CORRELATIONS BETWEEN OBJECTIVE AND SUBJECTIVE EVALUATIONS OF MUSIC SOURCE SEPARATION

Erika RUMBOLD (erumbold@uvic.ca) [1], **George TZANETAKIS** (gtzan@uvic.ca) [1], and **Bryan PARDO** (pardo@northwestern.edu) [2]

[1] *Department of Computer Science*, **University of Victoria**, BC Canada
[2] *Interactive Audio Lab*, **Northwestern University**, IL United States

ABSTRACT

The standard quality metric for audio source separation is the signal-to-distortion ratio (SDR), despite correlating poorly to human perception. In this paper, we investigate the correlation between subjective listener evaluations and multiple objective quality metrics, including the SDR. Music tracks are separated using five source separation models, and the resulting separation quality is calculated using five objective evaluation metrics. A listener study was conducted to also obtain subjective evaluations of the music separation quality. It was found that none of the observed objective metrics correlated well to human perception, with the best correlation coefficient being 0.246. We also found that the objective metrics did not agree with each other regarding which is the best and worst performing music source separation models.

1. INTRODUCTION

Music source separation is the task of isolating instruments, or groups of instruments, from a mixture of multiple instruments playing at once. For example, the group, or stem, for vocals would include background vocals in addition to the lead singer. The target stems can be specified in many ways; the MUSDB18 dataset for music separation [1] features the stems `bass`, `drums`, `vocals`, and `other`. Given a mixture of these four stems, the goal is to generate four waveforms that correspond to each of the original stems. In addition to being a well-studied topic on its own, source separation is used in a variety of other music information retrieval (MIR) applications, including automatic music transcription, lyric recognition, and music enhancement.

Source separation is continually advancing, improving on prior work. As with many scientific endeavors, this advancement relies on an evaluation step. For music source separation, this evaluation typically measures how "clean" the separated audio is (i.e., whether or not there are extra elements present that wouldn't be in a recording of the instrument by itself), and how complete it is (i.e., whether there are missing elements in the separated signal). There are many methods of evaluating audio quality, which fall under two categories: human evaluation, and automated methods. The latter can be further distinguished as either a closed-form formula that calculates an amount of error between the separator's output signal and the expected signal, or a statistical model that estimates a "goodness" score with information learned about either a data distribution [2] or human evaluation data [3].

With audio data being the focus of MIR tasks, the general goal should be to produce something that sounds good to human listeners. Human evaluation is the most direct way to determine whether we think some audio sounds good, but it is expensive and requires a lot of time to collect sufficient data to inform algorithm development. In response to these shortcomings, several automated error calculation methods have been proposed and used in various MIR tasks, including music source separation.

The most commonly used metric for audio source separation is the signal-to-distortion ratio (SDR), 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 [4]. SDR is frequently cited in existing audio source separation work and serves as the baseline measure for many source separation challenges (e.g., Sony Demixing Challenge [5] and the MUSDB18 benchmark [1]).

Despite its popularity, SDR has been proven to correlate poorly to human perception [6,7], meaning an audio signal that a listener would deem as having poor quality may still get a good SDR value, or vice versa. With the objective of making audio that sounds good to listeners, the correlation between evaluation metrics and human perception should be prioritized. This discrepancy has been acknowledged in the speech domain, prompting the development of new evaluation methods that achieve better correlation to human perception [8–17]; but this has not yet been realized to the same degree in the music domain.

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

2. RELATED WORK

2.1 Subjective Evaluation of Audio Quality

Subjective evaluation of audio quality refers to the rating of audio stimuli by human subjects. Participants of a sub-

jective 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. Some are designed for the evaluation of individual stimuli one by one, while others have participants evaluate multiple stimuli in comparison with each other. These protocols also differ in the granularity of data received. Evaluations on a scale of 1 to 100, as in a MUSHRA study [18], tend to have tighter differences between responses than evaluations on a scale of 1 to 5.

In this paper, we conduct a Mean Opinion Score (MOS) study. Participants in a MOS assessment are asked to rate stimuli individually on a rational number scale, usually as follows: 1-Bad, 2-Poor, 3-Fair, 4-Good, and 5-Excellent. MOS is one of the most common subjective assessment methods in audio applications and is ideal to get a more general idea of human perception.

