ADVERSARIAL PERMUTATION INVARIANT TRAINING
FOR UNIVERSAL SOUND SEPARATION

Emilian Postolache*1,2, Jordi Pons*1, Santiago Pascual1, Joan Serra1
1Dolby Laboratories 2Sapienza University of Rome

ABSTRACT

Universal sound separation consists of separating mixes with arbitrary sounds of different types, and permutation invariant training (PIT) is used to train source agnostic models that do so. In this work, we complement PIT with adversarial losses but find it challenging with the standard formulation used in speech source separation. We overcome this challenge with a novel $I$-replacement context-based adversarial loss, and by training with multiple discriminators. Our experiments show that simply improving the loss (keeping the same model and dataset) we obtain a non-negligible improvement of 1.4 dB SI-SNR in the reverberant FUSSET dataset. We also find adversarial PIT to be effective at reducing spectral holes, ubiquitous in mask-based separation models, which highlights the potential relevance of adversarial losses for source separation.

Index Terms— Adversarial, PIT, universal source separation.

1. INTRODUCTION

Audio source separation consists of separating the sources present in an audio mix, as in music source separation (separating vocals, bass, and drums from a music mix [1–3]) or speech source separation (separating various speakers talking simultaneously [4–6]). Recently, universal sound separation was proposed [7]. It consists of building source agnostic models that are not constrained to a specific domain (like music or speech) and can separate any source given an arbitrary mix. Permutation invariant training (PIT) [8] is used for training universal source separation models based on deep learning [7, 9, 10].

We consider mixes $m$ of length $L$ with $K'$ arbitrary sources $s$ as follows: $m = \sum_{k=1}^{K'} s_k$, out of which the separator model $f_\theta$ predicts $K$ sources $\hat{s} = f_\theta(m)$. PIT optimizes the learnable parameters $\theta$ of $f_\theta$ by minimizing the following permutation invariant loss:

$$L_{\text{PIT}} = \min_{P} \sum_{k=1}^{K} \mathcal{L}(s_k, [P\hat{s}]_k),$$

where we consider all permutation matrices $P$, $P^*$ is the optimal permutation matrix minimizing Eq. 1, and $\mathcal{L}$ can be any regression loss. Since $f_\theta$ outputs $K$ sources, in case a mix contains $K' < K$ sources, we set the target $s_k = 0$ for $k > K'$. Note that a permutation invariant loss is required to build source agnostic models, because the outputs of $f_\theta$ can be any source and in any order. As such, the model must not focus on predicting one source type per output, and any possible permutation of output sources must be equally correct [7, 8]. A common loss $\mathcal{L}$ for universal sound separation is the $\tau$-thresholded logarithmic mean squared error [7, 9], which is unbounded when $s_k = 0$. In that case, since $m \neq 0$, one can use a different $\mathcal{L}$ based on thresholding with respect to the mixture [9]:

$$\mathcal{L}(s_k, \hat{s}_k) = \begin{cases} 10 \log_{10} \left( \frac{||s_k||^2 + \tau||m||^2}{||m||^2} \right) & \text{if } s_k = 0 \\ 10 \log_{10} \left( \frac{||s_k - \hat{s}_k||^2 + \tau||\hat{s}_k||^2}{||\hat{s}_k||^2} \right) & \text{otherwise}. \end{cases}$$

In this work, we complement PIT with adversarial losses for universal sound separation. A number of speech source separation works also complemented PIT with adversarial losses [11–14]. Yet, we find that the adversarial PIT formulation used in speech separation does not perform well for universal source separation (sections 3 and 4). To improve upon that, in section 2 we extend speech separation works with: a novel $I$-replacement context-based adversarial loss, by combining multiple discriminators, and generalize adversarial PIT such that it works for universal sound separation (with source agnostic discriminators dealing with more than two sources). Table 1 outlines how our approach compares with speech separation works.

