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

Yun-Ning Hung[1,2], Gordon Wichern[1], Jonathan Le Roux[1]
[1]Mitsubishi Electric Research Labs (MERL) [2]Georgia Tech

Introduction / Takeaway

Problem in current music source separation systems:

- Rely on separated stems for supervised training
- Lots of available songs do not have separated stems but have musical scores

Our solution

- Use a three-steps method to train source separation without signal ground truth
- Rely on weak labels (scores) to train music separation system

Experiments & Results

- Train and evaluate on Slakh dataset [1] for separation of three instruments (bass, guitar, and piano)
- Our proposed system outperforms baseline system [2]

Results

Table 1. Separation performance (note accuracy)

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

Table 2. Transcription performance (SI-SDR)

Training	Evaluated on	$\mathcal{L}_{\text{AML}}$	$\mathcal{L}_{\text{ATL}}$	Bass	Guitar	Piano
pre-train	mixture			0.85	0.44	0.58
fine-tune	mixture			0.84	0.42	0.54
fine-tune	mixture	✓		0.86	0.51	0.61
fine-tune	mixture		✓	0.85	0.50	0.60
pre-train	iso tracks			0.91	0.52	0.66
fine-tune	iso tracks			0.90	0.53	0.63
fine-tune	iso tracks	✓		0.91	0.58	0.68
fine-tune	iso tracks		✓	0.91	0.57	0.66

- Our proposed approach (using transcriptor) outperforms baseline system [2] (using classifier)
- Additional masking constraint can improve separation
- Adversarial fine-tuning improves both separation and transcription
- Compared to baseline system, we close a significant gap from the mixture SI-SDR to the supervised setting

Proposed training method

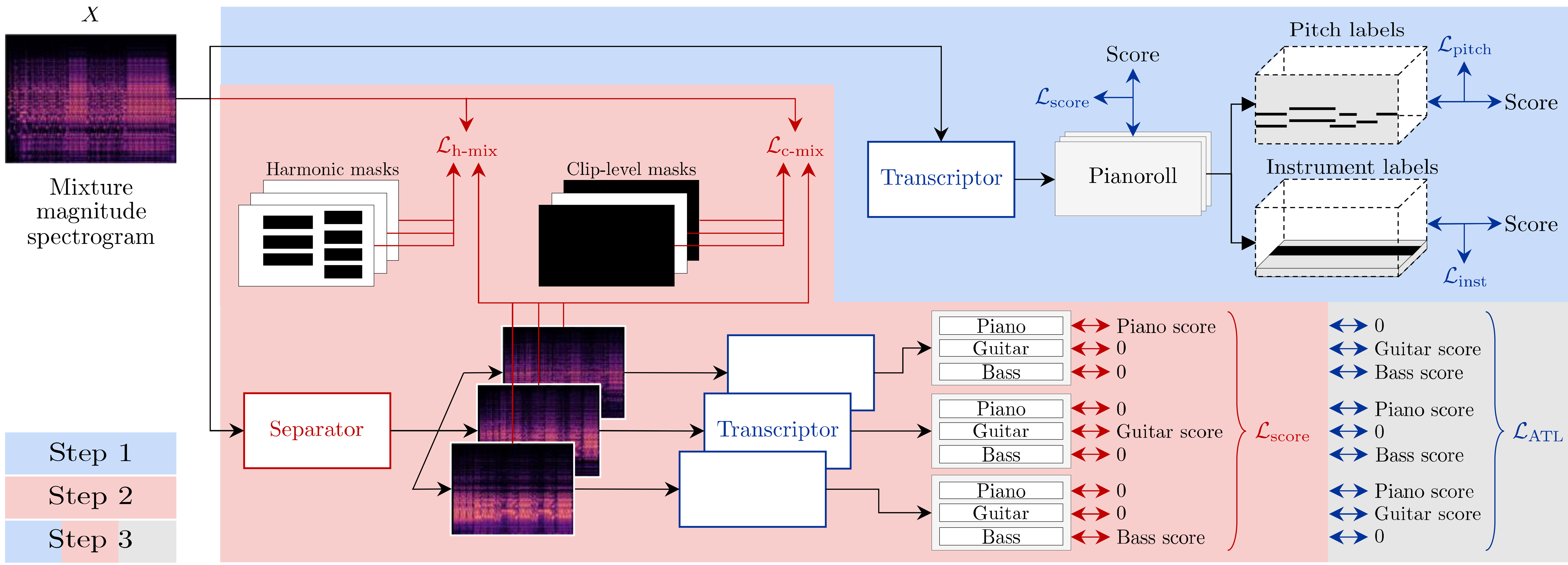

Fig 1. Diagram of our proposed training strategy

Step 1: Train a transcriptor (blue part in Fig. 1)

- Transcriptor learns to transcribe the score of individual instruments from a music mixture

Step 2: Train a separator (red part in Fig. 1)

