# MUSIC SEPARATION ENHANCEMENT WITH GENERATIVE MODELING

**Noah Schaffer**⋆1 **Boaz Cogan**⋆1 **Ethan Manilow**1
**Max Morrison**1 **Prem Seetharaman**2 **Bryan Pardo**1
1 Interactive Audio Lab, Northwestern University, Evanston IL, USA
2 Descript, Inc.

## ABSTRACT

Despite phenomenal progress in recent years, state-of-the-art music separation systems produce source estimates with significant perceptual shortcomings, such as adding extraneous noise or removing harmonics. We propose a post-processing generative model (the Make it Sound Good (MSG) post-processor) to enhance the output of music source separation systems. We apply our post-processing model to state-of-the-art waveform- and spectrogram-based music source separators, including separators unseen by MSG during training. Crowdsourced subjective evaluations demonstrate that human listeners prefer source estimates of bass and drums that have been post-processed by MSG. Further analysis of the errors produced by source separators with and without MSG demonstrate that certain errors are prevalent in certain types of separators (e.g., waveform-based or spectrogram-based separators) and that MSG generally improves the reconstruction of spectral rolloff and onsets in the source estimate.

## 1. INTRODUCTION

Audio source separation is the problem of isolating a sound producing source (e.g., a singer) or group of sources (e.g., a backing band) in an audio scene (e.g., a music recording). Source separation is a core problem in computer audition and facilitates music remixing and many Music Information Retrieval (MIR) tasks that are easier when the audio is separated into individual sources, such as music instrument labeling [1, 2] and transcription [3, 4].

Current state-of-the-art source separation systems often produce source estimates that contain perceptible artifacts, such as high-frequency noise, source leaking (e.g., drum hits heard in the bass source estimate), unnatural transients, or missing overtones. For many downstream tasks in MIR or music creation, it is preferable for source separators to minimize these errors. Given that we have observed these artifacts to be endemic to the separators themselves, we propose an additional post-processing step to clean up the initial outputs of these separators.

In this work, we introduce Make it Sound Good (MSG), a generative [1] post-processor for enhancing the quality of music source separation. MSG uses elements of off-the-shelf architectures from speech vocoding and denoising to enhance the output of pre-trained source separation models in both the waveform and spectrogram domains.

The main contributions of this work are:

- A source separation post-processor (MSG) that performs both imputation and denoising to enhance the output of both waveform and spectrogram models for music audio source separation.
- A subjective listener study that confirms MSG improves the perceptual quality of bass and drum source estimates on a set of five separation models, including one on which it was not trained.
- An in-depth exploration of the kinds of errors produced by different classes of source separators and how MSG affects these errors.

Audio examples and code can be found at master.d3neehbzr2rbmh.amplifyapp.com/.

## 2. RELATED WORK

Deep learning is the dominant approach for music source separation. For example, all entries to the 2021 Sony Music Demixing Challenge [5] were deep learning based separators. Most separators fall into one of two classes. *Waveform models* [6–9] take an audio waveform as input and produce an audio waveform for each separated source. *Spectrogram models* [10–18] take a mixture spectrogram as input and output a mask to apply to the spectrogram for each source being separated. Despite the recent successes of these deep learning methods, state of the art sys-

⋆ = Equal contribution

[1] We use the phrase "generative" to describe a model that performs a generative modeling task (e.g. inferring missing frequencies from a signal). We acknowledge that while our model does not fit under every definition of a generative model, we adopt language common in the audio domain.

