3. METHODS

In general, a limiter is used as the last signal processor in music mastering of commercial music [7, 8]. Since it tends to be used excessively in modern commercial music [1, 2, 4], it should be considered in music source separation for the development of robust application.

We assumed that the key differences between the real-world mastering-finished music, i.e. a limiter applied music, and the standard training examples for music source separation are *i)* the overall amplitude scale, and *ii)* the signal distortion caused by a limiter.

Here we introduce two ways to avoid each domain mismatch problem.

3.1 Input loudness normalization

First of all, the simple loudness normalization, which is a linear gain adjustment of network inputs to the pre-defined reference level, is the easiest technique to avoid the domain shift caused by *i)* overall amplitude scale mismatch. This can be categorized into two, *(i)* input loudness normalization during both the training and evaluation stages of networks, and *(ii)* normalization only at the evaluation stage.

The method *(i)* is already used in various studies with different ways. For example, the network such as *Demucs v3* [19], uses input standardization in time domain based on the mean and standard deviation of the training data. In *Open-unmix* [6], the network implicitly normalizes the input by utilizing trainable input scaling parameters that works in time-frequency domain. However, we assumed that explicitly adjusting the gain of the inputs based on the waveform, thereby minimizing the overall amplitude scale mismatch between train and test domain, can greatly reduce the performance degradation comes from the domain shift.

The method *(ii)*, normalization only at the evaluation stage, can be considered as a readily applicable workaround for the models that were trained without considering music loudness. This method is a simple idea to reduce the overall amplitude scale difference between train and real-world domain. Since the music source separation networks are non-linear systems, we hypothesized that linear scale difference of the inputs might affect the quality of final outputs. That is, input normalization only at the evaluation stage can be a simple, yet effective trick for models that were already trained without considering the domain mismatch.

Though these methods can mitigate the domain mismatch related to *i)* amplitude scale, there needs to be another strategies that can reflect *ii)* the signal distortion caused by a limiter.

3.2 *LimitAug*

The best way to consider the characteristics of the limiter applied source is simply using the limiter during the network training. Therefore, we propose the *LimitAug* data augmentation method, which utilizes an online limiter during the training examples construction process, to reduce the domain shift between the training examples and real-world commercial music. The *LimitAug* can be considered as creating train examples that forcibly reflect the distortion that comes from a limiter, which cannot be reflected by the simple input loudness normalization technique.

```
# mix, tgt : mixture and tgt sources.
# gain(src, tgt_lufs) : calculate lufs of source,
#  then adjust its gain targeting given lufs,
#  return output and adjusted gain
# gain_adj(src, adj_gain) : gain adjustment
#  of source with given adj_gain.
tgt_lufs = random_sampled_tgt_lufs
mix_loud, _ = gain(mix, tgt_lufs)
mix_loud_limited = limiter(mix_loud)
ratio = mix_loud_limited / mix
tgt_loud = tgt * ratio
if input_loud_norm:
  mix_loud_limited, adj_gain = gain(src, tgt_lufs)
  tgt_loud = gain_adj(tgt_loud, adj_gain)
estimates = network(mix_loud_limited)
loss = objective(estimates, tgt_loud)
```

Figure 3. `numpy`-like pseudocode of the proposed *LimitAug* method.

The pseudocode for *LimitAug* is shown in Figure 3. First, calculate the LUFS of the mixture created by the data sampling process. Second, adjust the gain of the input mixture source targeting the randomly sampled LUFS value. Then, it is followed by the online limiter to fit the waveform under 0 dBFS. Lastly, calculate the sample-wise (A sample refers to an each waveform value) ratio between the limiter applied mixture source and the original mixture source, then multiply the ratio to the original target source to get the ground truth target signal of the limiter applied mixture source.

When adjusting the gain and applying the limiter to the input mixture, for example, if the input mixture had -15 LUFS and the randomly sampled target LUFS was -10, note that the gain-scaled mixture by +5 dB does not have exact -10 LUFS due to the compression of a limiter and the nature of LUFS calculation, which considers frequency weighting [11, 12].

The proposed *LimitAug* can be used with other data augmentation methods; in our study, random gain scaling, channel swap and mixing of different instrumental stem sources [9, 10] were used. Also, additional loudness normalization can be applied after the *LimitAug* as depicted in conditional statement of Figure 3, so that the overall amplitude scale mismatch between training and evaluation stages be minimized.

