# Don’t Separate, Learn to Remix: End-to-End Neural Remixing with Joint Optimization

Haici Yang[1], Shivani Firodiya[1], Nicholas J. Bryan[2], Minje Kim[1]

[1] *Indiana University, Luddy School of Informatics, Computing, and Engineering, Bloomington, IN, USA*

[2] *Adobe Research, San Francisco, CA, USA*

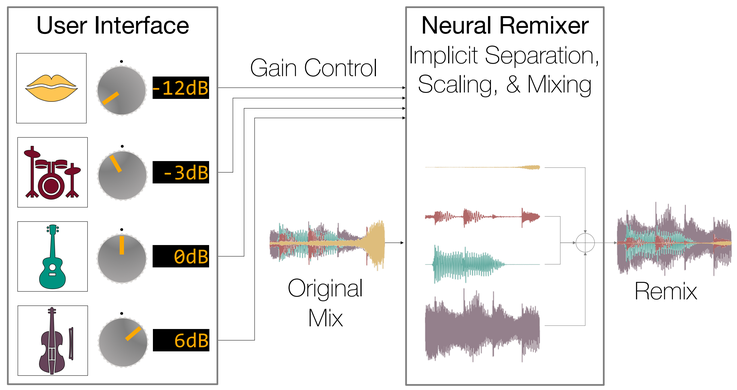
hy17@iu.edu

SAIGE

Paper ID 3310

## Introduction

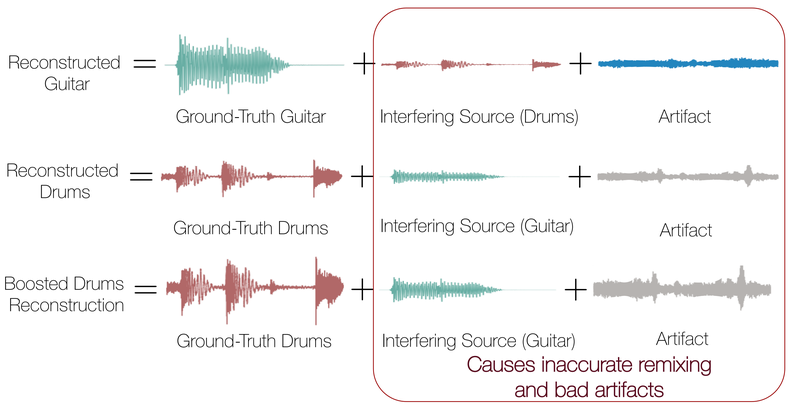
- **What is Remixing:**




  - Manipulate level and/or effects of individual instrument tracks.
  - Usually needs individual tracks, which are not always available.

- **Traditional methods – Separate, then Remix**

- **Traditional methods are problematic**




- **Our method - Do not Separate, Learn to Remix**
  - First end-to-end neural method to jointly learn MSS + remixing.
  - Higher-quality results for a wider range of volume changes.
  - Focus
    - Volume change range from -12 to 12 dB.
    - Can deal with up to five sources.

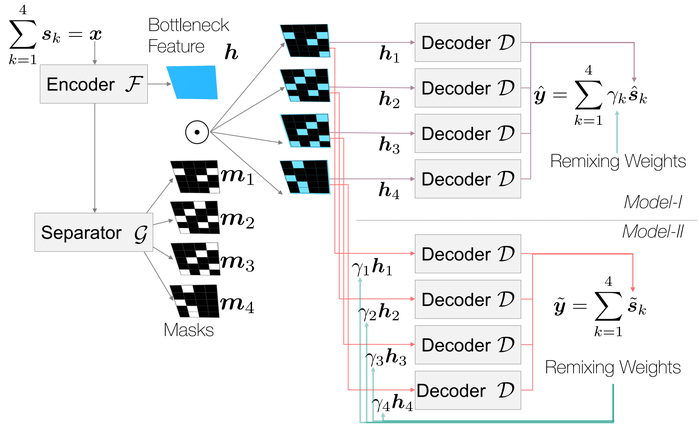
## Models and Methods

- **Conv-TasNet-like architecture** *[ Luo et.al 2019]*
  - Conv-TasNet provides a special setup that a latent space for masking-based separation is explicitly learned.
  - We replace the SISDR with regular SDR as the objective function to train the models, to make them scale sensitive for remixing task.
- **We propose two models based on the Conv-TasNet architecture:**
  - Model I jointly optimizes a separation and remix.

