INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Unsupervised Audio Source Separation using Generative Priors

Vivek Narayanaswamy[1], *Jayaraman J. Thiagarajan*[2], *Rushil Anirudh*[2] *and Andreas Spanias*[1]

[1]SenSIP Center, School of ECEE, Arizona State University, Tempe, AZ
[2]Lawrence Livermore National Labs, 7000 East Avenue, Livermore, CA

vnaray29@asu.edu, jjayaram@llnl.gov, anirudh1@llnl.gov, spanias@asu.edu

Abstract

State-of-the-art under-determined audio source separation systems rely on supervised end to end training of carefully tailored neural network architectures operating either in the time or the spectral domain. However, these methods are severely challenged in terms of requiring access to expensive source level labeled data and being specific to a given set of sources and the mixing process, which demands complete re-training when those assumptions change. This strongly emphasizes the need for unsupervised methods that can leverage the recent advances in data-driven modeling, and compensate for the lack of labeled data through meaningful priors. To this end, we propose a novel approach for audio source separation based on generative priors trained on individual sources. Through the use of projected gradient descent optimization, our approach simultaneously searches in the source-specific latent spaces to effectively recover the constituent sources. Though the generative priors can be defined in the time domain directly, e.g. WaveGAN, we find that using spectral domain loss functions for our optimization leads to good-quality source estimates. Our empirical studies on standard spoken digit and instrument datasets clearly demonstrate the effectiveness of our approach over classical as well as state-of-the-art unsupervised baselines.

Index Terms: audio source separation, unsupervised learning, generative priors, projected gradient descent

1. Introduction

Audio source separation, the process of recovering constituent source signals from a given audio mixture, is a key component in downstream applications such as audio enhancement and music information retrieval [1, 2]. Typically formulated as an inverse optimization problem, source separation has been traditionally solved using a broad class of matrix factorization methods [3, 4, 5], e.g., Independent Component Analysis (ICA) and Principal Component Analysis (PCA). While these methods are known to be effective in over-determined scenarios, i.e. the number of mixture observations is greater than the number of sources, they are severely challenged in under-determined settings [6]. Consequently, in the recent years, supervised deep learning based solutions have become popular for under-determined source separation [7, 8, 9, 10, 11, 12]. These approaches can be broadly classified into time domain and spectral domain methods, and often produce state-of-the-art performance on standard benchmarks. Despite their effectiveness, there is a fundamental drawback with supervised methods. In addition to requiring access to large number of observations, a supervised source separation model is highly specific to the given set of sources and the mixing process, consequently requiring complete re-training when those assumptions change. This motivates a strong need for the next generation of unsupervised separation methods that can leverage the recent advances in data-driven modeling, and compensate for the lack of labeled data through meaningful priors.

Utilizing appropriate priors for the unknown sources has been an effective approach to regularize the ill-conditioned nature of source separation. Examples include non-Gaussianity, statistical independence, and sparsity [13]. With the emergence of deep learning methods, it has been shown that choice of the network architecture implicitly induces a structural prior for solving inverse problems [14]. Based on this finding, Tian *et al.* recently introduced a *deep audio prior* (DAP) [15] that directly utilizes the structure of a randomly initialized neural network to learn time-frequency masks that isolate the individual components in the mixture audio without any pre-training. Interestingly, DAP was shown to outperform several classical priors.
Here, we consider an alternative approach for under-determined source separation based on *data priors* defined via deep generative models, and in particular using generative adversarial networks (GANs) [16]. We hypothesize that such a data prior will produce higher quality source estimates by enforcing the estimated solutions to belong to the data manifold. While GAN priors have been successfully utilized in inverse imaging problems [17, 18, 19, 20] such as denoising, deblurring, compressed recovery etc., their use in source separation has not been studied yet – particularly in the context of audio. In this paper, we propose a novel unsupervised approach for source separation that utilizes multiple source-specific priors and employs *Projected Gradient Descent* (PGD)-style optimization with carefully designed spectral-domain loss functions. Since our approach is an inference-time technique, it is extremely flexible and general such that it can be used even with a single mixture. We utilize the time-domain based WaveGAN [21] model to construct the source-specific priors, and interestingly, we find that using spectral losses for the inversion leads to superior quality results. Using standard benchmark datasets (spoken digit audio (SC09), drums and piano), we evaluate the proposed approach under the assumption that mixing process is known. From our rigorous empirical study, we find that the proposed *data prior* is consistently superior to other commonly adopted priors, including the recent deep audio prior [15]. The codes for our work are publicly available.[i]

