section consisted of the same questions, the only difference is whether they are related to interpersonal or non-interpersonal channels. Upon presenting a section to the participants, they were first asked if they had ever received recommendations through the respective channel. If they had not, they would skip the questions of that section and would move on to the next section.

The survey questions were divided into two parts: 1) evaluation criteria, and 2) usage behavior. Participants were asked to respond to the four evaluation criteria (i.e., relevance, diversity, novelty, and serendipity) using a 5-point Likert scale. Participants were then asked to respond to the usage behavioral questions of frequency and convenience in a similar 5-point Likert scale fashion. Additionally, participants were asked to estimate the adoption rate of the recommended music to their playlist (i.e., adding the music to their playlist) on a percentage level. As with the original survey, no actual recommendations were presented, as the survey solely focused on past experiences of received recommendations.

4. RESULTS

We recruited a total of 142 participants for this study in the context of three Master's course projects at TU Wien via open calls for participation advertised to other students in the course, as well as other peers and acquaintances in their extended networks, cf. [19–21]

Out of the disclosed and collected information of the total 142 participants, 45% identify as female, 55% as male, with 86% aged between 18 and 32. When it comes to the use of music subscription services, 69% of the participants indicated to use music subscription services for more than a year, and 16% shared music with others more than once a week (a side-by-side comparison with the original study can be found in Table 1). To investigate the reproducibility of Kim et al. [1], similar statistical tests were used. In terms of cultural background, 89% of reporting participants in our study state their origin to be in a European country. In contrast to the original study, where 100% of participants reported their background as Asian (specifically, Korean), in our study, only 7% report their origin to be in an Asian country (see Table 3).

4.1 Evaluation of music recommendation channels

Initial mean comparison show that diversity, novelty, and serendipity on average score higher for interpersonal music channels, while relevance, convenience, frequency, and adoption rate are higher for the non-interpersonal channels (see Table 4). When conducting significance testing on the four evaluation criteria, paired t-tests confirm that diversity ($p < .001$) and novelty ($p = .002$) are rated higher for interpersonal channels and relevance ($p < .001$) is rated higher for non-interpersonal channels (see Figure 1A).

An additional principal component analysis (PCA) was conducted to investigate the relationship between the four evaluation criteria. The correlation matrix show that some of the evaluation criteria appeared to be similar than others. The PCA show two components to be extracted from the data with a total accounted variance of 67%: component 1 accounts for 42% and component 2 accounts for 24%. PCA bi-plot (see Figure 2) show that relevance is mainly explained by component 1, and diversity, novelty, and serendipity by component 2. The bi-plot indicate that diversity, novelty, and serendipity are more closely aligned, while relevance is almost orthogonal to these three evaluation criteria.

4.2 Usage behavior of music recommendation channels

Non-interpersonal channels rated higher for convenience, frequency, and adoption rate (see Table 4). Significance testing by using paired t-tests show that frequency was significantly rated higher ($p < .001$) for non-interpersonal channels than for interpersonal channels (see Figure 1B). When it comes to the convenience of using a channel, non-interpersonal channels were rated higher ($p = .004$) as well (see Figure 1C). However, when it comes to the adoption rate, although the non-interpersonal channel scored slightly higher, there was no significant difference ($p = ns$) between non-interpersonal and interpersonal channels.

Additionally, a linear regression was conducted to investigate the effect of frequency on adoption rate. Frequency has a significant effect on adoption rate for interpersonal channels ($F(1, 110) = 4.416, p = .038$) as well as for non-interpersonal channels ($F(1, 90) = 4.971, p = .028$). Whereas convenience plays an additional role on adoption rate of non-interpersonal channels ($F(1, 90) = 4.312, p = .041$).

Furthermore, a univariate regression was conducted to investigate the four evaluation criteria on the adoption rate. For the interpersonal channels, diversity ($F(1.68) = 1.574, p = .003$), novelty ($F(1.68) = 2.220, p = .002$) show to have a significant effect on adoption rates, while for the non-interpersonal channels only novelty ($F(1.68) = 1.596, p = .06$) appeared to be significant.