- Separator should generate separated spectrogram for each instrument
- A pre-trained fixed transcriptor acts as a critic: transcribe the separated spectrogram into score that should be close to correct one
- Mixture loss: separated spectrograms should sum to the mixture spectrogram
- Clip-level mask: only active instruments should be used in mixture loss
- Harmonic mask: only harmonic frequencies should be used in mixture loss

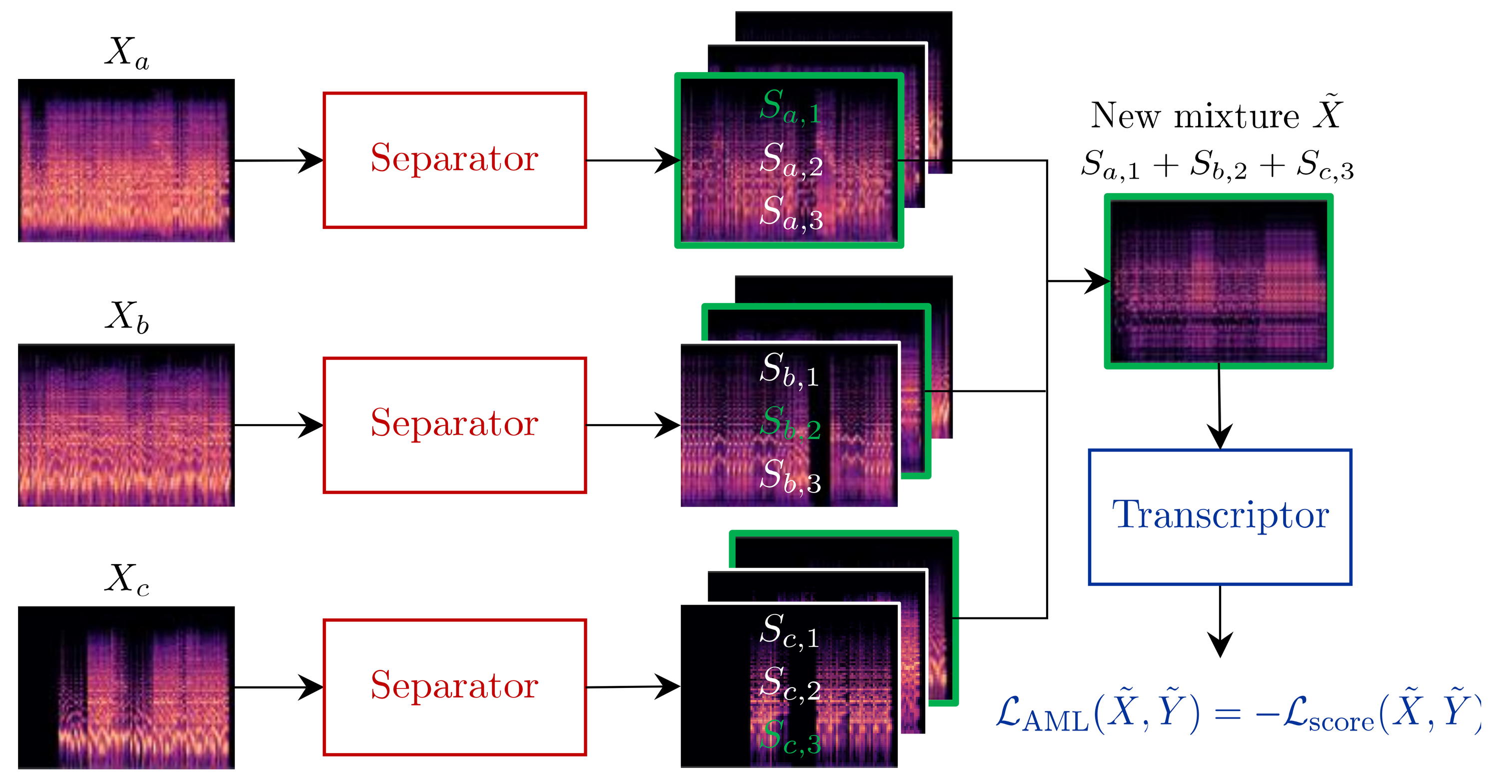

Fig 2. Diagram of the adversarial mixture loss

Step 3: Fine-tune separator and transcriptor together

- Load pre-trained transcriptor and separator to train together
- Adversarial transcription loss (ATL): transcriptor tries to detect the remaining interference in separated spectrogram (grey part in Fig. 1)
- Adversarial mixture loss (AML): transcriptor tries to detect errors in mixture created by separated spectrograms (Fig. 2)

Future work

- Semi-supervised learning: combine our proposed training method with supervised learning
- Using real-world data and include vocal and drum separation
- Alignment problem between audio and score

References

[1] 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.
[2] Fatemeh Pishdadian, Gordon Wichern, and Jonathan Le Roux. "Finding strength in weakness: Learning to separate sounds with weak supervision," IEEE/ACM TASLP, 2020.