$$\mathcal{L}_{Model-I} = \underbrace{\psi\mathcal{E}(\boldsymbol{y}||\hat{\boldsymbol{y}})}_{\text{Remixing loss}} + \lambda \underbrace{\sum_{k=1}^{K}\mathcal{E}(\boldsymbol{s}_k||\tilde{\boldsymbol{s}}_k)}_{\text{Separation loss}}$$

  - Model II is similar but mult. remixing weights direct in latent space.

$$\mathcal{L}_{Model-II} = \underbrace{\psi\mathcal{E}(\boldsymbol{y}||\hat{\boldsymbol{y}})}_{\text{Remixing loss}} + \lambda \underbrace{\sum_{k=1}^{K}\mathcal{E}(\gamma_k\boldsymbol{s}_k||\tilde{\boldsymbol{s}}_k)}_{\text{Separation loss}}$$




## Experimental Design

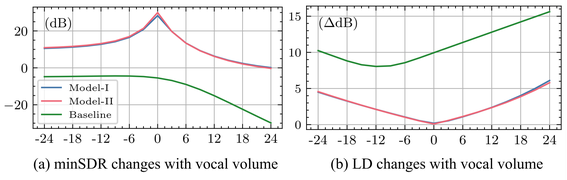
- **Datasets**
  - MUSDB18 and Slakh with cross dataset evaluation.
- **Baseline**
  - ConvTasNet-based separation + remix.
- **Evaluation Criterion**
  - minSDR: the minimum of SDR and SDSDR. *[Le Roux et.al 2018]*
  - Loudness difference: the difference between target loudness scale and the output ones for each instrumental source.
- **Ablation**
  - Different loss weight ratios: $(\psi{:}\,\lambda) = (1{:}\,1)\,, (4{:}\,1), (1{:}\,0)$
  - When $\lambda = 0$, model is solely trained towards the remix objective.

## Results

- **Overall remixing quality**
  - Model-I&II outperform baseline, especially on different cases.
  - The remix-only loss is preferred in most K = 2 and K = 3 cases.
  - When the task gets harder, more separation control is preferred.