2. Designing Priors for Inverse Problems

Despite the advances in learning methods for audio processing, under-determined source separation remains a critical challenge. Formally, in our setup, the number of mixtures or ob-

This work was supported in part by the ASU SenSIP Center, Arizona State University. Portions of this work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

[i] https://github.com/vivsivaraman/sourcesepganprior

http://dx.doi.org/10.21437/Interspeech.2020-3115

servations $m \ll n$, *i.e.* the number of sources. A common approach to make this ill-defined problem tractable is to place appropriate priors to restrict the solution space. Existing approaches can be broadly classified into the following categories:

(i) Statistical Priors. This includes the class of matrix factorization methods conventionally used in source separation. For example in ICA, we enforce the assumptions of non-Gaussianity as well as statistical independence between the sources. On the other hand, PCA enforces statistical independence between the sources by linear projection onto mutually orthogonal subspaces. KernelPCA [22] induces the same prior in a reproducing kernel Hilbert space. Another popular approach is Non-negative matrix factorization (NMF), which places a non-negativity prior on the estimated basis matrices [23]. Finally, a sparsity prior (ℓ_1) [13] placed either in the observed domain or in the expansion via an appropriate basis set or a dictionary has also been widely adopted to regularize this problem.

(ii) Structural Priors. Recent advances in deep neural network design have shown that certain carefully chosen networks have the innate capability to effectively regularize or behave as a prior to solve ill-posed inverse problems. These networks essentially capture the underlying statistics of data, independent of the task-specific training. These *structural priors* have produced state-of-the-art performance in inverse imaging problems [14] and recently, Tian *et al.* [15] utilized the structure of an U-Net [24] model to learn time-frequency masks that can isolate the individual components in the mixture audio.

(iii) GAN Priors. A third class of methods have relied on priors defined via generative models, e.g. GANs [16]. GANs can learn parameterized non-linear distributions $p(X; \mathbf{z})$ from a sufficient amount of unlabeled data X [21, 25], where $\mathbf{z}$ denotes the latent variables of the model. In addition to readily sampling from trained GAN models, they can be leveraged as an effective prior for X. Popularly referred to as *GAN priors*, they have been found to be highly effective in challenging inverse problems [19, 20]. In its most general form, when one attempts to recover the original data $\mathbf{x}$ from its corrupted version $\tilde{\mathbf{x}}$ (observed), one can maximize the posterior distribution $p(X = \mathbf{x}|\tilde{\mathbf{x}}; \mathbf{z})$ by searching in the latent space of a pre-trained GAN. Since this posterior distribution cannot be expressed analytically, in practice, we utilize an iterative approach such as *Projected Gradient Descent* (PGD) to estimate the latent features $\hat{\mathbf{z}}$ followed by sampling from the generator, *i.e* $p(X; \mathbf{z} = \hat{\mathbf{z}})$.

Proposed Work. In this work, we propose to utilize GAN priors to solve the problem of under-determined source separation. Existing solutions with data priors utilize a single GAN model to perform the inversion process [20]. However, by design, source separation requires the simultaneous estimation of multiple disparate source signals. While one can potentially build a generative model that can jointly characterize all sources, it will require significantly large amounts of data. Hence, we advocate the use of source-specific generative models and generalizing the PGD optimization with multiple GAN priors. In addition to reducing the data needs, this approach provides the crucial flexibility of handling new sources, without the need for retraining the generative models for all sources. From our study, we find that utilizing multiple GAN priors $\{\mathcal{G}_i | i = 1 \dots K\}$ to be highly effective for under-determined source separation. In particular, we choose a popular waveform synthesis model WaveGAN [21] as our GAN prior $\mathcal{G}_i$ as we found the generated samples to be of high perceptual quality. While we utilize time domain GAN prior models, we find that spectral domain loss functions are critical in source estimation using PGD.