4. EXPERIMENTS

Here we briefly summarize our following experiments.

In Section 5.1, we quantitatively evaluated various music source separation networks that were trained without considering the loudness and heavy dynamic range compression, and confirmed the performance degradation on *musdb-L* and *musdb-XL*, compared to the origianl *musdb-hq* evaluation dataset.

network	extra train data	test data	SDR median (mean) [dB]				
			vocals	*bass*	*drums*	*other*	*avg*
Open-unmix [6]	-	*hq*	6.16 (2.54)	5.03 (2.67)	6.00 (5.46)	4.22 (3.46)	5.35 (3.53)
		L	6.33 (1.63)	4.81 (2.71)	5.82 (5.38)	4.11 (3.42)	5.27 (3.28)
		XL	5.98 (0.89)	4.76 (2.59)	4.97 (4.89)	4.04 (3.29)	4.94 (2.92)
TFC-TDF -U-net [20]	-	*hq*	7.18 (4.26)	5.59 (3.35)	5.76 (5.30)	4.04 (3.18)	5.64 (4.02)
		L	7.03 (3.65)	5.41 (3.08)	5.52 (5.09)	3.67 (3.00)	5.41 (3.71)
		XL	6.95 (3.14)	5.48 (2.90)	5.11 (4.68)	3.55 (2.82)	5.27 (3.39)
Demucs v3-A [19]	-	*hq*	**8.11 (5.22)**	**9.34** (6.21)	**8.57 (8.01)**	**5.51 (5.03)**	**7.88 (6.12)**
		L	7.54 (5.15)	9.32 **(6.22)**	8.26 (7.65)	5.51 (5.01)	7.66 (6.01)
		XL	7.30 (4.86)	9.19 (6.14)	7.62 (6.78)	5.37 (4.97)	7.37 (5.69)
Open-unmix [6]	✓	*hq*	7.02 (4.93)	5.91 (4.06)	7.18 (6.91)	4.94 (4.76)	6.26 (5.17)
		L	6.83 (5.12)	6.23 (4.09)	7.07 (6.92)	4.94 (4.78)	6.27 (5.23)
		XL	6.70 (4.77)	6.16 (3.87)	6.80 (6.48)	4.89 (4.61)	6.14 (4.93)
Spleeter [21]	25000+	*hq*	6.51 (4.42)	4.77 (3.57)	6.00 (6.09)	4.22 (4.12)	5.38 (4.55)
		L	6.18 (3.90)	4.73 (3.34)	5.67 (5.94)	4.37 (4.03)	5.24 (4.30)
		XL	6.03 (3.38)	4.80 (3.13)	5.55 (5.52)	4.24 (3.91)	5.15 (3.98)
Demucs v3-B [19]	200+ including *musdb-hq* test set	*hq*	**9.24 (7.05)**	**11.65 (9.58)**	**11.73 (11.34)**	**7.83 (8.03)**	**10.11 (9.00)**
		L	9.05 (6.91)	11.61 (9.55)	11.05 (10.27)	7.83 (7.91)	9.88 (8.66)
		XL	8.76 (6.41)	11.56 (9.29)	9.22 (8.78)	7.52 (7.51)	9.26 (8.00)

Table 2. Performance of various music source separation networks trained with *musdb-hq* [14] on the test using *musdb-hq*, *musdb-L* and *musdb-XL*.

In Section 5.2, we simply normalized the network inputs in the evaluation stage, to observe the performance degradation caused by the signal distortion caused by a limiter, not an overall amplitude mismatch between the training and the evaluation data. Then, we carefully analyzed the results on *Demucs v3* [19], which took the 1st place in the 2021 Sony Music Demixing (MDX) Challenge leaderboard A and the 2nd place in the leaderboard B [22]. The leaderboard A was the competition that only allowed training on the only *musdb-hq* train subset, and the leaderboard B allowed the extra training data.

In Section 5.3, we conducted a comparative study on various training data construction strategies. Unfortunately, we were unable to train the current state-of-the-art *Demucs v3*, due to the constraint of our GPU experimental environments. Considering the training efficiency and reasonable performance, we used *TFC-TDF-U-Net* [20] — a backbone architecture of KUIELab-MDX-Net [23] with slight modifications, which took the 2nd place in the 2021 MDX Challenge leaderboard A — in this experiment.

