MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Location as supervision for weakly supervised multi-channel source separation of machine sounds

Falcon Perez, Ricardo; Wichern, Gordon; Germain, Francois; Le Roux, Jonathan

TR2023-119 September 14, 2023

Abstract

In this work, we are interested in learning a model to separate sources that cannot be recorded in isolation, such as parts of a ma- chine that must run simultaneously in order for the machine to function. We assume the presence of a microphone array and knowledge of the source locations (potentially obtained from schematics or an auxiliary sensor such as a camera). Our method uses the source lo- cations as weak labels for learning to separate the sources, since we cannot obtain the isolated source signals typically used as training targets. We propose a loss function that requires the directional features computed from the separated sources to match the true direction of arrival for each source, and also include a reconstruction loss to ensure all frequencies are taken into account by at least one of the separated sources output by our model. We benchmark the performance of our algorithm using synthetic mixtures created using machine sounds from the DCASE 2021 Task 2 dataset in challenging reverberant conditions. While reaching lower objective scores than a model with access to isolated source signals for training, our proposed weakly-supervised model obtains promising results and applies to industrial scenarios where collecting isolated source signals is prohibitively expensive or impossible.

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2023

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

LOCATION AS SUPERVISION FOR WEAKLY SUPERVISED MULTI-CHANNEL SOURCE SEPARATION OF MACHINE SOUNDS

Ricardo Falcon-Perez[1,2], *Gordon Wichern*[1], *François G. Germain*[1], *Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Acoustics Lab, D.I.C.E., Aalto University, Espoo, Finland

ABSTRACT

In this work, we are interested in learning a model to separate sources that cannot be recorded in isolation, such as parts of a machine that must run simultaneously in order for the machine to function. We assume the presence of a microphone array and knowledge of the source locations (potentially obtained from schematics or an auxiliary sensor such as a camera). Our method uses the source locations as weak labels for learning to separate the sources, since we cannot obtain the isolated source signals typically used as training targets. We propose a loss function that requires the directional features computed from the separated sources to match the true direction of arrival for each source, and also include a reconstruction loss to ensure all frequencies are taken into account by at least one of the separated sources output by our model. We benchmark the performance of our algorithm using synthetic mixtures created using machine sounds from the DCASE 2021 Task 2 dataset in challenging reverberant conditions. While reaching lower objective scores than a model with access to isolated source signals for training, our proposed weakly-supervised model obtains promising results and applies to industrial scenarios where collecting isolated source signals is prohibitively expensive or impossible.

Index Terms— Multichannel source separation, weak supervision, directional features, machine sound

1. INTRODUCTION

When monitoring machine performance and health, highly skilled human operators often use their ears to listen to the sounds produced during machine operation. As automation increases, algorithms that use microphones to monitor machine sounds are becoming more and more important. In light of this, there has been a surge of recent interest in anomalous sound detection [1–5], where automated algorithms determine whether sound produced during machine operation is normal or anomalous. In particular, recent public challenges [6–8] have spurred research in increasingly difficult problem setups where only normal data is available for training, and domain shift causes changes in the sound signal unrelated to the presence or absence of anomalous sound.

However, much of the existing literature on machine sound analysis treats the recorded sound as being produced by a single machine part, when in practice most industrial machinery is composed of multiple sound-producing components or parts, and we may want to monitor the health of each component individually. In these situations, audio source separation could be a useful preprocessing step to isolate the sound from each machine part, and the separated sound signals from each part could then be used as input to downstream processing such as anomalous sound detection or other types of audio monitoring. While there has been tremendous progress in deep-learning-based audio source separation, particularly in the areas of speech enhancement [9], speech separation [10], music source separation [11], and general sound separation [12, 13], the vast majority of these approaches use a fully supervised framework, where a database of isolated sound signals is used to create artificial mixture signals that the source separation model is trained to separate using the ground-truth isolated signals as targets. However, for separating machine parts, collecting a database of isolated signals from individual parts may not be possible, as all parts of the machine may need to operate simultaneously for the machine to run, and we must thus consider approaches with limited supervision.