2.1.1 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, efforts have been made to adapt subjective quality assessments to an online crowdsourced format [19–22].

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 [18] in order to eliminate the possibility of a participant's evaluation being affected by extraneous noise or a lack of understanding of the task. However, conducting tests in a lab costs a significant amount of both time and money; and subjective assessment trials that take place in a lab attract fewer participants, leading to results that are less statistically significant [21].

These drawbacks have led to the rise of subjective evaluations that take place online. Many efforts have been made to adapt subjective assessment protocols to an online, crowd-sourced environment [19–22], forgoing some of the strict participant eligibility criteria and environmental control in favor of acquiring a larger and more diverse set of results. 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 [21] 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. [22] also asked about listening device in their web-based MUSHRA assessment, 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).

An online assessment can also contain a hearing test to assess the participants' hearing capabilities. For example, the MUSHRA assessment from Cartwright, et al. [22] 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 [19]. 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 do.

2.2 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: closed form equations that compute an amount of error, or 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.

2.2.1 Closed Form Evaluation Methods

The Signal-to-distortion ratio (SDR) [4] 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. 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 $\hat{s}_i$ [6].

From these attributes, we are able to compute different energy ratios by the relation of these terms to the true source. SDR, or the overall measure of how good the source estimate sounds in comparison to the true source can be represented as:

$$\text{SDR} = 10\log_{10}\left(\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2}\right) \quad (1)$$

SDR is measured in decibels (dB), and higher values are better.

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 the proposal of a version of SDR that is not dependent on amplitude scaling, SI-SDR [23]. 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

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

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 source separation algorithms Demucs [24, 25] and Spleeter [26]. L1 can be observed as the absolute error, and L2 as the squared error.

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.

2.2.2 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 [1] [10] and MOSNet [14] 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 [3] 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.

An alternative approach to a "referenceless" model is given by the Fréchet Audio Distance (FAD) [2]. Inspired by the Fréchet Inception Distance (FID) [27], which was developed to evaluate generative models for images, FAD compares statistics computed on a set of estimated signals to reference statistics computed on a large set of studio recorded music.

FAD uses a VGGish [28] model to generate embeddings for the reference set and the evaluation set. Like how Fréchet Audio Distance, it 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. [29] define the Fréchet distance between two Gaussians as:

$$F(N_b, N_e) = ||\mu_b - \mu_e||^2 + tr(\Sigma_b + \Sigma_e - 2\sqrt{\Sigma_b\Sigma_e}) \quad (3)$$

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 also be effective for music source separation.

3. LISTENER STUDY

To make comparisons between objective and subjective evaluation methods, we performed source separation on a subset of MUSDB18 using five selected algorithms. We then conducted a crowdsourced listener study to collect MOS ratings, and measured the separation quality according to five objective metrics. These evaluations and the correlation between them are discussed further in Section 4.

3.1 Dataset

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

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

stereo streams corresponding to the mixture, drums, bass, other, and vocals. MUSDB18 is among the most-cited datasets for existing work in music source separation.

From the 150 songs in the MUSDB18 dataset, we curated a subset of 30 songs, mostly from the test set, that represented a wide range of musical genres and were balanced between male and female singers. We did not randomly select songs because MUSDB18 skews in favor of male singers and the Pop/Rock genre; manual curation makes it easier to represent the genres and vocals that are less common in the MUSDB18 data set as a whole.

Each of the 30 MUSDB18 songs was truncated to a 7-second segment. This duration is short enough to be separated efficiently while also being long enough for listeners to effectively evaluate. These segments were screened to ensure that the clip contained enough of each stem - bass, drums, and vocals - to be evaluated.

The source separation algorithms used in this experiment are referred to as **HDX** [25], **DMX** [24], **D3N** [30], **SPL** [26], and **SSW** [31]. These were chosen due to their rankings on the Papers with Code leaderboard for average SDR, seeking algorithms that represented the top, middle, and bottom tiers. 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`, `vocals`, and `other`. We disregarded the other track in this experiment due to its ambiguity. The `other` stem could contain a keyboard synthesizer or a harp - a solo saxophone or an entire string orchestra.

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.

3.2 Listening Assessment

We conducted a 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, "artifacts" were defined in the introduction as "extra sound that cannot be recognized as a musical instrument or voice."

3.2.1 Participants

Participants 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.

