Spatial Loss for Unsupervised Multi-channel Source Separation

Kohei Saijo\textsuperscript{1,2\*}, Robin Scheibler\textsuperscript{2}

\textsuperscript{1}Waseda University, Japan \textsuperscript{2}LINE Corporation, Japan

\texttt{saijo@pcl.cs.waseda.ac.jp}

Abstract
We propose a \textit{spatial loss} for unsupervised multi-channel source separation. The proposed loss exploits the duality of direction of arrival (DOA) and beamforming: the steering and beamforming vectors should be aligned for the target source, but orthogonal for interfering ones. The spatial loss encourages consistency between the mixing and demixing systems from a classic DOA estimator and a neural separator, respectively. With the proposed loss, we train the neural separators based on minimum variance distortionless response (MVDR) beamforming and independent vector analysis (IVA). We also investigate the effectiveness of combining our spatial loss and a \textit{signal loss}, which uses the outputs of blind source separation as the references. We evaluate our proposed method on synthetic and recorded (LibriCSS) mixtures. We find that the spatial loss is most effective to train IVA-based separators. For the neural MVDR beamformer, it performs best when combined with a signal loss. On synthetic mixtures, the proposed unsupervised loss leads to the same performance as a supervised loss in terms of word error rate. On LibriCSS, we obtain close to state-of-the-art performance without any labeled training data.

\textbf{Index Terms:} unsupervised learning, spatial loss, direction-of-arrival, beamforming, blind source separation

1. Introduction
Source separation has been used as front-ends of speech systems to recover sources corrupted by interference and background noise. On the one hand, blind source separation (BSS) such as independent component analysis (ICA) \cite{Comon1994}, independent vector analysis (IVA) \cite{Cardoso1998, Amari1995}, and independent low-rank matrix analysis (ILRMA) \cite{Saito2016} have been an area of intense research. On the other hand, supervised learning of deep neural networks (DNNs) for single-channel source separation has been eagerly investigated \cite{Rozenberg2016, Gay2015, Dai2016}. In the multi-channel setup, methods that exploit a single-channel separation network to estimate spatial cues such as DNN-based minimum variance distortionless response (MVDR) beamforming \cite{Haykin2005, Brandstein2011, Sainath2013} or BSS with a neural source model \cite{Papandreou2017, Colantoni2019, Staworowsky2021} has led to stunning improvements in performance. Such linear filtering has also been shown empirically to be better front-ends for ASR than non-linear ones \cite{Monaural2016}. However, supervised learning requires access to the massive amount of mixtures and corresponding ground-truth signals. Because such a dataset of natural recordings cannot be obtained, many prior works rely on simulation instead \cite{Monaural2016, Dhurandhar2019}.

Recently, unsupervised source separation with a \textit{signal loss}, which uses the outputs of BSS as pseudo-targets instead of ground-truth clean signals, has been proposed \cite{Ganapathi2018, Thies2019}. In \cite{Ganapathi2018}, time-frequency (TF) masks estimated by a blind spatial clustering technique were used to train a deep clustering model. \cite{Thies2019} proposed a loss function that evaluates Kullback-Leibler Divergence (KLD) between the posterior probability density function of the separated signals of BSS and that of the DNN-based separator to avoid overfitting to the errors in the BSS outputs. These unsupervised losses enforce consistency between the output signals of BSS and those of DNN-based separators.

In contrast, we propose a \textit{spatial loss} function for unsupervised multi-channel source separation that enforces consistency of the estimated spatial parameters. We exploit the duality of direction of arrival (DOA) and beamforming: since linear separation relies on the assumption that sources are mixed linearly, the mixing and the beamforming matrix should be inverse of each other. We train a neural separator so that the beamforming vector for a given source should have a large inner product with the corresponding steering vector, while being close to orthogonal to those of the other sources. The steering vectors are obtained by conventional DOA estimation, the multiple signal classification (MUSIC) algorithm \cite{Schmidt1986}. We also investigate the effectiveness of combining the spatial loss and the signal loss \cite{Saijo2022a, Saijo2022b}. We train two types of neural separators with our proposed loss function, auxiliary function-based IVA (AuxIVA) \cite{Scheibler2018} with a neural source model \cite{Scheibler2019}, and neural MVDR beamformer \cite{Scheibler2019b}. We evaluate our proposed method with the recorded dataset LibriCSS \cite{LibriCSS2022}. Synthetic mixtures are also used for a more detailed analysis that requires the groundtruth.