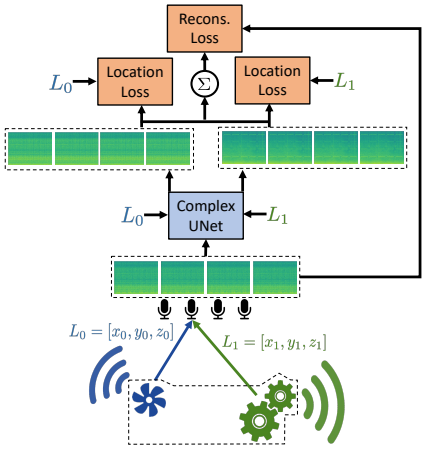

Figure 1: Illustration of using location as weak supervision. A microphone array observes a mixture of machine parts, and assumes the precise location of each part is known. The model is trained to output multi-channel complex spectrograms of the separated sources by encouraging the delays of the separated sources to align with the expected delays given the known location.

While unsupervised separation algorithms such as those based on MixIT [14] have achieved impressive results when all sources to be separated are independent, they struggle with correlated sources such as music signals. Alternatively, there is a class of weakly-supervised separation approaches where auxiliary information such as class activities [15], musical score [16], or video [17] is available and, when combined with an appropriate loss function, can supervise the separation process. In this work, we assume a microphone array records the sound from simultaneously operating machine parts, and the precise location of each machine part is available, either from sensors in a different modality (e.g., radar, vision) or from schematic diagrams.

This work was performed while R. Falcon-Perez was an intern at MERL.

979-8-3503-2372-6/23/$31.00 ©2023 IEEE

Recently, separation conditioned on source location has begun receiving significant attention for separating multi-channel speech signals in fully supervised setups [18–24], and these methods have vastly exceeded the performance of beamformers with oracle information. Furthermore, several works train unsupervised models for separating speech signals using multi-channel information [25–28]. Closest to our work, Saijo et al. [28] propose to train a neural separator that estimates a time-invariant linear demixing filter, using an unsupervised loss that matches estimated source locations and the spatial demixing matrices. In this work, we consider a related approach, but directly estimate the separated sources using a neural network instead of constraining ourselves to a time-invariant linear filter, and derive our loss function using the recently proposed directional features [29], which quantify how much the observed inter-channel phase difference between microphones at each time-frequency bin aligns with the expected phase difference computed using the known target location. As the main building block of our weakly supervised loss function, each separated source output by our model should have directional features well matched with the true source location, and we also include a mixture reconstruction loss to ensure the whole signal is accounted for (Fig. 1). Using a simulated dataset of mixtures of machine sounds from the DCASE 2021 Task 2 dataset [7] in difficult reverberation conditions, we demonstrate that our weakly supervised model learns to separate sounds that are impossible to record in isolation.

2. METHOD

2.1. Multi-channel Source Separation

Throughout this work, we assume a P-channel audio mixture signal $\mathbf{y} = \mathbf{s}_0 + \mathbf{s}_1 \in \mathbb{R}^{P\times N}$ with length N samples, which is the sum of reverberant signals $\mathbf{s}_0$ and $\mathbf{s}_1$, where we consider the source images at the microphones. We consider a class of models that take as input time-frequency (T-F) mixtures $\mathbf{Y} = \mathrm{STFT}(\mathbf{y}) \in \mathbb{C}^{P\times T\times F}$ and estimate $\hat{\mathbf{S}}_i \in \mathbb{C}^{P\times T\times F}$ for $i = 0, 1$. If isolated source signals are available, then we can use the complex T-F representations of the true source signals $\mathbf{S}_i$ as training targets. Following [30], our fully supervised loss function contains three mean square error terms for the real, imaginary, and magnitude losses between $\mathbf{S}_i$ and $\hat{\mathbf{S}}_i$.