4.3 Gender differences

We conducted an additional multivariate test to investigate the role of gender between interpersonal and non-interpersonal channels. Gender differences were found when it comes of frequency of the receiving recommendations through non-interpersonal channels ($F(1, 90) = 4.475, p = .035$), indicating that males receive more frequent recommendations through non-interpersonal channels than females. However, no significant gender differences ($p = ns$) were found within interpersonal channels.

Furthermore, a marginal significant effect was found of gender on novelty within interpersonal channels ($F(1, 68) = 3.138, p = .061$), indicating that males in particular receive more novel recommendations through interpersonal channels such as friends and acquaintances than females. No significant effects ($p = ns$) of gender were found within non-interpersonal channels.

Evaluation criteria	Relevance	Is the music recommend through the channel consistent with the context and theme of the playlist?
	Diversity	Does the music recommended through the channel diversify genre or mood of the playlist?
	Novelty	Is the music recommended through the channel new or novel?
	Serendipity	Is the music recommended through the channel unexpected but good in terms of the context and theme of the playlist?
Usage behavior	Frequency	How often do you receive music recommendations through the channel?
	Convenience	How convenient is the process of listening to the music recommended through the channel?
	Adoption rate	What percentage of the recommended music through the channel is added to the playlist?

Table 2. Survey questions. Same questions were used for both interpersonal and non-interpersonal channels. Adopted from Kim et al. [1].

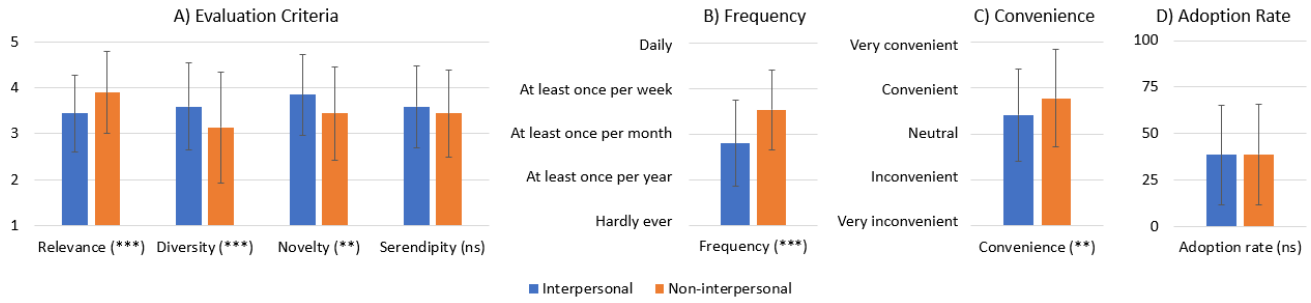

Figure 1. Mean scores with standard deviation error bars ($*\, p < 0.01$, $**\, p < 0.001$, $***\, p < 0.0001$).

Continent	Percentage
Europe	89%
Asia	7%
Africa/Middle East	3%
Oceania	1%

Table 3. Distribution of participants of the current study according to disclosed origin per continent.

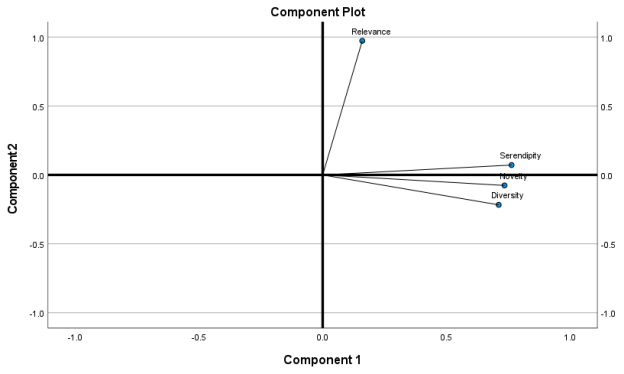

Figure 2. PCA biplot with the two extracted components.

5. DISCUSSION

In this section, we discuss the results obtained with regards to reproducibility, deviations from the original study, and additional analyses carried out to gain further insights.

5.1 Reproducibility

For this reproducibility study we followed similar procedures and methods of the original work by Kim et al. [1]. Following the same procedures we recruited 142 participants to fill in questions on their music consumption behaviors through interpersonal and non-interpersonal music channels.