The key contributions are summarized as follows. 1) We propose a spatial loss function for unsupervised multi-channel source separation. It gave the best performance for IVA, and improved the performance of MVDR when combined with the signal loss. 2) We conduct extensive experiments using both synthetic mixtures and real-world recordings, whereas most prior works only considered the former. The proposed loss leads to the same performance as a supervised loss in terms of word error rate on synthetic mixtures. On LibriCSS, the proposed method outperformed strong baselines \cite{Wang2022} and obtains close to state-of-the-art performance \cite{Wang2022} with no data besides that available in LibriCSS, nor any ground-truth.

2. Background
Assuming \( N \) sources are captured by \( M \) microphones, the observed signal in the short-time Fourier transform (STFT) domain is represented with a mixing matrix \( \mathbf{A}_f \in \mathbb{C}^{M \times N} \) as,

\[
\mathbf{x}_{f,t} = \mathbf{A}_f \mathbf{s}_{f,t} + \mathbf{b}_{f,t},
\]

where \( \mathbf{s}_{f,t} \) is the clean sources vector, \( \mathbf{b}_{f,t} \) is the background noise, and \( f = 1, \ldots, F \) and \( t = 1, \ldots, T \) are the frequency bin and the time frame index. Linear source separation is the problem of estimating a demixing matrix \( \mathbf{W}_f \in \mathbb{C}^{N \times M} \), that estimates the sources as,

\[
\mathbf{s}_{f,t} \approx \mathbf{W}_f \mathbf{x}_{f,t}.
\]

Thus, the optimal demixing matrix should satisfy \( \mathbf{W}_f \mathbf{A}_f \approx \mathbf{I} \), where \( \mathbf{I} \) is the identity matrix. In the following, \( \mathbf{I}^\dagger \) and \( \mathbf{I}^H \) denote the transpose and Hermitian transpose of vectors or matrices.
2.1. DOA Estimation

Knowing the location of each microphone, \( d_m \in \mathbb{R}^3 \), the steering vector of the \( n \)th source is,

\[
a_f(q_n) = M^{-\frac{1}{2}} \left[ e^{j\omega_f a_n^1}, \ldots, e^{j\omega_f a_n^M} \right]^\top,
\]

where \( \omega_f = 2\pi \frac{f}{c} \) with sound speed \( c \). \( q_n \in \mathbb{R}^3 \) is a unit-length vector pointing towards the \( n \)th source, and is represented with elevation \( \phi \) and azimuth \( \theta \) as \( q_n = [\cos \phi \sin \theta, \sin \phi \sin \theta, \cos \phi]^\top \).

Here we consider the MUSIC method [18]. MUSIC assumes that \( a_f(q_n) \) are orthogonal to the subspaces spanned by the \( M - N \) least eigenvectors of the covariance matrix of (1), i.e., the noise subspace. Let \( E_f \) be a matrix containing a basis for the noise subspace in its columns. Then, \( q_n \) is estimated by finding local maxima of the following cost function,

\[
G(q_n) = \frac{1}{F} \sum_f \left| (E_f^\top a_f(q_n))^2 \right|^{-1}.
\]

Hereafter, we denote \( a_f(q_n) \) as \( a_{n,f} \).

2.2. Source Separation

Here we mention two separation methods, AuxIVA and MVDR. AuxIVA estimates the demixing matrix by likelihood maximization assuming the prior distribution of sources \( p(y_{n,f,t}) \), where \( y_{n,f,t} \) is the \( n \)th separated signal. The demixing matrix is estimated by minimizing the following negative log-likelihood,

\[
J = -2T \sum_f \log |\det W_f| + \sum_{n,f,t} \frac{|w_{n,f}^H r_{n,f,t}|^2}{r_{n,f,t}},
\]

where \( w_{n,f} \) is the \( n \)th row vector of \( W_f \) and \( r_{n,f,t} \) the variance of the source prior \( p(y_{n,f,t}) \). \( W_f \) can be updated with the techniques such as iterative projection [19] or iterative source steering (ISS) [23]. While conventional IIVA uses a fixed source prior, replacing the source model with a trained DNN has recently demonstrated high performance [13].

