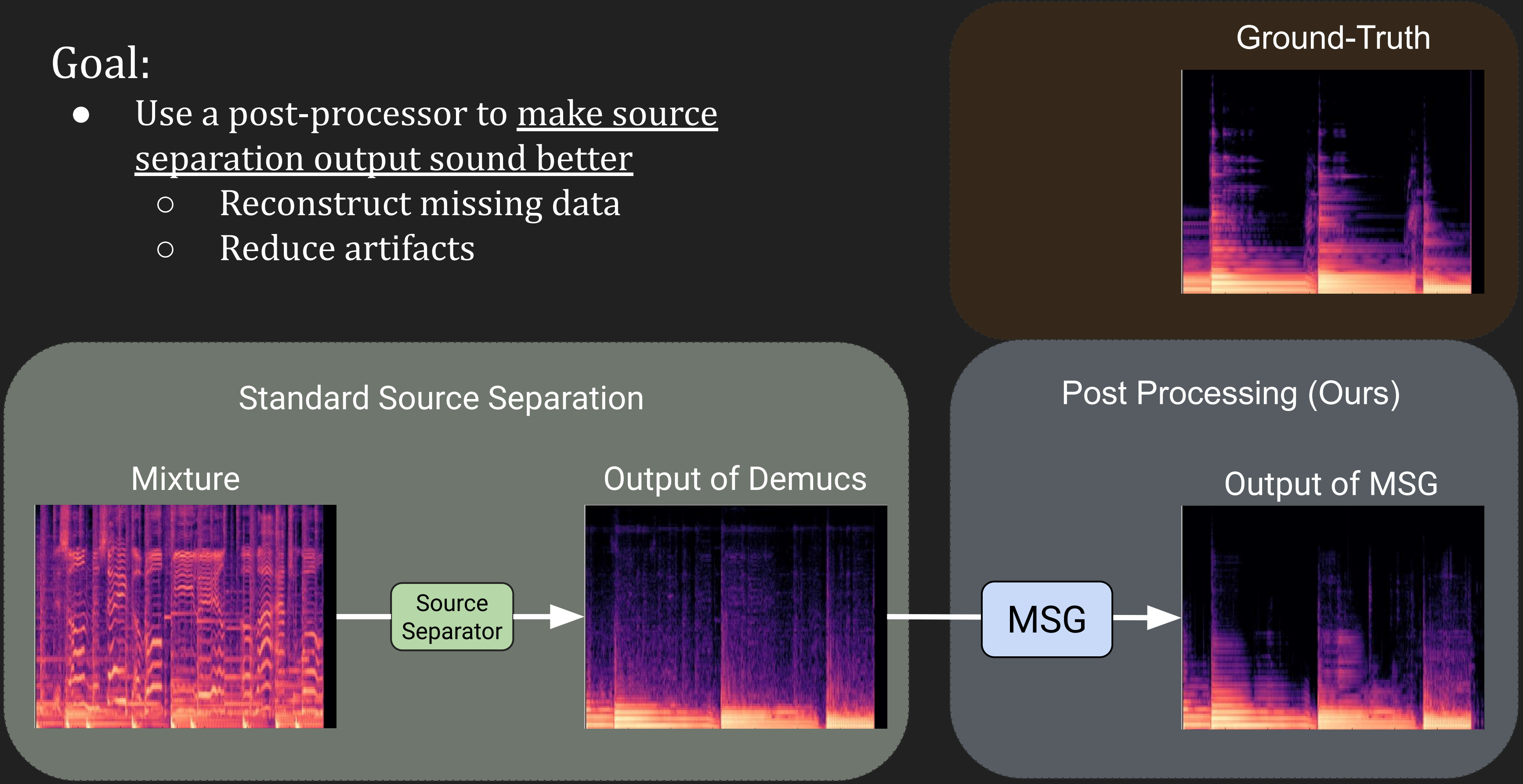
MSG: Make it Sound Good

Boaz Cogan,🍜 Noah Schaffer,🍜 Ethan Manilow, Bryan Pardo

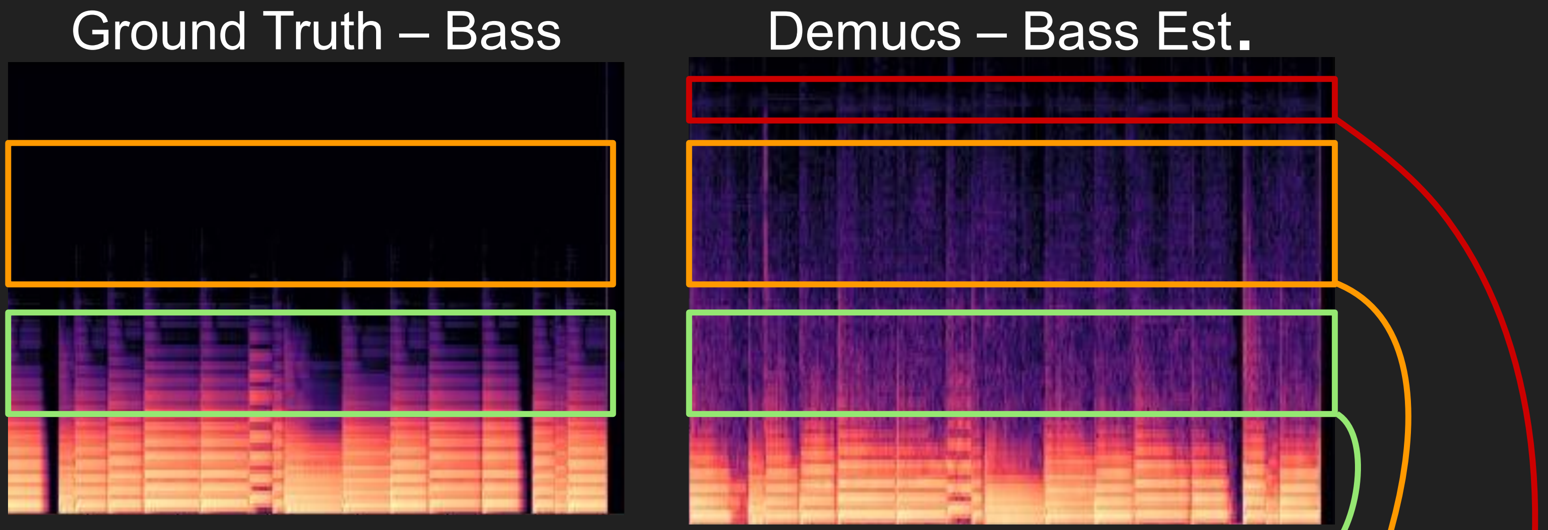
🍜 = equal contribution

Goal:

- Use a post-processor to make source separation output sound better
  - Reconstruct missing data
  - Reduce artifacts



SOTA Source Separation is Imperfect!


Ground Truth – Bass

Demucs – Bass Est.