2.2. Directional Features

In addition to using the complex multi-channel spectrogram $\mathbf{Y}$ as input to the separation model, we also consider other input features. In multi-channel scenarios [18, 20, 29], interaural phase difference (IPD) are commonly used as spatial features, and defined as

$$\text{Real-IPD}^p_{t,f}(\mathbf{Y}) = \angle\mathbf{Y}^{p_0}_{t,f} - \angle\mathbf{Y}^{p}_{t,f} \in \mathbb{R}, \tag{1}$$

where p_0 is the reference microphone, and the IPD is computed for each of the non-reference microphones, i.e., $p = 1, \ldots, P-1$. To mitigate discontinuities caused by phase wrapping, IPD features are typically mapped to a complex number, i.e.,

$$\text{IPD}^p_{t,f}(\mathbf{Y}) = \cos(\text{Real-IPD}^p_{t,f}(\mathbf{Y})) + j\sin(\text{Real-IPD}^p_{t,f}(\mathbf{Y})) \in \mathbb{C}. \tag{2}$$

When the source location $L_i = [x_i, y_i, z_i]$ (defined relative to reference microphone p_0) is known, we define $\tau(L_i, p)$ as the pure time delay in seconds between a signal from a point source located at L_i traveling to microphone p and the signal traveling to p_0. The target phase difference (TPD) for source i is defined as

$$\text{TPD}^p_f(L_i) = \cos(2\pi f\tau(L_i, p)) + j\sin(2\pi f\tau(L_i, p)) \in \mathbb{C}. \tag{3}$$

The directional feature (DF), which serves as location conditioning in terms of an input feature in [20], indicates whether a T-F bin in spectrogram $\mathbf{Y}$ is dominated by a source at location L_i, and is defined as

$$d_{t,f}(\mathbf{Y}, L_i) = \sum_{p=1}^{P-1} \text{TPD}^p_f(L_i)\overline{\text{IPD}}^p_{t,f}(\mathbf{Y}) \in \mathbb{C}, \tag{4}$$

where $\overline{\text{IPD}}^p_{t,f}$ is the complex conjugate of $\text{IPD}^p_{t,f}$. We then use as network input the multi-channel complex spectrogram, IPD, DF, and frequency positional encodings [31] concatenated along the feature dimension.

2.3. Weakly Supervised Loss Function

When isolated sources are not available for training, one simple objective is to ensure the separated sources output by the network reconstruct the mixture. In this work, the reconstruction loss is composed of two terms: the spectral loss,

$$\mathcal{L}_{\text{spec}} = \left\|\mathbf{Y} - \hat{\mathbf{Y}}\right\|_2^2 + \left\||\hat{\mathbf{Y}}| - |\mathbf{Y}|\right\|_2^2, \tag{5}$$

and the time domain spatial covariance loss,

$$\mathcal{L}_{\text{spat}} = \left\|\mathbf{y}\mathbf{y}^\mathsf{T} - \hat{\mathbf{y}}\hat{\mathbf{y}}^\mathsf{T}\right\|_2^2, \tag{6}$$

where $\hat{\mathbf{y}} = \text{iSTFT}(\hat{\mathbf{S}}_0) + \text{iSTFT}(\hat{\mathbf{S}}_1)$ is the estimated mixture.