4.1 Training

In Section 5.1 and Table 2, we used official pre-trained weights of each model except *TFC-TDF-U-Net* since there were no official weights for the 4 stems of *musdb-hq*. For fair comparison in Section 5.3, we trained *TFC-TDF-U-Net* for 300 epochs with early stopping, based on official training framework of *Open-unmix* [6] and official network implementation of *TFC-TDF-U-net*. We used default network hyper-parameters introduced in its webpage [5] .

For the *LimitAug*, we used the limiter implemented in `pedalboard` [24]. The threshold parameter was set to 0 dBFS, and the release parameter was randomly sampled from uniform distribution of (30, 200) millisecond in data sampling process. Also, we used `pyloudnorm` [25] for loudness calculation.

[5] https://github.com/ws-choi/ISMIR2020_U_Nets_SVS

4.2 Evaluation

In all of the evaluations, Signal-to-Distortion Ratio (SDR) [26] was calculated using `museval` python library [18]. Also, in following experimental results, both *median* and *mean* SDR scores were presented for the detailed comparison. Note that we only used the *musdb-hq* train subset for training the networks. *musdb-L* and *XL* are only for evaluation.

In the original *musdb-hq* test subset, as stated in the official webpage [6] , 'PR - Oh No' track's mixture signal is panned to the right channel, which causes the inconsistency between linear summation of stems and mixture source. Since this causes the limiter to be operated in unmusical way while the construction of *musdb-L* and *XL*, which also can be hardly found in popular music, we used the linear summation of stems as a mixture only for this track. This may results in the slight difference of SDR scores between the Table 2 and the official scores of each network.

5. RESULTS

5.1 Performance degradation on *musdb-L* and *XL*

Here we quantitatively evaluated the performance of state-of-the-art networks on *musdb-hq* [14], *L* and *XL* datasets and confirmed that the domain shift in perspective of music loudness and compression negatively affect the performance, indeed. As shown in the Table 2, all networks showed significant performance degradation on the evaluations with *musdb-L* and *musdb-XL* datasets. The amount of decrease on *Demucs v3* [19] was slightly larger than the others. Overall, we concluded that every networks are somewhat overfitted to the *musdb-hq* data, making the networks vulnerable to loud and heavily compressed music. Therefore, it is highly required to consider these characteristics for the robust music source separation. Note that

[6] https://sigsep.github.io/datasets/musdb

network	extra train data	test data	SDR median (mean) [dB]				
			vocals	*bass*	*drums*	*other*	*avg*
Demucs v3-A [19]	-	*hq*	**8.11** (5.22)	**9.34 (6.21)**	**8.57 (8.01)**	5.51 (**5.03**)	**7.88 (6.12)**
		L	8.05 (**5.23**)	9.25 (6.20)	8.47 (7.92)	**5.53** (5.02)	7.82 (6.09)
		XL	7.93 (5.03)	9.27 (5.92)	7.74 (7.44)	5.55 (4.91)	7.62 (5.82)
Demucs v3-B [19]	200+ including *musdb-hq* test set	*hq*	**9.24 (7.05)**	**11.65 (9.58)**	**11.73 (11.34)**	**7.83 (8.03)**	**10.11 (9.00)**
		L	9.19 (7.04)	11.64 (9.55)	11.68 (11.21)	7.82 (8.02)	10.08 (8.95)
		XL	9.13 (6.90)	11.56 (9.33)	11.32 (10.75)	7.74 (7.95)	9.94 (8.73)

Table 3. Performance of *Demucs v3* [19] trained with *musdb-hq* [14]. On the test using *musdb-L* and *musdb-XL*, each of the input sources were loudness normalized targeting the LUFS of its original *musdb-hq* sources.

network	extra train data	SDR median [dB]		
		hq	*L*	*XL*
Open-unmix [6]	-	5.35	5.32	5.25
TFC-TDF-U-Net [20]	-	5.64	5.62	5.51
Demucs v3-A [19]	-	**7.88**	**7.82**	**7.62**
Open-unmix [6]	✓	6.26	6.25	6.18
Spleeter [21]	✓	5.38	5.33	5.21
Demucs v3-B [19]	✓	**10.11**	**10.08**	**9.94**

Table 4. Performance of networks with loudness normalized input at the evaluation stage. Each score represents the average score across the 4 stems. Note that the networks were trained without considering music loudness or dynamic range compression explicitly.

we stated the scores on different test datasets in each block to emphasize the performance degradation between the domains.

