NORTHWESTERN UNIVERSITY

A Critical Analysis of Objective Evaluation Metrics for Music Source Separation Quality

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

MASTER OF SCIENCE

Field of Computer Science

By

Erika Jianyue Rumbold

EVANSTON, ILLINOIS

September 2022

ABSTRACT

A Critical Analysis of Objective Evaluation Metrics for Music Source Separation Quality

Erika Jianyue Rumbold

Despite our world becoming more and more noisy, humans have retained from ages of evolution the ability to process multiple sounds occurring at once and focus on one that is most important. The process of parsing audio scenes in this way is a combination of multiple auditory tasks, and many researchers have taken it upon themselves to fully understand the systems involved in these tasks and to engineer systems that can replicate the process that is innate to ourselves. These studies fall into the category of *computer audition* - the study of how machines can parse audio like humans do.

A major subject under the umbrella of computer audition is *audio source separation*, or isolating sounds from a mixed audio scene. In order for audio source separation to progress, it is essential for researchers to be able to evaluate their work throughout the process; they would need a way to quantify how well the audio is isolated in order to make improvements on their source separation system.

Human evaluation is the most direct way to determine whether humans think some audio sounds good, but it is significantly expensive and time inefficient to collect enough

data. In response to these shortcomings, several error calculation and deep learning methods have been proposed and used in various computer audition tasks, including source separation. The de facto standard metric for audio source separation is signal-to-distortion ratio (SDR), which is frequently cited in audio source separation research and serves as the baseline measure for many source separation challenges (e.g., Sony Demixing Challenge [**25**]).

Despite its popularity, SDR has been proven to correlate poorly to human perception [**38, 1**]. With the objective of making audio that sounds good to listeners, it is crucial that these evaluation methods correlate well to human perception. This discrepancy has been acknowledged in the speech domain, prompting the development of new evaluation methods that achieve better correlation to human perception [**7, 8, 10, 11, 12, 13, 21, 23, 24, 40**]; but this progress has not yet been realized in the same capacity in the music domain. **In this work, I investigate existing evaluation methods for music source separation to determine if any achieves a strong enough correlation to human perception to be a reliable alternative to subjective human evaluation.**

My approach is as follows. First, I conduct a subjective listening study to acquire Mean Opinion Score data. I recruited study participants online to rate the quality of source separated audio on a scale of 1-5. Then I made observations on this subjective evaluation data, determining how well existing objective evaluation metrics correlated to the listener opinion data. Finally, I make a comparison of music source separation systems by different ranking criteria. The Papers With Code leaderboard for music source separation is ordered by average SDR output. I use the collected evaluation data to determine whether it is reliable to rank by SDR.

Acknowledgements

I would like to thank Bryan Pardo for being an incredible advisor. I have learned so much in even the brief amount of time that I have been working with you. Not only have you supported me in my research, you have shown me what it takes to *be* a good researcher. I will take your lessons with me throughout my academic journey.

Thank you to Han Liu, my thesis committee member. Thank you for taking the time out of your busy schedule to give me valuable feedback on my work.

Thank you to my current and former labmates - Ethan Manilow, Max Morrison, Patrick O'Reilly, Hugo Flores García, Boaz Cogan, Noah Schaffer, Aldo Aguilar, and Andreas Bugler. Thank you for welcoming me to the Interactive Audio Lab with open arms, helping me with my research, and sharing with me the work you are passionate about. I have learned so much from each and every one of you.

Thank you to my undergraduate professors - Doug Turnbull, John Barr, and Paul Dickson. Thank you for guiding me in my fledgling years as a scientist and believing in my future in this field. I wouldn't have made it this far without the foundation you laid before me.

I would like to thank my mother, Barbara Rumbold, the strongest person I know. Thank you for raising me with more love and care than I knew possible. Thank you for encouraging me to explore the world and follow my dreams. Thank you for supporting me, no matter where life takes me.

Thank you to my dear friends - David Kaucic, Molly North, Danny Akimchuk, Joey Garrett, and Matt Lucas. You have always been there to cheer me on, make me laugh, and let me breathe. I wouldn't have been able to do this without you. Thank you!

List of abbreviations

CNN: Convolutional Neural Network

FAD: Fréchet Audio Distance

MOS: Mean Opinion Score

MTurk: Amazon Mechanical Turk

MUSHRA: Multiple Stimuli with Hidden Reference and Anchor

SAR: Signal-to-artifacts ratio

SIR: Signal-to-interference ratio

SDR: Signal-to-distortion ratio

SI-SDR: Scale-invariant signal-to-distortion ratio

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Despite our world becoming more and more noisy, humans have retained from ages of evolution the ability to process multiple sounds occurring at once and focus on one that is most important. To illustrate this phenomenon, imagine being in a crowded stadium. You would likely hear chatter from other people – ones nearby talking normally to each other and on the opposite side of the stadium as they shout down to the referees. You could hear popcorn being popped and soda cans being opened. You could also hear a vendor calling out their cotton candy and people tossing their trash in bins. You're aware of all of these sounds around you, but if you're focused on the announcer over the loudspeaker, you might not fully register everything. However, you'd still be able to react and change focus if someone called your name.

The process of parsing audio scenes in this way is a combination of multiple auditory tasks (e.g., timbre recognition, source localization), and many researchers have taken it upon themselves to fully understand the systems involved in these tasks and to engineer systems that can replicate the process that is innate to ourselves. These studies fall into the category of *computer audition* - the study of how machines can parse audio like humans do.

A major subject under the umbrella of computer audition is *audio source separation*, or isolating sounds from a mixed audio scene. Recordings of audio scenes with multiple overlapping sounds are often referred to as *mixtures*. Common applications of source

separation include separating lead singing vocals from the rest of the musical mixture, and isolating a primary speaker from a mixture of people talking.

In order for audio source separation research to progress, it is essential for researchers to be able to evaluate their work throughout the process. Consider a research team that is developing a system that isolates lead singing vocals from the rest of the mixture, for example. A perfect system would output a new audio signal with only vocal sound, and a second signal with everything else. However, an imperfect system would result in some of the other instrument sounds appearing on the "vocals only" signal, also known as *bleeding*. To mitigate as much bleeding as possible, the research team would need a way to quantify how well the audio is isolated in order to make improvements on their system.

There are many methods of evaluating audio quality that have been implemented in existing audio research. The two main forms of evaluation method are 1) human evaluation, and 2) automated methods. The latter can be further distinguished as (a) a closed-form formula that calculates an amount of error between the separator's output signal and the expected signal, and (b) a statistical model that estimates a "goodness" score with information learned about either a data distribution [**19**] or human evaluation data [**30**]. Each type of evaluation has its own benefits and detriments, which will be discussed throughout this work. The main focus of this work is a comparison of existing evaluation metrics for audio source separation quality.

1.1. Problem Statement

A common goal in audio research fields of study is to make something that sounds good to human listeners. For example, we could give sound designers tools to make a film

sound better; or we could enhance the audio quality of President Franklin D. Roosevelt's "fireside chats"[1] so they are more intelligible and are better suited to stand the test of time.

Human evaluation is the most direct way to determine whether humans think some audio sounds good, but it is significantly expensive and time inefficient to collect enough data. In response to these shortcomings, several error calculation and deep learning methods have been proposed and used in various computer audition tasks, including source separation.

The de facto standard metric for audio source separation is **signal-to-distortion ratio (SDR)**, which I discuss further in Section 1.4.3. It is an error calculation that finds the ratio of unwanted sound (i.e., distortion) that occurs in an audio signal to the entirety of a target signal [**37**]. SDR is frequently cited in audio source separation research and serves as the baseline measure for many source separation challenges (e.g., Sony Demixing Challenge [**25**] and the MUSDB18 leaderboard on Papers With Code[2]).

Despite its popularity, SDR has been proven to correlate poorly to human perception [**1, 38**]. In other words, an audio signal that a human listener would deem poor quality may still get a good SDR value, or vice versa. With the objective of making audio that sounds good to listeners, it is crucial that these evaluation methods correlate well to human perception. This discrepancy has been acknowledged in the speech domain, prompting the development of new evaluation methods that achieve better correlation to human perception [**7, 8, 10, 11, 12, 13, 21, 23, 24, 40**]; but this progress has not yet been realized in the same capacity in the music domain.

[1] A series of evening radio addresses given by Franklin D. Roosevelt between 1933 and 1944.
[2] `https://paperswithcode.com/sota/music-source-separation-on-musdb18`

In this work, I investigate existing evaluation methods for music source separation to determine if any achieves a strong enough correlation to human perception to be a reliable alternative to subjective human evaluation.

1.2. Contributions

In this thesis, I critically analyze existing evaluation metrics for music source separation. The main contributions of this work include:

- A dataset of Mean Opinion Scores (MOS) for music tracks in the MUSDB18 music source separation dataset that can be used by others to train new separation models and separation quality evaluators[3] (Chapter 2),
- Observations of how human opinions relate to each other. In other words, I investigate whether one human's perception correlates well to that of another, and the general reliability of human opinion data (Chapter 2),
- A correlation analysis of existing music source separation evaluation metrics against human opinion, including metrics that have not previously been evaluated for correlation to human opinion (Chapter 3),
- A comparison of music source separation algorithms when ranked by different criteria. Specifically, I determine whether the rankings of these algorithms according to their SDR output is the same when ranking according to human opinion (Chapter 3).

[3]This MOS dataset can be found at `https://erumbold.github.io/nu-thesis`

1.3. Broader Impact

Reliable audio source separation can open doors for many practical applications. It would help a music producer, for example, edit an instrument that was recorded on the same microphone as another instrument, or at a live concert with a cheering audience. It could also be used by a film audio engineer to remove unwanted sounds (e.g., birds chirping, refrigerator hum) from important dialogue.

Even within the research field of computer audition, there are subfields that would benefit from higher performance of audio source separation. Music transcription is the process by which a machine takes in musical audio data and transcribes it into a visual, written form (e.g., piano roll, MIDI notation, traditional sheet music). An improved method of isolating each instrument in a mixture would in turn improve the accuracy of the notation as there would be less of other instruments bleeding into the isolated instrument.

Audio source separation affects more than just the audio industry and audio research; it can be implemented in things that everyday people use as well. For example, the experience of wearing hearing aids could be significantly improved with audio source separation research. Existing hearing aids simply make all sounds around the wearer louder. However, this can be extremely frustrating for the wearer because environmental sounds (e.g. cars passing, dogs barking) are amplified the same amount as a person to whom they want to pay attention, leading to hearing aid wearers tuning out of conversations they cannot follow. This situation could be alleviated by having the hearing aids isolate and increase the volume of only sounds that matter (e.g., a person speaking to you), and lowering the volume of everything else.

Another real-world device that would benefit from improved audio source separation is automatic speech recognition. Voice assistant systems (e.g. Amazon Echo, Apple Siri) listen for a command word (e.g. "Okay, Google"), parse your request, and execute the request to the best of their ability. These systems, however, often get confused when multiple people are talking at once. They could be improved by being able to identify unique speakers and only listen to the one who spoke the command word.

In order for audio source separation to improve these real-world applications, evaluation must occur during development and it must be consistent with the opinions of end users (i.e., humans). A team researching hearing aid attention, for example, would need to evaluate their work in progress and iterate until it is the best it can be before it can be implemented in hearing aids for the general public. Even though the effects of reliable evaluation are not directly felt by end users, it is a crucial aspect of the research and development process for these previously mentioned applications of audio source separation.

1.4. Related Work

In this section, I present three important concepts that are relevant to this work - music source separation, subjective evaluation of audio quality, and objective evaluation of audio quality. For each, I give a general overview of the subject as well as prior work related to these concepts.

1.4.1. Music Source Separation

Music source separation is the task of decomposing a musical mixture into its individual components. For the MUSDB18 dataset [**27**] that I use in this work, these components

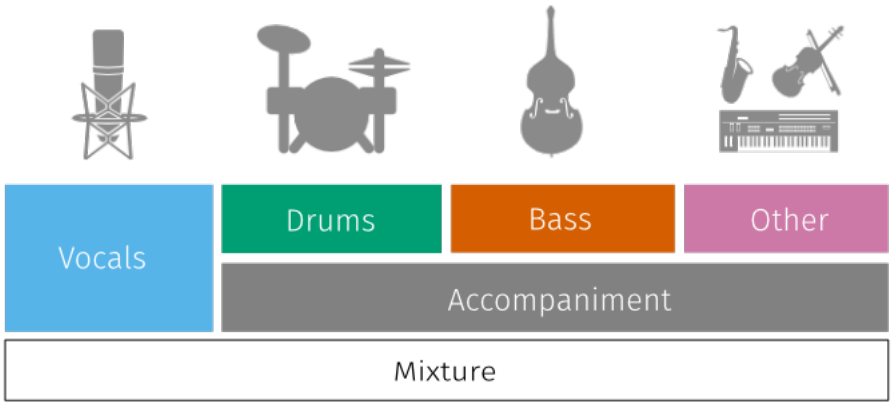
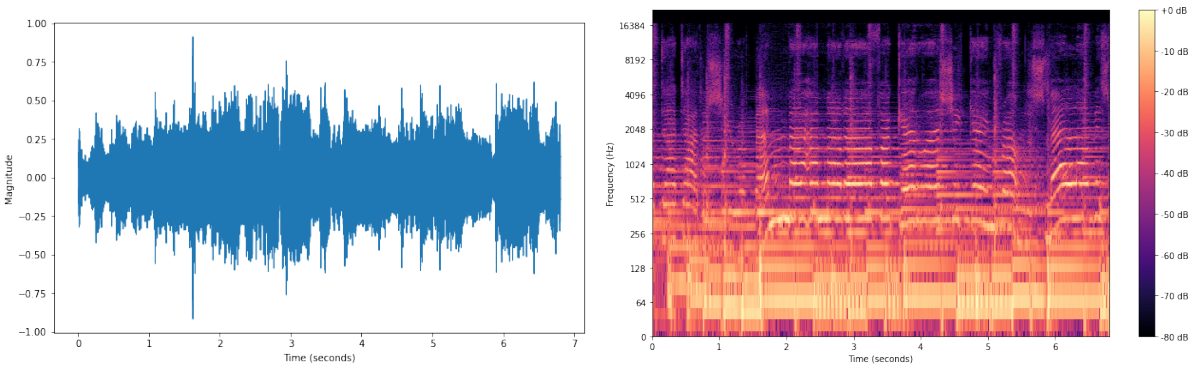


Figure 1.1. A representation of the components of an audio mixture relevant to music source separation

are defined as **bass**, **drums**, **vocals**, and **other**. These components are also referred to as *stems*, or general groupings of similar instruments that appear a mix. For example, the **vocals** stem would include any background vocals in addition to the lead singer. Given a mixture of these four stems, the goal is to generate four audio files, or *waveforms*, that correspond to each of the original stems.