2. ADVERSARIAL PIT

Adversarial training, in the context of source separation, consists of simultaneously training two models: $f_\theta$ producing plausible separations $\hat{s}$, and one (or multiple) discriminator(s) $D$ assessing if separations $\hat{s}$ are produced by $f_\theta$ (fake) or are ground-truth separations $s$ (real). Under this setup, the goal of $f_\theta$ is to estimate (fake) separation that is as close as possible to the (real) ones from the dataset, such that $D$ misclassifies $\hat{s}$ as $s$ [15, 16]. We propose combining variations of an instance-based discriminator $D_{\text{inst}}$ with a novel $I$-replacement context-based discriminator $D_{\text{ctx}, I}$. Each $D$ has a different role and is applicable to various domains: waveforms, magnitude STFTs, or masks. Without loss of generality, we present $D_{\text{inst}}$ and $D_{\text{ctx}, I}$ in the waveform domain and then show how to combine multiple discriminators operating at various domains to train $f_\theta$.

Instance-based adversarial loss — The role of $D_{\text{inst}}$ is to provide adversarial cues on the realness of the separated sources without context. This is, $D_{\text{inst}}$ assesses the realness of each source individually:

$$[s_1] / [\hat{s}_1] \ldots [s_K] / [\hat{s}_K].$$

Throughout the paper, we use brackets $[\cdot]$ to denote the $D$‘s input and left/right for real/fake separations (not division). Hence, individual real/fake separations (instances) are input to $D_{\text{inst}}$, which learns to classify them as real/fake (Fig. 1). $D_{\text{inst}}$ is trained to maximize

$$L_{\text{inst}} = \frac{1}{K} \sum_{k=1}^{K} (L_{\text{real}, k} + L_{\text{fake}, k}),$$

where $L_{\text{real}, k}$ and $L_{\text{fake}, k}$ correspond to the hinge loss [17]:

$$L_{\text{real}, k} = \min (0, -1 - D_{\text{inst}}(s_k)),$$

$$L_{\text{fake}, k} = \min (0, -1 - D_{\text{inst}}(\hat{s}_k)).$$

Previous works also explored using $D_{\text{inst}}$. However, they used source specific setups where each $D_{\text{inst}}$ was specialized in a source type, e.g., for music source separation each $D_{\text{inst}}$ was specialized in bass, drums, and vocals [1, 18], or for speech source separation $D_{\text{inst}}$ was specialized in speech [12, 19]. Yet, each $D_{\text{inst}}$ for universal sound separation is not specialized in any source type (are source agnostic) and assesses the realness of any audio, regardless of its source type.

*Equal contribution.
I-replacement context-based adversarial loss — The role of \( D_{\text{ctx},I} \) is to provide adversarial cues on the realness of the separated sources considering all the sources in the mix (the context):

\[
[s_1, \ldots, s_K] / [\hat{s}_1, \ldots, \hat{s}_K].
\]

Here, all the separations are input jointly to provide context to \( D_{\text{ctx},I} \), which learns to classify them as real / fake. \( D_{\text{ctx},I} \) can also be conditioned on the input mix \( m \):

\[
[m, s_1, \ldots, s_K] / [m, \hat{s}_1, \ldots, \hat{s}_K].
\]

Fake inputs contain entries \( \hat{s}_k \), obtained by randomly getting \( I < K \) source indices without replacement \( \Lambda_I \subset \{1, \ldots, K\} \), and substituting the selected estimated sources \( \hat{s}_k \) with their ground-truth \( s_k \):

\[
\hat{s}_k = \begin{cases} 
    s_k & \text{if } k \in \Lambda_I, \\
    \hat{s}_k & \text{otherwise},
\end{cases}
\]

where \( \hat{s}_k = [P^*\hat{s}]_k \) and \( P^* \) is the optimal permutation matrix minimizing Eq. 1 with \( \mathcal{L} \) as in Eq. 2. Note that finding the right permutation \( P^* \) is important to replace the selected source \( \hat{s}_k \) with its corresponding ground-truth \( s_k \), because the source agnostic estimations do not necessarily match the order of the ground-truth (see Fig. 2).

