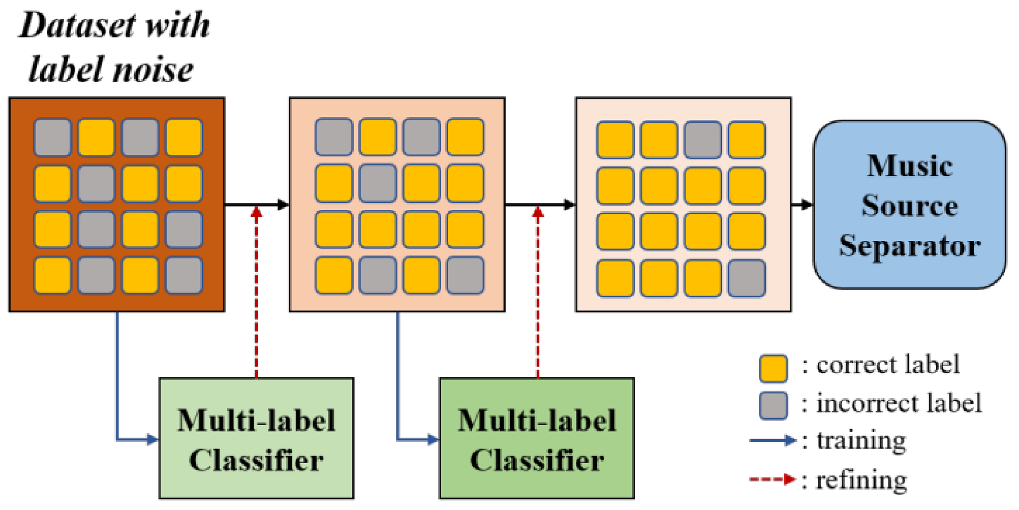
# Self-Refining of Pseudo Labels for Music Source Separation with Noisy Labeled Data

**Junghyun Koo*, Yunkee Chae*, Chang-Bin Jeon, Kyogu Lee**
*Music and Audio Research Group, Seoul National University, Seoul, Republic of Korea*
**{dg22302, yunkimo95, vinyne, kglee}@snu.ac.kr**

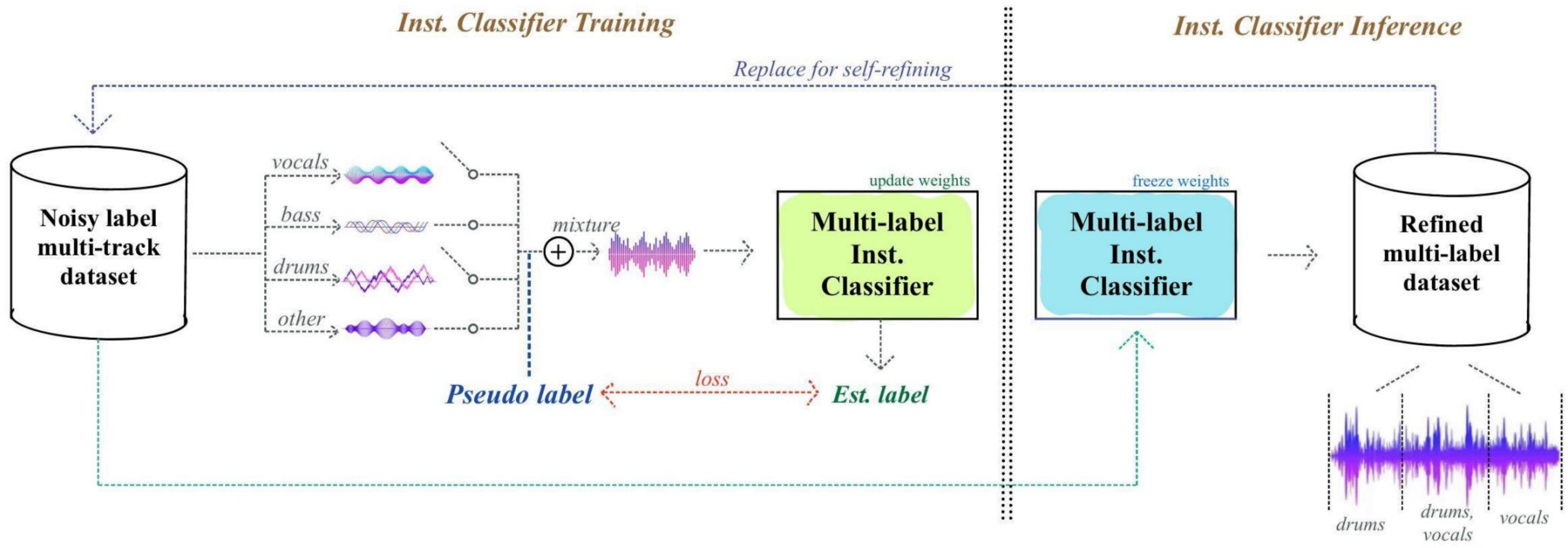
MUSIC & AUDIO RESEARCH GROUP
ISMIR 2023 Milan, Italy Nov. 5-9, 2023 POLITECNICO MILANO 1863

