

AMICABLE EXAMPLES FOR INFORMED SOURCE SEPARATION

Naoya Takahashi, Yuki Mitsufuji

Sony Group Corporation, Japan

ABSTRACT

This paper deals with the problem of informed source separation (ISS), where the sources are accessible during the so-called *encoding* stage. Previous works computed side-information during the encoding stage and source separation models were designed to utilize the side-information to improve the separation performance. In contrast, in this work, we improve the performance of a pre-trained separation model that does not use any side-information. To this end, we propose to adopt an adversarial attack for the opposite purpose, i.e., rather than computing the perturbation to degrade the separation, we compute an imperceptible perturbation called amicable noise to improve the separation. Experimental results show that the proposed approach selectively improves the performance of the targeted separation model by 2.23 dB on average and is robust to signal compression. Moreover, we propose multi-model multi-purpose learning that control the effect of the perturbation on different models individually.

Index Terms— informed source separation, adversarial example

1. INTRODUCTION

Audio source separation has been intensively studied over the last decade [1–5] owing to its wide range of applications such as karaoke, remixing, spatial audio, and many other downstream tasks. Although the separation performance has greatly improved thanks to recent advances in deep neural network (DNN)-based methods, it remains far from perfect in many challenging scenarios including the separation of music containing many instrumental sounds mixed in a stereo format [6–12]. In some cases, such as music production, sources can be assumed to be known during the mixing stage. Informed source separation (ISS) takes advantage of this and computes side-information during the so-called *encoding* stage. Side-information can be either embedded into mixtures by using watermark approaches [13, 14] or simply transmitted along with the mixtures [15]. Separation models are designed to utilize the side-information to improve the performance.

In this work, different from previous works, we adopt a pre-trained DNN-based separation model that is trained without any side-information for ISS. Rather than modifying the pretrained separation model to use the side-information, we compute an imperceptible perturbation that is carefully designed to improve the separation of the model and add it to the mixture. The proposed approach is closely related to adversarial examples, which were originally discovered in image classification [16], that is, imperceptibly small perturbations can significantly alter DNN predic-

tions. The proposed method in this paper can be seen as an application of adversarial attacks for the opposite purpose, namely, the perturbation is computed to improve the separation rather than degrade it. In this analogy, we refer to samples computed by the proposed method as *amicable examples*. However, the effectiveness of amicable examples is unclear as they can have potentially different properties from adversarial examples. This is because (i) while (untargeted) adversarial examples only need to be apart from targets and many possible perturbation can degrade the separation, amicable examples have concrete targets; thus, amicable examples may be difficult to find or less effective; (ii) if the loss curve becomes flat around the target y but becomes steep away from the target, the improvement of an amicable example may be not as significant as that of an adversarial example; (iii) amicable examples may be more prone to stacking with local optima. In our experiment, we provide both quantitative and qualitative evaluations of the effect of amicable examples.

An advantage of the proposed method is that since the separation model is not modified to use side-information, the model can be used for both standard mixtures and amicable examples in a unified manner. Moreover, we show that amicable examples are robust against audio signal compression, which allows us to transmit amicable examples at low bit rates. As shown in our experiment, amicable examples can selectively improve the performance of the targeted separation model and have very limited effects on untargeted separation models. Although this is often a desirable property, having explicit control of the effects of amicable examples on multiple models is more desirable. To this end, we propose the use of multi-model multi-purpose perturbation learning (MMPL) to control the effect of perturbation in both positive and negative ways depending on the separation model.

The contributions of this work are summarized as follows:

1. We propose amicable examples, the opposite optimization problem to adversarial examples, and apply them to ISS. The proposed method allows us to use the same separation model universally under both informed- and non-informed conditions.
2. We investigate the effectiveness of amicable examples on targeted and untargeted models and show the selective effectiveness for the targeted model. We further show the robustness of amicable examples against distortions caused by signal compression.
3. We further propose MMPL to control the effects of the perturbation against multiple models individually.
4. We show that, by using MMPL, amicable and adversarial examples can co-exist, namely, a perturbation can significantly