3.2.2 Pre-Screening

The subjective assessment started with a hearing screening similar to the screening defined by Cartwright, et al. [22] 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.

3.2.3 Procedure

Following the hearing screening, a description of the rating system was given to the participants who passed the hearing test. It was explained that they would give two ratings for each audio example on a 1-5 scale - one for presence of artifacts and another for presence of sounds from other instruments. A rating of 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 presented with example audio clips and descriptions of what was 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 assessment versions. 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 unpublished 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.

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. At the end of the assessment, participants were asked to report the type of 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.

4. RESULTS & DISCUSSION

In this section, we review the responses received from the listener study described in the previous section, and com-

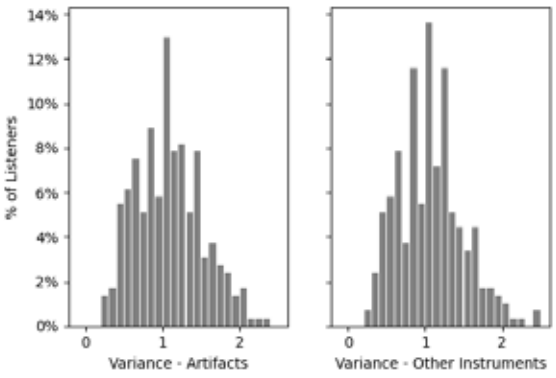

Figure 1. Distribution of variance between subjective ratings and the MOS of each audio example.

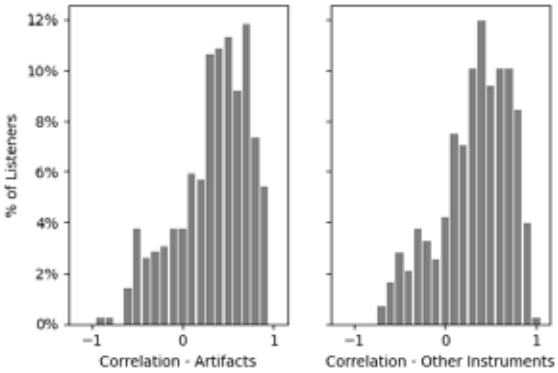

Figure 2. Distribution of Spearman rank correlation coefficients between subjective ratings and the MOS of each audio example.

pare those results to the evaluation of selected objective metrics.

4.1 Listener Responses

Fig. 1 shows the distribution of variance between an individual rating of an audio example and the MOS of that example. For both types of distortion, artifacts and other instruments, most individual rater scores varied, on average, by about 1.0 from the MOS. 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.

We also investigated the Spearman rank correlation between the ratings a listener gave and the mean opinion scores for the same songs. The distribution of these correlation coefficients is shown in Fig. 2. Most listeners had a correlation coefficient around 0.5 to the MOS of the examples they rated. It also shows a significant number of assessments that were negatively correlated to the others. Specifically, 78 out of 450 assessments had negatively correlated ratings; 25 bass assessments, 28 drum assessments, and 25 vocals assessments.

	SDR	SI-SDR	L1	L2	FAD
Bass	-0.301	-0.235	0.238	0.246	0.076
Drums	0.003	-0.021	-0.006	-0.027	-0.020
Vocals	-0.006	-0.055	0.035	0.023	0.204

Table 1. Spearman rank correlation coefficients for each objective metric compared to MOS for presence of artifacts

	SDR	SI-SDR	L1	L2	FAD
Bass	-0.088	-0.043	-0.026	-0.000	-0.122
Drums	0.017	0.027	0.034	0.023	-0.036
Vocals	0.027	0.000	0.049	0.053	-0.003

Table 2. Spearman rank correlation coefficients for each objective metric compared to MOS for presence of other instruments

We considered excluding data that had a negative correlation. 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.

We also noticed that many submissions from unique participants 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, according to average SDR, responses like these are highly unlikely. We investigated the correlation on a subset of the responses that excluded those that gave all of the same or two adjacent ratings. However, excluding such data did not significantly affect the distributions of variance or correlation. Given these observations, the rest of this paper uses the full set of subjective evaluation responses.

4.2 Correlation Between Objective and Subjective Evaluation

We have chosen five existing objective metrics to examine: SDR [4], scale-invariant SDR (SI-SDR) [23], L1 loss, L2 loss, and Fréchet Audio Distance (FAD) [2]. SDR serves as the baseline for these experiments, being the current standard metric for music source separation. SI-SDR was chosen 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. FAD is a music evaluation metric that does not require the ground-truth target signal.