## Introduction



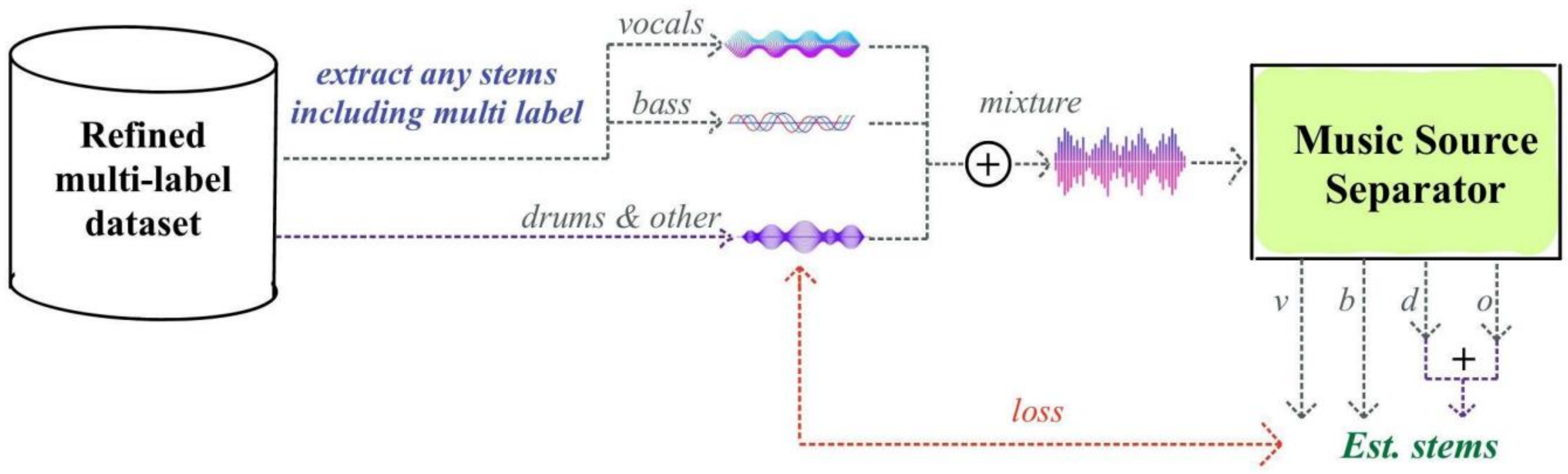

- Obtaining clean and accurately labeled individual instrument tracks for training Music Source Separation (MSS) models is challenging.
- We propose a technique for refining mislabeled instrument tracks in partially noisy-labeled datasets.
- In classification task, our self-training approach results in only a 1% accuracy degradation for multi-label instrument recognition compared to clean-labeled datasets.
- Notably, MSS models trained on self-refined datasets outperform models refined with a classifier trained on clean labels.

## Multi-label Instrument Recognition with Self-Refining




- Multi-label instrument classifier is trained with mixtures that are synthesized by randomly selecting each stem from the noisy labeled dataset.
- Similar yet different from self-training, our approach learns directly from noisy labeled data and re-labels the training data. We call this procedure *self-refining.*
- **Random mixing**: not only creates various multi-labeled mixtures, but also brings the chance to generate correct pseudo label from mislabeled stems.
- Additional data augmentation: dynamic range compression, algorithmic reverb, stereo imaging, loudness manipulation.

## Music Source Separation with Refined Dataset




- Our refined dataset contains sources labeled with multiple stems, which is unsuitable for ordinary MSS methods.
- First, we determine whether to include the multi-stem source for each input mixture sample with some probability.
- If we decide not to include the multi-labeled source, we can train the MSS model in a conventional manner.
- Otherwise, we select a multi-labeled source and choose the remaining stems from a pool of single-labeled sources.
  - Ex) select *bass+drums* → select remaining sources (*vocals, others*) from single-labeled sources
- After inference, we add the estimated stems corresponding to the multi-stem source of the input mixture.