5.1.1 Extra training data

Though the extra training data was of help, it did not work as the fundamental solution to the domain shift. *Open-unmix* [6] with extra training data showed more robust performance than that of the network without extra data, but performance degradation on *Demucs v3-B* was more significant than that of *Demucs v3-A*. Since *Demucs v3-B* was trained with extra training data including *musdb-hq* test subset, it is reasonable to guess that the network is highly overfitted to the *musdb-hq* data. Nevertheless, the amount of performance decrease especially on *drums* comparing the scores between *musdb-hq* and *XL*, 2.5 dB, is somewhat high. Considering *Demucs v3* uses an input standardization technique based on mean and standard deviation of waveform values both at training and evaluation stages, this implies that there needs to be another solution for the robust music source separation.

5.1.2 Performance degradation on drums and vocals

It is noteworthy that the degradation on *drums* and *vocals* were more significant than the others. Due to the percussive nature of *drums*, in general, they have the biggest momentary energy in music. Also, since *vocals* are important ingredients in modern commercial music, they usually consist of not only a single singing voice but also doubling, harmony and chorus. Therefore, we assumed *drums* and *vocals* are most affected by the limiter, which is activated when loud input sources that are above the threshold are given. To quantitatively confirm the signal distortion by the limiter, we calculated Scale-invariant Signal-to-Distortion Ratio (SI-SDR) [27] between *musdb-hq* and *musdb-XL* for each stem. As a result, *drums* and *vocals* scored each 19.97 and 23.69 dB, on the other hand, *bass* and *other* scored each 25.12 and 25.48 dB on average. That is, the signal distortion caused by a limiter is more significant on *drums* and *vocals*.

Unfortunately, since the networks in the Table 2 have never seen these kinds of distorted *drums* or *vocals* as training examples, we assumed that the degradation on *drums* and *vocals* were significant compared to the rest. This result strongly supports the necessity of considering the heavy dynamic range compression from the training stage, i.e. the *LimitAug*, which forcibly makes the distorted and compressed training examples for training purposes, thereby minimizing the domain shift.

5.2 Analysis on the input normalization at the evaluation stage

If the key differences between the real-world music and the training examples of music source separation networks are *i)* overall amplitude scale, and *ii)* the signal distortion caused by a limiter, as stated in Section 3, then what if we give the loudness normalized *musdb-L* or *XL* data as inputs to the networks in Table 2? Due to the non-linear nature of deep neural networks, we assumed that simply normalizing the amplitude scale of the networks on the evaluation stage may affect the performance.

The inference was conducted with following procedures, *(i)* reducing the loudness of *musdb-L* or *XL* input so that its loudness becomes same with that of the corresponding original *musdb-hq* data, *(ii)* inference with loudness normalized input, and *(iii)* increasing the scale of output as much as reduced in step *(i)*.

Comparing the scores between Table 2 and Table 4, we confirmed that the performance degradation was greatly alleviated by just the simple loudness normalization of the inputs only at the evaluation stage. Note that the networks were not trained with loudness normalized inputs. This result shows that the input loudness normalization at the evaluation stage can be a quick and easy solution to get robust results from the pre-trained music source separation networks.

Especially, on *drums* of *Demucs v3-B*, it should be noted that the amount of decrease on *median* SDR score between the test using *musdb-hq* and *XL* was sharply reduced from 2.5 dB in Table 2 to 0.4 dB in Table 3. This implies that the network is overfitted not only to the contents of the signal, but also to the scale or loudness, especially