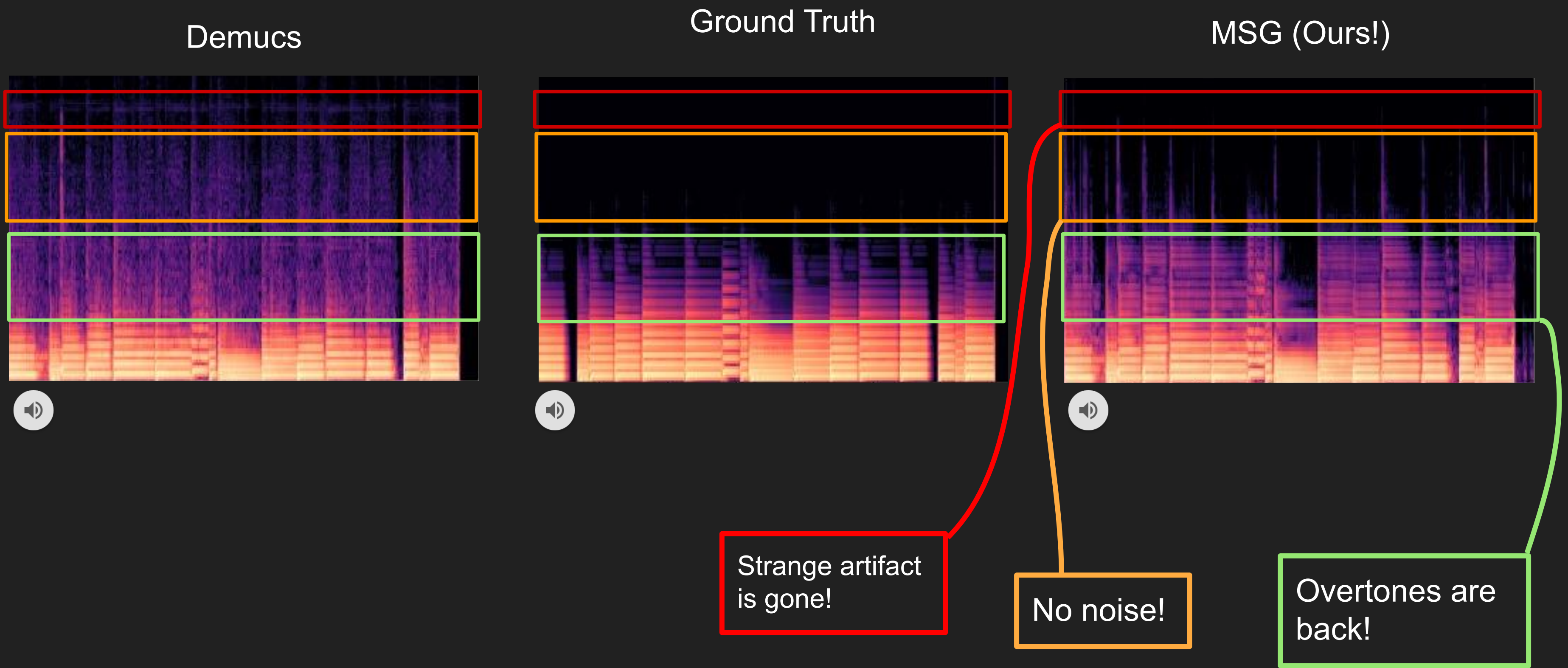
Missing Frequencies!

Added Noise!

Strange Artifacts!

- Source estimates that contain added noise and artifacts.
- How can we make this sound better?

Clean up source separation to Make it Sound GOOD!



Demucs

Ground Truth

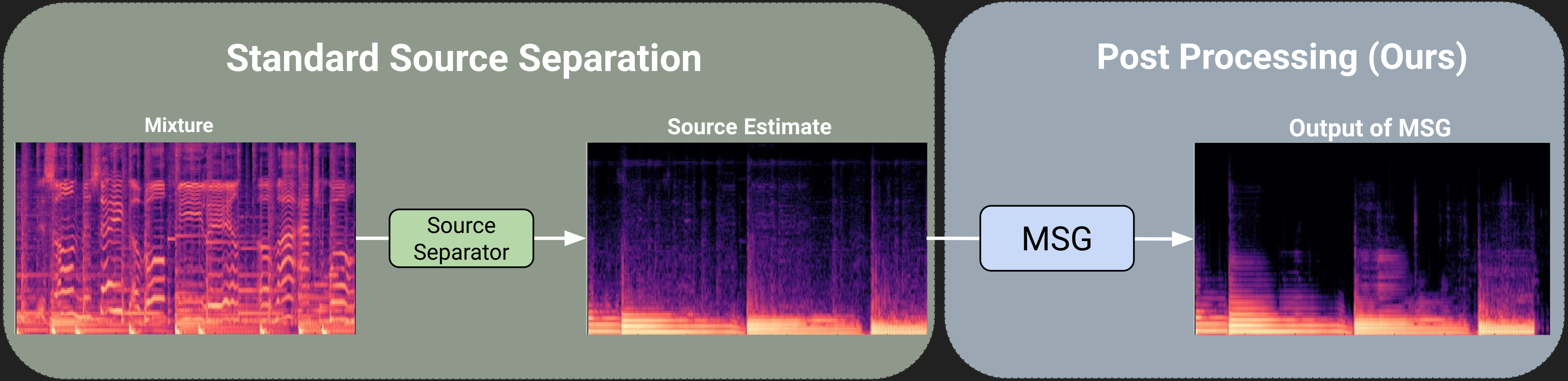
MSG (Ours!)