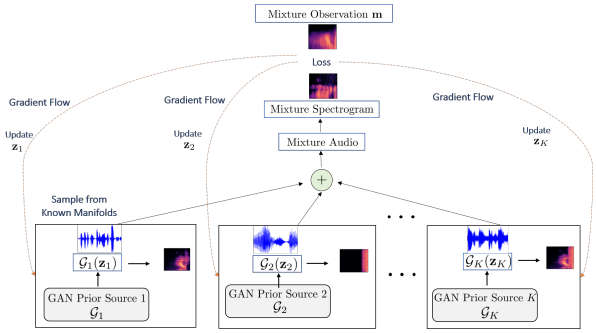

Figure 1: *An overview of the proposed unsupervised source separation system.*

3. Approach

Audio source separation involves the process of recovering constituent sources $\{\mathbf{s}_i \in \mathbb{R}^d | i = 1 \cdots K\}$ from a given audio mixture $\mathbf{m} \in \mathbb{R}^d$, where K is the total number of sources and d is the number of time steps. In this paper, without loss of generality, we assume the source and mixtures to be mono-channel and the mixing process to be a sum of sources i.e., $\mathbf{m} = \sum_{i=1}^{K} \mathbf{s}_i$. Figure 1 provides an overview of our proposed approach for unsupervised source separation. Here, we sample the source audio from the respective priors and perform additive mixing to reconstruct the mixture *i.e*, $\hat{\mathbf{m}} = \sum_{i=1}^{K} \mathcal{G}_i(\mathbf{z}_i)$. The mixture is then processed to obtain the corresponding spectrogram. In addition, we also compute the source level spectrograms. We perform source separation by efficiently searching the latent space of the source-specific priors $\mathcal{G}_i$ using *Projected Gradient Descent* optimizing a spectral domain loss function $\mathcal{L}$. More formally, for a single mixture $\mathbf{m}$, our objective function is given by,

$$\{\mathbf{z}_i^*\}_{i=1}^{K} = arg \min_{\mathbf{z}_1, \mathbf{z}_2 \dots \mathbf{z}_K} \mathcal{L}(\hat{\mathbf{m}}, \mathbf{m}) + \mathcal{R}(\{\mathcal{G}_i(\mathbf{z}_i)\}), \quad (1)$$

where the first term measures the discrepancy between the true and estimated mixtures and the second term is an optional regularizer on the estimated sources. In every PGD iteration, we perform a projection $\mathcal{P}$, where we constrain the $\{\mathbf{z}_i\}_{i=1}^{K}$ to their respective manifolds. Upon completion of this optimization, the sources can be obtained as $\hat{\mathbf{s}}_i^* = \mathcal{G}_i(\mathbf{z}_i^*), \forall i$. Here, we reformulate the process of source separation by first estimating the source-specific latent features $\mathbf{z}_i^*$ followed by sampling from the respective generators. There are two key ingredients that are critical to the performance of our approach: (i) choice of a good quality *GAN Prior* for every source and (ii) carefully chosen loss functions to drive the PGD optimization. We now elaborate our methodology in the rest of this section.

3.1. WaveGAN for Data Prior Construction

WaveGAN [21] is a popular generative model capable of synthesizing raw waveform audio. It has exhibited success in producing audio from different domains such as speech and musical instruments. Both the generator and discriminator of the WaveGAN model are similar in construction to DCGAN [25] with certain architectural changes to support audio generation. The generator $\mathcal{G}$ transforms the latent features $\mathbf{z} \in \mathbb{R}^{d_z}$ where $d_z = 100$ from a uniform distribution in $[-1, 1]$, to produce waveform audio $\mathcal{G}(\mathbf{z})$ of dimension $d = 16384$ which is approximately of 1s duration at a sampling rate of 16kHz. The discriminator $\mathcal{D}$ regularized using phase shuffle learns to distinguish between the real and synthesized samples. The WaveGAN is

Algorithm 1: Proposed Approach.