Also, note that the \( I = 0 \) case corresponds to the standard context-based adversarial loss used for speech source separation (without I-replacement, see Table 1). Thus, our work generalizes adversarial PIT for universal sound separation by proposing a novel I-replacing schema that explicitly uses the ground-truth to guide the adversarial loss. \( D_{\text{ctx},I} \) is trained to maximize

\[
\mathcal{L}_{\text{ctx},I} = \mathcal{L}_{\text{real}}^{\text{adv}} + \mathcal{L}_{\text{fake}}^{\text{adv}},
\]

where we again use the hinge loss [17]:

\[
\mathcal{L}_{\text{real}}^{\text{adv}} = \min(0, -1 + D_{\text{ctx},I}(s_1, \ldots, s_K)), \\
\mathcal{L}_{\text{fake}}^{\text{adv}} = \min(0, -1 - D_{\text{ctx},I}(\hat{s}_1, \ldots, \hat{s}_K)).
\]
a richer set of adversarial loss cues to train \( f_0 \) [18, 20, 21], such that both \( D_{\text{inst}} \) and \( D_{\text{ctx},i} \) can provide different perspectives with respect to the same signal in various domains: waveform, magnitude STFT, and mask. Hence, in addition to train each \( D \) alone, one can train multiple combinations, e.g., \( D_{\text{inst}}^\text{wave} + D_{\text{ctx},i}^\text{wave} + D_{\text{ctx},i}^\text{mask} + D_{\text{ctx},i}^\text{sep} \), or any combination of the discriminators above. However, the more discriminators used, the more computationally expensive it is to run the loss, and the longer it takes to train \( f_0 \) (but does not affect inference time). To the best of our knowledge, training with multiple discriminators has never been considered for source separation before.

**Separator loss** — In adversarial training, \( f_0 \) is trained such that its (fake) separations \( \hat{s} \) are misclassified by the discriminator(s) as ground-truth ones \( s \) (real). To do so, during every adversarial training step, we first update the discriminator(s) (without updating \( f_0 \)) based on \( \mathcal{L}_{\text{inst}}, \mathcal{L}_{\text{ctx},i} \), or any combination of the losses above. Then, \( \mathcal{L}_{\text{sep}} \) is minimized to train \( f_0 \) without updating the discriminator(s). For example, when using \( D_{\text{inst}}^\text{wave} \) (with \( D_{\text{inst}}^\text{wave} \) frozen) we minimize

\[
\mathcal{L}_{\text{sep}} = - \frac{1}{K} \sum_{k=1}^{K} D_{\text{inst}}^\text{wave}(\hat{s}_k),
\]

when using \( D_{\text{ctx},i}^\text{wave} \) (with \( D_{\text{ctx},i}^\text{wave} \) frozen) we minimize

\[
\mathcal{L}_{\text{sep}} = - D_{\text{ctx},i}^\text{wave}(\{\hat{s}_1, \ldots, \hat{s}_K\}),
\]

or when using \( D_{\text{sep}}^\text{wave} \) and \( D_{\text{ctx},i}^\text{wave} \) conditioned on \( m \) (with \( D_{\text{sep}}^\text{wave} \) and \( D_{\text{ctx},i}^\text{wave} \) frozen) we minimize

\[
\mathcal{L}_{\text{sep}} = - \frac{1}{K} \sum_{k=1}^{K} D_{\text{sep}}^\text{wave}(\hat{s}_k|m),
\]

\[
\mathcal{L}_{\text{sep}} = - D_{\text{ctx},i}^\text{wave}(\{\hat{s}_1, \ldots, \hat{s}_K\}|m).
\]

Again, note that we use the hinge loss [17]. While we are not presenting all possible loss combinations for brevity, from the above examples one can infer all the combinations we experiment with in section 4. Finally, we can also add an \( \mathcal{L}_{\text{PIT}} \) term (as in Eqs. 1 and 2) to adversarial PIT: \( \mathcal{L}_{\text{sep}} + \lambda \mathcal{L}_{\text{PIT}} \), where \( \lambda \) scales \( \mathcal{L}_{\text{PIT}} \) such that it is of the same magnitude as \( \mathcal{L}_{\text{sep}} \) [22]. All previous adversarial PIT works for speech source separation used \( \mathcal{L}_{\text{sep}} + \lambda \mathcal{L}_{\text{PIT}} \) (Table 1). Yet, in section 4 we show that our adversarial training setup allows dropping \( \mathcal{L}_{\text{PIT}} \) while still obtaining competitive results, possibly because of the strong cues provided by \( D_{\text{ctx},i} \) (with \( i \)-replacement) and the multiple discriminators. To the best of our knowledge, we are the first to report results similar to \( \mathcal{L}_{\text{PIT}} \) with a purely adversarial setup (cf. [12]).