Source separation typically is performed on one of two representations of audio data - waveform and spectrogram. Audio waveform data, as shown in Figure 1.2a, presents the audio's amplitude, or loudness, as a function of time. A spectrogram, as shown in Figure 1.2b, is a type of time-frequency representation that shows the magnitude, or power, of each frequency channel at each time interval.

Furthermore, there are two common source separation methods - building a mask on a spectrogram and estimating waveforms for each stems [**28**]. Consider, for example, the task of separating the **vocals** from the rest of the mixture. In a masking context, an



(a) Waveform (b) Spectrogram

Figure 1.2. One audio example shown in two forms of representation for audio data: (a) waveform and (b) spectrogram

array M is made in the same shape as the mixture audio data, usually represented as a spectrogram. The values of M are 0 where **vocals** are not present, or 1 where **vocals** are present in the mixture; this is known as a *ideal binary mask*. One could also make an *ideal ratio mask*, which consists of values 0.0-1.0 corresponding to the power of the **vocals**. The separated output is produced by multiplying the mixture by this mask M, resulting in frequencies other than **vocal** frequencies being reduced, or *masked* out.

Modeling methods of audio source separation employ neural networks to predict the power spectrogram or waveform for each stem. Many network architectures have been used before, including simple fully connected networks [**36**], convolutional networks [**34**], and U-Net architectures [**14**]. Neural networks like these are able to take data as either waveforms or spectrograms, although models operating in the waveform domain generally do not perform as well as those in the spectrogram domain [**6**].

1.4.1.1. Source Separation Models Used in this Work. I used five different source separation models to create separated stem data to be evaluated by listeners and by objective methods of evaluation. These models were selected from the Papers With Code

leaderboard for separation on the MUSDB18 dataset. The current rankings of these separators is shown in Table 1.1. The application of these models in this work is further described in Section 2.1.2.

Source Separation Models Rankings on MUSDB18		
Rank	**Model**	**SDR**
1	Hybrid Demucs	7.68
3	Demucs-Extra	6.79
10	D3Net	6.01
11	Spleeter	5.91
18	Source Separation Wavenet	3.5

Table 1.1. Music source separation models used in this work, ranked by their performance on the MUSDB18 dataset according to average SDR over all stems, as shown on the Papers With Code leaderboard

The first type of model I will discuss is spectrogram-to-spectrogram models, meaning they take input data in as spectrograms and output spectrograms for the separated stems. Spleeter [**14**] is a collection of three music source separation models, each optimized to separate mixtures into different types of stems. These models can be referred to as 2-stem (vocals and accompaniment), 4-stem (vocals, bass, drums, and other), and 5-stem (vocals, bass, drums, piano, and other). To remain consistent with the rest of this work, we will only consider the 4-stem model. The Spleeter models are U-nets [**18**], or an encoder/decoder Convolutional Neural Network (CNN) architecture with skip connections, with 6 layers each for the encoder and decoder. This network is tasked with estimating a ratio mask for each stem. It's trained with an Adam optimizer and L1 normalization between the masked spectrograms of the input mixture and the target spectrogram for each stem. Spleeter is the only music source separation model that I used that was not trained on the MUSDB18 dataset.

The second spectrogram-to-spectrogram model is D3Net [**35**], or the (**D**)ensely connected multi(**d**)ilated (**D**)enseNet for music source separation. The base of D3Net is the DenseNet [**17**], or densely connected convolutional networks. The DenseNet architecture connects every other layer of a CNN in a feed-forward fashion. D3Net combines the DenseNet with dilated convolution, which is a convolution where the filter is applied over an area larger than its length by skipping input values with a certain step [**26**]. Dilated convolutions allow the base network to cover a large receptive field with a small number of layers; this is important because audio data can have long time and wide frequency dependencies. The D3Net architecture is comprised of nested dilated dense blocks in order to apply different dilation factors multiple times and ensure a sufficient depth is achieved by the network.

There are fewer music source separation models that work in the waveform domain. Regardless of the lack of availability, I have chosen two waveform-to-waveform separation models for this work. The first model was developed by Lluís, et al. [**20**] and was adapted from the generative model for raw audio Wavenet [**26**] for the task of music source separation. Throughout this thesis, I will refer to the network from Lluís, et al. [**20**] as "Source Separation Wavenet." The architecture starts with a 3x1 CNN layer that linearly projects the input waveform to k channels. This projection is then processed by a series of dilated CNN layers. Two final, non-dilated CNN layers of size 3x1 adapt the resulting feature map dimensions. The output layer linearly projects this feature map into three channels, one for each of **bass**, **drums**, and **vocals**. The **other** stem is computed by subtracting the three estimated stems from the original mixture. The architecture of Source Separation Wavenet is shown in Figure 1.3.

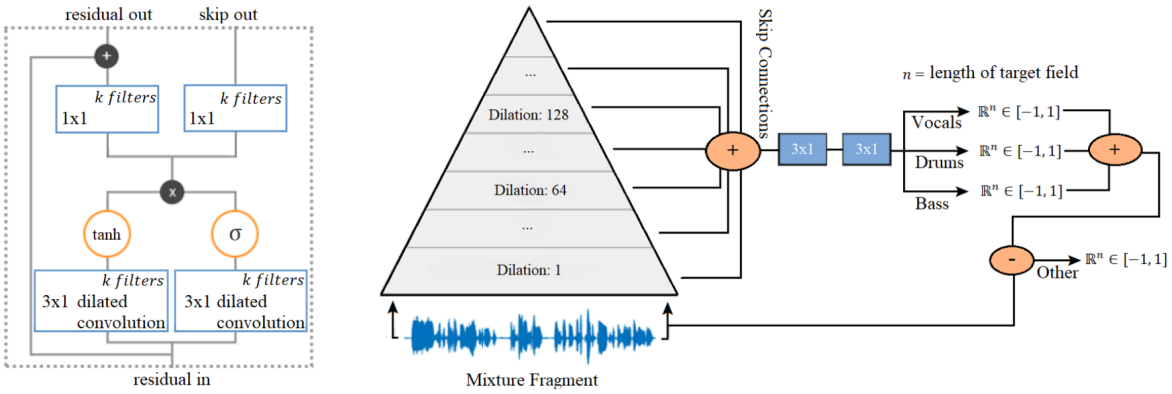


Figure 1.3. Source Separation Wavenet model architecture from [**20**]. **Left** - Residual layer. **Right** - Overview of the model.

Demucs [**6**] is the next waveform-to-waveform separation model that I used. The architecture was adapted from Conv-Tasnet [**22**], a model that was originally designed for monophonic source separation of speech. In addition to the base model, Demucs is inspired by models for music synthesis rather than masking. It is a U-net architecture with a convolutional encoder and decoder based on wide transposed convolutions with large strides. Demucs also includes bidirectional LSTM between the encoder and decoder. In order to adapt the original Conv-Tasnet for stereophonic music source separation, Défossez, et al. needed to increase the receptive field of the network. Conv-Tasnet had a receptive field of 1.5 seconds of audio sampled at 8 kHz. For reference, music audio data is commonly sampled at 44.1 kHz or 48 kHz. This increase in receptive field is achieved by increasing the kernel size of the encoder and decoder, resulting in the same receptive field at 44.1 kHz. While Conv-Tasnet was designed for short sentences of no more than a few seconds, Demucs achieved its best performance when source separating input audio that was 8 seconds long. The model architecture for Demucs is shown in Figure 1.4. In

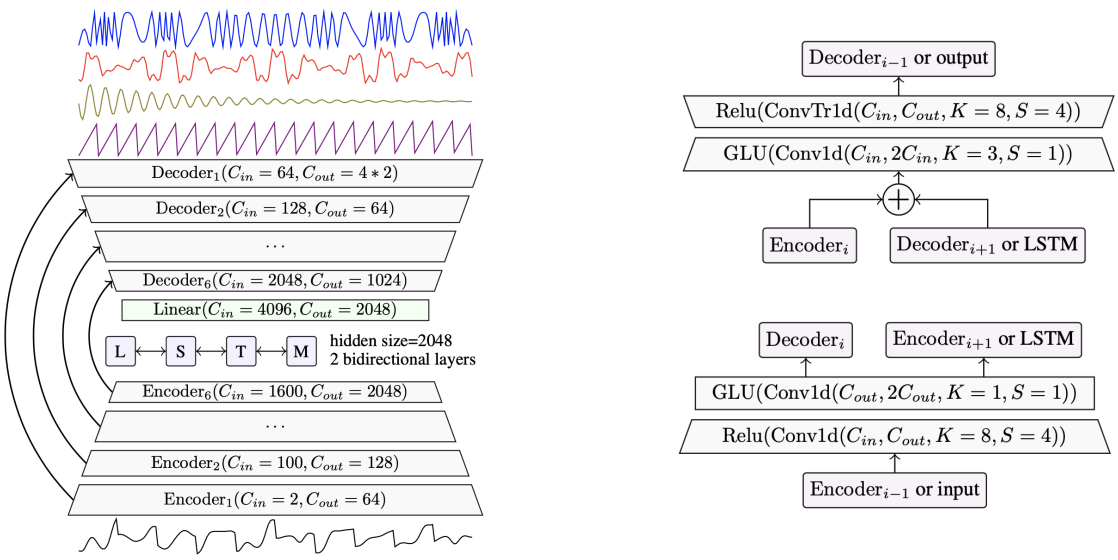


Figure 1.4. Demucs model architecture from [**6**]. **Left** - full architecture. **Right** - detailed representation of the encoder and decoder layers.

this work, I use the Demucs-Extra model version, which has the same architecture but was trained on MUSDB18 + additional data. I chose to use this version because of its ranking on the MUSDB18 leaderboard.

The fifth and final music source separation model I used in this work is a hybrid, spectrogram and waveform model and is an improvement upon the Demucs model described above. Referred to as Hybrid Demucs [**5**], this model is a dual U-net that is comprised of a temporal branch, a spectral branch, and shared layers. The temporal branch takes in waveform data and handles it like the Demucs-Extra model. The spectral branch takes in spectrogram data and reduces the frequency dimension by applying the same convolutions as in the temporal branch, but along the frequency dimension. The temporal and spectral representations are then summed before being passed through a shared encoder/decoder layer. The output of this shared decoder layer is passed as the input to the separate

temporal and spectral decoders. The summation of these decoders' outputs is the final model prediction. The hybrid design allows the model to use whichever representation is better for different parts of the signal, even within one source. The Hybrid Demucs model architecture is shown in Figure 1.5.