Strange artifact is gone!

No noise!

Overtones are back!

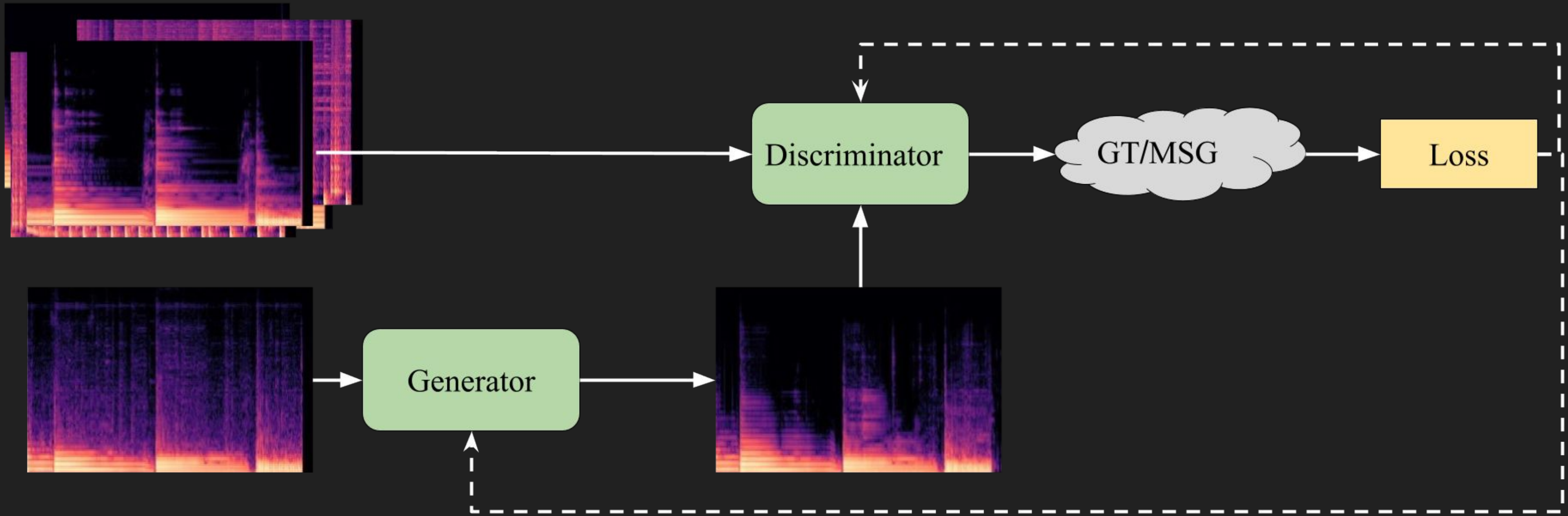
What’s a good way to do that?



- Treat source separator as a black box
- Use a post-processor to enhance its output

Use a GAN!

Similar to speech denoising (e.g., HiFi-GAN [Su et. al.])