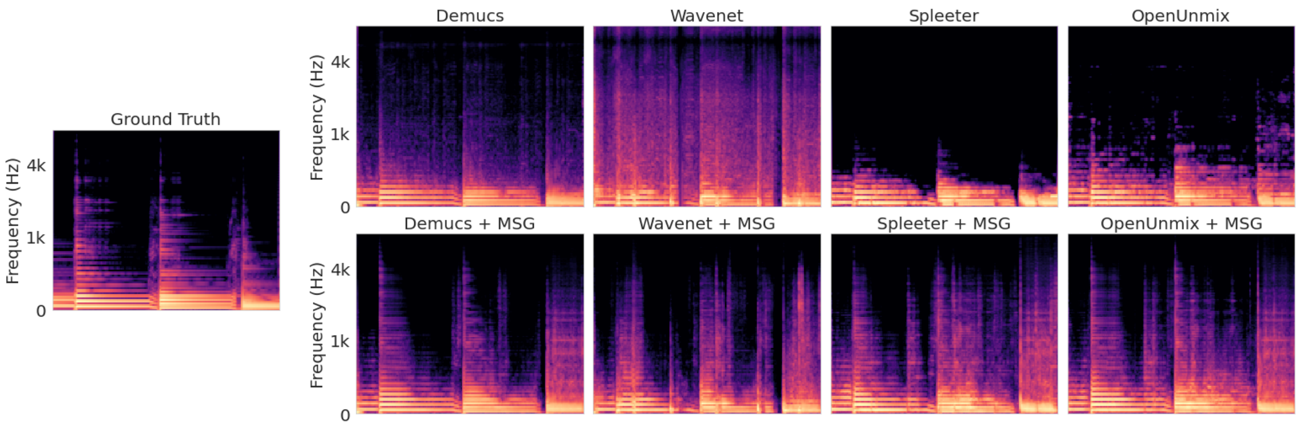


**Figure 1**: Spectrograms of ground-truth (left), source estimates (top), and MSG output (bottom) for the bass source. MSG is able to simultaneously infer missing frequencies and remove noise from the output of common source separation systems.

tems continue to exhibit perceptible artifacts in their outputs. We show in Section 5 that waveform models tend to introduce more high-frequency noise, while spectrogram models tend to lose transients and high frequency content.

Recent works in adversarial audio synthesis [19–25] and end-to-end speech enhancement [26–32] show that Generative Adversarial Networks (GAN) architectures [33] are effective in generating high-fidelity audio. Although GAN systems have been effective in audio synthesis and denoising, no previous work has explored using these systems for enhancing source separation output. Recent work has also shown that adversarial loss is effective for training source separation systems [34, 35], however this work does not look at using adversarial loss to enhance existing separation output.

While many recent works for speech enhancement have been proposed, music enhancement (e.g., denoising and artifact removal) is less common, with some recent works in music denoising [36–38] and bandwidth expansion [39, 40]. Existing work has also looked at the cause of source separation artifacts [41] and methods to fix those artifacts [42], however, this is from a purely theoretical perspective and has not been explored in pre-trained networks. Our work is most similar to Kandpal et. al [43], who proposed a generative model that can enhance the audio quality of a low-quality music recording taken on a consumer device. We are unaware of a prior system for enhancing the output of pre-trained music separation systems.

## 3. "MAKE IT SOUND GOOD" POST-PROCESSOR

Here we describe our "Make it Sound Good" (MSG) post-processor, which perceptually improves a source estimate by removing artifacts that the separator introduced and imputing elements the separator omitted. Due to the wide success of using Generative Adversarial Networks (GANs) [33] to denoise many types of audio, we design our MSG post-processor model using a GAN architecture.

The generator of MSG is a waveform-to-waveform U-Net with 1D convolutions. This is very similar to the Demucs v2 [7] architecture with the exception that Demucs has two BLSTM layers at the bottleneck, which we omit.

We train the generator using three loss functions. The first is the LSGAN [44] generator loss,

$$L_G(D) = \mathbb{E}\left[(D(G(\hat{s})) - 1)^2\right], \tag{1}$$

where $\hat{s}$ is the raw source estimate from the separator, $D$ is the discriminator, and $G$ is the generator. We use multiple discriminators (described later in this section). We evenly weight the contribution of each discriminator to the loss.

Next is deep feature matching loss [45], which is the $L_1$ distance between the intermediate activations of the discriminators on corresponding real and generated data. The last loss function we use is a multi-scale Mel-spectrogram reconstruction loss [46], which is the average Mel reconstruction loss over three different Short term Fourier transforms (STFTs), each of which uses different parameters for the number of STFT bins, window lengths, and hop sizes.

We use two types of discriminators: the multi-period discriminators from HiFi-GAN [24] and the multi-resolution spectrogram discriminators [25] from UnivNet. The multi-period discriminators operate on the waveform, and reshapes the waveform to a 2D tensor with a prime-valued stride (e.g., $[2, 3, 5, 7, 11]$) before processing the reshaped waveform with 2D convolutional layers. The multi-resolution spectrogram discriminators process a spectrogram with different STFT window sizes (see Section 4.3). Each discriminator uses the LSGAN [44] loss,

$$L_D(G) = \mathbb{E}\left[(D(s) - 1)^2 + (D(G(\tilde{s})))^2\right], \tag{2}$$

where $\tilde{s}$ is the cleaned up source estimate from the MSG generator and $s$ is the ground-truth source audio. Further details on the discriminator architectures are provided in the source papers [24, 25].