network	methods	linear gain increase	*LimitAug*	input loud-norm	target LUFS	SDR median (mean) [dB]			
						hq	*L*	*XL*	*avg*
TFC-TDF -U-Net [20]	baseline	-	-	-	-	5.64 (4.02)	5.41 (3.71)	5.27 (3.39)	5.44 (3.71)
	(1)	✓	-	-	$\mathcal{N}(\mu_L, \sigma_L^2)$	**5.90** (4.31)	5.86 (4.33)	5.73 (4.15)	5.83 (4.26)
					$\mathcal{N}(\mu_{XL}, \sigma_{XL}^2)$	5.32 (3.43)	5.36 (3.62)	5.28 (3.49)	5.32 (3.51)
	(2)	-	✓	-	$\mathcal{N}(\mu_L, \sigma_L^2)$	5.79 (4.30)	**5.90 (4.41)**	5.74 (4.25)	5.81 (4.32)
					$\mathcal{N}(\mu_{XL}, \sigma_{XL}^2)$	5.69 (3.93)	5.72 (4.22)	5.57 (4.15)	5.66 (4.10)
	(3)	-	-	✓	-14	5.89 **(4.38)**	5.87 (4.35)	**5.82 (4.25)**	**5.86 (4.33)**
	(4)	-	✓	✓	$\mathcal{N}(\mu_L, \sigma_L^2)$, -14	5.87 (4.25)	5.85 (4.21)	5.76 (4.16)	5.83 (4.21)
					$\mathcal{N}(\mu_{XL}, \sigma_{XL}^2)$, -14	5.78 (4.27)	5.78 (4.26)	5.73 (4.20)	5.76 (4.24)

Table 5. Performance of *TFC-TDF-U-Net* [20] trained with various training data construction strategies. Each score represents the average score across the 4 stems.

network	methods	target LUFS	test data	SDR median (mean) [dB]				
				vocals	*bass*	*drums*	*other*	*avg*
TFC-TDF -U-Net [20]	baseline	-	*hq*	7.18 (4.26)	5.59 (3.35)	5.76 (5.30)	4.04 (3.18)	5.64 (4.02)
			L	7.03 (3.65)	5.41 (3.08)	5.52 (5.09)	3.67 (3.00)	5.41 (3.71)
			XL	6.95 (3.14)	5.48 (2.90)	5.11 (4.68)	3.55 (2.82)	5.27 (3.39)
	(3) loud-norm	-14	*hq*	7.35 **(4.76)**	**5.93** (3.61)	**5.91 (5.37)**	4.39 (3.79)	**5.89 (4.38)**
			L	7.32 (4.72)	5.91 (3.61)	5.85 (5.29)	4.39 (3.78)	5.87 (4.35)
			XL	7.26 (4.64)	5.91 **(3.62)**	5.68 (4.99)	4.42 (3.78)	5.82 (4.25)
	(4) LimitAug, loud-norm	$\mathcal{N}(\mu_L, \sigma_L^2)$, -14	*hq*	**7.59** (4.64)	5.75 (3.25)	5.63 (5.28)	4.50 (3.82)	5.87 (4.25)
			L	7.58 (4.61)	5.69 (3.21)	5.62 (5.22)	4.50 (3.82)	5.85 (4.21)
			XL	7.48 (4.55)	5.67 (3.29)	5.36 (4.99)	**4.51 (3.82)**	5.76 (4.16)

Table 6. Stem-wise performance of *TFC-TDF-U-Net* [20] trained with the method *(3)* input loudness normalization, and *(4)* input loudness normalization after the proposed *LimitAug*.

on *drums*. Note that there was no distinction between the given input sources to the networks except the linear gain difference.

Although the performance degradation was reduced by the input loudness normalization, still there exists the performance degradation due to the signal distortion caused by a limiter. Similar to Section 5.1.2, this result also supports the necessity of the *LimitAug* for robust music source separation.

5.3 Analysis on various training strategies

In this section, we trained the *TFC-TDF-U-net* [20] with various training data creating methods; *(1)* linear gain increasing, *(2)* the proposed *LimitAug*, *(3)* input loudness normalization, and *(4)* input loudness normalization after the *LimitAug*. Of course, the methods *(3)* and *(4)* includes the input loudness normalization at the evaluation stage for the consistency between train and test domains. We compared the results to check which one is the most powerful way for training robust music source separation networks. For the input normalization, we chose the target reference LUFS value as -14.

In Table 5, we confirmed that all of the methods were effective for robust music source separation; every methods showed relatively robust performance on *musdb-L* and *XL*, compared to the baseline. Especially, except the method *(1)* targeting LUFS of a normal distribution following statistics of *musdb-XL*, $\mathcal{N}(\mu_{XL}, \sigma_{XL}^2)$, all methods showed greater performance on *musdb-hq* than the baseline. This result implies that these methods are useful not only for the domain shift, but also for the standard benchmark data.