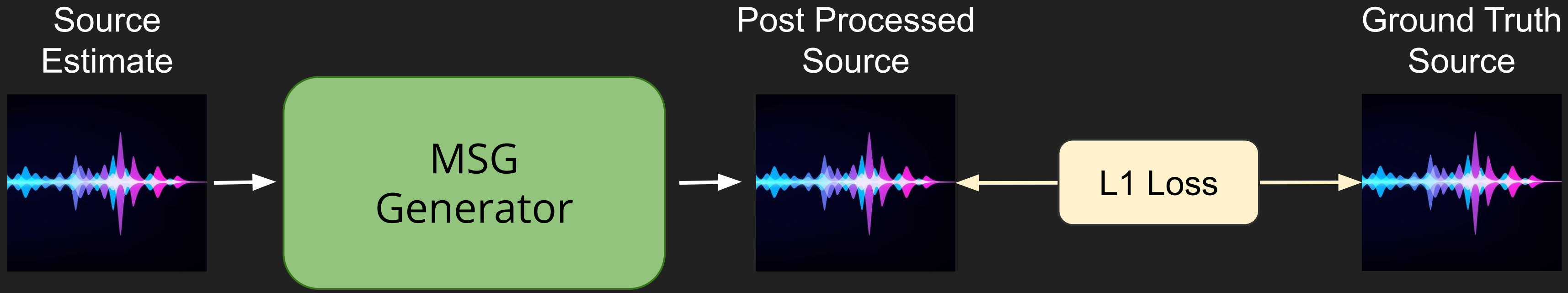
MSG Training Procedure

1. Supervised pretraining using L1 waveform Loss

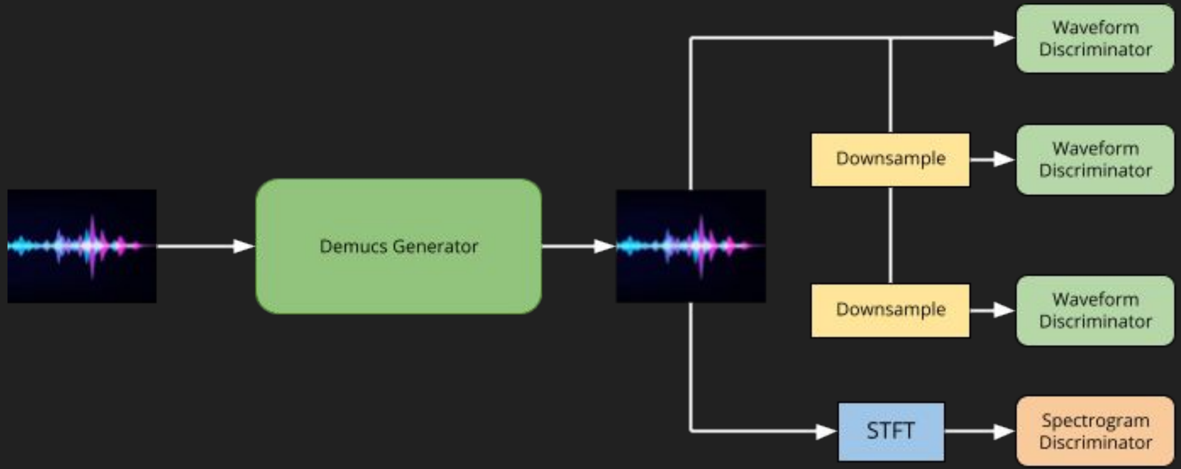
2. Add adversarial loss w/ discriminators

Supervised Pretraining

- 50 epochs using L1 waveform Loss against GT sources
- More stable GAN training → No generator collapse



Add Adversarial Loss



- 150 total training epochs
- Spectral discriminator weighting = combined waveform discriminator weighting
- Discriminator structure similar to HiFi-GAN [Su et. al. 2020]