Looking at the mean values, the overall strength of the mean values suggest that similar trends are found (see Table 4) in the current study as in the original study of Kim et al. [1] between interpersonal channels and non-interpersonal channels. By additionally conducting paired t-tests, we tested whether there are significant differences between the interpersonal and non-interpersonal channels. Also here, we found in general similar results aside of that novelty returned significant, but serendipity and adoption rate became not significant. These results indicate that the sample in this reproducibility study showed less pronounced difference in the areas of serendipity and adoption rate between interpersonal and non-interpersonal channels, but more differences in novelty.

When looking at the relationship between the four evaluation criteria (i.e., relevance, diversity, novelty, and serendipity), we also found similar results as in the original study. The PCA results show the extraction of two components with a total of 67% of the variance accounted for: component 1 accounting for 43% of the variance and component 2 accounting for 24% of the variance. Our results show that the four evaluation criteria account for a

	Kim et al. [1]			Current study		
	Interpersonal ($n = 155$)	Non-Interpersonal ($n = 143$)	p	Interpersonal ($n = 118$)	Non-Interpersonal ($n = 100$)	p
Relevance	$3.47(0.86)$	**4.05**(0.70)	< 0.0001	$3.45(0.88)$	**3.90**(0.88)	< 0.0001
Diversity	**3.92**(0.77)	$3.59(0.90)$	< 0.001	**3.59**(0.95)	$3.14(1.2)$	< 0.001
Novelty	**4.09**(0.71)	$3.94(0.82)$	ns	**3.85**(0.88)	$3.44(1.0)$	< 0.01
Serendipity	**3.65**(0.96)	$3.23(0.98)$	< 0.001	**3.59**(0.85)	$3.44(0.94)$	ns
Convenience	$2.64(0.86)$	**3.05**(0.83)	< 0.0001	$3.41(1.0)$	**3.78**(1.0)	< 0.01
Frequency	$2.03(0.71)$	**2.65**(0.89)	< 0.01	$2.82(0.89)$	**3.54**(0.83)	< 0.001
Adoption rate	$38.34(22.47)$	**45.76**(22.51)	< 0.01	$38.57(26.80)$	**38.68**(27.24)	ns

Table 4. Mean values of each factor with standard deviations and significance levels. Boldfaced values indicate the higher mean values between interpersonal vs. non-interpersonal.

little less of the variance compared to the original study (75.2% with component 1 accounting for 49% and component 2 accounting for 26.2%). Although the accounted variance of the PCA is lower than of the original study, the extracted components still account for a large part of the variance.

Further linear regression show similar results as in the original study in which frequency plays a significant role in both interpersonal and non-interpersonal channels when it comes to adoption rates. Additional univariate testing show that adoption rates are influenced by diversity and novelty through interpersonal channels, while the novelty factor only plays a role on adoption rates in non-interpersonal channels. In this respect, our sample differs significantly compared to the Korean sample of the original study in which relevance and novelty factors play a role on adoption rates in interpersonal channels and relevance and diversity influenced adoption rates in non-interpersonal channels.

5.2 Deviations

Although in general similar results were found in line with the original study, there are obvious differences regarding the effect of certain evaluation criteria on the adoption rate. Where the original study found relevance and novelty to play a role in interpersonal channels and relevance and diversity in non-interpersonal channels on adoption rates, the current study found that it is especially diversity and novelty for interpersonal channels and novelty for non-interpersonal channels on adoption rates. This indicates that for the participants in the current study, they believe that non-system recommendations are able to provide them with more diverse music to listen to. Hence, music that gets recommended through friends or acquaintances consist of better ways to find new music to listen to in order to deepen or broaden people's music taste. This would suggest that people feel that music recommended through a system or service is more homogeneous (e.g., recommending music from the short tail). The need to include more diversified recommendations (or at least the perception of diversification) in those systems may play an important role on the satisfaction of users [22].

The found differences need to have further exploration to find the root-cause of occurring. A possible cause could be that cultural differences play a role [23]. The sample in the current study consisted of European participants whereas the original study only had Korean participants. Another difference in the demographics of the sample in the current study is the low amount of participants that is sharing music with others (merely 16% opposed to 62% in the original study), which could contribute to the differences in findings.

