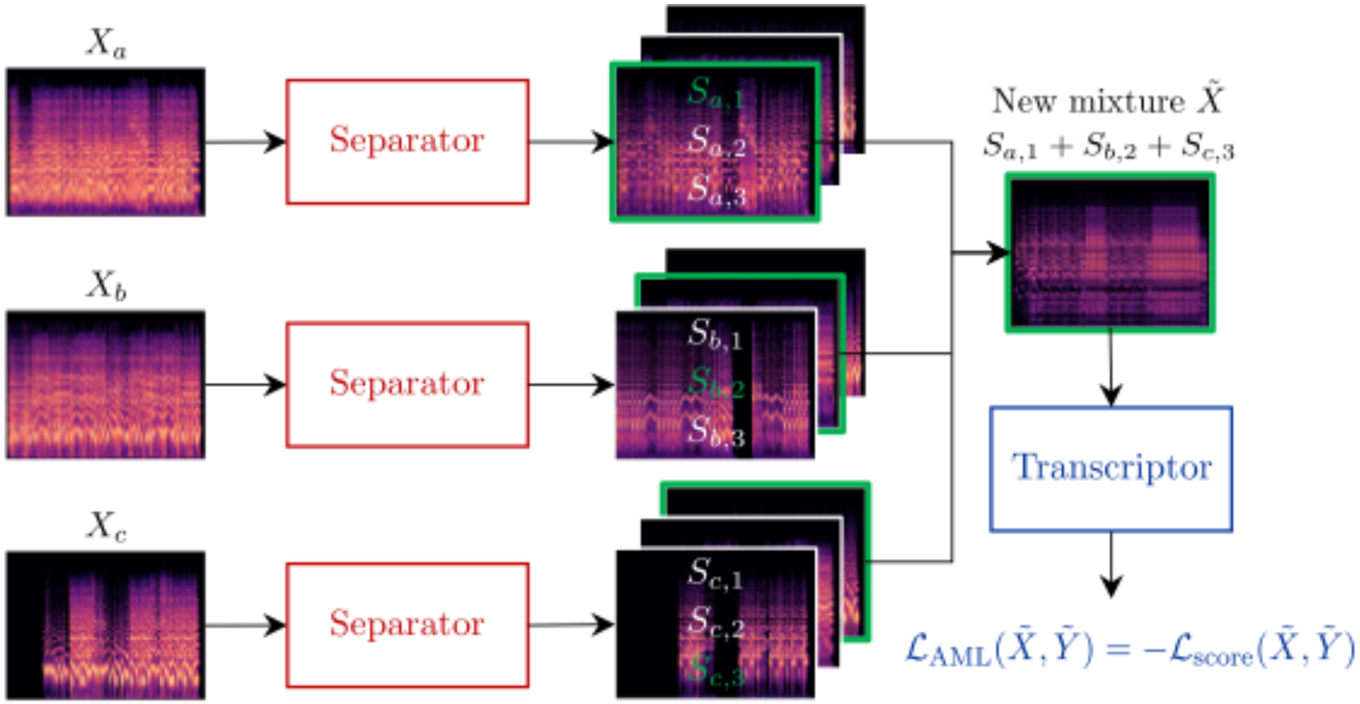
- Overview
- Load the pre-trained model in step 2 and fine-tune both transcriptor and separator together

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

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 matric

- Scale invariant signal to distortion ratio (SI-SDR)

paper id: 2698

Separation Results

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

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

paper id: 2698

Separation Results

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

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

- Our proposed system, using transcriptor, out-performs baseline system, using classifier

paper id: 2698

Separation Results

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

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

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

paper id: 2698

Separation Results

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

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

- 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])

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

- 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

paper id: 2698

Listening demo!

Thank you!

Paper id 2698

MITSUBISHI ELECTRIC

Changes for the Better