Objective Results – MUSDB18 test set

Source	Demucs	Demucs + MSG (Ours)
Bass	6.91	6.78
Drums	7.25	7.03

Median SDR (dB) – ↑ Higher is better

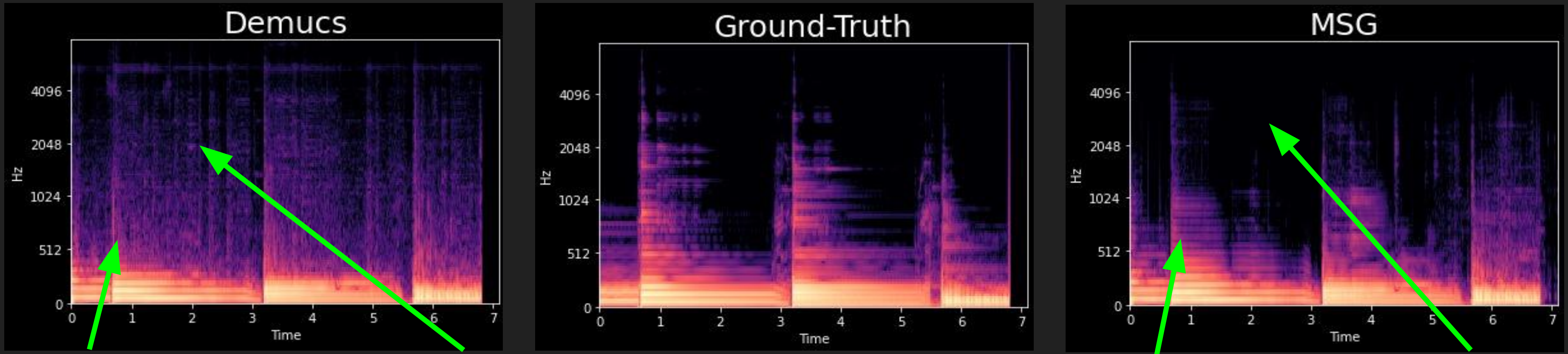
Slightly worse SDR!

...how do they sound?

Illustrative Results – Bass

SDR: 7.40

SDR: 6.88



Missing Overtones

Added Noise

Reconstructed Overtones

Removed Noise

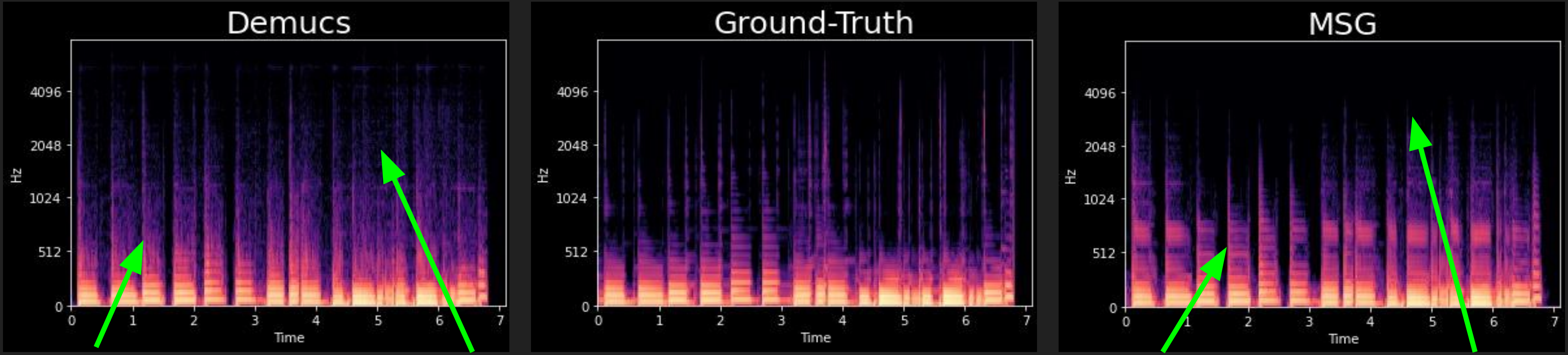
But which one *sounds* better?