The reconstruction loss only ensures that the sum of combined outputs of the network is equal to the input mixture, that is, the mixtures are fully reconstructed and there is no loss of audio content. However, to ensure separation, another term is required. We use a loss based on the DF from (4), but instead of computing the IPD using the mixture $\mathbf{Y}$ as in (1), we use the separated source estimates $\hat{\mathbf{S}}_i$. Our location loss is defined as

$$\mathcal{L}_{\text{loc}} = \sum_{i=1}^{2}\sum_{f}\sum_{t}\Bigg(\left\|\Re(d_{t,f}(\hat{\mathbf{S}}_i, L_i)) - P\right\|_2^2 + \left\|\Im(d_{t,f}(\hat{\mathbf{S}}_i, L_i)) - 0\right\|_2^2\Bigg), \tag{7}$$

where $d_{t,f}(\hat{\mathbf{S}}_i, L_i)$ is the DF computed using the ith estimated source and true location L_i. This loss ensures that the separated sources match the true locations, where $d_{t,f}(\hat{\mathbf{S}}_i, L_i) := P + 0j$ when the IPDs of the separated sources match the expected phase delays of the known source locations. The output order of the estimated sources is unambiguously determined by the order of source locations input to the network.

The total loss is given by:

$$\mathcal{L} = \beta_{\text{spec}}\mathcal{L}_{\text{spec}} + \beta_{\text{spat}}\mathcal{L}_{\text{spat}} + \beta_{\text{loc}}\mathcal{L}_{\text{loc}}, \tag{8}$$

where β_{spec}, β_{spat} and β_{loc} are hyperparameters that weight each term of the loss function. We experimentally found that $\beta_{\text{spec}} = 1.0$, $\beta_{\text{spat}} = 1e^{-3}$, and $\beta_{\text{loc}} = 5e^{-2}$ provide the best results. An ablation of these weights is shown in Section 4.

3. EXPERIMENTAL SETUP

3.1. Dataset

To train and evaluate our approach, we create a synthetic dataset that emulates machines in realistic conditions. We consider a scenario

Table 1: Simulation constraints for array and source placement.

Parameter	Range
Distance between sources	$[0.5, 1.5]$
Distance between sources and mic array center	$[0.75, 2.0]$
Distance between sources or mic, and room surface	$[0.5, \infty]$
Angle between mic array normal and sources	$[0°, 30°]$

where a machine with two sound generating sources are present in a room and the microphone array can be positioned arbitrarily, but with some constraints. This represents situations where it might not always be possible to place the microphone array in the ideal position (close to the machine, between the two sources), for instance, due to obstacles in the room. We use an 11 element microphone array harmonically spaced, similar to [30], but with spacing in cm of [16.8, 8.4, 4.2, 2.1, 2.1, 2.1, 2.1, 4.2, 8.4, 16.8] for a total span of 67.2 cm. For each data example, we first create a virtual shoebox room, with dimensions for each wall between 2 and 5 meters for small rooms, and between 4 and 10 meters for large rooms. The height for the ceiling is drawn between 3 and 5 meters. The acoustic properties for each surface are randomly drawn, using multi-band materials with absorption coefficients between 0.1 (mostly reflective) and 0.9 (mostly absorptive) defined at 7 octave bands (125-8000 Hz). The absorption of each band is drawn independently from other bands. This is an indirect method to control reverberation time instead of setting a desired T60. Nevertheless, this creates more realistic reverberant conditions, where the T60 is not uniform across the whole frequency range. Once the room has been created, the positions for the sources and microphone array are randomly selected. These positions can be placed anywhere in room under certain constraints, as detailed in Table 1.

Afterwards, the room impulse responses are generated using the image source method with a high order to create the early reflections, and a ray tracing method for the late reverberation, using the PyRoomAcoustics toolbox [32] for the simulations. The multichannel room impulse responses are then convolved with single-channel source signals to simulate omnidirectional point sources.

In total, we generate 4 versions of the dataset: anechoic and reverberant, with two sets of machines as fixed sources, such that s_0 and s_1 are always the same source type: "gearbox" and "slider" for SetA, and "pump" and "valve" for SetB respectively. For each example, we draw a random file of each class, and apply loudness normalization between [-17, -12] LUFS independently for each source. In total, we generate 24,000 mixtures of 10 seconds length, at 16 kHz sampling rate for a total of 66.7 hours. The mixtures are split into train/validation/test subsets with 15000/6000/3000 examples. All subsets have similar acoustic conditions but the source files are different, by drawing from different sections as defined in the DCASE2021 task 2 dataset. We use sections 1,2,3, for the train set, sections 4,5 for validation, and section 6 for test. We consider only normal files and discard any anomalies.