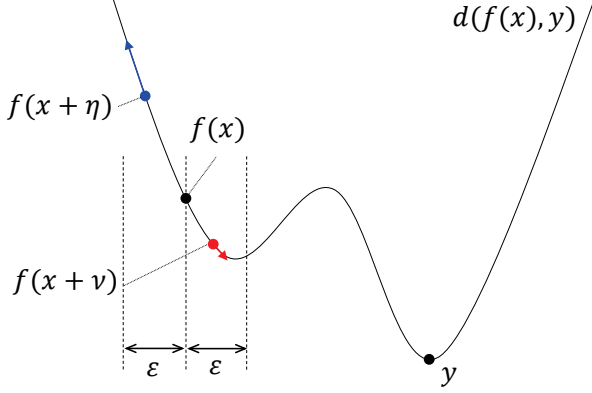


Fig. 1. Simplified visualization of a loss curve for the separation $f(x)$. For simplicity, we consider the l_∞ norm for the constraint ($\|\nu\|_\infty < \epsilon$). For the adversarial example $x + \eta$, the input x is perturbed towards the direction to increase the criterion $d()$ within an ϵ -ball while the perturbation ν for the amicable example decreases it. The minimum loss can be obtained when $x + \nu = y$, which may be outside of the ϵ -ball.

improve the performance of some models and significantly degrade the performance of others.

2. RELATED WORKS

Side-information approaches: In [13], the modified discrete cosine transform (MDCT) coefficients are used to form the side-information, and separation is performed by estimating a mask from the encoded MDCT coefficients at each time-frequency point by assuming the sparsity of sources. In [14, 15], local Gaussian models (LGM) are adopted to solve ISS problems. Bläser *et al.* applied non-negative tensor factorization (NTF) to ISS, where factorized matrices are compressed and transmitted as the side-information and reconstructed matrices are used to calculate the parameters of the Wiener filter [17]. Another line of works that closely related to ISS is to use extra information such as a music score or text [18–20], or to use an extra model trained for other tasks such as automatic speech recognition [21] to leverage knowledge from other domains. Existing ISS approaches heavily rely on the side-information, and separation models are specifically designed to use the side-information. Therefore, such models cannot perform separation or perform poorly if the side-information is not available.

Adversarial examples: Since adversarial examples can be a crucial problem for many DNN-based systems, they have been intensively investigated from different aspects including attack methods [22], defense methods [23, 24], transferability [25, 26], and the cause of network vulnerabilities [27, 28]. Recently Takahashi *et al.* investigated adversarial examples on audio source separation and reported that some attack methods are effective with limited transferability [29].

3. AMICABLE-EXAMPLE-BASED INFORMED SOURCE SEPARATION

Given N sources $y = [y_1, \dots, y_N]$ and a mixture $x = \sum_{i=1}^N y_i$, a DNN-based separation model $f_\theta()$ is trained to minimize the expectation of a training criterion $d()$ across data D as

$$\min_{\theta} \mathbb{E}_{(x,y) \in D} [d(f_\theta(x), y)], \quad (1)$$

where θ denotes network parameters. A typical choice for $d()$ is l_1 or l_2 distance. We use l_2 distance in this work. Unlike conventional ISS, where the separation model is designed to use the side-information $\psi_\omega(y)$ encoded from y and optimize the parameters θ, ω as $\min_{\theta, \omega} d(f_\theta(x, \psi_\omega(y)), y)$, we fix the separation model parameters θ and instead compute a perturbation ν that minimizes the criterion under a constraint \mathcal{C} on the perturbation as

$$\min_{\nu \in \mathcal{V}} d(f_\theta(x + \nu), y), \quad \mathcal{V} = \{\nu \mid \mathcal{C}(\nu) < \epsilon\}. \quad (2)$$

The perceptibility of the perturbation ν highly depends on the input mixture x to be added, for example, low-level noise can be perceptible when the mixture is also low-level, and high-level noise can be hardly perceptible when the mixture is also high-level. To incorporate the masking effect, we use short-term power ratio (STPR) regularization [29] as the constraint, i.e.,

$$\mathcal{C}_{\text{STPR}}(\nu) = \|\vartheta(\nu, l) / \vartheta(x, l)\|_1, \quad (3)$$