Illustrative Results – Bass

🔊 SDR: 12.27

🔊

🔊 SDR: 12.04



Missing Overtones

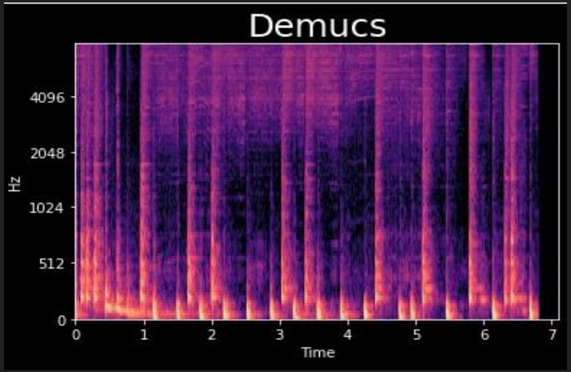
Added Noise

Reconstructed Overtones

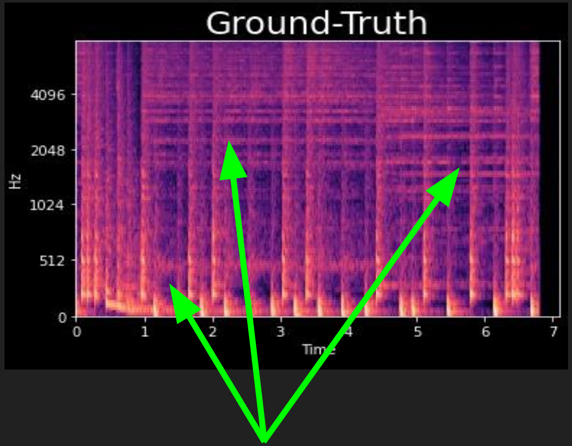
Removed Noise

But which one *sounds* better?

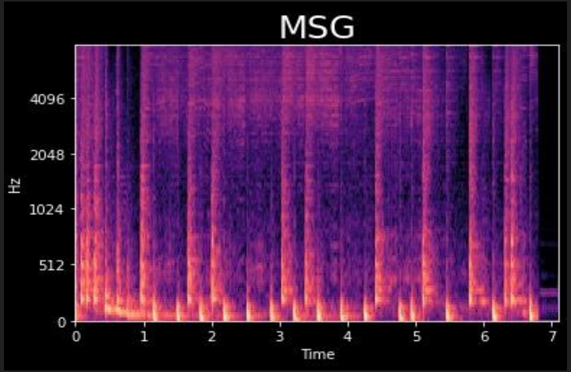
Failure case! – Drums



SDR: 8.63



Can’t reconstruct cymbal frequencies if they’re totally removed



SDR: 8.54

Discussion

- Mismatch between perceptual & SDR eval metrics → GANs get knocked on SDR
- GANs can add sounds to source estimates that hurt objective metrics...but they still might sound good.
- So what is the goal of source separation? Do well on SDR? Or to make it sound good?

Next steps / Additional Ideas

- Listening studies!