MSG post-processes the separation result of a music source separator. We use an adversarial loss typical of generative adversarial networks [33] but do not condition on a random input vector. Since we are using MSG to generate missing frequencies not present in source separation output, we are performing a generative modeling task, despite

MSG itself being a deterministic model. This mild misuse of the terminology is common in the audio processing literature [20, 24–32].

## 4. EVALUATION

We evaluate our MSG post-processor using a crowdsourced subjective evaluation to determine whether MSG improves the audio quality of music source separators (Section 4.4). We use this approach, rather than reporting the Signal to Distortion Ratio (SDR) [47, 48], because SDR and similar metrics are imperfect proxies [49] for human opinion. To help understand the listener study results, we then perform objective evaluation using spectral rolloff and onset detection results (Section 5).

### 4.1 Models

We train MSG post-processors using four existing source separation models: two waveform-based separators (Demucs v2 [7] and Wavenet [8]) and two spectrogram-based separators (Spleeter [17] and OpenUnmix [16]). In order to demonstrate the generality of our approach across separators, we create one enhancement model that is trained and evaluated on all separators instead of creating separator-specific models. We use pre-trained separator models provided by the authors of each model. We refer readers to these respective papers for details on their architectural and training details. To show that MSG can also enhance the source estimates of unseen separators, we evaluate our post-processor on the four seen separators as well as a fifth separator that it was never trained on: Hybrid Demucs (v3) [50]. All separation models use the trained weights released by the authors of the models, with no fine-tuning or alteration by us.

### 4.2 Data

We use the MUSDB18 dataset [51] for all experiments. MUSDB18 contains 150 songs (100 in the training set and 50 in the test set). Each song in MUSDB18 has a full mixture and isolated source audio stems for vocals, bass, drums and a fourth catch-all category called "other". We omit this catch-all category because we find that attempting to enhance many instruments at once with the same model greatly increases the difficulty of the task. We perform source separation on every song in MUSDB18 using all five of our source separators, producing source estimates of bass, drums and vocals.

We peak normalize our input audio before passing it through the network. Since Wavenet operates at 16 kHz we chose this sample rate. We downsample all systems to 16 kHz instead of upsampling all systems to 44.1 kHz because if we upsampled, our model would then have to jointly perform the tasks of enhancement and bandwidth extension for separation output from separators that operate under 44.1 kHz. We chose to focus on solely enhancement and left the task of bandwidth extension to future work. The output of MSG on all systems is at a sample rate of 16 kHz.

### 4.3 Training

We train one MSG model on the MUSDB18 training set for each source class (bass, drums, or vocals). Each model is trained using source estimates from four separators (Demucs v2, Wavenet, Spleeter, and OpenUnmix) as input and the ground-truth sources as training targets. We segment the audio into 1-second clips and reject silent clips where the ground-truth source has an RMS below -60dB FS, resulting in over 100,000 training examples per model. On each training iteration, we randomly swap the input data with ground-truth with a $10\%$ probability. This encourages the model to leave high-quality audio unaltered.

We compute the three resolutions of STFTs passed to our multi-resolution spectrogram discriminators (Section 3) using window sizes of 512, 1024, and 2048 samples and hop sizes of 128, 256, and 512 samples, respectively. We use one Adam optimizer [52] for the generator and another for the discriminator. We use a learning rate of 2e-4 and beta values of $(.5, .9)$. During the first $1,000$ iterations of training, we rescale loss weights for each term by solving a least-squares equation before every gradient update.

### 4.4 Subjective evaluation

We evaluated whether MSG enhances the perceptual quality of source separation output through a crowdsourced perceptual evaluation. For evaluation data, we used one seven-second segment from each of the 50 songs from the MUSDB18 test set. We performed source separation on each seven-second segment using each of our five source separation systems (Section 4.1) to create source estimates of bass, drums, and vocals. Each output was then processed with MSG, resulting in 50 matched pairs for each combination of separator and source class: the raw output, and the output processed by MSG.

There are 15 unique combinations of the five separators and three sources (bass, drums, and vocals). For each combination, we performed a two-way forced-choice listening test between the raw output and the output processed by MSG. We recruited 20 participants for each test. Each participant evaluated 25 randomly-selected pairs from the 50 examples for that combination of source and separator. We omitted responses from participants that failed a prescreening listening test. If the results of a particular test showed insufficient statistical power (i.e., $p < 0.05$) using a binomial statistical test) we recruited an additional 10 participants.