5.3 Additional findings

Aside of testing the reproducibility of the work of Kim et al. [1], we took this opportunity to further investigate other effects within interpersonal and non-interpersonal channels. Our findings show that the usability of non-interpersonal channels plays a significant role on the adoption rate of the music. In this case adding the recommended music to the user's playlist. Hence, even though improving algorithms is given a lot of attention, the usability still plays a vital role on whether recommendations of systems are being adopted by its users.

We furthermore found that gender plays a role in some aspects of interpersonal and non-interpersonal channels. In particular, males were found to receive more frequent music recommendations than females. This suggest that males in general might be using music recommendations systems more frequently than females. Additionally we found an effect of gender on receiving novel music recommendations through interpersonal channels. These results suggest that males believe that music recommendations received through friends or acquaintances are more often new or novel than females believe that they receive through interpersonal channels. Hence, it seems that the friends and acquaintances of males share and recommend more often music that is new or novel with each other.

6. ANALYSIS USING THE PRIMAD MODEL

In order to further analyze and systematically place the conducted reproducibility study, we apply the PRIMAD model [18]. The goal of PRIMAD is to systematically vary ("prime") the factors (P)latform, (R)esearch Objective, (I)mplementation, (M)ethod, (A)ctor, and (D)ata, and to categorize the various types of insights gained via the

resulting reproducibility characteristics by analyzing the variations made.

One of the core principles of the model is the underlying observation that a simple replication of experiments under completely identical conditions does not lead to any real knowledge gains, save from confirming a deterministic behavior of the process. A substantial knowledge gain, however, originates from specific priming and variation of factors. In the concrete case, we are *foremost interested in the effect of priming (D)ata*, with obvious deviations regarding (A)ctor, i.e. the persons executing the experiments, and (P)latform and (I)mplementation, which can be considered robust due to the well-established and replicable method, and are hence of lesser interest. By extending the analyses, we also vary the (M)ethod to gain further insights with regards to the unchanged overall (R)esearch Objective.

The repetition of user experiments using a different data sample with a different background provides insights on the generalization capabilities of the analysis. This offers insights on whether the original observations also hold in (slightly) different data settings. Simply using the same data but priming only the parameter settings, on the other hand, uncovers the parameter sensitivity of an analysis. While such a sensitivity evaluation should actually be performed by the initial investigation we frequently observe that studies report on the process of meta-search for the optimal parameter setting, yet sometimes failing to report on the performance variations observed. These, however, are crucial for understanding whether an insight holds as general observation and can thus be deployed, or whether a highly sensitive parameter that causes huge performance variations on minor changes to that parameter basically would limit application to predefined test settings.

More generally, the question of which factors to prime is difficult to answer. Priming of (D)ata (parameters, raw data) should be a given to derive conclusions that are not based on singular events and configurations. (M)ethod priming is essential for true confirmation of findings—and is finding additional support from the area of explainable AI, where (in principle) interpretable surrogate models are used to understand and explore the behavior of more powerful black-box models. Priming of the (P)latform, at the least, leads to information on the correctness of the (I)mplementation and completeness of the description. [2]

In the study investigated, the process was sufficiently documented and even could be reproduced without availability of the original code for the user study or for subsequent data analysis. In general, for user studies, the standards for description of study design, method description, and participants are high and expected to suffice for reproducing the study. Furthermore, as the number of settings and parameters to be documented and taken into account is much smaller than, e.g., in machine learning experiments, this should give further incentive to increasingly reproduce user studies and gather even further insights by systematically "priming" factors.

[2] Unfortunately, this aspect seems to see decreasing relevance through the trend of publicly sharing code as extensive testing of code before reuse appears to have lower priority. An investigation of the potentially harmful effects of code sharing on scientific practice are still pending, however.

7. CONCLUSION

In this work, we reproduced the study by Kim et al. [1] on differences of music recommendations obtained through non-interpersonal channels, i.e., recommender systems, and recommendations obtained through interpersonal channels, i.e. friends or acquaintances, with newly collected data (sample size $n = 142$). By re-implementing the analyses as carried out in the original study, we could largely confirm the original results and support their validity. Possibly caused by the differences in music sharing behavior and/or cultural differences (predominantly European vs. Korean study participants), we also found differences in our results mostly with regards to novelty in recommendations and their impact on adoption rate. From this we conclude that the aspect of cultural differences in music consumption should be accounted for more explicitly in future studies, i.e., by repeating and extending existing studies at larger, global scale, or specifically addressing cultures. As MIR systems ultimately are designed for users and user studies (if carried out) are the departing point for the design of these systems, findings of one existing study might not be reliable enough and generalize beyond the group it was conducted with.