Figure 2 shows the distributions of key acoustic parameters for the reverberant dataset, mainly T60, early decay time (EDT), directo-to-reverberant ratio (DRR), and clarity index (C50) that can have a large impact in the generalization of deep learning models to diverse audio tasks [33]. Although most of the rooms have moderate reverberation up to about 0.5 s, we include examples up to 4 seconds long. Moreover, due to the constraints set on the distance between microphone and sources, most examples have relatively high DRR and C50 with noticeable exceptions. Overall, the dataset presents diverse and challenging acoustical conditions.

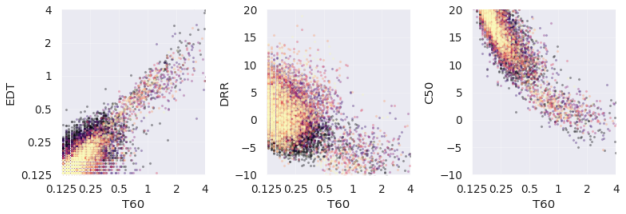

Figure 2: Joint distributions of selected room acoustic parameters of the simulated dataset. T60 and EDT in seconds, C50 and DRR in dB. Colored by frequency band from low (dark) to high (bright).

3.2. Network Architecture and Training

Our model architecture closely follows the complex dense UNET architecture from [20], which uses an encoder composed of alternating complex convolution layers and complex dense blocks, where each dense block has a skip connection to the corresponding block in the decoder. A complex bidirectional LSTM operates between the encoder and decoder, and all convolution layers use a stride of one in the time dimension and two in the frequency dimension such that an input of shape $C \times T \times F$ becomes $256 \times T \times 1$ at the input of the LSTM, where the number of frequency bins is $F = 257$, and the number of channels C depends on the number of microphones and input features, e.g., 11 microphones, 10 IPDs, 2 DFs (one channel per source), and 10 frequency positional encoding channels sums to $C = 33$. One minor modification we make to the architecture from [20] is that each dense block contains a skip connection and batch normalization as in [34]. We also use a multi-head decoder as is commonly done in music source separation [35], since our model outputs two sources instead of a single source as in the speech enhancement setup from [20]. We output multi-channel complex spectrograms instead of single-channel complex masks.

We train the model on $T = 2.0$ second chunks of STFT frames with a batch size of 8 for 100,000 steps using the ranger optimizer [36] with a learning rate of 1e-2. All STFT operations use a window size of 512, hop size of 128, and Hann window. We evaluate performance in terms of scale-invariant signal-to-distortion ratio (SI-SDR) [37] computed using `fast_bss_eval` [38] between the reverberant reference source and corresponding estimate at the reference microphone selected as the microphone at the center of the array. We report the results on the test set from the best performing model measured by the validation loss.

4. RESULTS

Table 2 compares the SI-SDR of different approaches. We present the results of systems trained and evaluated on two sets of machines as sources, and two reverberation conditions, simulated as described in Section 3. For the models evaluated in anechoic datasets, we first notice that the fully supervised model outperforms the ideal binary masks (IBM). This is expected because the IBM ignores the phase information that can be modeled by the complex UNET. Second, the weakly supervised model is less successful, but still achieves separation. Both models have lower performance when trained on the reverberant dataset. The poor performance of the delay-and-sum beamformer is most likely due to spatial aliasing at high frequencies, poor directivity at low ones, and lack of precise null placement, considering that the sources are often very close to each other.

For the models evaluated in the reverberant datasets, the results are slightly different. The fully supervised model struggles to match

Table 2: Performance in terms of mean ± standard deviation of SI-SDR (dB) for different source separation approaches evaluated on datasets with 2 different sets of machines, and 2 different acoustical conditions. SetA = [s_0 = gearbox, s_1 = slider]; SetB = [s_0 = pump, s_1 = valve].