3. EXPERIMENTAL SETUP

**Dataset, evaluation metrics, and baseline** — We use the reverberant Fuzz dataset, a common benchmark for universal sound separation with 20 k / 1 k / 1 k (train / val / test) mixes of 10 s with one to four sources [9, 23, 24]. Metrics rely on the scale-invariant SNR [9]:

\[
\text{SI-SNR}_1 = \text{SI-SNR}(s_k, \hat{s}_k) - \text{SI-SNR}(s_k, m),
\]

where \( \alpha = s_k^T s_k^* + \epsilon \|s_k^*\|^2 + \epsilon, \) and \( s_k^* = [P^* s_k], \) with \( P^* \) being the optimal source-permutation matrix maximizing SI-SNR. Further, to account for inactive sources, estimate-target pairs that have silent target sources are discarded [10]. For mixes with one source, we compute \( \text{SI-SNR}_S = \text{SI-SNR}(s_k, \hat{s}_k) \), which is equivalent to \( \text{SI-SNR}(m, \hat{m}) \) since with one-source mixes the goal is to bypass the mix (the S sub-index stands for single-source\(^1\)). For mixes with two to four sources, we report the average across sources of the SI-SNRs.

1\(^{SI-SNR}_S \) is named as IS or SS in [9, 10] and SI-SNRs as MSI in [9, 10].
4. EXPERIMENTS AND DISCUSSION

First, in Table 2, we study various $D_{\text{ctx},I}$ configurations. We observe that the standard adversarial PIT ($I=0$, as in speech source separation) consistently obtains the worst results for universal sound separation. In contrast, the models trained with $I$-replacement ($I>0$) consistently obtain the best results. We hypothesize that with $I=0$, $f_0$ does not separate much. Instead, it tends to approximate the naive solution of bypassing the mix. We can see this with the SI-SNR$_R$ scores, which tend to be closer (if not the same) to the SI-SNR$_S$ of the lower and upper bounds in Table 3. Overall, we note that the $I$-replacement context-based adversarial loss seems key to generalize adversarial PIT for universal sound separation, where multiple heterogeneous sources are separated. This contrasts with adversarial PIT works for speech source separation, where two similar sources are separated. We argue that the universal sound separation case is more challenging, as speech separation discriminators can judge the realness of separations based on speech cues, but discriminators for universal sound separation cannot as sources can be of any kind. We hypothesize that the effectiveness of $D_{\text{ctx},I>0}$ can be attributed to:

- Replacing $\hat{s}_k$ with $s_k$ explicitly guides the adversarial loss to perform source separation. Note that $D_{\text{ctx},I=0}$ (and $D_{\text{inst}}$) focuses on assessing the realness of its input. Under this setup, a naive solution is to always bypass the mix, which looks as real as one-source mixes where the goal is to bypass the mix. To avoid this naive solution, some guidance like the $I$-replacement is required.

- It is more difficult for $D_{\text{ctx},I>0}$ to distinguish between real and fake separations, because fake ones contain replacements. Consequently, such replacements help defining a non-trivial task for the discriminator that results in a better adversarial loss to train $f_0$.

Next, we study the discriminators introduced in section 2 and their combination. In Table 2, we note that the $m$-conditioned $D_{\text{ctx},I=3}$ generally outperform the rest. Hence, and for simplicity, in Table 3 we only experiment with this setup. We note the following trends:

- Our best result using adversarial PIT improves the state-of-the-art by 1 dB (from 12.8 to 13.8 dB) and improves the $L_{\text{PIT}}$ baseline by 1.4 dB (from 12.4 to 13.8 dB). Informal listening also reveals that our best model separations more closely match the ground-truth sources and contain less spectral holes than the DCASE baseline (audio examples are available online\(^2\)). Spectral holes are ubiquitous across mask-based source separation models, and are the unpleasant result of over-suppressing a source in a band where other sources are present. Adversarial training seems appropriate to tackle this issue since it improves the realness of the separations by avoiding spectral holes (which are not present in the training dataset). We also compare our best model score (13.8 dB) against the lower and upper bounds (25.3 and 0 dB) to see that there is still room for improvement (also note this in our examples online\(^2\)).