## Results – Instrument Recognition

| Label Type | Training Data | Accuracy / F1 Score<br>Precision / Recall | | | | |
|---|---|---|---|---|---|---|
| | | *vocals* | *bass* | *drums* | *other* | *avg* |
| Single-Label | *clean* | 97.8% / 0.947<br>0.91 / 0.98 | 94.4% / 0.891<br>0.84 / 0.94 | 95.1% / 0.914<br>0.85 / 0.98 | 93.2% / 0.880<br>0.90 / 0.85 | 95.1% / 0.906<br>0.87 / 0.93 |
| | *noisy* | 93.6% / 0.860<br>0.76 / 0.97 | **90.0% / 0.821**<br>**0.73** / 0.93 | **93.7% / 0.893**<br>**0.81 / 0.98** | **92.6% / 0.865**<br>**0.92** / 0.81 | 92.5% / 0.860<br>**0.80** / 0.92 |
| | *refined* | **96.1% / 0.911**<br>**0.84 / 0.98** | 89.6% / 0.818<br>0.71 / **0.96** | 93.1% / 0.884<br>0.79 / **0.98** | 92.3% / 0.862<br>0.90 / **0.82** | **92.8% / 0.866**<br>**0.80 / 0.93** |
| Multi-Label | *clean* | 92.4% / 0.929<br>0.92 / 0.93 | 89.6% / 0.905<br>0.89 / 0.92 | 90.5% / 0.913<br>0.87 / 0.95 | 88.1% / 0.878<br>0.90 / 0.85 | 90.2% / 0.907<br>0.90 / 0.91 |
| | *noisy* | 87.9% / 0.895<br>0.83 / 0.96 | 87.5% / 0.888<br>**0.86** / 0.93 | 87.7% / 0.891<br>0.82 / **0.96** | 87.3% / 0.872<br>**0.88 / 0.87** | 87.6% / 0.887<br>0.85 / 0.93 |
| | *refined* | **91.9% / 0.928**<br>**0.88 / 0.97** | **87.8% / 0.894**<br>0.84 / **0.95** | **89.6% / 0.906**<br>**0.85 / 0.96** | **87.4% / 0.874**<br>**0.88 / 0.87** | **89.2% / 0.901**<br>**0.86 / 0.94** |

**Table 1**. Instrument recognition performance on single and multi-label instrument classifiers trained with different datasets. The training data of *clean*, *noisy*, and *refined* each represents the training subset of MUSDB18, MDX2023, and MDX2023 refined with the instrument classifier trained with MDX2023 $\Psi_{noisy}$, respectively.

- For single-labeled data, the classifier achieves the highest average performance on the *clean* dataset.
  - It can be considered an upper bound for the performance, as *clean* dataset does not contain noisy labels.
  - The average performance achieves better performance when trained on *refined* dataset than noisy dataset.
- For multi-labeled data, the *refined* dataset achieves superior performance comparable to the *clean* dataset.
  - Contrary to the evaluation with single-labeled data, the *refined* dataset generally demonstrates superior performance across all metrics in comparison to the *noisy* dataset.
  - Notably, the recall values are observed to be even higher than those of the *clean* dataset.

## Experimental Setups

- **Dataset**
  - **Dataset w/ label noise: *MDX2023 Challenge track1 dataset***
  - **Dataset w/o label noise (clean): *MUSDB18 dataset***
- **Multi-label classifier**
  - ConvNext's tiny version
  - Thresholds = 0.9
- **MSS models**
  - Hybrid Demucs
  - CrossNet-Open-Unmix

## Results – Music Source Separation

| Network | Training Data | SDR [dB] *vocals* | *bass* | *drums* | *other* | *avg* |
|---|---|---|---|---|---|---|
| *Demucs* [38] | *clean* | 5.92 | 6.16 | 5.58 | 4.43 | 5.52 |
| | *noisy* | 3.37 | 1.92 | 0.70 | 0.86 | 1.71 |
| | *w/* $\Psi_{clean}$ | 5.31 | **5.12** | 1.32 | 2.16 | 3.48 |
| | *w/* $\Psi_{noisy}$ | 4.15 | 4.58 | 1.62 | 2.85 | 3.30 |
| | *w/* $\Psi_{refined}$ | **5.36** | 5.04 | **3.09** | **3.13** | **4.16** |
| *X-UMX* [39] | *clean* | 5.76 | 4.44 | 5.47 | 3.65 | 4.83 |
| | *noisy* | 3.39 | 1.78 | 1.52 | 0.96 | 1.91 |
| | *w/* $\Psi_{clean}$ | 4.50 | 3.22 | 3.66 | 2.73 | 3.53 |
| | *w/* $\Psi_{noisy}$ | 4.72 | **4.11** | 3.22 | 2.89 | 3.74 |
| | *w/* $\Psi_{refined}$ | **4.99** | 3.93 | **5.00** | **3.18** | **4.28** |