For each pairwise comparison, participants were given the following instructions, where <*source*> is one of "bass", "drums" or "vocals":

> Listen to both recordings of a <*source*>. After listening to both, select the recording that sounds like a higher-quality <*source*>. The higher-quality recording is the one that is more natural sounding, or has fewer audio artifacts (e.g., noise, clicks, or other instruments).

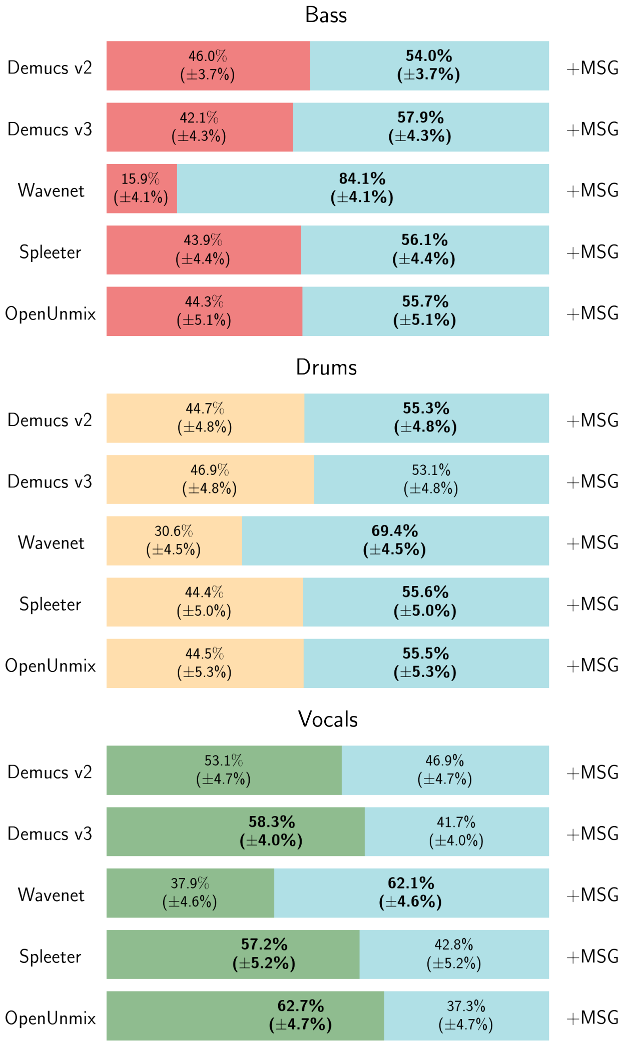


**Figure 2**: Subjective pairwise test results for bass, drums, and vocals. Each row contains the percent of listeners selecting that option as higher quality in a two-way forced choice listening test. A bold-faced value indicates a statistically significant difference.

We used Reproducible Subjective Evaluation (ReSEval) [53] to set up our listener studies. We recruited participants via Amazon Mechanical Turk (MTurk). Our participants were US residents at least 18 years old that completed 20 or more tasks on MTurk with an approval rating of at least 97%. Participants who passed the listening test and completed our evaluation were paid \$3.00.

#### *4.4.1 Subjective evaluation results*

Each listening test evaluates one combination of source class and source separator. For each of the 15 tests, we collected between 308 and 696 pairwise evaluations from between 15 and 30 participants who passed the prescreening listening test. The number of evaluations is not a strict multiple of 25 because a few participants did not finish all 25 examples in their set of pairwise evaluations.

Figure 2 shows the results for each of the 15 tests. Listeners preferred the MSG output to the raw source separator output in 11 out of 15 combinations of separator and source. This difference was statistically significant (using a binomial test) in 10 of the 11 combinations. Listeners preferred the quality of separators with MSG on bass source estimates and had a moderate preference for MSG on drums. For vocals, listeners had a slight preference for the source estimate without MSG.