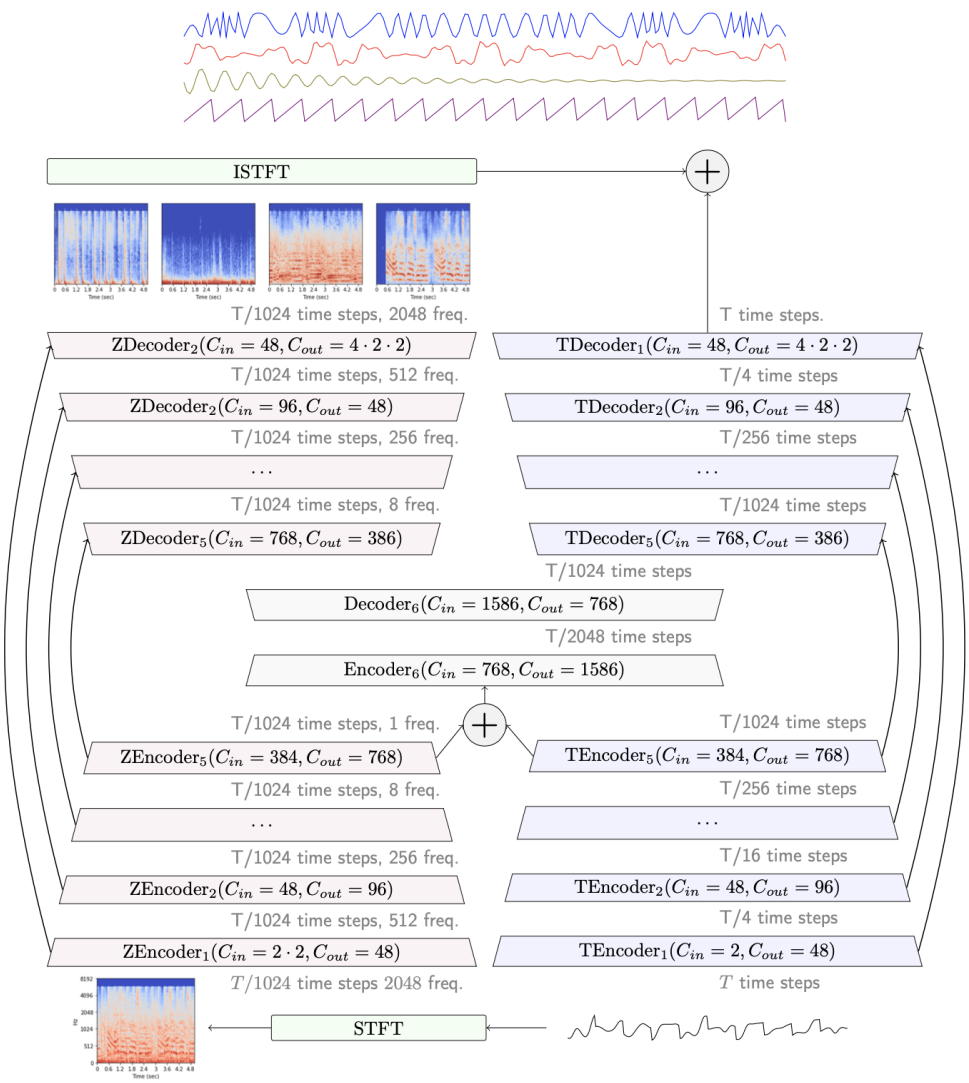


Figure 1.5. Hybrid Demucs architecture from [**5**]. The Z prefix is used for spectral layers, and T prefix for the temporal ones.

1.4.2. Subjective Evaluation of Audio Quality

Subjective evaluation of audio quality refers to the rating of audio stimuli by human subjects. Participants of a subjective listening study may be asked to rate the audio generally (i.e., "How good does the audio sound?"), or they could be asked to evaluate a specific attribute (e.g., level of interference, intelligibility, or musical intonation). In addition to the types of questions that can be asked, there are several evaluation protocols that are commonly used in audio research. These protocols are characterized by two main attributes - 1) whether the participant is asked to rate an audio stimulus individually or in comparison to other stimuli, and 2) whether the data is observed as a numerical rating for each stimulus, or a count of how many times a stimulus was selected out of a group of stimuli.

1.4.2.1. Subjective Assessment Protocols. One of most common subjective assessment method in audio applications is **Mean Opinion Score (MOS)**. Participants in an MOS assessment are asked to rate stimuli on a rational number scale, usually one as follows: 1-Bad, 2-Poor, 3-Fair, 4-Good, and 5-Excellent. At first glance, this rating system seems straightforward. However, the quantization into of the five discrete values imposes limits and a given participant may interpret the value "Good" differently than another [**4**]. Chen, et al. [**4**] also point out that the 1-5 point scale is assumed to be on an interval scale, but it more realistically acts like an ordinal scale. People tend to have a different cognitive distance between 1-Bad and 2-Poor than they do between 4-Good and 5-Excellent [**39**]. To validate this claim, it would be necessary to perform an experiment that directly compares the accuracy of the MOS rating method to pairwise comparison, whether crowdsourced or not. To my knowledge, this has not been done.

Pairwise comparison, or **AB testing**, asks participants to select one of two presented audio stimuli that more accurately fits the evaluation criteria (e.g., sounds better, appears to have less noise). This may also be modified to be an **ABX test**, which provides a reference stimulus in addition to the two being compared. An ABX test typically asks which of the two examples is most similar to the reference.

Pairwise comparison is easier to understand than the MOS 1-5 point scale because participants aren't required to mentally assign meaning to five different ratings [**4**]. Pairwise comparison also allows for easy consistency validation through the transitive property. If a participant that rates stimulus A better than B and B better than C, one could assume that they would in turn rate A better than C [**4, 2**]. This makes it easier to eliminate erroneous data from malicious or negligent selections since the transitive property is fairly quick to assess.

There are many scenarios in which MOS, AB testing, or ABX testing are optimal. However, due to the simplicity of these rating systems, these protocols are not ideal for applications in which fine-grain, nuanced data are required. One protocol that is also used in audio evaluation that achieves this level of detail is MUltiple Stimuli with Hidden Reference and Anchor, or **MUSHRA** [**32**]. MUSHRA presents anywhere between 3 and 12 stimuli for participants to rate comparatively, including a *reference* that is unlabeled and hidden among the other stimuli, as well as a hidden, unlabeled *anchor* stimulus that is an intentionally bad sound. Participants are asked to rate each stimulus using a set of sliders on a 1-100 point scale. It is expected that the *reference* be rated high and the *anchor* be rated low. Standard MUSHRA is highly regulated, calling for a listening

environment that meets certain criteria, specifying the training procedure, and requiring participants to meet certain qualifications.

1.4.2.2. Crowdsourcing. Studies to acquire subjective data traditionally have been conducted in a laboratory, in which researchers can ensure a controlled environment, such as the required environmental features of standard MUSHRA. However, recent efforts have been made to adapt subjective quality assessments to an online crowdsourced format [**4, 2, 29, 3**].

In this work, I conducted an online, crowdsourced MOS study, details of which can be found in Section 2.2. For the purposes of these experiments, it was best to collect rating data that was scored on a numbered scale. Furthermore, it was not necessary to conduct a MUSHRA assessment, which is more complex and intensive than MOS, since I was not concerned with analyzing the minute differences in music audio quality, but rather more general evaluations of quality.

Typically, lab-based audio tests require participants to complete assessments in rooms that meet specific acoustic qualities, using the same technology as all other participants (e.g. headsets, operating systems, etc.). Some protocols additionally restrict participants to those that meet certain criteria such as level of expertise in an audio-related field or not having been diagnosed with a hearing disorder [**32**]. These requirements eliminate the possibility of a participant's evaluation being affected by extraneous noise or a lack of understanding of the task. On the downside, conducting tests in a lab costs a significant amount of both time and money. Researchers must dedicate time to supervising trials done by each participant, and each participant must be compensated for their time. The

availability of time and money also limits how many assessments can be acquired. Furthermore, subjective assessment trials that take place in a lab attract fewer participants, leading to results that are less statistically significant [**29**].

These drawbacks led to the rise in subjective evaluations taking place online. Many efforts have been made to adapt subjective assessment protocols to an online, crowdsourced environment [**29, 2, 4, 3**], forgoing some of the strict participant eligibility criteria and environmental control in favor of acquiring a larger and more diverse results set. Moving these assessments online is also more cost effective; and it takes less time to acquire data than it takes in a lab. Traditional MUSHRA trials, for example, can take several hours for each participant to complete, depending on the number of trials completed. Online evaluation tasks, on the other hand, are designed to be completed much more quickly. Lab-based assessments typically have a small number of people evaluate a lot of things, whereas online assessments are taken by many more people, but each usually completes only a few tasks.

A shortcoming of online assessments is that they cannot be directly monitored, making it possible for results to be affected by the listening environment, the equipment used, or the participants' integrity. These effects cannot be screened ahead of time, but there are a few methods that can be implemented to filter online study results. For example, the CrowdMOS platform for crowdsourcing MOS studies [**29**] asks participants for the type of listening device they used during the study (e.g., headphones, laptop speakers). It is expected that a person listening on speakers would not be able to hear finer details of audio as acutely as listeners using headphones. Cartwright, et al. also asked about listening device in their web-based MUSHRA assessment [**3**], as well as the quietness of

the room in which the participant completed the study. Asking these questions allows the researchers to eliminate data from participants that do not fit their criteria (e.g., using headphones, being in a quiet environment). The online assessment could also contain a hearing test to assess the participants' hearing capabilities. For example, the MUSHRA assessment from Cartwright, et al. features two hearing tests which require listeners to report how many tones they hear in a sequence. This sequence always includes a tone pitched at 55 Hz and at 10 kHz tone with up to 6 other tones being between those pitches. It is expected that a listener completing the study in a noisy room or with an inadequate listening device would not be able to hear the 55 Hz or 10 kHz tone. Researchers can also hide anchor questions within the survey, as is already the practice in MUSHRA. The answer should be obvious, so researchers can easily identify participants that did not understand the directions or intentionally submitted inaccurate responses. If participants do not answer these anchor questions correctly, their data can be eliminated.

Despite the need to prune crowdsourced results, crowdsourced assessments can achieve comparable results to those of their lab-based equivalents while costing significantly less and being quicker to execute [**4**]. More time may be necessary to screen crowdsourced results, but this is usually done computationally and does not take a significant amount of time from the researchers like in-lab assessments.

1.4.2.3. Choosing the Right Protocol. Each subjective assessment protocol has its benefits and best use cases. For example, one could observe more minute differences between stimuli with MUSHRA data; or if a researcher need only compare two stimuli, they may opt for an AB assessment instead. A researcher could also use the same protocol to answer different questions. For example, they could present an MOS assessment in

which participants must answer "How good did the audio example sound?" or they could ask "How clearly could you hear the vocals?" Each question could result in different ratings, with the former question being much more broad and up to interpretation by the participant.

1.4.3. Objective Evaluation of Audio Quality

Objective evaluation of audio is achieved without human subject data. Therefore, objective evaluation metrics are more practical for researchers to use, being significantly quicker and cheaper to execute than a subjective evaluation study. There are many methods of evaluation that can be applied to music source separation, and they typically take one of two forms: 1) closed form equations that compute an amount of error, or 2) models that predict an audio quality rating. Objective evaluation methods can also be classified by whether or not they require a ground truth signal to which the source separation model's output can be compared; every closed form equation method requires a ground truth.

1.4.3.1. Closed Form Evaluation Methods. Signal-to-distortion ratio (SDR) [**37**] is the current standard evaluation metric for music source separation. An estimate of source $\hat{s}_i$ is assumed to be composed of four separate components,

$$\hat{s}_i = s_{target} + e_{interf} + e_{noise} + e_{artif}$$

where $\hat{s}_i$ is the true source, and e_{interf}, e_{noise}, and e_{artif} are terms for interference, noise, and artifacts, respectively [**37**]. From these attributes, we are able to compute four energy ratios by the relation of these terms to the true source. Cano, et al. [**1**] represent the four measures as follows: signal-to-artifacts ratio (SAR), or the amount of unwanted artifacts

present in a source estimate in relation to the true source,

$$\text{SAR} = 10\log_{10}\left(\frac{||s_{target} + e_{interf} + e_{noise}||^2}{||e_{artif}||^2}\right) \tag{1.1}$$

signal-to-interference ratio (SIR), or the amount of other source that can be heard in the source estimate,

$$\text{SIR} = 10\log_{10}\left(\frac{||s_{target}||^2}{||e_{interf}||^2}\right) \tag{1.2}$$

and SDR, or the overall measure of how good the source estimate sounds in comparison to the true source.

$$\text{SDR} = 10\log_{10}\left(\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2}\right) \tag{1.3}$$

These four evaluation measures together are known as the Blind Source Separation Evaluation Toolkit (BSSEval). Each of these measures is in decibels (dB), and higher values are better. These equations also assign equal weights to the different error terms. So it is assumed that each type of distortion contributes equally to the overall quality of the source $\hat{s}_i$ [**1**].

Since the original proposal of SDR, several issues with the metric have been discovered, including an easy way to boost one's scores by changing the amplitude scaling of source estimates. This prompted Le Roux, et al. [**31**] to propose a version of SDR that is not dependent on amplitude scaling, SI-SDR. They first rescale the target s by finding the orthogonal projection of the estimate $\hat{s}$ on the line spanned by s. The scaled reference is denoted as e_{target}, which allows us to break down the estimate $\hat{s}_i$ as $\hat{s}_i = e_{target} + e_{res}$.

From this, we can define SI-SDR by the equation

$$\text{SI-SDR} = 10\log_{10}\left(\frac{||e_{target}||^2}{||e_{res}||^2}\right) \tag{1.4}$$

As with SDR, SI-SDR is measured in decibels (dB), and higher values are better. And despite the potential improvements SI-SDR has over SDR, SDR remains the standard evaluation metric for the task of music source separation.

Although SDR is the established standard, most loss functions that are normally used in neural network training can also be used to evaluate the quality of audio source separation. For example, L1 and L2 losses can be used to evaluate the similarity between an estimated signal and the target signal. These loss functions have been previously implemented in music source separation, being parts of the training architectures for Demucs [**6, 5**] and Spleeter [**14**]. L1 can be observed as the absolute error, and L2 as the squared error. Given a target signal s and an estimate signal $\hat{s}$, the two loss functions can be expressed as the following:

$$L1 = |s - \hat{s}| \tag{1.5}$$

$$L2 = (s - \hat{s})^2 \tag{1.6}$$

In the context of music source separation, these calculations are typically done on the power spectrograms of the target and estimate signals, and lower values are better.