For conducting the various types of reproducibility studies in MIR, using the PRIMAD model seems to be a meaningful framework and we suggest its use in future studies. We should stress, however, that in many reproducibility studies the priming does not happen due to according study design, but to compensate for lack of information available, be it a lack of parameter descriptions, the lack of sharing the data sets involved in training and evaluating systems, failure to make (thoroughly tested) code available or have sufficiently detailed information on architecture and design of, e.g., deep neural network models. Such incomplete information forces the actors in reproducibility studies to make documented assumptions on the according setting, leading most likely to slight variations in the according setting compared to the original study design.

8. ACKNOWLEDGEMENTS

This research was funded in whole, or in part, by the Austrian Science Fund (FWF) [P33526]. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

9. REFERENCES

[1] H. J. Kim, S. Y. Park, M. Park, and K. Lee, "Do channels matter? illuminating interpersonal influence on music recommendations," in *Proceedings of the 14th ACM Conference on Recommender Systems*. Virtual Event, Brazil: ACM, Sep. 2020, p. 663–668.

[2] K. R. Page, B. Fields, D. D. Roure, T. Crawford, and J. S. Downie, "Capturing the workflows of music information retrieval for repeatability and reuse," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 435–459, 2013.

[3] W. Chen, J. Keast, J. Moody, C. Moriarty, F. Villalobos, V. Winter, X. Zhang, X. Lyu, E. Freeman, J. Wang, S. Cai, and K. Kinnaird, "Data Usage in MIR: History & Future Recommendations," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*. Delft, The Netherlands: ISMIR, Nov. 2019, pp. 25–32.

[4] R. Bittner, M. Fuentes, D. Rubinstein, A. Jansson, K. Choi, and T. Kell, "mirdata: Software for Reproducible Usage of Datasets," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*. Delft, The Netherlands: ISMIR, Nov. 2019, pp. 99–106.

[5] B. McFee, E. J. Humphrey, and J. Urbano, "A plan for sustainable mir evaluation," in *Proceedings of the 17th International Society for Music Information Retrieval Conference*, New York City, United States, Aug. 2016, pp. 285–291.

[6] M. Schedl, A. Flexer, and J. Urbano, "The neglected user in music information retrieval research," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 523–539, 2013.

[7] P. Knees, M. Schedl, B. Ferwerda, and A. Laplante, "User awareness in music recommender systems," in *Personalized Human-Computer Interaction*. De Gruyter Oldenbourg, 2019, pp. 223–252.

[8] S. J. Cunningham, N. Reeves, and M. Britland, "An ethnographic study of music information seeking: Implications for the design of a music digital library," in *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*. Houston, Texas: IEEE, May 2003, p. 5–16.

[9] J. H. Lee and R. Price, "Understanding Users of Commercial Music Services through Personas: Design Implications," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*. Málaga, Spain: ISMIR, Oct. 2015, pp. 476–482.

[10] K. Andersen and P. Knees, "Conversations with Expert Users in Music Retrieval and Research Challenges for Creative MIR," in *Proceedings of the 17th International Society for Music Information Retrieval Conference*. New York City, United States: ISMIR, Aug. 2016, pp. 122–128.

[11] E. L. M. Law, L. von Ahn, R. B. Dannenberg, and M. Crawford, "Tagatune: A game for music and sound annotation," in *Proceedings of the 8th International Conference on Music Information Retrieval*. Vienna, Austria: ISMIR, Sep. 2007, pp. 361–364.

[12] J. H. Lee, "Crowdsourcing Music Similarity Judgments using Mechanical Turk," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Utrecht, Netherlands: ISMIR, Aug. 2010, pp. 183–188.

[13] C. Inskip and F. Wiering, "In Their Own Words: Using Text Analysis to Identify Musicologists' Attitudes towards Technology," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*. Málaga, Spain: ISMIR, Oct. 2015, pp. 455–461.