**Table 2**. Source separation performance of Demucs v3 [38] and CrossNet-Open-Unmix [39] trained on different training datasets. Sub-items below *noisy* dataset indicate data refined with the respective instrument classifiers, denoted as $\Psi_{\bullet}$.

| Method | SDR [dB] *vocals* | *bass* | *drums* | *other* | *avg* |
|---|---|---|---|---|---|
| *proposed* | 4.99 | 3.93 | 5.00 | 3.18 | 4.28 |
| *threshold = 0.5* | 5.06 | 4.13 | 4.77 | 3.06 | 4.25 |
| *adaptive thresholds* | 4.70 | 3.72 | 3.70 | 2.62 | 3.68 |
| *train only w/ single-labeled* | 4.90 | 3.73 | 4.54 | 3.18 | 4.09 |
| *+ finetune w/ multi-labeled* | 4.33 | 4.33 | 4.19 | 3.14 | 4.00 |
| *self-refining* ×5 | 4.65 | 3.87 | 5.07 | 2.89 | 4.12 |

**Table 3**. Ablation studies on MSS performances with CrossNet-Open-Unmix.

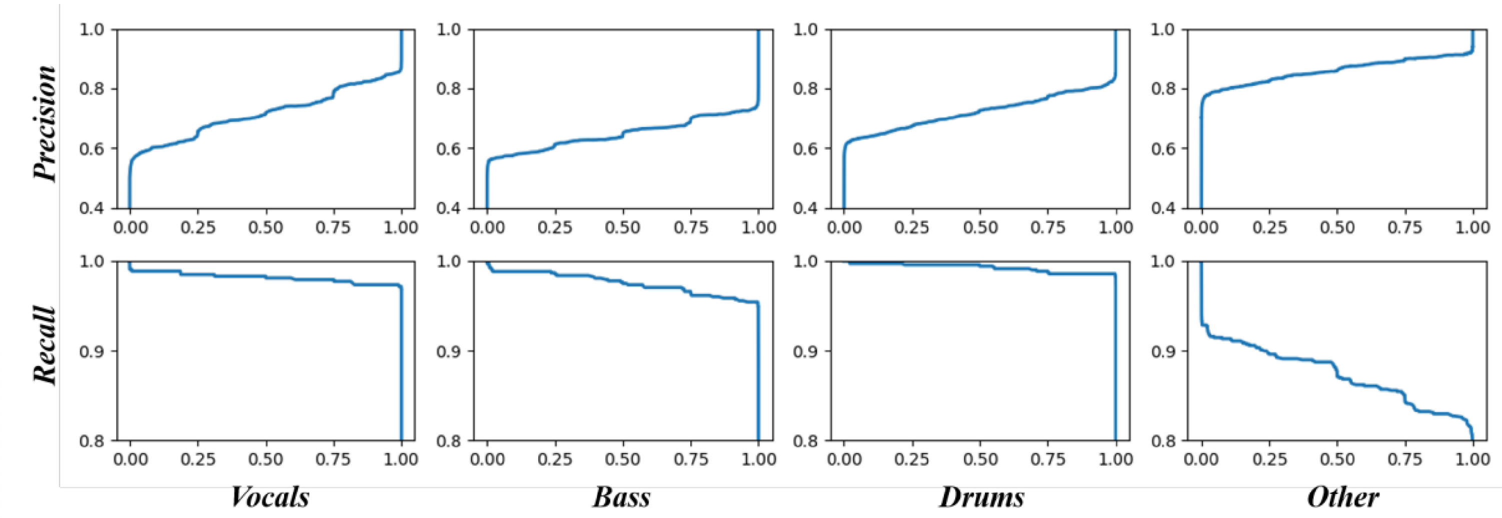



**Figure 4**. Precision and recall curves of the proposed classifier across different thresholds (x-axis) on each instrument. The curves are generated using the MUSDB18 test set (*clean*).

- **Baseline: MSS models trained on the *noisy* dataset.**
- Interestingly, the performance of $\Psi_{refined}$ exceeds the performance of $\Psi_{clean}$ , even though $\Psi_{clean}$ is trained with a noise-free labeled dataset.
- Additional factor to consider is the distinctive nature of the MSS model training framework in our approach.
  - If model receives a false-positive sample, it can simply needs to predict silence.
  - Conversely, if model receives false-negative sample, it confuses model seriously.
  - As a consequence, FN sample have a more significant impact on MSS compared to FP samples, highlighting the increased significance of the recall metric.