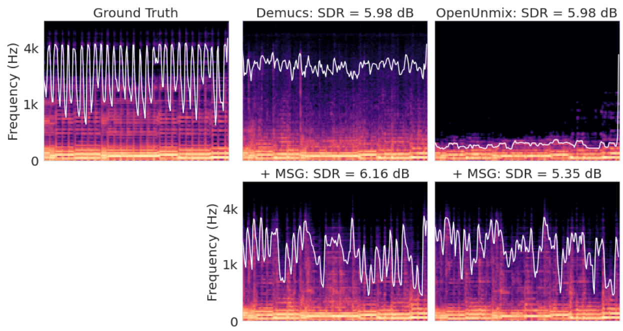


**Figure 3**: Spectrograms of the ground truth (left) and source estimates from Demucs v2 and OpenUnmix (top) and corresponding MSG output (bottom) for the bass source. The 98% spectral rolloff frequency is overlaid in white.

MSG performed best on the Wavenet separator, where it significantly improved the perceptual audio quality of all sources. MSG was also able to improve on the quality of source estimates of a separator not seen during training, Demucs V3, for bass sources—as well as an improvement on Demucs V3 drum sources that was not statistically significant. Note that Demucs V3 is a hybrid approach that operates in both the waveform and spectrogram domains. Our performance on vocals indicates that MSG is not able to enhance the quality of the source estimate of vocals. We note that we are not aware of any previous works that have proposed a model for enhancing the quality of music that demonstrates a statistically significant subjective improvement in quality specifically on singing—despite the many successes of speech denoising and enhancement. We are interested in exploring why singing enhancement is so much more difficult than speech, bass, or drum enhancement as a direction for future work.

## 5. FURTHER ANALYSIS

Our listener study indicates whether a separation is relatively “good” or “bad”, but it does not clarify *why* one separated source is better or worse than another. The widely-used Signal to Distortion Ration (SDR) [47, 48] is similarly agnostic to the different types of errors separators make. See, for example, the top row of Figure 3, which shows source estimates produced by Demucs v2 and OpenUnmix for the same bass source from the same mixture. Demucs adds additional high-frequency noise not present in the ground truth, while OpenUnmix removes many of the upper harmonics. Visually, the difference between these two systems is plain, however their SDR values (using `mus_eval` [54]) are equal to two decimal places: *5.98 dB!*

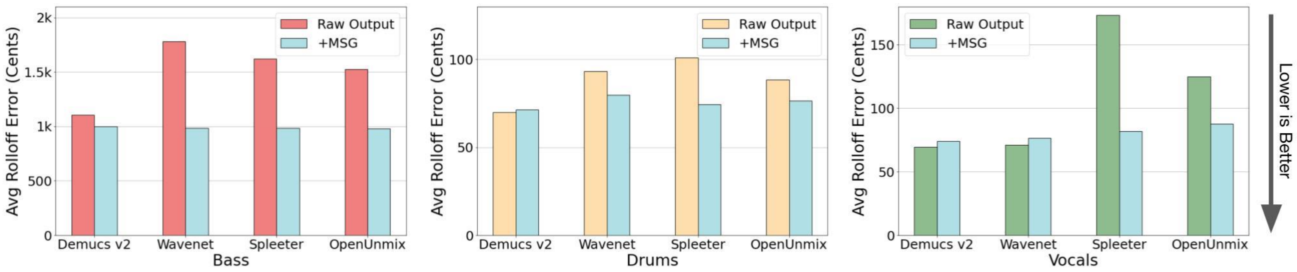


**Figure 4**: Mean spectral rolloff error for bass, drums, and vocals for separators with and without MSG post-processing.

In this section, we examine the output of the four state-of-the-art separation systems used in the training of MSG models: (Demucs v2 [7], Wavenet [8], Spleeter [17], and OpenUnmix [16]) as well as the MSG-processed outputs for those four systems. As before, we use the MUSDB18 [51] test set, and we omit the "other" source.

Anecdotally, we have noted that waveform separators tend to add extra high-frequency noise and spectrogram separators tend to remove high-frequency partials, especially in bass estimates (see Figure 1). Spectrogram separators also tend to smooth out transients. While these are not the only issues that current separation systems exhibit, the rest of this section will be dedicated to analysis of these two issues.

### 5.1 Added and Missing Frequency Content

Following our anecdotal observation that waveform separators tend to add extra high-frequency noise and spectrogram separators tend to remove high-frequency partials, we seek to formalize these notions.

One statistic that can be a good proxy for whether a source estimate has excess high-frequency content or is missing desirable high-frequency content is spectral rolloff. For a given time frame in a spectrogram, the spectral rolloff at $X\%$ is the frequency below which $X\%$ of the energy of the signal lies. For example, the white line on each spectrogram in Figure 3 shows the spectral rolloff at $98\%$.