Input: Unlabeled mixture $\mathbf{m}$, No. of sources K, Pre-trained *GAN Priors* $\{\mathcal{G}_i\}_{i=1\dots K}$
Output: Estimated sources $\{\hat{\mathbf{s}}_i^*\}_{i=1\dots K}$
Initialization: $\{\hat{\mathbf{z}}_i\}_{i=1\dots K} = \mathbf{0} \in \mathbb{R}^{d_z}$
for $n \leftarrow 1$ **to** N **do**
  $\hat{\mathbf{m}} = \sum_{i=1}^{K} \mathcal{G}_i(\hat{\mathbf{z}}_i)$
  Compute source level and mixture spectrograms
  Compute loss $\mathcal{L}$ using 6
  $\hat{\mathbf{z}}_i \leftarrow \hat{\mathbf{z}}_i - \eta \nabla_z(\mathcal{L}) \quad \forall i = 1 \dots K$
  $\hat{\mathbf{z}}_i \leftarrow \mathcal{P}(\hat{\mathbf{z}}_i)$ $\mathcal{P}$ projects $\{\mathbf{z}_i\}_{i=1\dots K}$ onto the manifold, i.e., clipped to $[-1, 1]$
end
return $\{\hat{\mathbf{s}}_i^*\} = \mathcal{G}_i(\mathbf{z}_i^*), \forall i$

trained to optimize the Wasserstein loss with gradient penalty (WGAN-GP) as prescribed in [26].

Given the ability of WaveGAN to synthesize high quality audio, the pre-trained generator of WaveGAN was used to define the *GAN Prior*. In our formulation, instead of using a single *GAN Prior* trained jointly for all sources, we construct K independent source-specific priors.

3.2. Losses

In order to obtain high-quality source estimates using GAN priors, we propose a novel yet intuitive combination of spectral-domain losses. Though one can utilize time-domain metrics such as the Mean-Squared Error (MSE) to compare the observed and synthesized mixtures, we find that even small variations in the phases of sources estimated from our priors can lead to higher error values. This in turn can misguide the PGD optimization process and may lead to poor convergence. This corroborates with the findings in [27].

3.2.1. Multiresolution Spectral Loss ($\mathcal{L}_{ms}$)

This loss term measures the ℓ_1-norm between log magnitudes of the reconstructed spectrogram and the input spectrogram at L spatial resolutions. This is used to enforce perceptual closeness between the two mixtures at varying spatial resolutions. Denoting $\mathbf{m}$ as the input mixture and $\hat{\mathbf{m}}$ as the estimated mixture, the loss $\mathcal{L}_{ms}$ is defined as

$$\mathcal{L}_{ms} = \sum_{l=1}^{L} \Big\| \log(1 + |STFT^l(\mathbf{m})|^2) - \log(1 + |STFT^l(\hat{\mathbf{m}})|^2) \Big\|_1, \tag{2}$$

where $|STFT^l(.)|$ represents the magnitude spectrograms at the l^{th} spatial resolution and $L = 3$. We compute the magnitude spectrogram at different resolutions by performing a simple average pooling operation with bilinear interpolation.

3.2.2. Source Dissociation Loss ($\mathcal{L}_{sd}$)

Minimizing this loss, defined as the aggregated gradient similarity between the spectrograms of the estimated sources, enforces them to be systematically different. Similar to [15, 28], we define this as a product of the normalized gradient fields of the log magnitude spectrograms computed at L spatial resolutions. In the case where there are K constituent sources, we compute

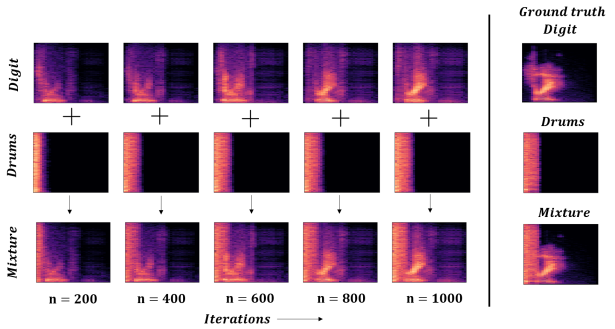

Figure 2: *Demonstration of our proposed approach using a digit-drum example. Through the use of multiple GAN Priors $\mathcal{G}_i$, our algorithm efficiently searches the source-specific latent spaces to estimate the underlying sources.*

this between every pair of sources. Formally,

$$\mathcal{L}_{sd} = \sum_{i=1}^{K} \sum_{j=i+1}^{K} \sum_{l=1}^{L} ||\Psi(\log(1 + |STFT^l(\mathcal{G}_i(\hat{\mathbf{z}}_i))|^2), \log(1 + |STFT^l(\mathcal{G}_j(\hat{\mathbf{z}}_j))|^2))||_F, \tag{3}$$

where $\Psi(x, y) = tanh(\lambda_1 |\nabla x|) \odot tanh(\lambda_2 |\nabla y|)$. ($\odot$ represents element-wise multiplication) and $L = 3$. The weights λ_1 and λ_2 are set at $\lambda_1 = \frac{\sqrt{|\nabla y|_F}}{\sqrt{|\nabla x|_F}}$ and $\lambda_2 = \frac{\sqrt{|\nabla x|_F}}{\sqrt{|\nabla y|_F}}$.