[14] J. H. Lee and J. S. Downie, "Survey Of Music Information Needs, Uses, And Seeking Behaviours: Preliminary Findings." in *Proceedings of the 5th International Conference on Music Information Retrieval*. Barcelona, Spain: ISMIR, Oct. 2004.

[15] J. H. Lee and S. J. Cunningham, "Toward an understanding of the history and impact of user studies in music information retrieval," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 499–521, 2013.

[16] S. Schmidt, "Shall we really do it again? the powerful concept of replication is neglected in the social sciences," *Review of General Psychology*, vol. 13, no. 2, pp. 90–100, 2009.

[17] J. Urbano, M. Schedl, and X. Serra, "Evaluation in music information retrieval," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 345–369, 2013.

[18] A. Rauber, V. Braganholo, J. Dittrich, N. Ferro, J. Freire, N. Fuhr, D. Garijo, C. Goble, K. Järvelin, B. Ludäscher, B. Stein, and R. Stotzka, "PRIMAD – information gained by different types of reproducibility," in *Reproducibility of Data-Oriented Experiments in e-Science (Dagstuhl Seminar 16041)*, J. Freire, N. Fuhr, and A. Rauber, Eds. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2016, vol. 6, no. 1, pp. 128–132.

[19] A. Resch, S. Strumbelj, and L. Tomandl, "A Reproducibility Analysis on: Do Channels Matter? Illuminating Interpersonal Influence on Music Recommendations," Jan. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4459832

[20] V. Bauer, J. Bobinac, and F. Y. Tang, "Reproducibility Study: Do Channels Matter? Illuminating Interpersonal Influence on Music Recommendations," Jan. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.5920550

[21] A. Ceranic, R. Dizdar, and A. Sokoli, "Reproduce experimental results from a paper," Jan. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.5921006

[22] B. Ferwerda, M. P. Graus, A. Vall, M. Tkalcic, and M. Schedl, "How item discovery enabled by diversity leads to increased recommendation list attractiveness,"

in *Proceedings of the Symposium on Applied Computing*. Marrakech, Morocco: ACM, Apr. 2017, p. 1693–1696.

[23] B. Ferwerda and M. Schedl, "Investigating the relationship between diversity in music consumption behavior and cultural dimensions: A cross-country analysis," in *Extended Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalisation*. Halifax, Canada: CEUR-WS, Jul. 2016.

MUSIC SEPARATION ENHANCEMENT WITH GENERATIVE MODELING

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

ABSTRACT

Despite phenomenal progress in recent years, state-of-the-art music separation systems produce source estimates with significant perceptual shortcomings, such as adding extraneous noise or removing harmonics. We propose a post-processing model (the Make it Sound Good (MSG) post-processor) to enhance the output of music source separation systems. We apply our post-processing model to state-of-the-art waveform-based and spectrogram-based music source separators, including a separator unseen by MSG during training. Our analysis of the errors produced by source separators shows that waveform models tend to introduce more high-frequency noise, while spectrogram models tend to lose transients and high frequency content. We introduce objective measures to quantify both kinds of errors and show MSG improves the source reconstruction of both kinds of errors. Crowdsourced subjective evaluations demonstrate that human listeners prefer source estimates of bass and drums that have been post-processed by MSG.

1. INTRODUCTION

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

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

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

The main contributions of this work are:

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

Audio examples and code can be found at `https://interactiveaudiolab.github.io/project/msg.html`.

2. RELATED WORK

Deep learning is the dominant approach for music source separation. For example, all entries to the 2021 Sony Music Demixing Challenge [5] were deep learning based separators. Most separators fall into one of two classes. *Waveform models* [6–9] take audio waveform input and produce an audio waveform for each separated source. *Spectrogram models* [10–18] take a mixture spectrogram as input and output a mask to apply to the spectrogram for each source being separated. Despite the recent successes of these deep learning methods, state of the art systems continue to exhibit perceptible artifacts in their outputs. We show in Section 5 that waveform models tend to introduce more high-frequency noise, while spectrogram models tend to lose transients and high frequency content.