First, we compare the MOS of each audio example against the corresponding score from each objective metric, as shown in Fig. 3 and 4. Clearly, there isn't a strong correlation in any of the plots, regardless of metric or stem type.

To confirm this notion, we can look at the Spearman's rank correlation coefficients, shown in Table 1 and Table 2. We can see that the strongest positive correlation occurred with L1 and L2 loss for the MOS of artifacts present in bass examples. With a correlation coefficient of 0.246, however, this is still not a strong relationship. This implies that, as MOS improves, only a quarter of L1 and L2 evaluations do

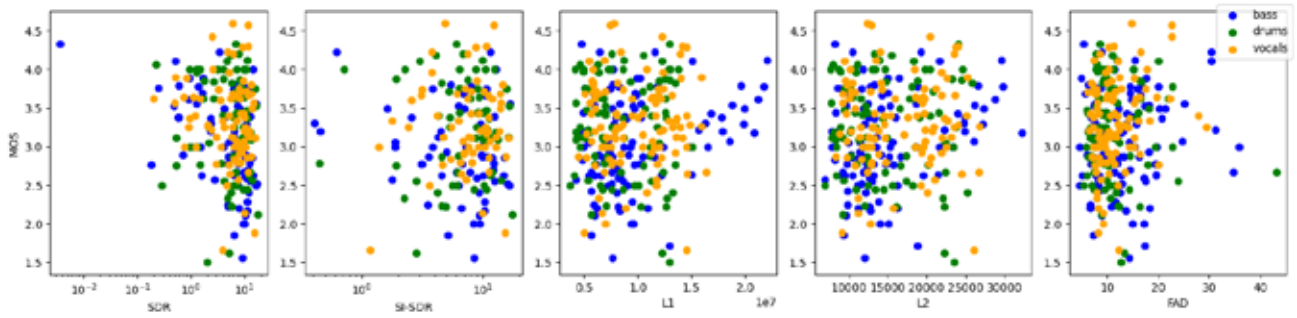

Figure 3. Objective metrics vs. MOS for presence of artifacts. The axes for SDR and SI-SDR are log-scaled.

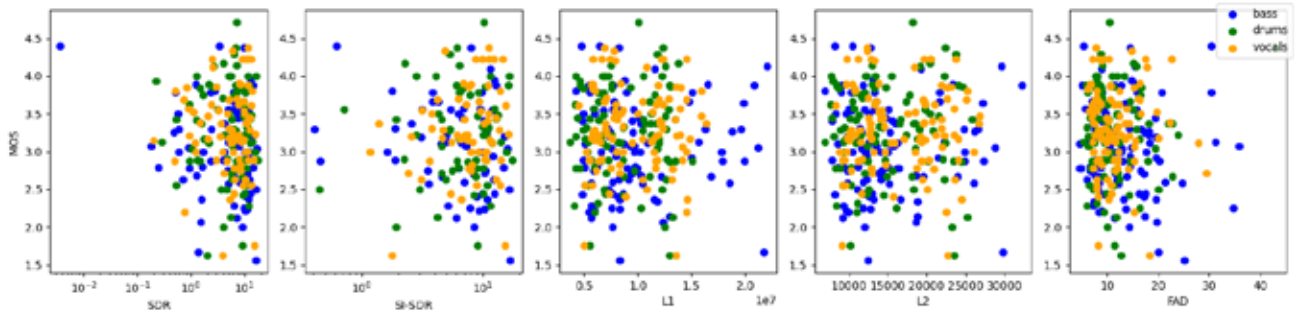

Figure 4. Objective metrics vs. MOS for presence of other instruments. The axes for SDR and SI-SDR are log-scaled.

so as well.

4.3 Ranking Source Separation Performance by Different Evaluation Criteria

In addition to correlation analysis, we considered how each evaluation metric would rank the performance of the five source separation algorithms. We also compare the ranking from the observed metrics to that from the MUSDB18 leaderboard [2] .

4.3.1 Rank Order from Objective Metrics

As shown in Table 3, no observed metric maintained the same exact rank order as the MUSDB18 leaderboard. This includes SDR, which the leaderboard uses. The difference between the average SDR from this experiment and that from the leaderboard is likely due to using a subset of the MUSDB18 instead of the full 50 song test set.