- Our best results are obtained when combining multiple discriminators with $L_{\text{PIT}}$ (over 13 dB). This shows that complementing the adversarial loss with $L_{\text{PIT}}$ is beneficial, and confirms that using multiple discriminators in various domains can help to improve the separations’ quality. We also note that $L_{\text{PIT}}$ alone and the best adversarially trained models (without $L_{\text{PIT}}$) obtain similar results (around 12.5 dB). Hence, purely adversarial models can obtain comparable results to $L_{\text{PIT}}$ alone even without explicitly optimizing for $L_{\text{PIT}}$, in Eq. 2, which is closely related to SI-SNR.

- When studying models trained with one $D$, we note that $D_{\text{ctx},I}$ alone tends to obtain better results than $D_{\text{inst}}$ alone. We argue that

\(\text{Table 3: Comparison of adversarial PIT variants and baselines. SI-SNR column: SI-SNR}_{R}/\text{SI-SNR}_{S} \text{db. All } D_{\text{ctx},I} \text{ above are } m\text{-conditioned with } I=3, \text{ since it outperforms other setups (Table 2). } \)

\(\text{All adversarial PIT ablations (rows 1-11 & Table 2) use the same } f_0.\)

\(\text{the replacements in } D_{\text{ctx},I} \text{ explicitly guide the separator to perform source separation, while for } D_{\text{inst}} \text{ this is not the case. In addition, } D_{\text{inst}}^\text{m} \text{ alone obtains a competitive score of 12.1 dB, which can be improved up to 12.9 dB if combined with 5 additional discriminators. Hence, results can be improved by using multiple discriminators but one can save computational resources by choosing the right } D \text{ without dramatically compromising the results. Finally, even though } D_{\text{inst}} \text{ alone under-performs the rest, we note that it can help improving the results when combined with } D_{\text{ctx},I}.\)

5. CONCLUSIONS

We adapted adversarial PIT for universal sound separation with a novel $I$-replacement context-based adversarial loss and by training with multiple discriminators. With that, we improve the separations by 1.4 dB SI-SNR$_R$ and reduce the unpleasant presence of spectral holes just by changing the loss, without changing the model or dataset. Even with the improved results, we also stress that the obtained separations can still be improved by an important margin.

6. ACKNOWLEDGMENTS

Work done during Emilian’s internship at Dolby. Emilian’s thesis is supported by the ERC grant no. 802554 (SPECGEO). Thanks to S. Wisdom and J. Hershey for your help with the metrics and baselines.

\(2\)http://jordipons.me/apps/adversarialPIT/
7. REFERENCES

[1] Daniel Stoller, Sebastian Ewert, and Simon Dixon, “Adversarial semi-supervised audio source separation applied to singing voice extraction,” in ICASSP, 2018.

[2] Francesc Luís, Jordi Pons, and Xavier Serra, “End-to-end music source separation: is it possible in the waveform domain?,” in Interspeech, 2019.

[3] Jordi Pons, Santiago Pascual, Giulio Cengarle, and Joan Serra, “Upsampling artifacts in neural audio synthesis,” in ICASSP, 2020.

[4] Berkant Kadoğlu, Michael Horgan, Xiaoyu Liu, Jordi Pons, Dan Darcy, and Vivek Kumar, “An empirical study of Conv-TasNet,” in ICASSP, 2020.

[5] Xiaoyu Liu and Jordi Pons, “On permutation invariant training for speech source separation,” in ICASSP, 2021.

[6] Vivek Narayanaswamy, Jayaraman J Thiagarajan, Rushil Anirudh, and Andreas Spanias, “Unsupervised audio source separation using generative priors,” arXiv, 2020.