1.4.3.2. Audio Quality Prediction Models. The second type of objective evaluation is an audio quality predictor, or in other words, a non-human system (i.e., neural network) that is trained using existing audio evaluation data to predict the evaluation of other audio

stimuli. These evaluation methods can be further distinguished by the type of data on which they are trained. The PEASS Toolkit[4] [**10**] and MOSNet [**21**] are two evaluation systems that are trained on human data - MUSHRA and MOS, respectively - to predict quality scores for the input audio. Alternatively, audio quality predictors can be trained on data that is another quality evaluation model's output. For example, Quality-Net [**11**] is a speech quality assessment model that is trained on PESQ[**30**] data and outputs a PESQ score prediction for an audio input. PESQ, or Perceptual Evaluation of Speech Quality, is a model developed for telephone networks and codecs that predicts Mean Opinion Score.

A shortcoming of audio quality prediction models operating in the music domain that is being addressed in the speech domain is the requirement of an available target signal, or ground truth separated signal. The previously mentioned speech models, Quality-Net and MOSNet, are two examples of evaluators that only take as training inputs estimated signals and their PESQ or MOS scores, respectively. PEASS, however, requires the target signal of each stem and the estimated mixture signal as inputs. This is a significant issue when no target audio is available.

To my knowledge, a prediction model that acts like Quality-Net or MOSNet does not exist for music source separation; that is, a model that only takes estimate signals and their ratings as inputs. However, an alternative approach to a "referenceless" model is given by Fréchet Audio Distance (FAD) [**19**]. Inspired by Fréchet Inception Distance (FID) [**16**], which was developed to evaluate generative models for images, FAD compares

[4]PEASS can operate in both the music and speech domains.

statistics computed on a set of estimate signals to reference statistics computed on a large set of studio recorded music.

FAD uses the VGGish [**15**] model to generate embeddings for the reference set and the evaluation set. Like how Fréchet Audio Distance is derived from Fréchet Inception Distance, VGGish is derived from the VGG image recognition architecture. Multivariate Gaussians are computed on both the evaluation set embeddings $N_e(\mu_e, \Sigma_e)$ and the reference embeddings $N_r(\mu_r, \Sigma_r)$; and Dowson, et al. [**9**] define the Fréchet distance between two Gaussians as:

$$F(N_b, N_e) = ||\mu_b - \mu_e||^2 + tr(\Sigma_b + \Sigma_e - 2\sqrt{\Sigma_b \Sigma_e}) \tag{1.7}$$

where tr is the trace of a matrix. As a distance measure, lower FAD scores are better. FAD was developed for the task of music enhancement, but the metric could still be effective at evaluating music source separation.

1.5. Limitations of Existing Work

With the general goal of producing audio that sounds good to human listeners, it's imperative that humans agree with the evaluations of these objective metrics. Subjective evaluation data will of course be similar to the opinions of human listeners because the data comes directly from human listeners. However, it is highly expensive and time inefficient to conduct a human listening study. Whether online or in a lab, assessment participants must be compensated for their time. And although it is cheaper to conduct studies online, the costs can quickly compound as more data becomes necessary. Regarding the cost of time, lab-based experiments can take hours for one participant to complete;

and for each hour of participant time, there are also hours taken away from the researcher or other proctors to monitor the assessment.

These costs can be alleviated by a considerable amount when transferred to an online setting, where a proctor is not required to dedicate time and participants can choose to stop taking surveys at their own discretion, earning payment accordingly. However, this may result in inconsistent data since online assessments tend to require more participants than in-lab assessments, meaning consistency between participants' answers becomes less likely. Furthermore, assessment participants, whether in a lab or online, are not necessarily obligated to complete an assessment with integrity; they could erroneously give random ansewrs to complete the assessment more quickly. A researcher could discourage this behavior by removing the incentive for participation (i.e., denying payment to those who clearly gave unreliable answers), but they would still need to spend more time conducting experiments in the lab, or publishing more assessments online.

An example of the costs associated with a subjective assessment study comes from the online study I conducted, which is fully discussed in Section 2.2. It took five days and $1,288.40 to acquire 4,500 responses. This is a significantly small dataset for a typical computer audition experiment.

On the other hand, objective audio quality evaluation data are considerably easier to acquire than subjective data; objective methods are more cost effective and require little to no human participation. However, Cano, et al. [**1**], and Ward, et al. [**38**] each have shown that existing objective evaluation metrics for music, such as SDR and PEASS, do not correlate well to human perception. This poor correlation between objective and subjective assessments indicates that objective evaluation methods for musical audio quality

are not entirely effective at achieving the goal of matching what humans think sounds good.

It's important to note that neither Cano, et al. nor Ward, et al. uses the most widely-used dataset for music source separation, MUSDB18 [**27**], in their studies. Because a significant number of source separation systems are trained on MUSDB18, these studies are not fully effective in showing the music source separation community how ineffective evaluation metrics like SDR really are. Furthermore, the only metrics observed by these studies were BSSEval, PEASS, and subjective listening assessments. I am not aware of a correlation analysis or comparison of metrics that includes newer metrics (e.g., Fréchet Audio Distance [**19**]).

Furthermore, the existing objective evaluation methods for music source separation discussed above require a target signal for comparison. However, this is not always feasible given the context. For example, source separation done on a band's live concert recording would not be able to be evaluated by metrics like PEASS or SDR because a clean recording of each instrument, played in the exact same way as they were performed live, would not exist. And although Fréchet Audio Distance does not have this issue, the correlation of FAD to human opinion has not been observed in previous work.

1.6. Approach

My approach to a critical analysis of objective source separation evaluation metrics has the following major steps:

(1) **A subjective listening study to acquire Mean Opinion Score data for the MUSDB18 dataset.** Using five selected music source separation models,

I source separated 30 songs from MUSDB18. I then recruited study participants on Mechanical Turk (MTurk) to rate how good these stems sounded on a scale of 1-5.

(2) **An analysis of these MOS data.** I investigate how well correlated are to each other by comparing study participant responses to the responses of others.

(3) **A correlation analysis of existing objective evaluation metrics to the subjective MOS data.** I determine how well each of the observed objective metrics ratings correlates to the MOS ratings collected from the listening study.

(4) **A comparison of music source separation systems by different ranking criteria.** The Papers With Code leaderboard for music source separation on MUSDB18 is ordered by average SDR output. I observe whether the source separation systems maintain the same order when ranked according to MOS ratings. If it does not, it would indicate that SDR is not a reliable ranking criteria for music source separation systems. Furthermore, if another metric ranks source separators more similarly to MOS, it would raise the question of whether it is a better ranking criteria for the leaderboard than SDR.

CHAPTER 2

Conducting a subjective evaluation study

This chapter covers the steps taken to collect Mean Opinion Score data for MUSDB18 songs separated by five different source separation models. First, I discuss the MUSDB18 songs I selected to be evaluated, the separators I used, and the reasoning for these choices 2.1. I then present the procedure for the subjective listening assessment I conducted to acquire these MOS data 2.2. Finally, I analyze the acquired data - discussing trends that appear and how human listeners relate to each other 2.3.

2.1. Audio Data

It is important that data used in training an evaluation metric meet the following criteria: 1) the data are suitable for the task, i.e., music source separation; and 2) there is a sufficient amount of data to train on. In these experiments, I also take into consideration the criteria that the data include examples from various genres of music and examples that feature an even distribution of male and female vocals. These additional criteria help ensure that the metric does not overfit to any given genre or voice type. In other words, a metric trained on music in genres such as Rock, Pop, Electronic, Bluegrass, Punk, Orchestral, etc. would be better suited to evaluate Ragtime music than one that is only trained on Pop and Rock. The accuracy of the metric is also affected by the types of voices on which it is trained. For example, a metric trained mostly on male singing

voices may mistake higher frequencies heard in a recording of a female singer as noise, rather than musical data, and thus inaccurately score the example as having low quality.

2.1.1. MUSDB18

In consideration of the above criteria, I have chosen to use a subset of MUSDB18 [**27**] - a corpus for music separation. MUSDB18 consists of 150 stereo mixtures of songs, about 10 hours of data, that span a variety of genres. The song files are encoded at 44.1kHz and in the Native Instruments stems format. This multitrack format is composed of five stereo streams corresponding to the mixture, drums, bass, other, and vocals. In addition to meeting the criteria defined above, MUSDB18 is among the top, most-cited datasets for existing work in music source separation; it is also the sole dataset for many source separation competitions like the Music Source Separation leaderboard on Papers with Code[1] and the Sony Music Demixing Challenge [**25**].

From the 150 songs in the MUSDB18 dataset, I curated a subset of 30 songs that met the aforementioned criteria of 1) representing a wide range of musical genres and 2) striking a balance between male and female singers. Details about these tracks are depicted in Appendix A. One may argue that it would be better to take a random selection to mitigate bias. However, MUSDB18 skews in favor of male singers and the Pop/Rock genre. As such, it would be likely that a random selection would take a similar form. Manual curation makes it easier to represent the genres and vocals that are less common in the MUSDB18 data set as a whole.

[1] `https://paperswithcode.com/sota/music-source-separation-on-musdb18`

28 of the chosen 30 songs were selected from the testing set of MUSDB18. The source separation algorithms I used in these experiments, described in Section 2.1.2, were all trained on the MUSDB18 training set. So it would be expected that separation of songs from the training set would be considerably better. Therefore, I have chosen to draw the majority of the sounds used in this study from the MUSDB18 testing set, which is not necessarily expected to be separated well. The two songs that I've included in the experimental data were chosen to provide representation of a genre or voice part that was lacking in the testing set. None of the songs in the testing set was Jazz, so I added the Jazz song "A Reason To Leave" by Patrick Talbot from the training set. Also, the curated data set was lacking fast tempo songs sung by a female singer. To account for this, I added the training set song "One Minute Smile" by Actions, as well. These additions are denoted in Table A.1.

2.1.2. Source Separation

Each of the 30 MUSDB18 songs was truncated to a 7 second segment. Audio clips that are 7 seconds long are short enough to be separated efficiently while also being long enough for listeners to effectively evaluate.

These segments were auditioned to ensure that the clip contained enough of each stem, **bass**, **drums**, and **vocals**, to be evaluated. The 30 songs were source separated using Hybrid Demucs [**5**], Demucs-Extra [**6**], D3Net [**35**], Spleeter [**14**], and Wavenet [**20**]. I chose these five separation algorithms due to their rankings on the Papers with Code leaderboard[2], seeking algorithms that represented the top, middle, and bottom

[2]Rankings listed are determined by the average SDR over all stems. SDR scores for each stem individually are also provided on Papers with Code.

tiers. At the time of writing, there were 19 separation algorithms on the leaderboard. Hybrid Demucs and Demucs-Extra held 1st and 3rd places - representing the top tier of algorithms. D3Net and Spleeter were mid-tier algorithms, placed at 10th and 11th; and Wavenet was ranked 18th, thus being the bottom-tier algorithm. In addition to comparing the correlations of existing metrics to human perception, we can determine the reliability of an SDR-based leaderboard when considering human perception.

Each of these separators outputs tracks corresponding to the stems **bass**, **drums**, **other** and **vocals**. I chose to disregard the **other** track in this experiment due to its ambiguity. **Other** could contain a keyboard synthesizer or a harp - a solo saxophone or an entire string orchestra. There are countless instruments and quantities of these instruments that could be placed on the **other** track, so it would be extremely difficult for a model to learn how to evaluate all of the different possibilities.

Ignoring the **other** track leaves us with 30 tracks separated by five source separating systems into three stems, or a data set of 450 stems to evaluate.

2.2. Listening Assessment

I sought to collect Mean Opinion Score (MOS) data as the training targets for the audio data described in Section 2.1. I conducted an MOS study in which participants were asked to rate two separate attributes of the audio that was presented - the level of **other instruments** present, and the level of **artifacts** present. To clarify the term for participants without audio training, I define **artifacts** in the subjective assessment study's introduction as "extra sound that cannot be recognized as a musical instrument or voice."

2.2.1. Participants

Participants in the subjective assessment study were recruited and paid through Amazon Mechanical Turk (MTurk), a platform for crowdsourcing user studies. They were paid $2.00 for each 10-question study they completed. Participants were required to be at least 18 years of age, which is enforced by MTurk, and they were strongly encouraged to complete the study with headphones or earbuds in a quiet environment.

2.2.2. Procedure

The subjective assessment started with a hearing screening similar to the screening defined by Cartwright, et al. [**3**] in their online MUSHRA assessment. Participants were first asked to adjust the volume of a 1000 Hz sine wave to a comfortable level and encouraged to not change the level afterward. They then listened to two 8 second audio clips and counted how many separate sine wave tones they heard. Each clip contained at least a 55 Hz and a 10 kHz tone, with the possibility of up to six more tones between 55 Hz and 10 kHz. It is expected that a participant in a suitable listening environment with an appropriate listening device should be able to hear the 55 Hz and 10kHz tones. Participants had three attempts to answer both screening questions correctly. Incorrect answers would be followed by a prompt for the participant to change their listening environment or device and try again. Failing this check three times would prompt the participant to submit their responses; they would not be able to view the rest of the study and they would not be compensated.