Furthermore, the methods *(3)* and *(4)*, which prevent the domain shift caused by overall amplitude scale mismatch by input loudness normalization, showed slightly better performances than others. In the stem-wise analysis as presented in Table 6, though we expected that the *LimitAug* would be of help for *vocals* and *drums*, the method *(4)* was better at *vocals* and *other* than the method *(3)*.

Overall, it seems obvious that considering the music loudness and heavy dynamic range compression from the training stage is beneficial for robust music source separation. For real world applications, it is highly recommended that not just using the single method we experimented, but using various training methods on different stems or using a bag of models trained with various methods.

6. CONCLUSIONS

In this study, we questioned the domain shift between the research and the real-world data for music source separation, from the viewpoint of music loudness and heavy dynamic range compression. To answer this, We first built new evaluation datasets, *musdb-L* and *musdb-XL*, which reflect dynamic range compressed music characteristics and heavy loudness. Then, we confirmed the significant performance degradation of state-of-the-art networks on our datasets. To alleviate this, we conducted various experiments on training data construction strategies, including the proposed *LimitAug* method, and confirmed that the methods using the input loudness normalization only or with the *LimitAug* greatly improved the robustness. We hope that our proposed methods and evaluation datasets could contribute to future music source separation research and application.

7. ACKNOWLEDGEMENTS

We appreciate Zafar Rafii, the author of *musdb*, for allowing us to reprocess the original musdb data. We thank Antoine Liutkus, also the author of *musdb*, for giving the creative suggestion on the distribution of our proposed datasets. We are grateful to Ben Sangbae Chon, Keunwoo Choi, and Hyeongi Moon from GaudioLab for their fruitful discussions on our proposed methods. Last but not least, we thank Donmoon Lee, Juheon Lee, Jaejun Lee, Junghyun Koo, and Sungho Lee for helpful feedbacks.

8. REFERENCES

[1] S. Dredge, "Pop music is louder, less acoustic and more energetic than in the 1950s," *The Guardian*, 25 Nov. 2013.

[2] G. Milner, "They really don't make music like they used to," *The New York Times*, 7 Feb. 2019.

[3] N. B. Croghan, K. H. Arehart, and J. M. Kates, "Quality and loudness judgments for music subjected to compression limiting," *The Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1177–1188, 2012.

[4] E. Vickers, "The loudness war: Background, speculation, and recommendations," in *Audio Engineering Society Convention 129*. Audio Engineering Society, 2010.

[5] E. F. Stikvoort, "Digital dynamic range compressor for audio," *Journal of the Audio Engineering Society*, vol. 34, no. 1/2, pp. 3–9, 1986.

[6] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix - a reference implementation for music source separation," *Journal of Open Source Software*, 2019. [Online]. Available: https://doi.org/10.21105/joss.01667

[7] B. Katz and R. A. Katz, *Mastering audio: the art and the science*. Butterworth-Heinemann, 2003.

[8] L. Bregitzer, *Secrets of recording: Professional tips, tools & techniques*. Routledge, 2008.

[9] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 261–265.

[10] L. Prétet, R. Hennequin, J. Royo-Letelier, and A. Vaglio, "Singing voice separation: A study on training data," in *ICASSP 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2019, pp. 506–510.

[11] R. ITU-R, "Itu-r bs. 1770-2, algorithms to measure audio programme loudness and true-peak audio level," *International Telecommunications Union, Geneva*, 2011.

[12] R. EBU-Recommendation, "Loudness normalisation and permitted maximum level of audio signals," *European Broadcasting Union*, 2011.

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

[14] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "Musdb18-hq - an uncompressed version of musdb18," Aug. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3338373

[15] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 45–49.

[16] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *International conference on latent variable analysis and signal separation*. Springer, 2017, pp. 323–332.

[17] R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "Medleydb: A multitrack dataset for annotation-intensive mir research." in *ISMIR*, vol. 14, 2014, pp. 155–160.

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

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

[20] W. Choi, M. Kim, J. Chung, D. Lee, and S. Jung, "Investigating u-nets with various intermediate blocks for spectrogram-based singing voice separation," in *21th International Society for Music Information Retrieval Conference, ISMIR, Ed*, 2020.

[21] 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.

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

[23] M. Kim, W. Choi, J. Chung, D. Lee, and S. Jung, "Kuielab-mdx-net: A two-stream neural network for music demixing," *arXiv preprint arXiv:2111.12203*, 2021.