MVDR computes the demixing vector that minimizes the noise variance under the distortionless constraint of the target source,

\[
w_{n,f} = \arg \min_{w_{n,f}} w_{n,f}^H R_{n,f}^{(n)} w_{n,f} \quad \text{s.t.} \quad w_{n,f}^H \hat{a}_{n,f} = 1,
\]

where \( R_{n,f}^{(n)} \) is the spatial covariance matrix of the noise. The steering vector \( \hat{a}_{n,f} \) and \( R_{n,f}^{(n)} \) can be obtained using DNNs [8].

2.3. Unsupervised Learning using Signal Loss

Recently, learning source separation using the separated signals from a BSS method as the pseudo-target signals has been proposed [16, 17]. Let \( \tilde{y} \) and \( \hat{y} \) be the separated signal of BSS and that of DNN-based separator. The training is simply done with the loss between the two signals, which we call signal loss, as

\[
L_{\text{sig}} = L(\tilde{y}, \hat{y}).
\]

3. Proposed Method

We propose a spatial loss function for unsupervised multi-channel source separation. We exploit the duality of DOA and beamforming. The beamforming vector for a given source should have a large inner product with the corresponding steering vector, while being close to orthogonal to those of competing sources. DOAs are estimated by a conventional MUSIC-based technique [24]. Different from the signal loss described in Sec.2.3, our proposed method explicitly imposes the spatial constraints on the demixing matrix.

3.1. Spatial Loss Function

We propose a spatial loss function that exploits DOAs estimated by MUSIC as the pseudo-targets. As shown in Fig.1, there is a duality between beamforming and DOA, i.e., the beam should be directed to the target source direction and the null to the interference direction. Let \( \hat{A} \) be the mixing matrix computed with the estimated DOAs and \( \hat{W} \) the demixing matrix estimated by the neural separator. Our proposed loss function is,

\[
L_{\text{dual}} = \min_{\Pi} \sum_{f,t} (\| \Pi - [\hat{W}_f \hat{A}_f] \|_F)^2
\]

where \( [\hat{W}_f \hat{A}_f] \) is normalized so that all the elements range from 0 to 1. \( \Pi \) is a vector whose all the elements are one and \( \Pi^\top \Pi \) computes the sum of all the elements of a matrix \( \Pi \). \( \Pi \) denotes a permutation matrix, where only one element in each row and each column is 1 and the others are 0. For example, if \( N = 2 \), \( \Pi \) is either \( \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \) or \( \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \). Here we consider two options of normalizing \([\hat{W}_f \hat{A}_f] \). One is to normalize rows and
columns of $|W_f|$ and $|A_f|$, respectively, and the other is to normalize the rows of $|W_f A_f|$. We denote the former as DOA$_1$ loss and the latter as DOA$_2$ loss, respectively.

### 3.2. DOA Estimator

We estimate DOAs using the MUSIC-based method [24]. When all the sources are highly active in a mixture, we simply obtain $q_n$ from the entire signal with length $T$. However, when some sources are not very active, we found out that such way of estimation often missed the DOAs of the less-active ones. In such cases, we obtain DOAs at regular intervals with sliding window clustering described in 3.2. We used different STFT parameters for WPE and separation. The window/shift size were set to 15.0 ms and 10.0 ms, respectively, and the other is to normalize the rows of $|W_f A_f|$. We denote the former as DOA$_1$ loss and the latter as DOA$_2$ loss, respectively.

#### 3.3. Neural Separator

As in Fig. 2, we first apply weighted prediction error (WPE) [27], and then the dereverberated mixture is separated by the neural separator. We consider two neural separators, AuxIVA with a neural source model (DNN-IVA) [13] and DNN-based MVDR beamforming (DNN-MVDR) [8–10]. In DNN-IVA, the DNN estimates $1/r_{n,f,t}$ in (5) in the form of a TF mask, and spatial model update is done with IIS [23]. The network is composed of three gate linear units [28] and a transpose convolution layer as in [13]. From [13], the size of the intermediate feature is changed to 256, and group normalization [29] with four groups is used instead of batch normalization. In DNN-MVDR, the DNN estimates TF masks of the target source and the noise for each source, which are used to compute the demixing matrix. The network consists of three bi-directional long short-term memory layers and two feed forward layers as in [10], where we changed the number of units to 512.