where $\vartheta(\nu, l) = [\|\nu_1\|_2, \dots, \|\nu_N\|_2]$ is the framewise l_2 norm function, which computes the norms of short frames $\nu_n = [\nu(t_n), \dots, \nu(t_n + l)]$ of length l starting from time index $t_n = (n - 1)l$. We use $l = 4096$ samples. Unlike adversarial examples, where the perturbation can be arbitrarily large without the constraint on the magnitude of the perturbation, amicable noise ν can be *self-regularized*; the perturbation may not become too large without any constraint because injecting too large perturbation in the mixture itself makes the separation difficult and thus, the amicable perturbation may inherently need to be small to minimize the separation error. Nevertheless, we found that the constraint is essential not only for regularizing the magnitude of the perturbation but also for robustly obtaining improvements in the separation.

By introducing a Lagrange weight λ , (2) can be solved by minimizing the loss function L using stochastic gradient descent:

$$L(\nu) = \|f(x + \nu) - y\|_2^2 + \lambda \mathcal{C}_{\text{STPR}}(\nu). \quad (4)$$

We omit θ for the sake of clarity.

If the negative of the first term is used instead, (4) results in the loss function for the adversarial example. However, optimization behavior can be different depending on the loss surface, as shown in Fig. 1. If the loss curve becomes flat around the (local) optimal point but becomes steep away from the optimal point, the improvement by amicable examples may not be as significant as that by adversarial examples and vice versa.

4. INCORPORATING MULTIPLE MODELS FOR MULTIPLE PURPOSES

An amicable example perturbs the mixture towards the direction where the separator *believes* it sounds more like the target sources.

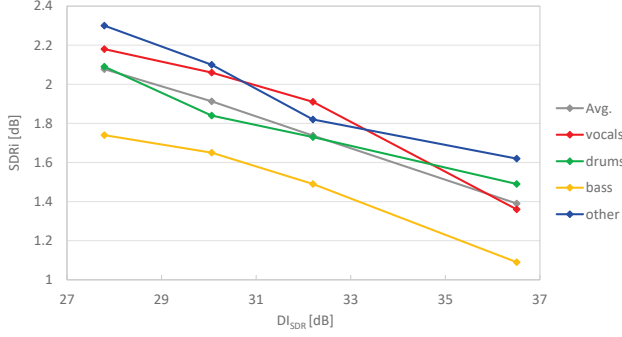


Fig. 2. SDR improvement with different amicable noise levels. Higher DI_{SDR} indicates lower noise level.

A natural question is “*Is the amicable example universal for other separation models?*”. As shown in our experiment in Sec. 5.3, we found that amicable example is specific to the separation model used to compute it, which we call the *targeted model*, and it does not markedly improve the performance of untargeted models. This property is useful when one wants to minimize the side-effect on other models or to design a system where a targeted model exclusively benefits from the amicable example. However, it is more useful to have individual control in the separation of multiple separation models. To this end, we propose MMPL as

$$L(\nu) = \sum_i \alpha_i \|f^i(x + \nu), y\|_2^2 + \lambda \mathcal{C}_{STPR}(\nu), \quad (5)$$

where f^i denotes the i th separation model and α_i the weight to control the effect on each model. Note that α_i can be a negative value and in such a case, the term promotes the perturbation to be an adversarial example for model f^i . When all α_i are negative and $\sum_i \alpha_i = -1$, the loss becomes similar to the adversarial attack against an ensemble of models [30]. However, MMPL is more general and flexible as it can produce both amicable and adversarial examples at the same time depending on the model, i.e., suppose $\Gamma = \{i | \alpha_i > 0\}$ and $\Lambda = \{i | \alpha_i < 0\}$, the perturbation ν acts as the amicable example for models $f^{i \in \Gamma}$ but acts as adversarial example for models $f^{i \in \Lambda}$. To the best of our knowledge, this is the first attempt to learn a perturbation that serves as both amicable and adversarial examples simultaneously.

5. EXPERIMENTS

5.1. Setup

Experiments are conducted on the *test* set of the MUSDB18 dataset [31], which contains 50 songs recorded in stereo format at 44.1 kHz. Four sources (*vocals*, *drums*, *bass*, *other*) and their mixture are available for each song. To speed up the evaluation for our extensive experiments, we crop 10s clips from each song and use them for the evaluation.