[24] Spotify, "pedalboard : A python library for manipulating audio." 2022. [Online]. Available: https://github.com/spotify/pedalboard

[25] C. J. Steinmetz and J. D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *150th AES Convention*, 2021.

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

[27] J. Le 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)*. IEEE, 2019, pp. 626–630.

TOWARDS QUANTIFYING THE STRENGTH OF MUSIC SCENES USING LIVE EVENT DATA

Michael Zhou
Columbia University
mgz2112@columbia.edu

Andrew McGraw
University of Richmond
amcgraw@richmond.edu

Douglas R. Turnbull
Ithaca College
dturnbull@ithaca.edu

ABSTRACT

There are many benefits for a community when there is a vibrant local music scene (e.g., increased mental & physical well-being, increased economic activity) and there are many factors that contribute to an environment in which a live music scene can thrive (e.g., available performance spaces, helpful government policies). In this paper, we explore using an estimate of the *live music event rate* (LMER) as a rough indicator to measure the strength of a local music scene. We define LMER as the number of music shows per 100,000 people per year and then explore how this indicator is (or is not) correlated with 28 other socioeconomic indicators. To do this, we analyze a set of 308,051 music events from 2019 across 1,139 cities in the United States. Our findings reveal that factors related to transportation (e.g., high walkability), population (high density), economics (high employment rate), age (high proportion of individuals age 20-29), and education (bachelor's degree or higher) are strongly correlated with having a high number of live music events. Conversely, we did not find statistically significant evidence that other indicators (e.g., racial diversity) are correlated.

1. INTRODUCTION

Certain American cities such as Los Angeles and New York City are famous for having great music scenes. However, there are hundreds of small, medium, and large cities around the country that support vibrant music scenes. These cities often have well-known venues like the Grand Ole Opry in Nashville, TN, or put on large music festivals such as SXSW in Austin, TX. They can be associated with artists who obtain some level of regional, national, or international fame such as Minneapolis, MN with Prince and Asbury Park, NJ with Bruce Springsteen. They sometimes become historically connected with specific genres as in New Orleans, LA with jazz, Seattle, WA with grunge, and Asheville, NC with bluegrass.

There have been many studies that detail the ways in which music scenes can benefit their communities by enhancing social bonding, improving emotional well-being [1,2], and increased economic activity [3–15]. Researchers have also studied factors that can help foster an environment in which a local music scene can develop and thrive [4, 5, 8, 9, 11, 12, 16–22]. These factors include having a rich music history, having strong support for music education, and government regulations that are favorable for live performance (see Section 2 for details).

Investment by government and non-government organizations (e.g., Chambers of Commerce, Arts Councils) are associated with strong music scenes ("music havens") to both further develop these havens as well as help cities with underdeveloped music scenes ("music deserts") [11]. Many organizations have produced extensive reports [19] about "music cities" based on interviews and surveys of cities around the world (e.g., Austin, USA [3], London, UK [23], Victoria, AU [24]). While these reports produce valuable and transferable knowledge, they tend to be narrow in their geographic focus (i.e., one city or region). To complement this body of work, we propose a quantitative approach that uses the live music event rate (LMER) to estimate the *relative* strength of a local music scene. We argue that this simple statistic is straightforward to calculate, easy to interpret, and useful.

In this work, we introduce a music event dataset that contains information for 308,051 live music events that took place in 1,139 American cities during 2019. Here we consider an event as a live performance by one or more artists at a venue (e.g., bar, concert hall, festival) on a given date. This dataset was collected for music event recommendation application and combines event information that was collected from BandsInTown [1] and Facebook [2]. While this dataset has a number of limitations (e.g., only music events with a digital footprint, data collected using snowball sampling), it allows us to calculate a rough estimate of LMER for each city to enable comparison.

In this paper, we explore how our estimated LMER is (or is not) correlated with 28 socioeconomic indicators across 6 different categories: transportation, population, economics, age, education, and race. These indicators are closely related to some of the factors (e.g., transportation convenience, population density) that other researchers have suggested are important factors for fostering healthy music scenes.

In Section 2, we explore existing research on the ben-

Attribution: M. Zhou, A. McGraw, and D. Turnbull, "Towards Quantifying the Strength of Music Scenes using Live Event Data", in *Proc. of the 23rd Int. Society for Music Information Retrieval Conf.,* Bengaluru, India, 2022.