⋆ Equal contribution

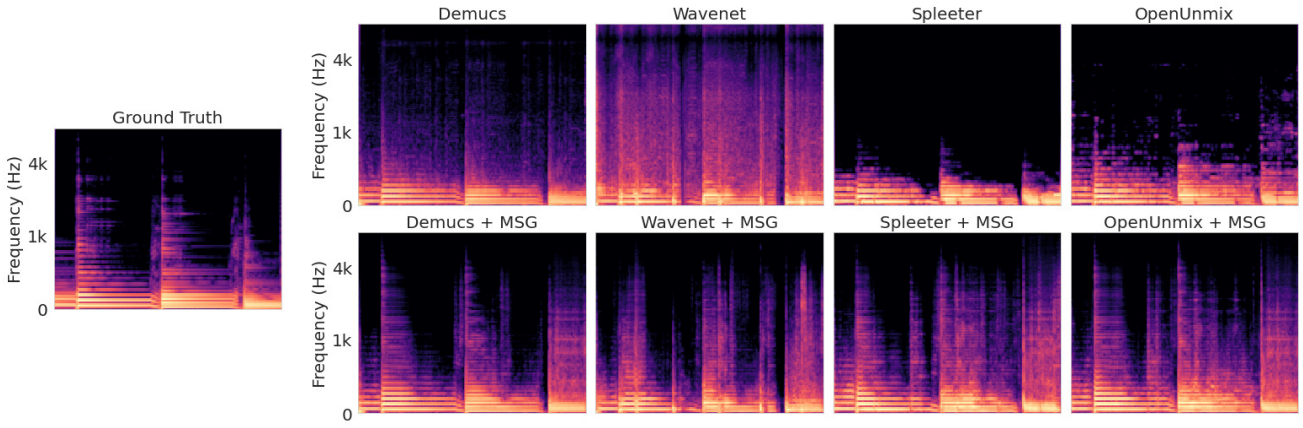

Figure 1: Spectrograms of ground-truth (left), source estimates (top), and MSG output (bottom) for the bass source. MSG is able to simultaneously infer missing frequencies and remove noise from the output of common source separation systems.

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

While many recent works for speech enhancement have been proposed, music enhancement (e.g., denoising and artifact removal) is less common. There are only a few recent works in music denoising [36–38] and bandwidth expansion [39, 40]. Some work has also looked at potential causes of [41] and remedies for [42] audio artifacts in untrained source separation networks. Our work is most similar to Kandpal et. al [43], who proposed a generative model that can enhance the audio quality of a low-quality music recording taken on a consumer device. We are unaware of a prior system for enhancing the output of trained music separation systems.

3. "MAKE IT SOUND GOOD" POST-PROCESSOR

Here we describe our "Make it Sound Good" (MSG) post-processor, which perceptually improves a source estimate by removing artifacts that the separator introduced and imputing elements the separator omitted. We use the adversarial loss of Generative Adversarial Networks (GANs) [33] due to its success denoising many types of audio.

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

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

$$L_G = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left(D_k(G(\hat{s}))-1\right)^2\right], \qquad (1)$$

where $\hat{s}$ is the raw source estimate from the separator, D_k is the k-th discriminator, K is the total number of discriminators, and G is the generator.

Next is deep feature matching loss [45], which is the L_1 distance between the intermediate activations of the discriminators on corresponding real and generated data. The last loss function we use is a multi-scale Mel-spectrogram reconstruction loss [46, 47], which is the average Mel reconstruction loss over three different Short-time Fourier transforms (STFTs), each of which uses different parameters for the number of STFT bins, window lengths, and hop sizes.

We use two types of discriminators: the multi-period discriminators from HiFi-GAN [24] and the multi-resolution spectrogram discriminators from UnivNet [25]. The multi-period discriminators operate on the waveform, and reshapes the waveform to a 2D tensor with a prime-valued stride before processing the reshaped waveform with 2D convolutional layers. The multi-resolution spectrogram discriminators process a spectrogram with different STFT window sizes (see Section 4.3). We use five multi-period discriminators with strides $[2, 3, 5, 7, 11]$, respectively. We also use three multi-resolution discriminators with FFT windows $[512, 1024, 2048]$, for a total of eight discriminators. Each discriminator uses the LSGAN [44] loss,

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

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

We use an adversarial loss typical of Generative Adversarial Networks, but do not condition on a random input vector. This produces a deterministic model that is not