| *minSDR* / LD | | Baseline | *Model-I* | | | *Model-II* | | |
|---|---|---|---|---|---|---|---|---|
| Train + Test | $K$ | $(\psi:\lambda)=(0:1)$ | $(\psi:\lambda)=(1:1)$ | $(\psi:\lambda)=(K:1)$ | $(\psi:\lambda)=(1:0)$ | $(\psi:\lambda)=(1:1)$ | $(\psi:\lambda)=(K:1)$ | $(\psi:\lambda)=(1:0)$ |
| Slakh + Slakh | 2 | 28.24 / 0.18 | 24.59 / 0.31 | 27.63 / 0.21 | **28.84 / 0.19** | 27.35 / 0.19 | 28.34 / 0.21 | 27.16 / 0.19 |
| | 3 | 18.72 / 0.67 | 19.88 / 0.8 | 19.7 / 0.87 | **21.26 / 0.67** | 20.09 / 0.69 | 19.81 / 0.77 | 19.26 / 0.81 |
| | 4 | 0.22 / 8.42 | 16.48 / 1.54 | 15.24 / 1.85 | 15.57 / 1.72 | 16.8 / 1.57 | 15.16 / 1.79 | **17.23 / 1.51** |
| | 5 | -4.08 / 11.31 | 7.92 / 3.87 | 12.2 / 3.2 | 11.71 / 3.34 | 8.24 / 3.86 | **12.44 / 3.15** | 11.5 / 3.45 |
| MUSDB18 + Slakh | 2 | 23.83 / 0.35 | 23.19 / 0.47 | 23.01 / 0.45 | 24.96 / 0.39 | 23.99 / 0.44 | 23.97 / 0.41 | **25.15 / 0.35** |
| | 3 | 11.88 / 1.64 | 14.13 / 1.72 | 13.37 / 1.94 | **15.3 / 1.6** | 15.2 / 1.56 | 14.76 / 1.49 | 15.15 / 1.68 |
| | 4 | -6.06 / 7.85 | 9.74 / 2.78 | 9.94 / 2.8 | 9.19 / 3.05 | 9.63 / 2.88 | **10.2 / 2.78** | 9.73 / 3.01 |
| MUSDB18 + MUSDB18 | 2 | 17.33 / 0.92 | 17.55 / 0.98 | 17.28 / 0.88 | 18.08 / 0.95 | 17.7 / 0.96 | 17.87 / 0.84 | **18.13 / 0.97** |
| | 3 | 11.82 / 1.94 | 13.37 / 1.93 | 12.52 / 2.29 | **14.49 / 1.72** | 14.17 / 1.71 | 14.13 / 1.64 | 14.15 / 1.94 |
| | 4 | -9.16 / 10.1 | 10.16 / 2.93 | **11.01 / 2.85** | 9.84 / 3.26 | 10.49 / 2.97 | 10.95 / 3.0 | 10.01 / 3.27 |
| Slakh + MUSDB18 | 2 | 12.26 / 1.61 | 14.54 / 1.31 | 14.54 / 1.39 | 14.71 / 1.36 | 14.25 / 1.42 | **15.1 / 1.29** | 13.43 / 1.56 |
| | 3 | 8.27 / 2.59 | 9.37 / 2.85 | 10.16 / 2.73 | 10.21 / 2.75 | 9.69 / 2.72 | 10.18 / 2.62 | **10.57 / 2.48** |
| | 4 | -6.33 / 9.88 | 7.46 / 3.77 | **8.44 / 3.66** | 8.34 / 3.65 | 7.75 / 3.68 | 8.29 / 3.68 | 8.06 / 3.76 |

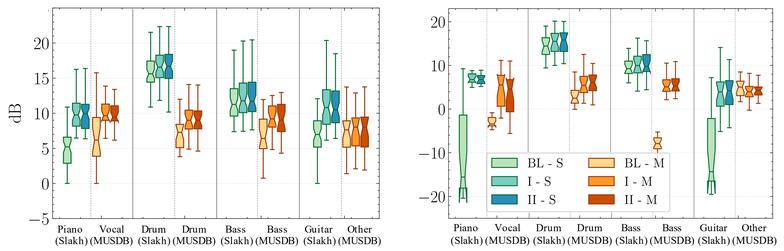
- **Remixing performance vs remixing weights**
  - Model I&II outperform the baseline in all the remixing weight choices.



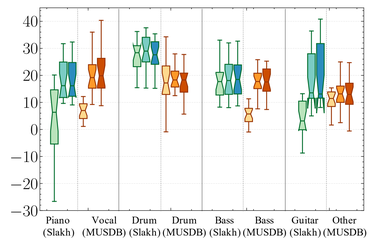

(a) minSDR changes with vocal volume (b) LD changes with vocal volume

  - Model I&II have distinctively better performance when the volume adjust amount is near 0 dB – more predictable.

- **Separation performance vs remixing results**




(a) ΔSDR (b) SAR




(c) ΔSIR

  - The baseline fails in recovering certain instruments.
  - The performance gap mostly comes from the SAR scores
  - Our neural remixer improve the remix quality by improving SAR and SIR.
  - By involving the remixing weights into the feed-forward process, Model-II can potentially associate separation behavior with the remix weights