			Anechoic				Reverberant			
	Trained on		SetA		SetB		SetA		SetB	
Approach	Set	Reverb	SI-SDR$_0$ ↑	SI-SDR$_1$ ↑	SI-SDR$_0$ ↑	SI-SDR$_1$ ↑	SI-SDR$_0$ ↑	SI-SDR$_1$ ↑	SI-SDR$_0$ ↑	SI-SDR$_1$ ↑
Mixture	n/a	n/a	−0.1±4.4	0.1±4.4	0.1±2.5	−0.1±2.5	0.0±2.6	0.0±2.6	0.1±2.6	−0.1±2.6
Delaysum	n/a	n/a	−3.0±7.1	−2.9±6.7	−2.4±5.8	−2.4±5.7	−3.2±5.1	−3.5±5.2	−3.1±5.1	−3.4±5.3
Ideal Binary Masks	n/a	n/a	8.7±5.4	8.8±5.3	8.8±3.4	8.2±3.3	9.0±3.3	8.9±3.2	9.1±3.3	8.7±3.2
Fully Supervised	A	✓	15.2±2.5	15.6±2.6	14.3±2.3	14.4±2.0	7.7±3.3	7.7±3.3	7.4±3.3	7.3±3.2
Fully Supervised	A	✗	21.4±2.8	23.6±3.2	18.9±3.7	21.3±3.3	3.8±5.0	4.2±5.0	3.6±4.8	4.0±4.7
WeakSup	A	✓	4.8±3.4	4.1±2.5	5.7±2.3	4.9±2.0	1.6±2.7	1.2±2.3	1.8±2.8	1.4±2.5
WeakSup	A	✗	7.0±3.4	7.1±3.3	7.7±2.2	7.5±2.3	3.2±2.8	3.2±2.6	3.2±2.9	3.2±2.6
Fully Supervised	B	✓	11.9±2.4	12.3±2.5	11.5±2.0	11.7±1.7	6.5±3.0	6.5±3.1	6.4±3.0	6.3±2.9
Fully Supervised	B	✗	19.0±2.5	19.4±2.7	18.3±2.6	18.8±2.0	4.2±4.4	4.1±4.4	4.1±4.3	4.0±4.3
WeakSup	B	✓	4.0±3.1	3.9±2.8	4.7±2.0	4.4±1.8	1.7±2.4	1.3±2.3	1.7±2.4	1.1±2.2
WeakSup	B	✗	3.9±3.9	3.7±2.6	5.4±2.4	4.8±2.1	1.9±2.7	1.5±2.1	2.3±2.7	1.6±2.2

the IBM. This is possibly due to the presence of large variance in reverberation time. In addition, the fully supervised model trained on anechoic data fails to generalize to reverberant data. On the other hand, the weakly supervised model also drops in performance, although the drop is less severe. The weakly supervised model trained on anechoic data generalizes better. This is most likely because this model focuses more on location rather than content. Finally, the set used for training has little impact.

Figure 3 shows the density scatter plots for the fully and weakly supervised models trained on reverberant data. Both models show the largest improvements on mixtures with low input SI-SDR. However, the weakly supervised model is less effective for those with high input SI-SDR. Moreover, the weakly supervised model particularly struggles with inputs with high T60. This is most likely because reverberation adds significant noise to the measured IPDs.