[1] https://www.bandsintown.com/
[2] https://www.facebook.com/

Benefits of Music Communities	Factors that Create a Strong Music Community
Job creation [3, 4, 11, 12]	Availability of performance spaces and venues [11]
Increased consumer spending [6, 8, 9]	Financial affordability [11, 20]
Patronages & sponsorships [5, 11, 12]	Music tourism [8, 9, 11]
Increased financial investments [3, 11]	Music education resources [11, 12, 22]
Lower poverty rates [25]	Music heritage and history [4, 11]
Strengthening social bonding [11]	Demographics (e.g. student populations) [4, 11, 12, 17]
Improved mental/physical health [1, 2]	Government support/regulation [5, 11, 12, 16]
Lower crime rates [7, 13, 15]	Transportation convenience [11, 20]
Better community reputation [3, 11]	Population density [19]
	Population growth [18, 21]

Table 1: Table summarizing the benefits of a music community and factors that stimulate a music community.

efits of and factors that support a strong music scene. In Section 3, we introduce our dataset of music event information and describe how we estimate LMER. Section 4 explores how LMER is correlated with different socioeconomic indicators some of which have been identified as being important factors in healthy music scenes. We then conclude in Section 5 with a discussion of how LMER can be useful for identifying potential music deserts.

2. RELATED WORK

In this section, we review both academic and music industry research that focuses on the following two questions:

1. What are the benefits of having a strong music scene?
2. What factors help foster a strong music scene?

A summary of our review is provided in Table 1.

2.1 Benefits of a Strong Music Community

Vibrant music scenes support local economies [3,4,11,12]. A live music event involves working musicians, booking and ticketing agents, sound and lighting technicians, bartenders, security guards, etc. Consumers spend money [6, 8, 9] on concert tickets, food and beverages, and artist merchandise. If a music community is strong enough, it also attracts tourism, where people from outside the community generate revenue for the local economy [8, 9, 11] through hotel stays and other local attractions (museums, parks, etc.). Live music scenes include many small businesses (music and record stores, recording studios, and private music teachers) and large institutions (music conservatories, theaters, and academic institutions). Finally, a vibrant music scene can generate economic knock-on benefits through patronage, sponsorship, cross-promotion, and cross-pollination relationships with other economic sectors [3, 5, 11, 12]. Such cultural-economic interactions can enhance a location's reputation and quality of life, attracting new residents and new participants to a scene.

The social and individual benefits of music are also well-studied. Music strengthens the social fabric of a community by "building bridges between cultures and languages, connecting people within a city, a region and across borders," as music "touches human beings" and "engages people" [11]. Vibrant music scenes are a component of the "social infrastructure" of healthy communities [26], spaces in which community members of various backgrounds can engage and interact with one another. They are physical spaces that afford and encourage face-to-face social interaction, the development of dense social networks, and the cultivation of shared values.

Many of these spaces gather together people across the social divisions of class, race, ethnicity, and faith. They are contexts in which people of different backgrounds can encounter one another in a common and often public space. According to the sociologist Robert Putnam, in the context of large pluralistic societies such as the USA, social activities such as music cultivate "bridging social capital." These are contexts in which we connect to people potentially unlike ourselves and through this cultivate a sense of "civic virtue" that widens "our awareness of the many ways in which our fates are linked" [27].

Empirical analyses of large social datasets suggest that engagement in arts events is associated with increased happiness and satisfaction measures [2]. Ethnographic case studies of particular music scenes suggest that participation in vibrant scenes is associated with subjective well-being [1, 11]. Industry analyses and censuses associate strong scenes with enhanced civic pride and cultural reputation [3, 11]. Our ongoing work will explore possible correlations between live, public music events across America and a wide range of well-being indicators, including political participation and the experience of belonging, trust, and reciprocity.

2.2 Factors that Create a Strong Music Community

Terrill et al. [11] identifies five essential elements of strong music scenes: a large number of active musicians, a community that supports diverse musical offerings, a variety of spaces for rehearsing, recording, and performance, along with a receptive and engaged audience, and a variety of music-related business. Other helpful factors include multi-level government support [5, 12, 16] (e.g., cultural zones, and accommodating noise ordinances and liquor laws [28]), city infrastructure (transportation, affordable housing), and music education [12, 22] from public school