The signal-to-distortion ratio (SDR) is used as the evaluation metric of separation performance. As in SiSEC 2018 [31], SDR is computed using the *museval* package and the median over all

Table 1. Subjective test on perceptibility of amicable examples.

DI_{SDR} [dB]	Accuracy
27.8	51.0%
30.1	49.0%

Table 2. SDRs of the separation of original mixtures and amicable example computed using UMX_{HQ} (in dB). The amicable example selectively improves the performance of the targeted model.

Model	input	vocals	drums	bass	other	Avg.
UMX_{HQ}	Original	6.25	6.24	5.07	3.40	5.24
Demucs		6.71	5.92	5.31	2.41	5.09
D3Net		7.08	6.79	5.08	3.56	5.63
UMX		6.65	5.91	4.86	3.39	5.20
UMX_{HQ}		8.44	8.03	6.76	5.61	7.21
Demucs	Amicable (UMX_{HQ})	6.66	5.96	5.49	2.51	5.16
D3Net		7.18	6.73	5.12	3.62	5.66
UMX		7.53	6.74	5.66	4.46	6.10

tracks of the median of each track is reported. To evaluate how much the mixture is distorted by the perturbation, we use the SDR between the mixture x and the perturbed mixture $x + \nu$,

$$DI_{SDR} = SDR(x, x + \nu), \quad (6)$$

which we call the degradation of input DI_{SDR} .

For the separation models, we use three open-source libraries, namely Open-Unmix (UMX) [32], D3Net [12], and Demucs [10], to ensure a variety of separation algorithms. UMX is based on bidirectional long-short term memory (BLSTM) layers and performs the separation in the frequency domain. D3Net is a convolutional neural network and also operates in the frequency domain. Demucs consists of both convolution and BLSTM layers and performs separation in the time domain. All models are trained on the MUSDB18 *train* set. In addition, UMX and Demucs have their variants: UMX_{HQ} is trained on the uncompressed MUSDB18 *train* set and $Demucs_{ex}$ is trained with 150 additional songs.

The initial perturbation is a uniform noise $[-\epsilon, \epsilon]$, $\epsilon = 0.01$, and is optimized using Adam for 300 iterations with the learning rate of 0.01.

5.2. Level of amicable noise and separation improvement

First, we investigate the relationship between the level of perturbation and separation performance improvement. We use UMX_{HQ} for the evaluation and set different λ values ([10, 20, 40, 100]) in (4) to control the perturbation level. Fig. 2 shows the SDR improvement SDR_i over the original mixture with different DI_{SDR} . As expected, the SDR improvement becomes more significant with increasing perturbation level. For 27.8 dB DI_{SDR} , an improvement of more than 2 dB is obtained on average. To evaluate the perceptibility of the amicable noise, we conduct a subjective test similarly to the double-blind triple-stimulus with hidden reference format (ITU-R BS.1116), where the reference is the

Table 3. SDRs for separation of perturbed samples computed using MMPL in two scenarios (α_i are both positive and α_i have opposite signs). Values in brackets indicate the SDR improvement over the separation of the original mixture.

Model	α_i	DI _{SDR}	vocals	drums	bass	other	Avg.
UMX _{HQ}	positive	29.21	7.90 (+1.65)	8.26 (d+2.02)	6.22 (+1.15)	5.08 (+1.68)	6.87 (+1.63)
Demucs _{ex}	positive		9.10 (+1.60)	9.89 (+2.01)	9.75 (+2.19)	5.87 (+2.56)	8.65 (+2.09)
UMX _{HQ}	positive	28.82	7.89 (+1.64)	8.26 (+2.02)	6.22 (+1.15)	5.08 (+1.68)	6.86 (+1.62)
Demucs _{ex}	negative		0.51 (-6.99)	0.5 (-7.38)	1.47 (-6.09)	-1.08 (-4.39)	0.35 (-6.21)

original mixture and either A or B is the same as the reference and the other is an amicable example. Evaluators are asked to identify which one of A or B is the same as the reference. We test two amicable noise levels and 40 audio engineers evaluated five songs of 10 s duration for each noise level. Table 1 shows that the accuracy of correctly identifying the reference is close to the chance rate (50%) at DI_{SDR} of 27.8 dB; thus, the amicable noise is imperceptible.