3.2.3. Mixture Coherence Loss ($\mathcal{L}_{mc}$)

Along with $\mathcal{L}_{ms}$, this loss, defined using gradient similarity between original and reconstructed mixtures, ensures that our PGD optimization produces meaningful reconstructions:

$$\mathcal{L}_{mc} = -\sum_{l=1}^{L} ||\Psi(\log(1 + |STFT^l(\mathbf{m})|^2), \log(1 + |STFT^l(\hat{\mathbf{m}}))|^2))||_F \tag{4}$$

3.2.4. Frequency Consistency Loss ($\mathcal{L}_{fc}$)

This loss helps improve perceptual similarity between the magnitude spectrograms of the input and synthesized mixtures by constraining components within a particular temporal bin of the spectrograms to remain consistent over the entire frequency range, i.e.,

$$\mathcal{L}_{fc} = \sum_{t=1}^{T} \sum_{f=1}^{F} \frac{\log(1 + |STFT(\mathbf{m})[t, f]|)}{\log(1 + |STFT(\hat{\mathbf{m}})[t, f]|)}. \tag{5}$$

The overall loss function for our source separation algorithm is thus obtained as:

$$\mathcal{L} = \beta_1 \mathcal{L}_{ms} + \beta_2 \mathcal{L}_{sd} + \beta_3 \mathcal{L}_{mc} + \beta_4 \mathcal{L}_{fc} \tag{6}$$

Through hyperparameter search we identified that $\beta_1 = 0.8, \beta_2 = 0.3, \beta_3 = 0.1, \beta_4 = 0.4$ to be effective in our experiments. Note, in our computations we obtain the spectrograms by computing the Short Time Fourier Transform (STFT) on the waveform in frames of length 256, hop size of 128 and FFT length of 256. The procedure for our approach is showed in Algorithm 1. Figure 2 illustrates the progressive estimation of the unknown sources using our approach.

Table 1: *Performance metrics averaged across* 1000 *cases for the Digit-Piano (K = 2) experiment (While higher Spectral SNR and SIR are better, lower RMS Env.Distance is better).*

Method	Spectral SNR (dB)		RMS Env. Distance		SIR (dB)	
	Digit	Piano	Digit	Piano	Digit	Piano
FastICA	-2.13	-13.45	0.22	0.61	-4.12	-0.66
PCA	-2.04	-12.01	0.22	0.54	-4.13	-1.44
Kernel PCA	-2.04	-3.30	0.22	0.26	-4.13	-1.61
NMF	-2.21	-5.80	0.23	0.26	-4.09	2.53
DAP	-1.77	**2.72**	0.22	0.22	2.20	-3.10
Proposed	**1.06**	**2.73**	**0.17**	**0.21**	**3.91**	**8.57**

Table 2: *Performance metrics averaged across* 1000 *cases for the Drums-Piano (K = 2) experiment.*

Method	Spectral SNR (dB)		RMS Env. Distance		SIR (dB)	
	Drums	Piano	Drums	Piano	Drums	Piano
FastICA	-5.25	-13.52	0.24	0.61	-6.51	-1.45
PCA	-5.19	-12.33	0.24	0.56	-6.53	-2.69
Kernel PCA	-5.19	-3.36	0.24	0.25	-6.53	-2.02
NMF	-5.39	-5.84	0.24	0.26	-6.59	3.84
DAP	-4.20	2.97	0.22	**0.21**	-21.62	**11.22**
Proposed	**0.84**	**3.06**	**0.10**	**0.21**	**11.70**	9.80

Table 3: *Performance metrics averaged across* 1000 *cases for the Digit-Drums (K = 2) experiment.*