Following the hearing screening, a description of the rating system was given to the participants who passed the hearing test. It was explained that audio clips were to be

rated on a 1-5 scale where **1** indicated **Bad** - a lot of sound from other instruments or artifacts, and **5** indicated **Excellent** - no sound from other instruments or artifacts. They were then presented with example audio clips and descriptions of what is meant by presence of other instruments and presence of artifacts.

Each assessment consisted of 10 audio clips of the same stem type - **bass**, **drums**, or **vocals**, and no audio clip was repeated across the published assessments. The assessments were released in batches grouped by stem type; so one batch would only contain audio clips of **drums**, for example. A participant could decide to complete each assessment in the batch, or just a few. MTurk does not have the capability to randomize the order in which assessments appear in a batch; so to ensure the latter assessments were taken enough times, an assessment in the batch was made unavailable when it had been completed by 10 participants. Assessments that were submitted with a failed hearing test were republished until it had been completed by 10 participants who passed the hearing test.

The minimum number of participants necessary would be 10, if each participant completed all 15 assessments in all three stem-grouped batches; and the maximum number of participants would be 450, if each participant completed only one of the 45 assessments. It would have been possible to add more than 10 questions to each assessment, thus requiring fewer individual participants. However, feedback I received before publishing the study indicated that participants might feel fatigued after 10 questions, resulting in inconsistent or inaccurate data. Therefore, I have decided to prioritize quality of responses over the ease of having fewer participants.

For each of the 10 questions on an assessment, participants were asked to listen to a 7 second audio clip in its entirety, then separately rate the level of **other instruments** and

artifacts present in the clip. An example of a question as it appeared on the assessment is shown in Figure 2.1. At the end of the assessment, participants were asked to report which listening device they used to complete the assessment and a rating on a 1-5 scale of how quiet their listening environment was throughout the assessment, where 5 meant no noise and 1 meant extremely noisy. They were provided spaces to report any changes to their listening environment that may have occurred, as well as any additional comments.

One assessment took on average 26 minutes and 37 seconds to complete. This is longer than I had expected since the test subjects who trialed the study before I published it were able to complete an assessment in about 15 minutes. One explanation for the longer assessment duration is that MTurk users often open multiple tasks at once; so the listening assessment could have be open in the background while participants completed other tasks, thus inflating the time it took to complete my assessment. There was also a more extensive declaration of research intent, rights, and consent at the beginning of the published study than in the trial versions. So participants of the published study may have spent more time reading this text.

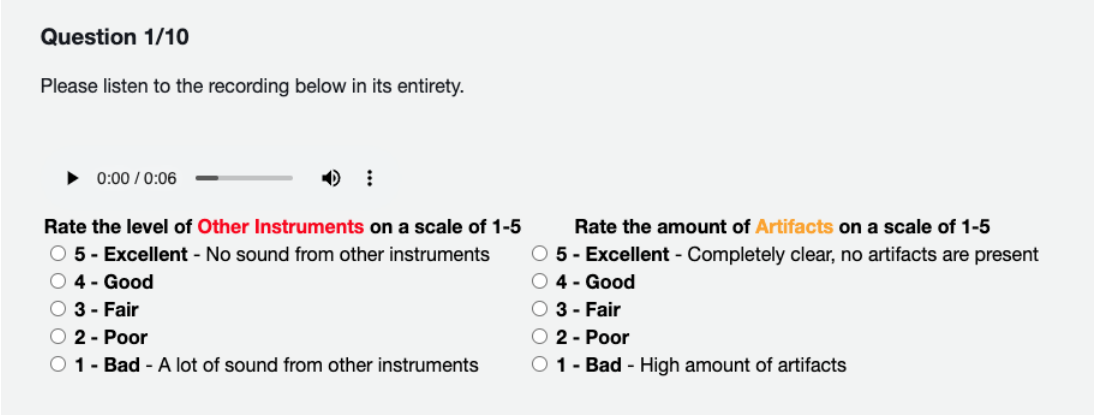


Figure 2.1. A listening question on the MOS study

Over all three stem-grouped batches, I collected surveys from 403 unique participants, 91 of whom were rejected for failing the hearing test. The listening devices and environment noise levels of the participants who passed the hearing test can be found in Table 2.1 and Table 2.2, respectively.

Listening Devices	
Headphones/Earbuds	420
Standalone Speakers	6
Built-In Speakers	24
Other	0

Table 2.1. Listening devices used by participants who passed the hearing test

Environment Noise Levels	
5 - No noise	209
4 - A little noise	117
3 - Somewhat noisy	77
2 - Very noisy	33
1 - Extremely noisy	14

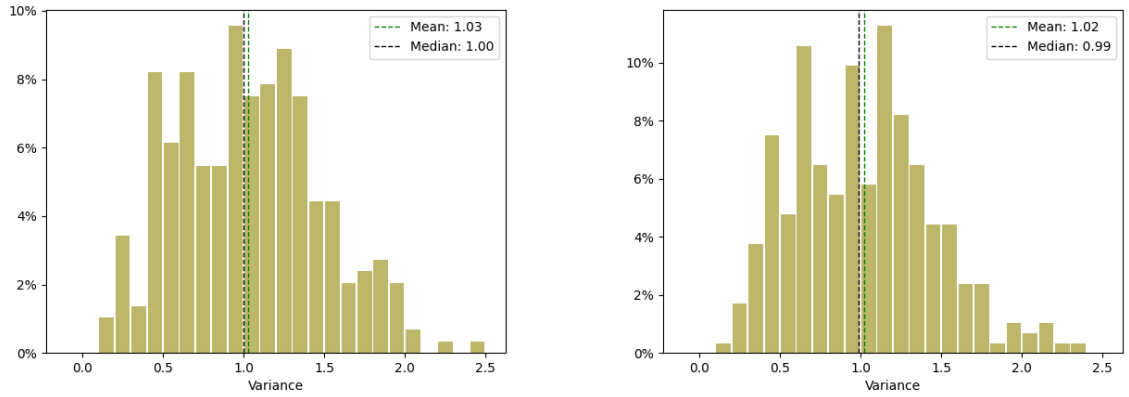
Table 2.2. Noise level rating of the participants' listening environments

2.3. Analysis of Subjective Data

It's important to note that not all humans rate things the same way. In order to illustrate the differences in individual participants' ratings, I analyzed the data acquired from the subjective listening assessment study described in Section 2.2.2 in comparison to the Mean Opinion Scores of that data.

2.3.1. Variance in Study Participant Data

The first feature I observe is the variance in participant data. In other words, I look at how far from an audio example's Mean Opinion Score each participant rated the audio example. For example, if audio example A had a Mean Opinion Score of 3.5, and participant P gave it a score of 4, then the variance would be 0.5.



(a) Artifacts (b) Other Instruments

Figure 2.2. Distribution of variance between subjective ratings and the Mean Opinion Score of each audio example

Figure 2.2 shows the distribution of amounts of variance between an individual rating of an audio example and the MOS of that audio example. The amount of variance is shown on the horizontal axis, and the vertical lines show the mean and median amounts of variance.

For both types of distortion - **Artifacts** and **Other Instruments**, most individual rater scores varied, on average, by about 1.0 from the Mean Opinion Score. This amount of variation seems logical since participants can only respond with integers within the small range of 1 through 5; and it would be unlikely for a listener to rate an audio example as a 5 when the majority rate it as a 1.

2.3.2. Correlation between Study Participant Responses

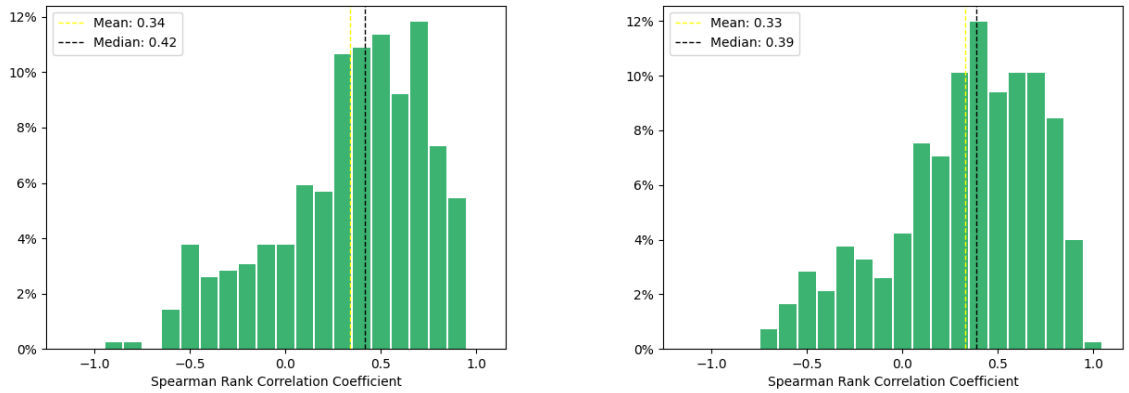
I suspected the correlation between participant responses and the MOS would be more indicative of a non-uniform relationship between an individual's hearing and that of the population. I calculated the Spearman rank correlation coefficient between the 10 ratings

of one subjective listening assessment submission and the Mean Opinion Scores of the 10 audio examples presented in that assessment. The Spearman rank correlation coefficient is set in the range of -1 to 1, where 1 indicates perfect correlation, 0 means there is no correlation, and -1 indicates perfect inverse correlation. For example, consider a set of 10 Mean Opinion Scores that steadily increases. A participant who gives ratings that also increase from example 1 to example 10 would have a positive correlation coefficient close to 1.0.

Figure 2.3 shows the distribution of these correlation coefficients, with the coefficients on the horizontal axis and the portion of assessments on the vertical axis. Most assessments achieved a correlation coefficient of around 0.5 to the Mean Opinion Scores of the audio examples they rated. It also shows a significant number of assessments that were negatively correlated to the others. I found that there were 78 out of 450 assessments that had negatively correlated ratings; 25 occurred in assessments with **bass** clips, 28 in **drums** assessments, and 25 in assessments of **vocals**.

2.3.3. Eliminating Unreliable Data

I considered excluding data that had a negative correlation to the MOS of those audio examples. However, it is possible that these participants genuinely heard the audio differently. Given the information acquired through the listening assessment, it would be impossible to prove that these participants, or which of them, were not completing the study with integrity. Instead, these results can be considered evidence that hearing is subjective and not all humans hear the same way.



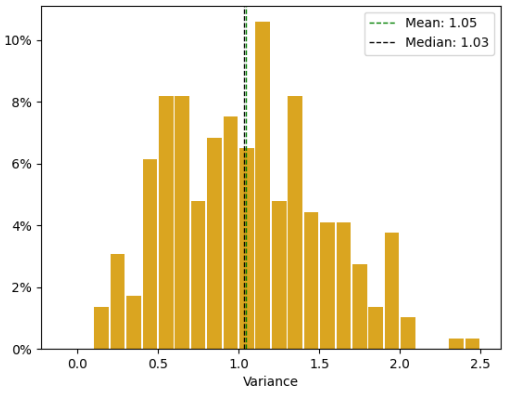
(a) Artifacts

(b) Other Instruments

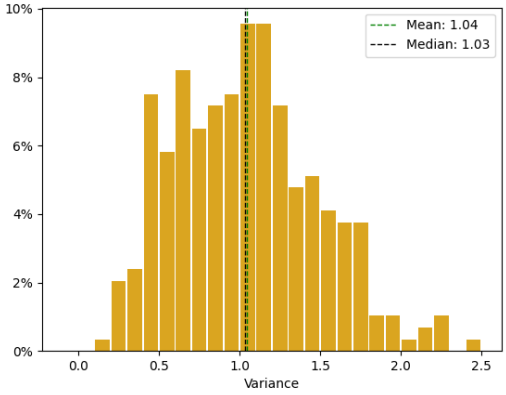
Figure 2.3. Distribution of Spearman rank correlation coefficients, measuring the correlation between each participants' responses and the MOS of each audio example they rated

In analyzing the subjective listening data, I also found many submissions from unique participants that had either all 4s and 5s as their ratings, or all 1s and 2s. Because the audio clips were created with source separation algorithms of varying quality, responses like these are highly unlikely. I decided to make further observations on a subset of the assessment response data that excluded assessments that had a rating of 0 or 1. For example, anyone who rated their 10 audio examples with the same value or the same two adjacent values was excluded.

I show the new distributions of rating variance and assessment correlation according to this exclusion criterion in Figures 2.4 and 2.5, respectively. Surprisingly, this did not significantly affect the distributions of variance or correlation. Given these observations, I continue to use the full subjective evaluation dataset through the rest of this work.