### 4. Experiments

#### 4.1. Datasets and Experimental Setup

**WSJ-mix:** We used six channel synthetic mixtures to evaluate the separation performance with speech metrics using the groundtruth. It consisted of speech from the WSJ0 [30] and WSJ1 [31], and noise from the CHIME3 dataset [32] sampled in 16 kHz. The reverberation times were chosen randomly from 200 ms to 600 ms. The number of sources was two, with relative power from –5 dB to 5 dB. The noise was scaled to attain an SNR between 10 dB to 30 dB. Training, validation, and test sets contained 37416, 503, and 333 mixtures, approximately 98.5, 1.33, and 0.5 hours of mixtures, respectively. To evaluate WER, we trained an ASR system with clean anechoic signals using the WSJ/AR1 recipe from the ESPNet framework [33]. WER for the clean, anechoic test set was 9.25%.

During training, the batch size was 16 and the input signal length was 7 seconds. Network parameters were optimized using the Adam optimizer [34] with learning rate $10^{-5}$. When combining signal and spatial loss, we took weighted sum of them, i.e., $L_{\text{loss}} = \alpha L_{\text{sp}} + L_{\text{sig}}$, where $\alpha$ was set to 0.2 for DNN-IVA and 1.0 for DNN-MVDR. When training DNN-IVA, three of six channels were used and the number of iterations was 15. DNN-MVDR was trained using all six channels. MUSIC and AuxIVA also used six channels. We assumed time-varying complex Gaussian prior for AuxIVA (Gauss-IVA). When using WSJ-mix, we obtain DOAs from the entire signal, not using clustering described in 3.2. We used different STFT parameters for WPE and separation. The window/shift size were set to 512/128 and 4096/1024, respectively. The number of iterations, the delay and the tap length of WPE were set to 3, 3 and 10.
In test of IVA, we evaluated the performance using two, three and six channels with 30, 25 and 15 iterations. When using more than two channels, two separated signals with the highest power were evaluated. For testing, we used the averaged model parameters of the five epochs with lowest unsupervised loss on the validation set.

**LibriCSS:** We used LibriCSS [20] to evaluate the performance on real-recorded speeches. Utterances taken from LibriSpeech test clean set were played back from loudspeakers and recorded by a seven-channel microphone with radius of 4.25 cm. LibriCSS had 10 hours of audio recordings and it was separated into 10 sessions. Each session contained six mini-sessions with different overlap ratios ranging from 0 to 40%, 0S, 0L, 10, 20, 30, and 40, where 0S/0L are overlap-free recordings with short/long inter-utterance silence. We conducted utterance-wise evaluation by evaluating WERs using the ASR systems given in [20]. We used sessions 1 to 9 as training set, and session 0 as test set.

Experimental setup was basically the same as that on WSJ-mix, but we used only three channels for Gauss-IVA and DNN-IVA to avoid degradation of WER due to over-separation. Note that MUSIC used all seven channels to estimate DOAs, where we used a seven-channel microphone with radius of 4.25 cm. LibriCSS used 10 hours of audio recordings and they were separated into 10 sessions. Each session contained six mini-sessions with different overlap ratios ranging from 0 to 40%, 0S, 0L, 10, 20, 30, and 40, where 0S/0L are overlap-free recordings with short/long inter-utterance silence. We conducted utterance-wise evaluation by evaluating WERs using the ASR systems given in [20]. We used sessions 1 to 9 as training set, and session 0 as test set.

4.3. Results on LibriCSS dataset

Table 3 shows the WERs of separated signals from LibriCSS dataset. WERs per overlap ratio and average of them are listed. Compared to the signals losses, our proposed spatial loss led to higher performance. Both SDR and KLD loss gave only the comparable performance to the supervised loss. In addition, our proposed method also outperformed Conformer-based MVDR beamforming [21] with less than one twentieth the amount of data. Furthermore, the performance was further improved by fine-tuning the supervised trained model on WSJ-mix to LibriCSS using the proposed unsupervised learning. Although our proposed method did not exceed state-of-the-art performance [22], we showed high performance with the small amount of data, which demonstrates the effectiveness of unsupervised learning in-domain data.

5. Conclusions

We proposed a spatial loss function that utilizes the DOAs estimated by classic techniques. It trains a neural separator so that it directs beam to the target source direction and null to the interference direction. We evaluated two neural separators, DNN-IVA and DNN-MVDR, using synthetic mixtures and real-recorded LibriCSS. In experiments with synthetic mixtures, we showed that the spatial loss worked well on both DNN-IVA and DNN-MVDR, which led to the same performance as the supervised loss. It also performed well for DNN-MVDR when combined with the signal loss. The spatial loss also gave high performance on LibriCSS. It outperformed the strong baselines with the small amount of training data without any labels.
6. References