Method	Spectral SNR (dB)		RMS Env. Distance		SIR (dB)	
	Digit	Drums	Digit	Drums	Digit	Drums
FastICA	2.91	-21.01	**0.13**	0.82	3.10	0.09
PCA	2.99	-20.00	**0.13**	0.77	3.12	0.02
Kernel PCA	2.99	-10.53	**0.13**	0.35	3.12	0.85
NMF	3.01	-13.75	**0.13**	0.39	3.20	-0.98
DAP	**3.59**	**0.92**	0.14	0.14	4.24	-11.48
Proposed	2.32	0.42	0.15	**0.10**	**25.91**	**23.68**

4. Empirical Evaluation

In this section, we evaluate our proposed approach on two source and three source separation experiments on the publicly available Spoken Digit (SC09), drum sounds and piano datasets. The SC09 dataset is a subset of the Speech Commands dataset [29, 21] containing spoken digits (0-9) each of duration $\sim$ 1s at 16kHz from a variety of speakers recorded under different acoustic conditions. The drum sounds dataset [21] contains single drum hit sounds each of duration $\sim$ 1s at 16kHz. The piano dataset [21] contains piano music (Bach compositions) each of duration ($>$ 50s) at 48kHz.

WaveGAN Training. Following [21], we train WaveGAN models on normalized 1s slices (*i.e* d =16384 samples) of the SC09 (Digit), Drums and Piano train datasets resampled to 16kHz respectively. All the models were trained using batches of size 128. The generator and discriminator were optimized using the WGAN-GP loss with the Adam optimizer and learning rate $1e^{-4}$ for 3000 epochs. The trained generator models were used to construct the GAN priors.

Setup. For the task of two source separation ($K = 2$), we conducted experiments on three possible mixture combinations: (i) Digit-Piano, (ii) Drums-Piano and (iii) Digit-Drums. In order to create the input mixture for every combination, we randomly sampled (with replacement) normalized 1s audio slices from the respective test datasets, and obtained 1000 mixtures through a simple additive mixing process. Similarly, we obtained 1000 mixtures for the case of $K = 3$, i.e., on the combination, Digit-Drums-Piano. In each case, we performed the PGD optimization using Eqn.6 for $N = 1000$ iterations with the ADAM optimizer and learning rate of $5e^{-2}$ to infer source specific latent features $\{\mathbf{z}_i^*\}_{i=1\ldots K}$. The estimated sources are then obtained as $\{\mathcal{G}_i(\mathbf{z}_i^*)\}_{i=1\ldots K}$. Though the choice of initialization for $\mathbf{z}_i$ is known to be critical for PGD optimization [20], we find that setting $\{\mathbf{z}_i\}_{i=1\ldots K} = \mathbf{0} \in \mathbb{R}^{d_z}$ to be effective.

Table 4: *Performance metrics averaged across* 1000 *cases for the Digit-Drums-Piano (K = 3) experiment.*

Metric	Source	FastICA	PCA	Kernel PCA	NMF	Proposed
Spectral SNR (dB)	Digit	-2.95	-2.47	-2.47	-2.47	**0.77**
	Drums	-10.8	-19.81	-8.1	-12.84	**0.64**
	Piano	0.27	0.1	-0.94	**4.94**	2.64
RMS Env. Distance	Digit	0.24	0.23	0.23	0.23	**0.17**
	Drums	0.4	0.75	0.28	0.37	**0.1**
	Piano	0.23	0.31	0.25	**0.15**	0.21
SIR (dB)	Digit	-4.73	-5.06	-5.06	-5.01	**3.02**
	Drums	-6.48	-5.51	-1.65	-5.69	**10.21**
	Piano	0.53	2.21	-3.87	2.60	**5.12**

Evaluation Metrics. Following standard practice, we used three different metrics - (i) mean spectral SNR [30, 31], a measure of the quality of the spectrogram reconstruction; (ii) mean RMS envelope distance [32] between the estimated and true sources; and (iii) mean signal-interference ratio (SIR) [33] to quantify the interference caused by one estimated source on another.