5.3. Effects on untargeted models

Next, we test the amicable example on untargeted models. The amicable example is computed using UMX_{HQ} and tested on different separation models. Table 2 shows the SDR values computed on the separations of the original mixture and amicable examples. By comparing the results of the original mixture and amicable example for each model, we observe that the amicable example significantly improves the SDRs of the targeted model UMX_{HQ} but only slightly improves the SDRs of Demucs and D3Net. This indicates that the loss surfaces (2) of these models are very different and thus the amicable noise is not generalized to different models. In contrast, the SDR improvement of UMX is more significant than that of Demucs and D3Net, probably because the architectures of UMX and UMX_{HQ} are identical and they were trained on very similar datasets (only their high-frequency components are different); thus, their loss surfaces are probably also similar.

5.4. Robustness against signal compression

As audio signals are often compressed to reduce the bandwidth or file size for transmission, it is important to assess the robustness of an amicable example against compression to verify its usability in realistic scenarios. To this end, we study how SDR improvements change by compressing the amicable example using an MP3 encoder with different compression levels. Fig. 3 shows that even after the amicable example is compressed with 256 kbps, the SDR improvement over the original mixture is nearly the same as that of the uncompressed example. Although more aggressive compression rates slightly degrade the effectiveness, the amicable example still improve the SDR by 1.57 dB on average at 128 kbps.

5.5. Amicable adversarial example with MMPL

Finally, we evaluate MMPL in two scenarios using UMX_{HQ} and Demucs_{ex}. In the first scenario, α_i in (5) is set to be positive for both UMX_{HQ} and Demucs_{ex}. In this case, the perturbation is computed to improve both models. In the second scenario, we use

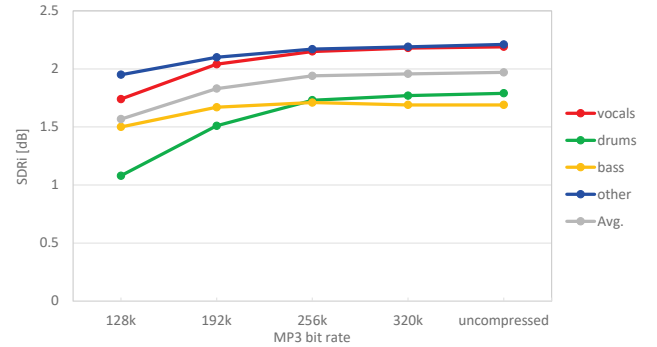


Fig. 3. SDR improvement with different MP3 compression bit rates.

a negative α_i for Demucs_{ex} instead. This makes the perturbation amicable for UMX_{HQ} but adversarial for Demucs_{ex}. To balance the magnitude of loss for both models, we set $(\alpha_{\text{UMX}}, \alpha_{\text{Demucs}})$ to be $[1, 100]$ for the former case and $[1, -100]$ for latter. The results are shown in Table 3. As observed, when we include both models to compute the amicable example, the performance of both models are improved, which is not the case when only one model is used, as shown in Sec. 5.3. More interestingly, in the second scenario, where we use opposite signs for the two models, we observe that the same perturbation significantly improves UMX_{HQ} but significantly degrades Demucs_{ex}. The results show that we can design a perturbation to be both an amicable example and an adversarial example depending on the model. We believe that this finding is important as it is closely related to security applications, e.g., even if a sample can be separated well with some models, it still can be an adversarial example for other models. We will further investigate this in the future.

6. CONCLUSION

We propose amicable example-based informed source separation, where an imperceptible perturbation from the mixture is computed to improve the separation. Experimental results show that amicable examples selectively improve the performance of the targeted model and are robust against the audio compression. We further propose multi-model multi-purpose learning (MMPL) to individually control the effect of the perturbation for multiple models. MMPL is shown to be capable of computing a perturbation that works as both an amicable example and an adversarial example depending on the model.