For every song in the MUSDB18 [51] test set, we compute the spectral rolloff at $98\%$ every 32 ms (a hop size of 512 samples at 16 kHz) for every ground truth isolated source, every estimate produced by one of the four training separators and every MSG-enhanced source estimate. To calculate our statistics, we omit any frames that have an RMS less than $-40$ dBFS, in the ground-truth source, so as not to examine rolloff in relatively silent regions. We report the error between a source estimate's rolloff and a ground-truth source's rolloff in cents, which is $1200 \times (\log_2 x - \log_2 y)$, where $x$ and $y$ are rolloff frequencies in Hz. We chose to use cents over Hz because it better correlates to how humans perceive audio.

In Figure 4, we show the mean error, in cents, between the ground truth rolloff and the source estimate's rolloff. We see that the source estimates of vocals and drums have spectral rolloff errors on the order 100-200 cents, whereas source estimates of bass have errors of roughly 1000 cents.

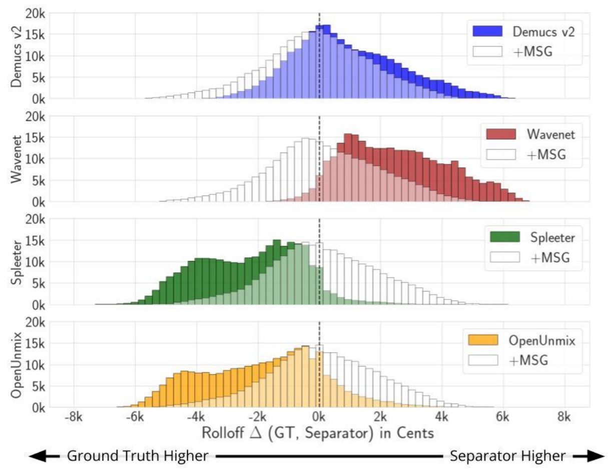


**Figure 5**: Histogram of the difference in spectral rolloff values between a given separator's bass estimate and ground truth bass source over the MUSDB18 test set. The vertical dotted line shows the desired difference of 0. MSG reduces the difference between the rolloff values of source estimate and ground-truth.

MSG reduces this error for all separators on bass sources, for three of the four separators on drums, and two of the four separators on vocals.

Because the bass estimates have such large errors, we examine them further in Figure 5, where we show a histogram of the per-frame differences between the spectral rolloff of source estimates and ground truth. The top two rows of the histogram show results for the two waveform separators (i.e., Demucs v2 and Wavenet), which each show an error distribution that is strongly skewed towards positive error. This corroborates our observation of high-frequency noise introduced by waveform separators, as shown in the bass spectrogram in Figure 3. The bottom two rows show the error distribution for two spectrogram separators, Spleeter and OpenUnmix. Both exhibit an error distribution for spectral rolloff that is strongly skewed toward negative values. This quantifies the effect illustrated in Figures 3 and 1, where the spectrogram separators remove the higher partials of the bass source.

We further observe that, when MSG is applied to the output of all four separators, the resulting error distribution is less biased and, as was already shown in Figure 4, reduces the mean error magnitude. Figure 1 illustrates the effect of MSG on a single bass example, showing improved spectral rolloff reconstruction for both waveform and spectrogram models.

### 5.2 Improving Transient Reconstruction

While listening to the source estimates from spectrogram separators, we noticed that the transients of source estimates for drums and bass did not sound as clear as in the ground truth source estimates. To quantify these observations, we measure the location and strength of onsets in the estimated sources relative to the ground-truth. We use librosa's [55] `onset_strength()` function [56], which computes the spectral flux onset strength envelope at every frame in a spectrogram. We approximate an onset by identifying every frame with a strength above a certain threshold. We select an onset strength threshold via manual tuning. We set the threshold value to a constant value of 0.75 for both bass and drums on the MUSDB18 dataset. We manually tuned this threshold to find a value that best corresponds with our perception of relevant peaks in the signal. We chose to threshold `onset_strength()` instead of using librosa's `onset_detect()` because we found that matching up onsets between two signals using the latter method was hard to correctly tune.