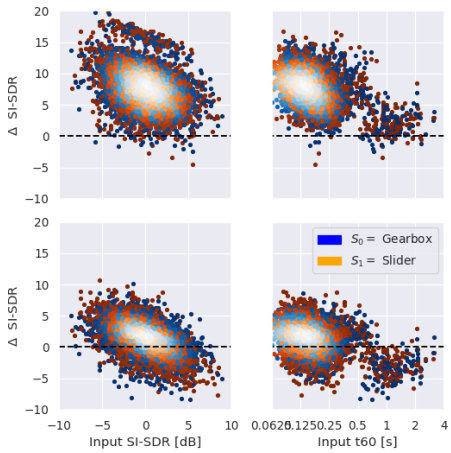

Figure 3: Density scatter plots of ΔSI-SDR compared against (left col.) input SI-SDR and (right col.) mean T60 at the reference mic. (top row) Fully supervised , (bottom row) weakly supervised.

4.1. Ablations

To understand the impact of each component of our approach, we present a limited ablation study. Table 3 shows the impact of the weight for the location loss β_{loc}, while keeping the other weights fixed. The best value for both sources lies between 10^{-3} and 10^{-2}, while most other values fail to do any separation.

Table 3: Mean SI-SDR (dB) using different location loss weights.

	β_{loc}					
Metric	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}
SI-SDR$_0$ ↑	0.0	1.0	1.6	**2.5**	−36.0	−21.3
SI-SDR$_1$ ↑	−0.2	0.9	**1.2**	0.8	−2.6	−20.0

Table 4 shows how different input features affect both the fully supervised and the weakly supervised approaches. The presence of the directional feature as input is crucial to have any separation. Otherwise the network has no other conditioning mechanism to inform it on which source locations we want to separate. Moreover, the presence of other input features such as IPDs or frequency encodings has little impact for the supervised model, and a modest impact in the weakly supervised case.

Table 4: Ablation of input features. All models trained and evaluated on the reverberant SetA dataset detailed in Section 3.1.

	Features				Metrics	
Approach	STFT	IPDs	DF	Enco	SI-SDR$_0$ ↑	SI-SDR$_1$ ↑
Fully Supervised	✓	✗	✗	✗	2.0±3.8	1.9±3.9
Fully Supervised	✓	✗	✓	✗	**7.0**±3.1	**7.0**±3.1
Fully Supervised	✓	✓	✗	✗	1.6±3.8	1.4±3.6
Fully Supervised	✓	✓	✓	✓	6.9±3.1	7.0±3.2
Weak Sup	✓	✗	✗	✗	−0.7±2.4	0.0±2.8
Weak Sup	✓	✗	✓	✗	**1.6**±2.7	1.2±2.3
Weak Sup	✓	✓	✗	✗	−17.1±2.0	−19.6±2.5
Weak Sup	✓	✓	✓	✓	1.4±2.6	**1.7**±2.4

5. CONCLUSION

We presented an approach for learning to separate machine sounds when the location (relative to a microphone array) of the source is known, but isolated training targets are unavailable. Through experiments on a difficult simulated dataset, we demonstrated promising results, especially when reverberation conditions are favorable. In the future, we plan to extend our approach to few-shot learning setups, and incorporate more realistic sound propagation models.

6. REFERENCES

[1] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, *et al.*, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks," in *Proc. ICASSP*, 2015.

[2] E. Cakır and T. Virtanen, "Convolutional recurrent neural networks for rare sound event detection," in *Proc. DCASE*, 2017.

[3] Y. Kawaguchi and T. Endo, "How can we detect anomalies from subsampled audio signals?" in *Proc. MLSP*, 2017.

[4] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, *et al.*, "Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 212–224, 2018.

[5] T. Hayashi, T. Komatsu, R. Kondo, T. Toda, *et al.*, "Anomalous sound event detection based on WaveNet," in *Proc. EUSIPCO*, 2018.

[6] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, *et al.*, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2020.

[7] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, *et al.*, "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proc. DCASE*, 2021.

[8] K. Dohi, K. Imoto, N. Harada, D. Niizumi, *et al.*, "Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. DCASE*, 2022.

[9] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, *et al.*, "The INTERSPEECH 2020 Deep Noise Suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, 2020.

[10] J. R. Hershey, Z. Chen, and J. Le Roux, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016.