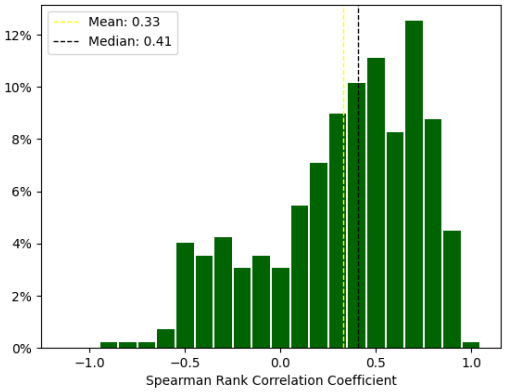


(a) Artifacts

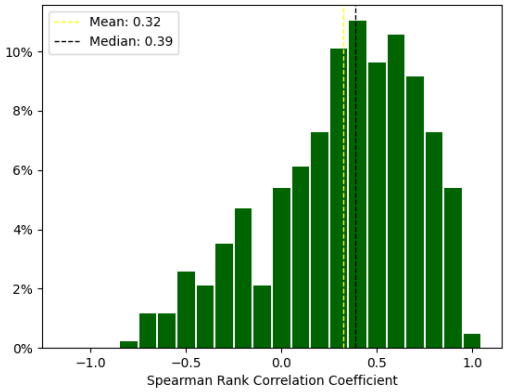


(b) Other Instruments

Figure 2.4. Distribution of variance between subjective ratings and the Mean Opinion Score of each audio example, using only ratings from study participants who gave a range of ratings greater than 1



(a) Artifacts



(b) Other Instruments

Figure 2.5. Distribution of Spearman rank correlation coefficients, measuring the correlation between each participants' responses, which had a range of ratings greater than 1, and the MOS of each audio example they rated

CHAPTER 3

Analysis of existing metrics

In this chapter, I do a comparative analysis between objective evaluation metrics for music audio quality and human perception 3.1. I also observe how five music source separation systems are ranked according to different evaluation metrics. 3.2.

3.1. Comparing Objective and Human Evaluation

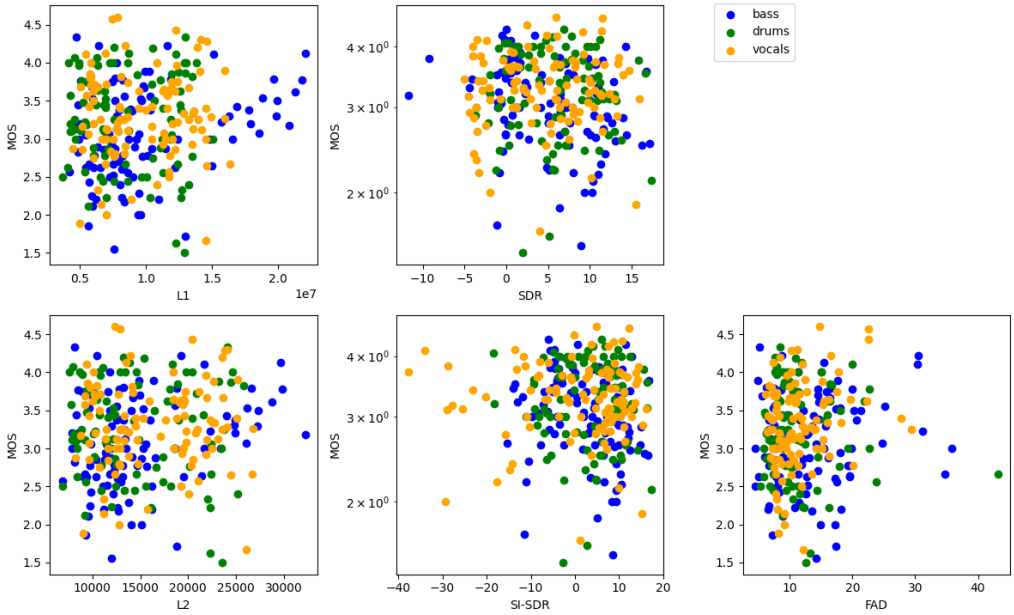
I have chosen five existing metrics to examine: signal-to-distortion ratio (SDR) [**37**], scale-invariant signal-to-distortion ratio (SI-SDR) [**31**], L1 loss, L2 loss, and Fréchet Audio Distance (FAD) [**19**]. SDR is the current standard metric for music source separation. These metrics are described in detail in Section 1.4.3. SDR serves as the baseline for these experiments, being the current standard metric for music source separation. I chose to observe SI-SDR to see if the scale-invariant aspect affects the outcome. L1 and L2 losses are typically used in training prediction models, but not in the final evaluation of audio quality. I chose to observe these loss functions to see if they would be valid evaluation methods, seeing as they are already used as training evaluators. Finally, I chose to observe FAD because it is the only evaluation metric for music that does not require the ground-truth, target signal. If FAD is shown to be a reliable evaluation method, it would be highly beneficial to the field of music source separation, specifically for experiments on audio data for which the ground-truth signals are not available.

The first method of analysis is to plot the Mean Opinion Score of each audio example against the score of each objective metrics. In this context, an audio example is a stem from one song that was separated by one of the five source separation systems described in Section 2.1.2. Figure 3.1 shows these plots, where each data point represents one separated stem. The horizontal axis shows the stem's score according to the labeled objective metric, and the vertical axis shows the stem's Mean Opinion Score. **Bass** examples are shown in blue, **drums** in green, and **vocals** in orange. The vertical axes on the SDR and SI-SDR plots have been log-scaled because they are measured in decibels, a logarithmic unit, but the other metrics are linear.

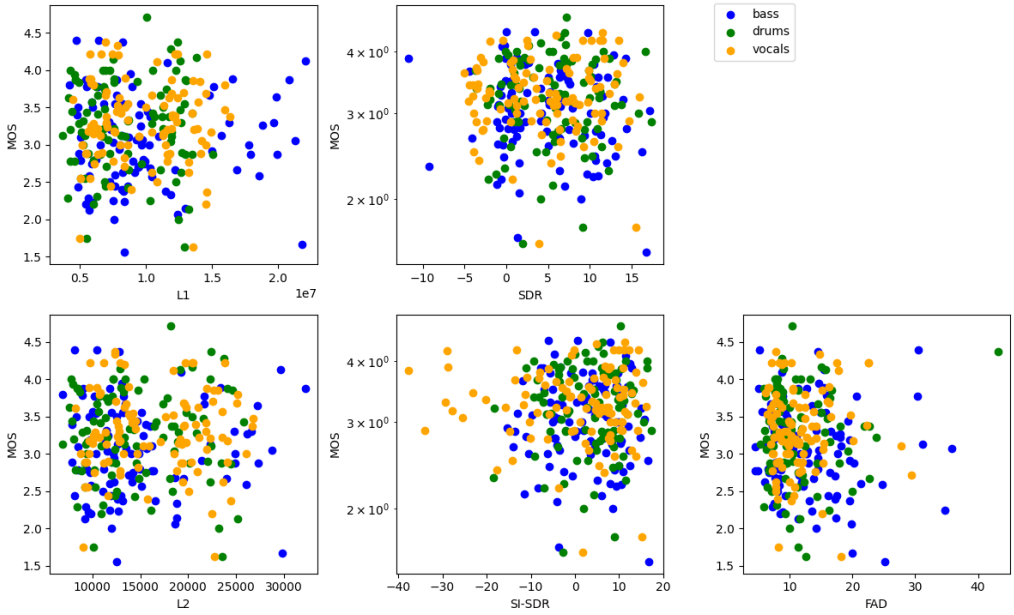
Perfect correlation between MOS and an objective metric would look like a diagonal line. With this in consideration, it's hard to claim there's any strong correlation in any of the plots in Figure 3.1, regardless of metric or stem type.

To confirm this notion, we can look at the Spearman's rank correlation coefficients. Spearman's rank correlation coefficients are set on the range of [-1, 1]. A positive coefficient indicates that the objective rating tends to increase as the MOS increases; and a negative coefficient indicates that the objective rating tends to decrease as the MOS increases. A coefficient of 0 indiates that there is no tendency for the objective rating to increase or decrease as MOS increases. When the objective ratings and MOS are perfectly monotone increasing, the coefficient is 1; and the coefficient is -1 when they are perfectly monotone decreasing.

Figure 3.2 shows the correlation coefficients for each stem type - **bass**, **drums**, and **vocals**. The horizontal axis denotes the objective metric and the vertical axis shows the correlation coefficient. We can see that the strongest positive correlation occurred with

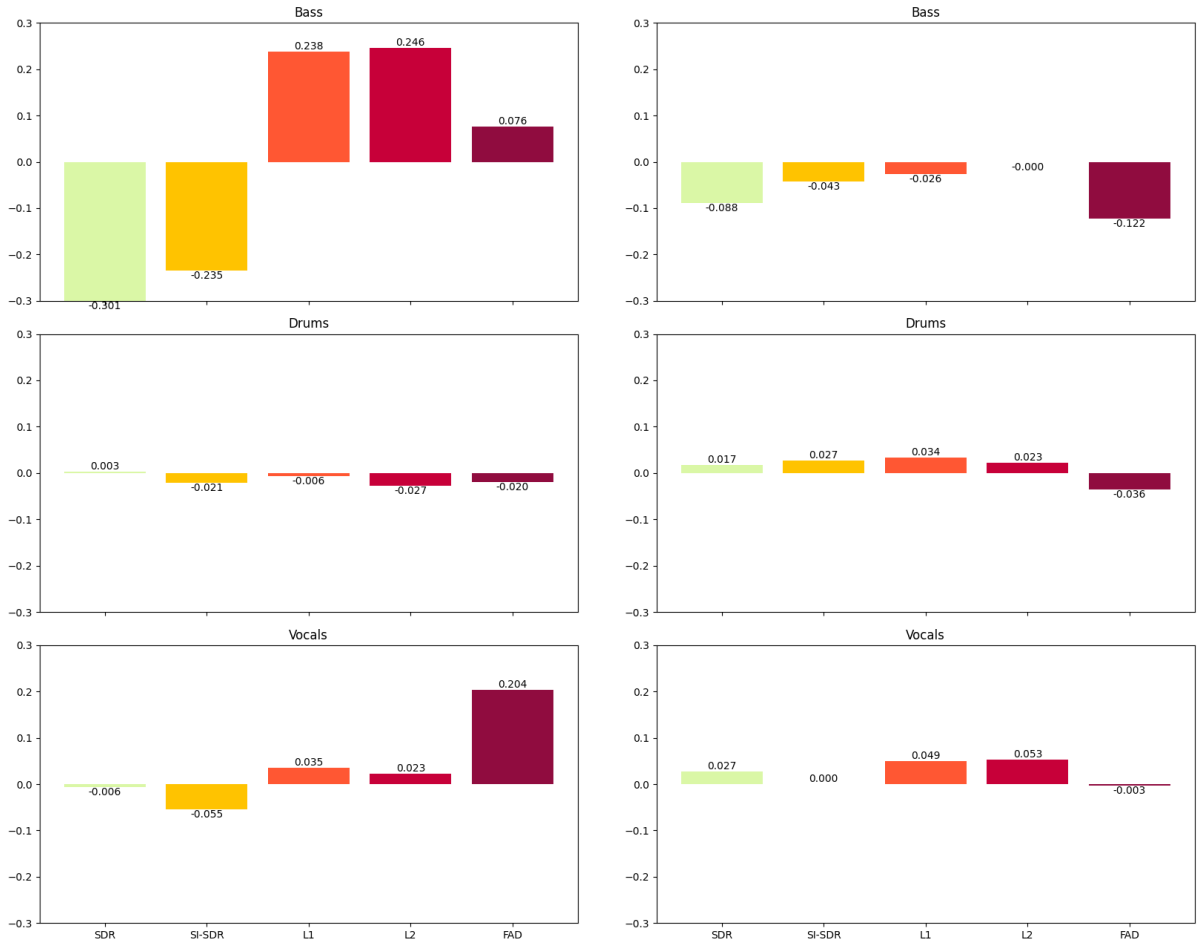


(a) Artifacts



(b) Other Instruments

Figure 3.1. Mean Opinion Scores vs. objective metric scores



(a) Artifacts (b) Other Instruments

Figure 3.2. Spearman's rank correlation coefficients between Mean Opinion Score and each objective metric score

L1 and L2 loss against the Mean Opinion Scores of artifacts present in **bass** examples. With a correlation coefficient of 0.257, however, this is still not a significant relationship. This implies that, as MOS increases, only a quarter of L1 and L2 evaluations do so as well.

3.2. Observations on Ranking by Different Evaluation Criteria

SDR is the most commonly used metric for training new music source separation models. As such, there is a leaderboard for how well these models source separated music from the MUSDB18 dataset, ranked by their overall SDR output. The first place, best separation model is currently the Hybrid Demucs [**5**] model, and last place is held by Wave-U-Net [**33**]. However, these rankings may not be the same when ranked by criteria other than SDR.

3.2.1. Comparing Rankings to the MUSDB18 Leaderboard

In Tables 3.1 to 3.5, I show the five source separation models ranked by the five objective evaluation metrics I observed. Higher values are better for SDR and SI-SDR, and lower values are better for L1 loss, L2 loss, and Fréchet Audio Distance. **In each table, model names that are in bold indicate that they maintain the same relative rank position as on the MUSDB18 leaderboard.**

These tables show that no observed metric maintained the same exact rankings as the MUSDB18 leaderboard; L2 loss was the most similar, with 3 models being ranked the same. The lack of consistency is even present in measuring by average SDR, just as the leaderboard is measured. This is most likely due to the dataset that was used in these experiments. The MUSDB18 leaderboard displays the average SDR of a model's output based on the entire MUSDB18 test set of 50 songs, but I only used 30 songs. This raises the question of how reliable it is to use an average as a ranking criterion. If its rank order only holds true for the same exact dataset, how valuable would that information be for

researchers who want to choose the best separator for their experiments using a different dataset?