[7] Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R Hershey, “Universal sound separation,” in WASPAA, 2019.

[8] Dong Yu, Morten Kolhebek, Zheng-Hua Tan, and Jesper Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in ICASSP, 2017.

[9] Scott Wisdom, Hakan Erdogan, Daniel PW Ellis, Romain Serizel, Nicolas Turpault, Eduardo Fonseca, Justin Salamon, Prem Seetharaman, and John R Hershey, “What’s all the fuss about free universal sound separation data?,” in ICASSP, 2021.

[10] Scott Wisdom, Ilyas Izzel, Hakan Erdogan, Ron Weiss, Kevin Wilson, and John Hershey, “Unsupervised sound separation using mixture invariant training,” NeurIPS, 2020.

[11] Chenxing Li, Lei Zhu, Shuang Xu, Peng Gao, and Bo Xu, “CBLDNN-based speaker-independent speech separation via generative adversarial training,” in ICASSP, 2018.

[12] Lianwu Chen, Meng Yu, Yamin Qian, Dan Su, and Dong Yu, “Permutation invariant training of generative adversarial network for monaural speech separation,” in Interspeech, 2018.

[13] Ziqiang Shi, Huibin Lin, Liu Liu, Ruige Liu, Shoji Hayakawa, and Jiqing Han, “Furcax: End-to-end monaural speech separation based on deep gated (de) convolutional neural networks with adversarial example training,” in ICASSP, 2019.

[14] Chengyun Deng, Yi Zhang, Shiqian Ma, Yongtao Sha, Hui Song, and Xiangang Li, “Conv-TasSAN: Separative adversarial network based on conv-tasnet,” in Interspeech, 2020, pp. 2647–2651.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, “Generative adversarial networks,” Communications of the ACM, vol. 63, no. 11, pp. 139–144, 2020.

[16] Santiago Pascual, Antonio Bonafonte, and Joan Serra, “SEGAN: Speech enhancement generative adversarial network,” arXiv, 2017.

[17] Jae Hyun Lim and Jong Chul Ye, “Geometric GAN,” arXiv, 2017.

[18] Enric Gusó, Jordi Pons, Santiago Pascual, and Joan Serrà, “On loss functions and evaluation metrics for music source separation,” in ICASSP, 2022.

[19] Y Cem Subakan and París Smaragdis, “Generative adversarial source separation,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 26–30.

[20] Mikolaj Brńkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan, “High fidelity speech synthesis with adversarial networks,” arXiv, 2019.

[21] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gstein, Wei Zhen Teoh, Jose Soteló, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville, “MelGAN: Generative adversarial networks for conditional waveform synthesis,” NeurIPS, 2019.

[22] Patrick Esser, Robin Rombach, and Bjorn Ommer, “Taming transformers for high-resolution image synthesis,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12873–12883.

[23] Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitriy Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra, “Freesound datasets: a platform for the creation of open audio datasets,” in ISMIR, 2017.

[24] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, “FSD50k: an open dataset of human-labeled sound events,” IEEE/ACM TASLP, vol. 30, pp. 829–852, 2021.

[25] Scott Wisdom, Aren Jansen, Ron J Weiss, Hakan Erdogan, and John R Hershey, “Sparse, efficient, and semantic mixture invariant training: Taming in-the-wild unsupervised sound separation,” in WASPAA, 2021.

[26] “DCASE2020 challenge: Sound event detection and separation in domestic environments,” https://dcase.community/challenge2020/task-sound-event-detection-and-separation-in-domestic-environments.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[28] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde, “Singing voice separation with deep U-net convolutional networks,” ISMIR, 2017.

[29] Scott Wisdom, John R Hershey, Kevin Wilson, Jeremy Thorpe, Michael Chinen, Brian Patton, and Rif A Saurous, “Differentiable consistency constraints for improved deep speech enhancement,” in ICASSP, 2019.

[30] Yi Luo and Nima Mesgarani, “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” IEEE/ACM transactions on audio, speech, and language processing, vol. 27, no. 8, pp. 1256–1266, 2019.