[11] Y. Mitsufuji, G. Fabbro, S. Uhlich, F.-R. Stöter, *et al.*, "Music demixing challenge 2021," *Front. Signal Process.*, vol. 1, 2022.

[12] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, *et al.*, "Improving universal sound separation using sound classification," in *Proc. ICASSP*, 2020.

[13] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, *et al.*, "Listen to what you want: Neural network-based universal sound selector," in *Proc. Interspeech*, 2020.

[14] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, *et al.*, "Unsupervised sound separation using mixture invariant training," in *Proc. NeurIPS*, 2020.

[15] F. Pishdadian, G. Wichern, and J. Le Roux, "Finding strength in weakness: Learning to separate sounds with weak supervision," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2386–2399, 2020.

[16] Y.-N. Hung, G. Wichern, and J. Le Roux, "Transcription is all you need: Learning to separate musical mixtures with score as supervision," in *Proc. ICASSP*, 2021.

[17] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proc. CVPR*, 2019.

[18] Z.-Q. Wang and D. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 457–468, 2018.

[19] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, *et al.*, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *Proc. SLT*, 2018.

[20] R. Gu, S.-X. Zhang, Y. Zou, and D. Yu, "Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain," *IEEE Signal Process. Lett.*, vol. 28, pp. 1370–1374, 2021.

[21] T. Jenrungrot, V. Jayaram, S. Seitz, and I. Kemelmacher-Shlizerman, "The cone of silence: Speech separation by localization," in *Proc. NeurIPS*, 2020.

[22] K. Tesch and T. Gerkmann, "Nonlinear spatial filtering in multichannel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1795–1805, 2021.

[23] D. Marković, A. Défossez, and A. Richard, "Implicit neural spatial filtering for multichannel source separation in the waveform domain," in *Proc. Interspeech*, 2022.

[24] K. Tesch and T. Gerkmann, "Spatially selective deep non-linear filters for speaker extraction," in *Proc. ICASSP*, 2023.

[25] L. Drude, J. Heymann, and R. Haeb-Umbach, "Unsupervised Training of Neural Mask-Based Beamforming," in *Proc. Interspeech*, 2019.

[26] M. Togami, Y. Masuyama, T. Komatsu, and Y. Nakagome, "Unsupervised training for deep speech source separation with kullback-leibler divergence based probabilistic loss function," in *Proc. ICASSP*, 2020.

[27] Y. Bando, K. Sekiguchi, Y. Masuyama, A. A. Nugraha, *et al.*, "Neural full-rank spatial covariance analysis for blind source separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 1670–1674, 2021.

[28] K. Saijo and R. Scheibler, "Spatial loss for unsupervised multi-channel source separation," in *Proc. Interspeech*, 2022.

[29] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, *et al.*, "Multi-modal multi-channel target speech separation," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 3, pp. 530–541, 2020.

[30] Z.-Q. Wang, G. Wichern, and J. Le Roux, "On the compensation between magnitude and phase in speech separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 2018–2022, 2021.

[31] U. Isik, R. Giri, N. Phansalkar, J.-M. Valin, *et al.*, "PoCoNet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," in *Proc. Interspeech*, 2020.

[32] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, 2018.

[33] D. Emmanouilidou and H. Gamper, "The effect of room acoustics on audio event classification," in *Proc. Int. Congr. Acoust. (ICA)*, 2019.

[34] E. Moliner and V. Välimäki, "A two-stage U-net for high-fidelity denoising of historical recordings," in *Proc. ICASSP*, 2022.

[35] O. Slizovskaia, G. Haro, and E. Gómez, "Conditioned source separation for music instrument performances," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2083–2095, 2021.

[36] L. Wright, "Ranger - a synergistic optimizer." https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer, 2019.

[37] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. ICASSP*, 2019.

[38] R. Scheibler, "SDR—medium rare with fast computations," in *Proc. ICASSP*, 2022.