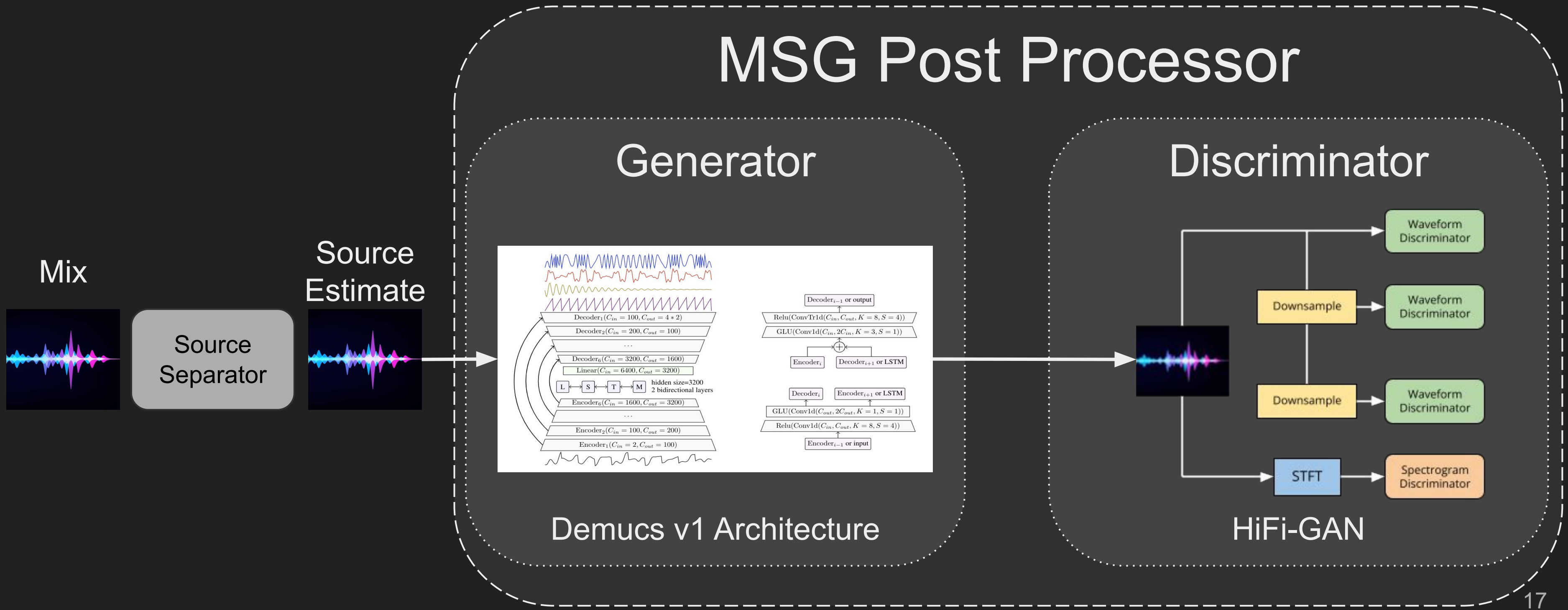
- Can we also improve SDR?

- Condition MSG on the mix
    - Restore content that separation erases

- GAN Loss during separation training
    - Integrate MSG post-processor and separation

- Multi-source MSG models

Full System Architecture


Mix
Source Separator
Source Estimate
MSG Post Processor
Generator
Decoder$_1$($C_{in} = 100, C_{out} = 4 * 2$)
Decoder$_2$($C_{in} = 200, C_{out} = 100$)
...
Decoder$_6$($C_{in} = 3200, C_{out} = 1600$)
Linear($C_{in} = 6400, C_{out} = 3200$)
L
S
T
M
hidden size=3200
2 bidirectional layers
Encoder$_6$($C_{in} = 1600, C_{out} = 3200$)
...
Encoder$_2$($C_{in} = 100, C_{out} = 200$)
Encoder$_1$($C_{in} = 2, C_{out} = 100$)
Decoder$_{i-1}$ or output
Relu(ConvTr1d($C_{in}, C_{out}, K = 8, S = 4$))
GLU(Conv1d($C_{in}, 2C_{in}, K = 3, S = 1$))
Encoder$_i$
Decoder$_{i+1}$ or LSTM
Decoder$_i$
Encoder$_{i+1}$ or LSTM
GLU(Conv1d($C_{out}, 2C_{out}, K = 1, S = 1$))
Relu(Conv1d($C_{in}, C_{out}, K = 8, S = 4$))
Encoder$_{i-1}$ or input
Demucs v1 Architecture
Discriminator
Waveform Discriminator
Downsample
Waveform Discriminator
Downsample
Waveform Discriminator
STFT
Spectrogram Discriminator
HiFi-GAN