Results. Tables 1, 2, 3 and 4 provide a comprehensive comparison of the proposed approach against the standard baselines (FastICA, PCA, KernelPCA, NMF) [34] as well as with the state-of-the-art unsupervised Deep-Audio-Prior [15]. It can be observed that our approach significantly outperforms all the baselines in most cases, except for the Digits-Drums experiment where our method is in par with DAP. These results indicate the effectiveness of our unsupervised approach on complex source separation tasks. We find that the spectral SNR metric, which is relatively less sensitive to phase differences [27, 30], is consistently high with our approach, indicating high perceptual similarities between estimated and the ground truth audio. We also find lower envelope distance estimates, further emphasizing the perceptual quality of our estimated sources. Finally, we attribute the significant improvements in the SIR metric to the *source dissociation loss* ($\mathcal{L}_{sd}$), which enforces the estimated sources from the priors to be systematically different.

5. Conclusions

In summary, we find that source-specific *GAN Priors* are effective in recovering the constituents of an unlabeled mixture, often significantly outperforming unsupervised state-of-the-art benchmarks. Additionally, we find that such generative priors can be further improved with PGD-style optimization using carefully designed spectral domain loss functions. Our approach is highly flexible because it is entirely an inference-time technique, and as a result can efficiently deal with varying number of known sources in a given mixture. This is in contrast with standard supervised approaches which require re-training or extensive fine-tuning. Future extensions to our work include performing source separation when the mixing process is unknown, and dealing with mixtures that contain novel sources.

6. References

[1] A. Spanias, T. Painter, and V. Atti, *Audio signal processing and coding*. John Wiley & Sons, 2006.

[2] A. Spanias, "Advances in speech and audio processing and coding," *6th IEEE International Conference on Information, Intelligence, Systems and Applications (IISA)*, pp. 1–2, July 2015.

[3] S. Makino, S. Araki, R. Mukai, and H. Sawada, "Audio source separation based on independent component analysis," in *IEEE International Symposium on Circuits and Systems*, vol. 5, pp. May, 2004.

[4] J. Karhunen, L. Wang, and R. Vigario, "Nonlinear pca type approaches for source separation and independent component analysis," *International Conference on Neural Networks (ICNN)*, vol. 2, pp. 995–1000, 1995.

[5] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias, "Mixing matrix estimation using discriminative clustering for blind source separation," *Digital Signal Processing*, vol. 23, no. 1, pp. 9–18, 2013.

[6] L. Wang, J. D. Reiss, and A. Cavallaro, "Over-determined source separation and localization using distributed microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1573–1588, 2016.

[7] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[8] Y. Luo and N. 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.

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

[10] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," pp. 106–110, 2018.

[11] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," pp. 1577–1581, 2018.

[12] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Demucs: Deep extractor for music sources with extra unlabeled data remixed," *arXiv preprint arXiv:1909.01174*, 2019.

[13] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective." *ICMC*, pp. 231–234, 2003.

[14] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9446–9454, 2018.

[15] Y. Tian, C. Xu, and D. Li, "Deep audio prior," *arXiv preprint arXiv:1912.10292*, 2019.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[17] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," *34th International Conference on Machine Learning (ICML)*, vol. 70, pp. 537–546, 2017.

[18] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *IEEE international conference on computer vision (ICCV)*, pp. 2223–2232, 2017.

[19] V. Shah and C. Hegde, "Solving linear inverse problems using gan priors: An algorithm with provable guarantees," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4609–4613, 2018.

[20] R. Anirudh, J. J. Thiagarajan, B. Kailkhura, and P.-T. Bremer, "Mimicgan: Robust projection onto image manifolds with corruption mimicking," *International Journal of Computer Vision*, pp. 1–19, 2020.

[21] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *ICLR*, 2019.

[22] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, "Kernel pca and de-noising in feature spaces," *Advances in neural information processing systems*, pp. 536–542, 1999.

[23] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with nmf: Divergences, constraints and algorithms," *Audio Source Separation*, pp. 1–24, 2018.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, 2015.

[25] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[26] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, pp. 5767–5777, 2017.

[27] A. Défossez, N. Zeghidour, N. Usunier, L. Bottou, and F. Bach, "Sing: Symbol-to-instrument neural generator," *Advances in Neural Information Processing Systems*, pp. 9041–9051, 2018.

[28] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4786–4794, 2018.

[29] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[30] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," *Proceedings of the 12th International Conference on Digital Audio Effects*, 2009.

[31] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[32] P. Morgado, N. Nvasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360 video," *Advances in Neural Information Processing Systems*, pp. 362–372, 2018.

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

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.