7. REFERENCES

- [1] A. Ozerov and C. Févotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.
- [2] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, “Kernel additive models for source separation,” *IEEE Trans. Signal Processing*, vol. 62, no. 16, pp. 4298–4310, 2014.
- [3] J. LeRoux, J. R. Hershey, and F. Weninger, “Deep NMF for speech separation,” in *Proc. ICASSP*, 2015, p. 66–70.
- [4] D. Fitzgerald, A. Liutkus, and R. Badeau, “PROJET - spatial audio separation using projections,” in *Proc. ICASSP*, 2016, pp. 36–40.
- [5] N. Takahashi and Y. Mitsufuji, “Multi-scale multi-band DenseNets for audio source separation,” in *Proc. WASPAA*, 2017, pp. 261–265.
- [6] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, “Singing voice separation with deep u-net convolutional networks,” in *Proc. ISMIR*, 2017, pp. 745–751.
- [7] N. Takahashi, N. Goswami, and Y. Mitsufuji, “MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation,” in *Proc. IWAENC*, 2018.
- [8] J. H. Lee, H.-S. Choi, and K. Lee, “Audio query-based music source separation,” in *Proc. ISMIR*, 2019.
- [9] J.-Y. Liu and Y.-H. Yang, “Dilated convolution with dilated GRU for music source separation,” in *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.
- [10] A. Défossez, N. Usunier, L. Bottou, and F. Bach, “Music source separation in the waveform domain,” *arXiv preprint arXiv:1911.13254*, 2019.
- [11] N. Takahashi, P. Sudarsanam, N. Goswami, and Y. Mitsufuji, “Recursive speech separation for unknown number of speakers,” in *Proc. Interspeech*, 2019.
- [12] N. Takahashi and Y. Mitsufuji, “Densely connected multidilated convolutional networks for dense prediction tasks,” in *Proc. CVPR*, 2021.
- [13] M. Parvaix, L. Girin, and J.-M. Brossier, “A watermarking-based method for informed source separation of audio signals with a single sensor,” *Trans. Audio, Speech, and Language Processing*, 2010.
- [14] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, “Informed source separation through spectrogram coding and data embedding,” in *Signal Processing, Elsevier*, 2012.
- [15] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, “Informed source separation: source coding meets source separation,” in *Proc. WASPAA*, 2011.
- [16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in *Proc. ICLR*, 2014.
- [17] M. Bläser, C. Rohlfing, Y. Gao, and M. Wien, “Adaptive coding of non-negative factorization parameters with application to informed source separation,” in *Proc. ICASSP*, 2018, pp. 751–755.
- [18] M. Miron, J. Janer, and E. Gómez, “Monaural score-informed source separation for classical music using convolutional neural networks,” in *Proc. ISMIR*, 2017.
- [19] E. Manilow and B. Pardo, “Bespoke neural networks for score-informed source separation,” in *Proc. ISMIR*, 2020.
- [20] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, “Text-informed speech enhancement with deep neural networks,” in *Proc. Interspeech*, 2015.
- [21] N. Takahashi, M. K. Singh, S. Basak, P. Sudarsanam, S. Ganapathy, and Y. Mitsufuji, “Improving voice separation by incorporating end-to-end speech recognition,” in *Proc. ICASSP*, 2020.
- [22] J. Su, D. V. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” *IEEE Trans. Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.
- [23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *Proc. ICLR*, 2018.
- [24] Y. Bai, Y. Feng, Y. Wang, T. Dai, S.-T. Xia, and Y. Jiang, “Hilbert-based generative defense for adversarial examples,” in *Proc. ICCV*, 2019.
- [25] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, “Boosting adversarial attacks with momentum,” in *Proc. CVPR*, 2018.
- [26] D. Wu, Y. Wang, S.-T. Xia, J. Bailey, and X. Ma, “Skip connections matter: On the transferability of adversarial examples generated with ResNets,” in *Proc. ICLR*, 2020.
- [27] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *Proc. ICLR*, 2015.
- [28] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, “Adversarial examples are not bugs, they are features,” in *Proc. NeurIPS*, 2019.
- [29] N. Takahashi, S. Inoue, and Y. Mitsufuji, “Adversarial attacks on audio source separation,” in *Proc. ICASSP*, 2021.
- [30] Y. Liu, X. Chen, C. Liu, and D. Song, “Delving into transferable adversarial examples and black-box attacks,” in *Proc. ICLR*, 2017.
- [31] A. Liutkus, F.-R. Stöter, and N. Ito, “The 2018 signal separation evaluation campaign,” in *Proc. LVA/ICA*, 2018.
- [32] 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.