Models Ranked by Average SDR		
	Model	**Avg. SDR**
1	Hybrid Demucs	10.962
2	Demucs-Extra	7.793
3	Spleeter	6.076
4	SS Wavenet	-0.175
5	D3Net	-1.067

Table 3.1. Rankings according to average signal-to-distortion ratio. Higher values are better.

Models Ranked by Average SI-SDR		
	Model	**Avg. SI-SDR**
1	Hybrid Demucs	10.854
2	Demucs-Extra	7.250
3	Spleeter	5.038
4	SS Wavenet	-5.007
5	D3Net	-11.151

Table 3.2. Rankings according to average scale-invariant signal-to-distortion ratio. Higher values are better.

Models Ranked by Average L1 Loss		
	Model	**Avg. L1 Loss**
1	D3Net	6,606,019.436
2	Hybrid Demucs	6,782,265.619
3	Demucs-Extra	7,839,315.456
4	Spleeter	11,038,175.683
5	SS Wavenet	14,260,813.344

Table 3.3. Rankings according to average L1 loss. Lower values are better.

Models Ranked by Average L2 Loss		
	Model	**Avg. L2 Loss**
1	Hybrid Demucs	11,136.053
2	D3Net	11,442.753
3	Demucs-Extra	12,885.587
4	Spleeter	19,223.974
5	SS Wavenet	21,660.467

Table 3.4. Rankings according to average L2 loss. Lower values are better.

Models Ranked by Average FAD		
	Model	**Avg. FAD**
1	D3Net	8.291
2	Spleeter	9.950
3	Hybrid Demucs	10.098
4	Demucs-Extra	13.158
5	SS Wavenet	17.416

Table 3.5. Rankings according to average Fréchet Audio Distance. Lower values are better.

3.2.2. Comparing Rankings to Rankings by Human Opinion

I return to the idea that a goal of music source separation is to create audio that sounds good to human listeners. The observed evaluation metrics were not consistent with each other nor the rankings of the official MUSDB18 source separation leaderboard. I now explore whether any of these metrics is consistent with the ranking of source separation models according to human evaluation.

In Tables 3.6 and 3.7, I show the rankings of the five source separation models according to Mean Opinion Score. These are separated by presence of **Artifacts** and presence

of **Other Instruments** since the study participants were asked to rate these attributes separately for the audio examples they were given. For MOS, higher values are better.

Models Ranked by MOS - Artifacts		
	Model	**MOS**
1	Source Separation Wavenet	3.389
2	D3Net	3.249
3	Spleeter	3.119
4	Hybrid Demucs	3.057
5	Demucs-Extra	3.015

Table 3.6. Rankings according to Mean Opinion Score of **Artifacts** present. Higher values are better.

Models Ranked by MOS - Other Instruments		
	Model	**MOS**
1	Source Separation Wavenet	3.196
2	Spleeter	3.152
3	D3Net	3.108
4	Hybrid Demucs	3.103
5	Demucs-Extra	3.087

Table 3.7. Rankings according to Mean Opinion Score of **Other Instruments** present. Higher values are better.

We can see that the rank order according to MOS is completely different from the rank order according to objective evaluation metrics. For example, the top two models according to SDR and SI-SDR, Hybrid Demucs and Demucs-Extra, are shown at the bottom of the MOS rankings. And the lowest performing model according to SDR and SI-SDR, Source Separation Wavenet, is shown as the top performing model.

A key observation from these MOS rankings is that the range of Mean Opinion Scores was quite small. Between the top and lowest ranking models, there was only a 0.374 difference in MOS for **Artifacts** and a 0.109 difference for **Other Instruments**. This

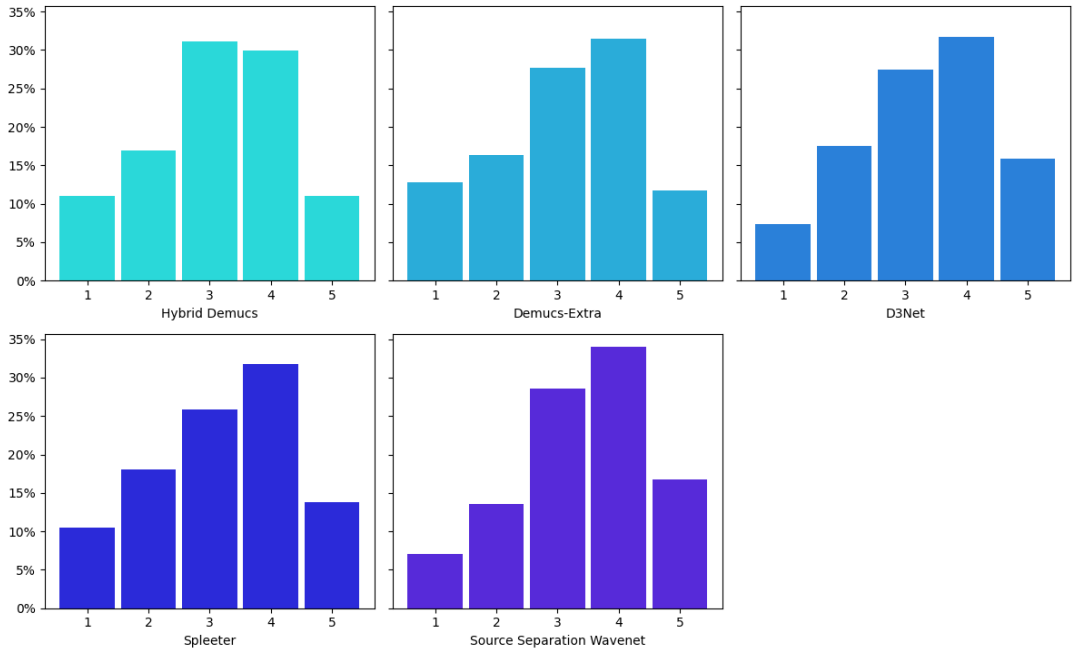
may indicate that human listeners wouldn't be necessarily good at discerning outputs of one source separation model from another.

To further illustrate the responses of human listeners, I show the distribution of ratings in Figure 3.3. The horizontal axis shows the score that a study participant could choose, any integer in the range [1,5], and the vertical axis shows the portion of all ratings that received that score. Across all of these plots, the distribution shape is extremely similar, with most scores being 3s or 4s. This would align logically with Tables 3.6 and 3.7, which shows every separator's Mean Opinion Score being between 3.0 and 4.0.

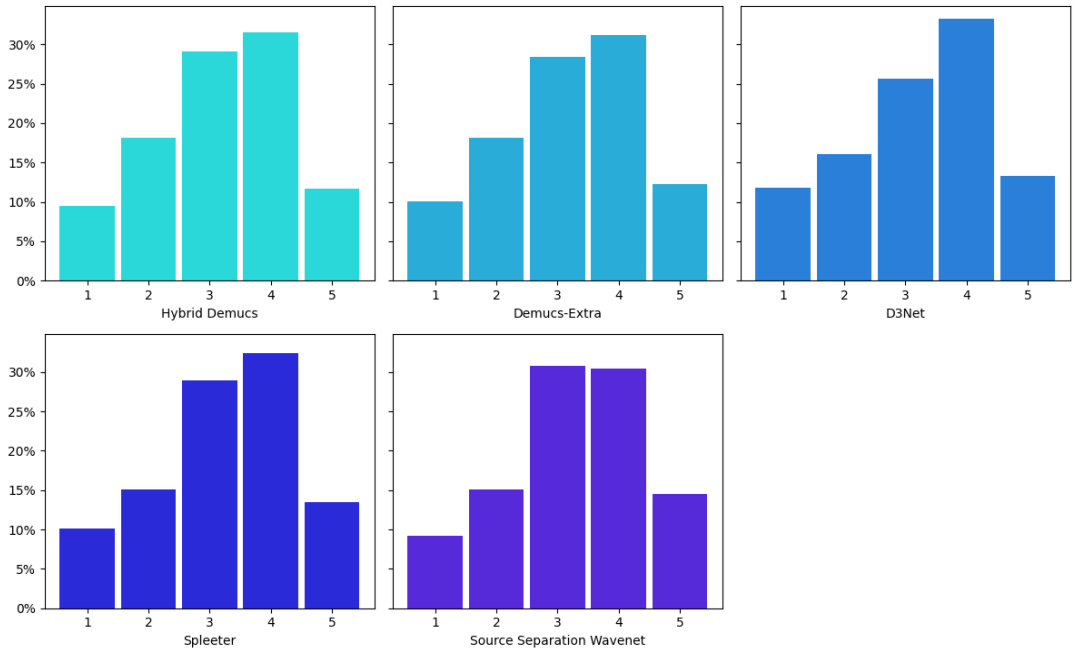
These results may tell us that the outputs of these source separation models are more similar than SDR and other metrics would indicate. I listened to the separated audio from the five source separation models to verify this notion[1]. However, I do not agree with this theory; for example, audio examples from Source Separation Wavenet had a lot of artifacts and other instrument sound present, whereas Hybrid Demucs examples had very little extra noise.

There are many factors that could contribute to the results differing from my own perception. Despite providing examples of what to listen for, participants may not have fully understood what they should be evaluating. Listeners could have also been affected by the environment in which they completed the study. As shown in Table 2.2, 124 out of 450, or 27.55% of participants completed the study in an environment that was "somewhat noisy" or worse. There is also the possibility that participants who reported a good level of environmental noise could not have been genuine. In both the question of environmental noise and the assessment itself, it is difficult to determine whether a

[1]Separated audio examples are available at `https://erumbold.github.io/nu-thesis`.



(a) Artifacts



(b) Other Instruments

Figure 3.3. Distribution of ratings for each source separation model

participant was authentic in their responses. It is also possible that listeners would be able to hear the differences in audio quality more acutely if the audio examples were presented in comparison with each other, instead of one at a time. This could be accomplished at a later date with a MUSHRA study.

CHAPTER 4

Conclusion

In this work, I have critically analyzed existing evaluation metrics for music source separation. My approach to this analysis is as follows. First, I curated a subset of the MUSDB18 dataset for music source separation that featured a wide range of musical genres and representation of both male and female singing voices. I then source separated these songs into the stems **bass**, **drums**, and **vocals**, using five source separation models that appear on the MUSDB18 leaderboard, representing the top, middle, and low ranks.

I developed a subjective listening assessment to obtain Mean Opinion Score data for the audio examples separated by the five source separation models. I recruited participants on Amazon Mechanical Turk and collected 4,500 individual ratings. Using this data, I observed how individuals listen in relation to each other. I found that, on average, listeners agreed with each other's rating opinions about 1/3 of the time, achieving an average correlation coefficient of 0.33.

I then made observations of how existing objective evaluation metrics for music source separation relate to the Mean Opinion Scores of human listeners. I found that the five metrics I observed, including the standard metric for music source separation, were not consistent with the opinions of human listeners. The best correlation coefficient between an objective metric and human evaluation was 0.257, which does not indicate a strong relationship.

Finally, I compare the ranking of source separation models according to different criteria - the average evaluation from the five objective metrics used in previous parts of this work. Only SDR and SI-SDR shared the same rank order; every other observed metric ranked the models in a different order from the rest. I also examined the ranking of source separation models according to the subjective Mean Opinion Scores, which differed even further from the objective metric-based ranks.

Determining whether an objective evaluation metric exists that is similar to human opinion was the central goal of this work. From observing the correlation of objective evaluation scores to subjective ratings, and comparing the ranking of separators by these different scores, I found that no existing objective evaluation metric correlates to the opinion of human listeners. On the MUSDB18 leaderboard, there was a 4.1 dB difference in SDR, and a 17-place difference in ranking, between the best and worst ranked separators that I used in this work. So it was expected that there would also be a significant difference in the ratings from human listeners. However, human listeners evaluated all five separators very similarly, with the MOS of the best and worst separators being no more than 0.374 points apart, and most of the audio examples, regardless of separator, being rated a 3 or 4 out of 5 for audio quality.

Furthermore, I found that the average SDR for each separator differed depending on the data used. The leaderboard is ranked according to the models' average SDR output for the full 50-song test set of MUSDB18. But the average SDR values for these separators were different when evaluated on only 30 songs. This would indicate that the MUSDB18 leaderboard is not a valid representation of separator quality for experiments using a subset of MUSDB18 or an entirely different dataset. In other words, the MUSDB18

leaderboard could be a reliable source to help choose a separator, but only when using the full MUSDB18 data set. This also suggests that ranking separators by an average score, whether average SDR or another metric's average, is not useful for every music source separation experiment. Instead, an evaluation of the separator as a whole, and how it performs on multiple datasets would be more representative of separation quality.

4.1. Limitations

As is expected with conducting a subjective listening assessment, the availability of time and money was a significant limitation of this work. The subjective listening study cost more than $1,000 and I only acquired 4,500 data points. This is a significantly small dataset for a computer audition experiment. Furthermore, it took five days to acquire those 4,500 data points, during which I had to verify each submission and republish assessment forms that were completed by workers who did not meet certain eligibility criteria. It could have taken even longer if more and more people failed to meet these requirements.