Among the five objective metrics, the order most similar to the MUSDB18 leaderboard was given by L2 loss, placing **HDX** as the best separator and **SPL** and **SSW** as the worst. The least similar rankings to the leaderboard were from FAD, which only agreed with **SSW** being at the bottom.

4.3.2 Rank Order from Listener Study

We also looked at the rank orders from the objective metrics in comparison to the collected mean opinion scores, also shown in Table 3. We can see that the rank order according to MOS is completely different from the order according to both the MUSDB18 leaderboard and the observed objective evaluation metrics.

These subjective results may tell us that the outputs of these source separation models are more similar than SDR and other objective metrics would indicate. However, there are other factors that could have affected the listeners' responses. Despite being provided with examples of what to listen for, participants may not have fully understood the task. They could have also been affected by the environment in which they completed the study. The responses at the end of the assessment showed that 27.55% of the surveys were completed in an environment that was "somewhat noisy" or worse. There is also the possibility that participants who reported an appropriate level of environmental noise could not have been genuine. These are both caveats that come with conducting an online crowdsourced study.

5. CONCLUSION

We have investigated the relationship between objective and subjective evaluation of music source separation. We collected mean opinion score data for separated stems from five different source separation algorithms, and compared those results to evaluations from five different objective metrics.

With the goal of producing audio that sounds good to listeners, the objective metrics used in MIR work would ideally correlate to human evaluation. We found that none of the observed objective metrics correlated well with the listener opinion data, with the best correlation being 0.246. We also found that no evaluation metric agreed on which source separation algorithm was the best, worst, or in be-

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

Rank	MUSDB18	SDR	SI-SDR	L1	L2	FAD	MOS (Artif.)	MOS (Inst.)
(Best) 1	HDX	HDX	HDX	D3N	HDX	D3N	SSW	SSW
2	DMX	DMX	DMX	HDX	D3N	SPL	D3N	SPL
3	D3N	SPL	SPL	DMX	DMX	HDX	SPL	D3N
4	SPL	SSW	SSW	SPL	SPL	DMX	HDX	HDX
(Worst) 5	SSW	D3N	D3N	SSW	SSW	SSW	DMX	DMX

Table 3. Relative rank order of source separation algorithms according to the MUSDB18 leaderboard, the five observed objective metrics, and the mean opinion score (MOS) when listening for presence of "Artifacts" or "Other Instruments."

tween. On the MUSDB18 leaderboard, there was a 4.1 dB difference in average SDR between the best and worst separator used in this paper. However, human listeners evaluated all five separators very similarly and in a vastly different order.

5.1 Future Work

We confirm that the most commonly used evaluation metric for music source separation does not achieve a strong correlation to the opinions of listeners, and none of the other observed metrics 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 listener evaluation data would be required to achieve this goal. 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. If so, a MUSHRA assessment would be worth pursuing.

6. REFERENCES

[1] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-a corpus for music separation," 2017.

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

[3] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "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)*, vol. 2, 2001, pp. 749–752 vol.2.

[4] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

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

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

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

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

[9] ——, "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)*. IEEE, 2020, pp. 911–915.

[10] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[11] S. Fu, Y. Tsao, H. Hwang, and H. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," *CoRR*, vol. abs/1808.05344, 2018. [Online]. Available: http://arxiv.org/abs/1808.05344

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

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

[14] C. Lo, S. Fu, W. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H. Wang, "Mosnet: Deep learning based objective assessment for voice conversion," *CoRR*, vol. abs/1904.08352, 2019. [Online]. Available: http://arxiv.org/abs/1904.08352

[15] P. Manocha, A. Finkelstein, R. Zhang, N. J. Bryan, G. J. Mysore, and Z. Jin, "A differentiable perceptual

audio metric learned from just noticeable differences," 2020.

[16] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, "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.

[17] Z. Zhang, P. Vyas, X. Dong, and D. S. Williamson, "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)*. IEEE, 2021, pp. 316–320.

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

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

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

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

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

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

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

[25] A. Défossez, "Hybrid spectrogram and waveform source separation," *arXiv preprint arXiv:2111.03600*, 2021.

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

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

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

[29] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of multivariate analysis*, vol. 12, no. 3, pp. 450–455, 1982.

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

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