We run this onset strength thresholding on both the ground-truth source and a source estimate and then calculate the F1 between the binary threshold arrays of the raw source estimate and the MSG post-processed estimate as a proxy for how well a separator preserves transients. A true positive (TP) is when a detected onset exists at the same spectrogram frame in the ground-truth and source estimate, a false positive (FP) is when an onset is detected at a frame in the source estimate but not the ground truth, and a false negative is when an onset is detected (FN) at a frame in the ground truth but not in the source estimate. We report the F1 score of onset reconstruction, $TP/(TP + \frac{1}{2}(FP + FN))$.

We report the F1 scores for onset detection on bass, drums, and vocals in Table 1. The results for vocals are not in favor of MSG for 3 of the 4 separators. We include the vocals results for completeness. Evaluating the transients for vocals might be slightly unusual, but the observed results align with the listener studies. In contrast with vocals, bass and drums both see improved F1 scores across multiple separators: MSG improves the F1 score in 7 out of the 8 combinations of source and separator, with the sole exception of drums separated by Demucs v2. This indicates that the ability to represent transients is generally improved by applying MSG-based post processing on bass and drums.

| | Model | Type | Onset Strength F1 Raw | + MSG |
|---|---|---|---|---|
| Bass | Demucs v2 | Wave | 0.52 | **0.54** |
| | Wavenet | Wave | 0.36 | **0.44** |
| | Spleeter | Spec | 0.36 | **0.52** |
| | OpenUnmix | Spec | 0.39 | **0.49** |
| Drums | Demucs v2 | Wave | **0.84** | 0.82 |
| | Wavenet | Wave | 0.73 | **0.74** |
| | Spleeter | Spec | 0.78 | **0.82** |
| | OpenUnmix | Spec | 0.78 | **0.79** |
| Vocals | Demucs v2 | Wave | **0.58** | 0.57 |
| | Wavenet | Wave | **0.51** | 0.49 |
| | Spleeter | Spec | **0.71** | 0.66 |
| | OpenUnmix | Spec | 0.41 | **0.57** |

**Table 1**: F1 scores for thresholded onset strength for bass, drums, and vocals for four separators with and without MSG post-processing. "Raw" means that F1 is computed between the separator's raw output and ground truth. "+MSG" means that MSG post-processing is applied to the raw source estimates. According to this measure, MSG is able to better preserve onsets in 7 out of 8 cases between the bass and drums sources, which most clearly demonstrate artifacts with transients.

## 6. CONCLUSION

State-of-the-art music source separators create audible perceptual degredations, such as missing frequencies and transients. In this work, we propose Make it Sound Good (MSG), a post-processing generative model that enhances the perceptual quality of music source separators. In listening studies, users prefer bass and drum source estimates produced with MSG post-processing—even on a state-of-the-art separator not seen during training. We analyzed the errors of waveform-based and spectrogram-based separators with and without MSG. Without MSG, we show that waveform-based separators induce high-frequency noise and spectrogram-based separators fail to reconstruct high-frequencies in the bass source, and have trouble reconstructing transients. We measure these artifacts via spectral rolloff and onset detection and show that, for both bass and drums, MSG generally improves reconstruction of spectral rolloff and onsets of the source estimate relative to the ground-truth sources.

In future work, we aim to further improve results on bass and drums as well as the more difficult vocal and "other" sources in MUSDB18. Some recent spectrogram models [57, 58] operate on complex spectrograms or complex masks. A study of the kinds of errors exhibited by these systems and application of an MSG like system to these newer frameworks is another promising direction for future work.

## 7. REFERENCES

[1] E. Vincent and X. Rodet, "Instrument identification in solo and ensemble music using independent subspace analysis," in *International Society for Music Information Retrieval (ISMIR)*, 2004.

[2] J. F. Woodruff, B. Pardo, and R. B. Dannenberg, "Remixing stereo music with score-informed source

separation." in *International Society for Music Information Retrieval (ISMIR)*, 2006.

[3] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti, and M. B. Sandler, "Automatic music transcription and audio source separation," *Cybernetics & Systems*, vol. 33, no. 6, pp. 603–627, 2002.

[4] E. Manilow, P. Seetharaman, and B. Pardo, "Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[5] Y. Mitsufuji, G. Fabbro, S. Uhlich, and F.-R. Stöter, "Music demixing challenge 2021," *International Society for Music Information Retrieval (ISMIR)*, 2021.

[6] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[7] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv e-prints*, pp. arXiv–1911, 2019.

[8] F. Lluís, J. Pons, and X. Serra, "End-to-end music source separation: is it possible in the waveform domain?" *Interspeech*, 2019.