MUSDB18 is the most widely used dataset for music source separation. However, it is not the perfect, most ideal dataset for the task. The perfect dataset would equally feature a wide variety of genres, tempi, and instrumentations. The songs of MUSDB18 are categorized into 11 genres, with 72 out of 151 songs being listed as "Pop/Rock." Furthermore, the genre labels are not consistent with each other. For example, "Reggae" is one of the listed genres, but the song "Reggae" by Music Delta is listed as Rock. In my experiments, I chose to use only a subset of MUSDB18 that consisted of 30 songs

specifically chosen to even out the distribution of genres, tempi, and singing voice types. However, I recognize that such a small dataset is not able to wholly represent all music.

Finally, the choice to conduct a Mean Opinion Score assessment made some aspects of the process easier, but created significant discrepancies in the data. One reason I chose to conduct an MOS study was the simplicity of the rating system. I assumed this would make it easier for participants understand the task. However, the results shown in Section 3.2.2 may indicate that their understanding was less than expected.

4.2. Future Work

There is a lot of future work that can be done based on the data and observations shown throughout this thesis. First, the issues that may have been due to the format of the Mean Opinion Score assessment could be alleviated by conducting a MUSHRA assessment instead. In MUSHRA, audio examples are presented at once so the participant can listen to them in comparison to each other. This could help participants understand what types of distortion they should be listening for when they can quickly and easily listen to multiple examples.

In this work, I also show that the most commonly used evaluation metric for music source separation does not achieve a strong correlation to the opinions of human listeners, and no other metric does either. One could use these insights to develop a new evaluation metric with the intent of correlating well to human perception. While possible, a robust dataset of audio as well as human evaluation data would be required to achieve this goal.

References

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

[2] Cartwright, M., Pardo, B., and Mysore, G. J. Crowdsourced pairwise-comparison for source separation evaluation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), pp. 606–610.

[3] Cartwright, M., Pardo, B., Mysore, G. J., and Hoffman, M. Fast and easy crowdsourced perceptual audio evaluation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), IEEE, pp. 619–623.

[4] Chen, K.-T., Wu, C.-C., Chang, Y.-C., and Lei, C.-L. A crowdsourceable qoe evaluation framework for multimedia content. In *Proceedings of the 17th ACM international conference on Multimedia* (2009), pp. 491–500.

[5] Défossez, A. Hybrid spectrogram and waveform source separation. *arXiv preprint arXiv:2111.03600* (2021).

[6] Défossez, A., Usunier, N., Bottou, L., and Bach, F. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254* (2019).

[7] Dong, X., and Williamson, D. S. A classification-aided framework for non-intrusive speech quality assessment. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2019), IEEE, pp. 100–104.

[8] Dong, X., and Williamson, D. S. An attention enhanced multi-task model for objective speech assessment in real-world environments. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), IEEE, pp. 911–915.

[9] Dowson, D., and Landau, B. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis 12*, 3 (1982), 450–455.

[10] Emiya, V., Vincent, E., Harlander, N., and Hohmann, V. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing 19*, 7 (2011), 2046–2057.

[11] Fu, S., Tsao, Y., Hwang, H., and Wang, H. Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM. *CoRR abs/1808.05344* (2018).

[12] Fu, S.-W., Liao, C.-F., Tsao, Y., and Lin, S.-D. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning* (2019), PMLR, pp. 2031–2041.

[13] Fu, S.-W., Yu, C., Hsieh, T.-A., Plantinga, P., Ravanelli, M., Lu, X., and Tsao, Y. Metricgan+: An improved version of metricgan for speech enhancement. *arXiv preprint arXiv:2104.03538* (2021).

[14] Hennequin, R., Khlif, A., Voituret, F., and Moussallam, M. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software 5*, 50 (2020), 2154.

[15] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (2017), IEEE, pp. 131–135.

[16] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems 30* (2017).

[17] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 4700–4708.

[18] Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., and Weyde, T. Singing voice separation with deep u-net convolutional networks.

[19] Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466* (2018).

[20] Lluís, F., Pons, J., and Serra, X. End-to-end music source separation: is it possible in the waveform domain? *arXiv preprint arXiv:1810.12187* (2018).

[21] Lo, C., Fu, S., Huang, W., Wang, X., Yamagishi, J., Tsao, Y., and Wang, H. Mosnet: Deep learning based objective assessment for voice conversion. *CoRR abs/1904.08352* (2019).

[22] Luo, Y., and Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing 27*, 8 (2019), 1256–1266.

[23] Manocha, P., Finkelstein, A., Zhang, R., Bryan, N. J., Mysore, G. J., and Jin, Z. A differentiable perceptual audio metric learned from just noticeable differences, 2020.

[24] Manocha, P., Jin, Z., Zhang, R., and Finkelstein, A. Cdpam: Contrastive learning for perceptual audio similarity. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), pp. 196–200.

[25] Mitsufuji, Y., Fabbro, G., Uhlich, S., and Stöter, F.-R. Music demixing challenge 2021. *arXiv preprint arXiv:2108.13559* (2021).

[26] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[27] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., and Bittner, R. Musdb18-a corpus for music separation.

[28] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., FitzGerald, D., and Pardo, B. An overview of lead and accompaniment separation in music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 26*, 8 (2018), 1307–1335.

[29] Ribeiro, F., Florêncio, D., Zhang, C., and Seltzer, M. Crowdmos: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2011), IEEE, pp. 2416–2419.

[30] Rix, A., Beerends, J., Hollier, M., and Hekstra, A. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)* (2001), vol. 2, pp. 749–752 vol.2.

[31] Roux, J. L., Wisdom, S., Erdogan, H., and Hershey, J. R. Sdr – half-baked or well done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), pp. 626–630.

[32] Series, B. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly* (2014).

[33] Stoller, D., Ewert, S., and Dixon, S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185* (2018).

[34] Takahashi, N., and Mitsufuji, Y. Multi-scale multi-band densenets for audio source separation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2017), IEEE, pp. 21–25.

[35] Takahashi, N., and Mitsufuji, Y. D3net: Densely connected multidilated densenet for music source separation. *arXiv preprint arXiv:2010.01733* (2020).

[36] Uhlich, S., Giron, F., and Mitsufuji, Y. Deep neural network based instrument extraction from music. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), IEEE, pp. 2135–2139.

[37] Vincent, E., Gribonval, R., and Fevotte, C. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing 14*, 4 (2006), 1462–1469.

[38] Ward, D., Wierstorf, H., Mason, R. D., Grais, E. M., and Plumbley, M. D. Bss eval or peass? predicting the perception of singing-voice separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), IEEE, pp. 596–600.

[39] Watson, A., and Sasse, M. A. Measuring perceived quality of speech and video in multimedia conferencing applications. In *Proceedings of the sixth ACM international conference on Multimedia* (1998), pp. 55–60.

[40] Zhang, Z., Vyas, P., Dong, X., and Williamson, D. S. An end-to-end non-intrusive model for subjective and objective real-world speech assessment using a multi-task framework. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), IEEE, pp. 316–320.

APPENDIX A

MUSDB18 Data

This appendix includes a table of data used in my subjective assessments experiments from the MUSDB18 dataset [**27**]. I curated 30 songs, 28 from the testing set and 2 from the training set, that cover a range of musical genres and tempi. Because male singers are more represented in the overall MUSDB18 dataset, I ensured the curated set included a significant number of female singers. The 30 songs are shown in Table A.1.

A.1. Designation of Genres

MUSDB18 provides genre labels for each song. However, some songs are labeled inconsistently (e.g. two songs that sound very similar are given different labels) or can be labeled more specifically (e.g. a song labeled **Pop Rock** can be labeled as **Electronic**, **Pop**, or **Pop Punk**, which are related genres or subgenres of **Pop Rock**). To denote these discrepancies, Table A.1 indicates both the label assigned by MUSDB18 and the label considered in this work. Table A.2 shows the number of songs that belong to each genre as defined by the MUSDB18 labels and the new labels.

Artist	Song Title	Genre - MUSDB18	Genre - New	Tempo	Vocals
Actions	One Minute Smile*	Pop Rock	Pop Punk	Fast	Female
Al James	Schoolboy Facination	Pop Rock	Pop	Slow	Male
Angels In Amplifiers	I'm Alright	Rock	Singer-Songwriter	Medium	Male
Arise	Run Run Run	Reggae	Reggae	Medium	Male
Ben Carrigan	We'll Talk About It All Tonight	Pop Rock	Singer-Songwriter	Medium	Male
BKS	Too Much	Pop Rock	Rock	Medium	Male
Buitraker	Revo X	Pop Rock	Rock	Medium	Male
Carlos Gonzalez	A Place For Us	Pop Rock	Pop Rock	Slow	Male
Enda Reilly	Cur An Long Ag Seol	Pop Rock	Singer-Songwriter	Slow	Male
Forkupines	Semantics	Pop Rock	Pop Punk	Fast	Male
Hollow Ground	Ill Fate	Heavy Metal	Heavy Metal	Medium	Male
Juliet's Rescue	Heartbeats	Pop Rock	Pop Rock	Fast	Female
Little Chicago's Finest	My Own	Rap	Rap	Medium	Male
Louis Cressy Band	Good Time	Rock	Funk	Slow	Male
Lyndsey Ollard	Catching Up	Pop Rock	Singer-Songwriter	Slow	Female
Motor Tapes	Shore	Pop Rock	Pop Rock	Slow	Male
Nerve 9	Pray For The Rain	Pop Rock	Pop Rock	Slow	Female
Patrick Talbot	A Reason To Leave*	Jazz	Jazz	Slow	Male
Raft Monk	Tiring	Pop Rock	Grunge	Slow	Male
Sambasevan Shanmugam	Kaathaadi	Pop Rock	Bollywood	Slow	Female
Secretariat	Over The Top	Pop Rock	Rock	Fast	Male
Side Effects Project	Sing With Me	Rap	Rap	Medium	Male
The Doppler Shift	Atrophy	Pop Rock	Rock	Fast	Male
The Easton Ellises	Falcon 69	Pop Rock	Electronic	Medium	Male
The Long Wait	Dark Horses	Pop Rock	Country	Slow	Female
The Sunshine Garcia Band	For I Am The Moon	Reggae	Reggae	Slow	Female
Timboz	Pony	Heavy Metal	Heavy Metal	Fast	Male
Triviul feat. The Fiend	Widow	Pop Rock	Hip Hop	Medium	Female
We Fell From The Sky	Not You	Heavy Metal	Heavy Metal	Fast	Male
Zeno	Signs	Pop Rock	Pop Rock	Medium	Female

* denotes a song is from the training set of MUSDB18

Table A.1. MUSDB18 songs used in subjective assessment experiments

A.2. Distribution of Tempo

Song tempi are categorized as **Slow**, **Medium**, or **Fast**. I define these tempo labels by ranges of beats per minute (bpm) as follows:

Genre	Count (MUSDB18 Labels)	Count (New Labels)
Bollywood	0	1
Country	0	1
Electronic	0	1
Funk	0	1
Grunge	0	1
Heavy Metal	3	3
Hip Hop	0	1
Jazz	1	1
Pop	0	1
Pop Punk	0	2
Pop Rock	20	5
Rap	2	2
Reggae	2	2
Rock	2	4
Singer-Songwriter	0	4

Table A.2. Number of songs in each genre

- **Slow:** 85 bpm or lower
- **Medium:** 86 bpm to 125 bpm
- **Fast:** 126 bpm or higher

Table A.3 displays the number of songs in each tempo category, separated by genre.

A.3. Male vs. Female Vocals

Male singers are dominant in the MUSDB18 dataset. In order to mitigate bias due to singing range, I made sure to select songs that achieved a more balanced ratio of male to female singers while maintaining representation of a wide variety of genres. This resulted in 30% of the selected songs featured female singing voices. Of the full MUSDB18 dataset, 28% of the songs feature female singers. Although an increase of 2% is marginal, there are a couple factors that lead to why the representation of female singers in the chosen

Genre	Count (MuOSNet Labels)	Slow	Medium	Fast
Bollywood	1	1	0	0
Country	1	1	0	0
Electronic	1	0	1	0
Funk	1	1	0	0
Grunge	1	1	0	0
Heavy Metal	3	0	1	2
Hip Hop	1	0	1	0
Jazz	1	1	0	0
Pop	1	1	0	0
Pop Punk	2	1	0	1
Pop Rock	5	3	1	1
Rap	2	0	2	0
Reggae	2	1	1	0
Rock	4	0	2	2
Singer-Songwriter	4	2	2	0
Total	30	13	11	6

Table A.3. Number of songs in each tempo category, by genre

subset was not greater; 1) I prioritized representation of genre higher than that of gender, and 2) I sought to keep the subset to just MUSDB18's test set as much as possible, but it would require most of the training set songs featuring female singers to create an even ratio of male to female singers.

The separation of voice types in each genre is illustrated in Table A.4.

Genre	Count (New Labels)	Male	Female
Bollywood	1	0	1
Country	1	0	1
Electronic	1	1	0
Funk	1	1	0
Grunge	1	1	0
Heavy Metal	3	3	0
Hip Hop	1	0	1
Jazz	1	1	0
Pop	1	1	0
Pop Punk	2	1	1
Pop Rock	5	2	3
Rap	2	2	0
Reggae	2	1	1
Rock	4	4	0
Singer-Songwriter	4	3	1
Total	30	21	9

Table A.4. Number of songs featuring male or female singers, by genre