[9] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *International Society for Music Information Retrieval (ISMIR)*, 2018.

[10] A. M. Reddy and B. Raj, "Soft mask estimation for single channel speaker separation," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA)*, 2004.

[11] R. J. Weiss and D. P. Ellis, "Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition (SAPA)*, 2006.

[12] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[13] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[14] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[15] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *International conference on acoustics, speech and signal processing (ICASSP)*, 2017.

[16] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix-a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[17] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020.

[18] N. Takahashi and Y. Mitsufuji, "D3net: Densely connected multidilated densenet for music source separation," *arXiv e-prints*, pp. arXiv–2010, 2020.

[19] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *International Conference on Learning Representations (ICLR)*, 2018.

[20] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "Melgan: generative adversarial networks for conditional waveform synthesis," in *Neural Information Processing Systems (NeurIPS)*, 2019.

[21] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," *International Conference on Learning Representation (ICLR)*, 2020.

[22] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "Gansynth: Adversarial neural audio synthesis," *International Conference on Learning Representations (ICLR)*, 2019.

[23] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Neural Information Processing Systems (NeurIPS)*, 2020.

[25] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," *Interspeech*, 2021.

[26] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *Interspeech*, 2017.

[27] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning (ICML)*, 2019.

[28] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Interspeech*, 2020.

[29] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[30] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

[31] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," in *Interspeech*, 2021.

[32] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: a unified framework for neural vocoding, bandwidth extension and speech enhancement," *arXiv e-prints*, pp. arXiv–2203, 2022.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Neural information processing systems (NeurIPS)*, 2014.

[34] E. Gusó, J. Pons, S. Pascual, and J. Serrà, "On loss functions and evaluation metrics for music source separation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 306–310.

[35] D. Stoller, S. Ewert, and S. Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2391–2395.

[36] Y. Li, B. Gfeller, M. Tagliasacchi, and D. Roblek, "Learning to denoise historical music," *International Society for Music Information Retrieval (ISMIR)*, 2020.

[37] E. Moliner and V. Välimäki, "A two-stage u-net for high-fidelity denoising of historical recordings," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[38] J. Imort, G. Fabbro, M. A. M. Ramírez, S. Uhlich, Y. Koyama, and Y. Mitsufuji, "Removing distortion effects in music using deep neural networks," *arXiv preprint arXiv:2202.01664*, 2022.

[39] S. Sulun and M. E. Davies, "On filter generalization for music bandwidth extension using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 132–142, 2020.

[40] S. Kim and V. Sathe, "Bandwidth extension on raw audio via generative adversarial networks," *arXiv e-prints*, pp. arXiv–1903, 2019.

[41] J. Pons, S. Pascual, G. Cengarle, and J. Serrà, "Upsampling artifacts in neural audio synthesis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3005–3009.

[42] J. Pons, J. Serrà, S. Pascual, G. Cengarle, D. Arteaga, and D. Scaini, "Upsampling layers for music source separation," *arXiv e-prints*, pp. arXiv–2111, 2021.

[43] N. Kandpal, O. Nieto, and Z. Jin, "Music enhancement via image translation and vocoding," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[44] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *International Conference on Computer Vision (ICCV)*, 2017.

[45] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International Conference on Machine Learning (ICML)*, 2016.

[46] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *Interspeech*, 2020.

[47] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[48] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[49] E. Cano, D. FitzGerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1758–1762.

[50] A. Défossez, "Hybrid spectrogram and waveform source separation," *ISMIR Workshop on Music Demixing (MDX)*, 2021.

[51] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2014.

[53] M. Morrison, B. Tang, G. Tan, and B. Pardo, "Reproducible subjective evaluation," in *ICLR Workshop on ML Evaluation Standards*, April 2022.

[54] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Surrey, UK*, 2018, pp. 293–305.

[55] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Python in Science (SciPy)*, 2015.

[56] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Digital Audio Effects (DAFx)*, 2013.

[57] W. Choi, M. Kim, J. Chung, D. Lee, and S. Jung, "Investigating u-nets with various intermediate blocks for spectrogram-based singing voice separation," in *International Society for Music Information Retrieval (ISMIR)*, 2020.

[58] H. Liu, Q. Kong, and J. Liu, "Cws-presunet: Music source separation with channel-wise subband phase-aware resunet," *ISMIR Workshop on Music Demixing (MDX)*, 2021.