Interspeech 2022
18-22 September 2022, Incheon, Korea

Independence-based Joint Dereverberation and Separation with Neural Source Model

Kohei Saijo[1,2*], *Robin Scheibler*[2]

[1]Waseda University, Japan [2]LINE Corporation, Japan

saijo@pcl.cs.waseda.ac.jp

Abstract

We propose an independence-based joint dereverberation and separation method with a neural source model. We introduce a neural network in the framework of time-decorrelation iterative source steering, which is an extension of independent vector analysis to joint dereverberation and separation. The network is trained in an end-to-end manner with a permutation invariant loss on the time-domain separation output signals. Our proposed method can be applied in any situation with at least as many microphones as sources, regardless of their number. In experiments, we demonstrate that our method results in high performance in terms of both speech quality metrics and word error rate (WER), even for mixtures with a different number of speakers than training. Furthermore, the model, trained on synthetic mixtures, without any modifications, greatly reduces the WER on the recorded dataset LibriCSS.

Index Terms: source separation, dereverberation, memory-efficient gradient computation, deep neural network

1. Introduction

Source separation and dereverberation have been studied as pre-processing for speech systems such as automatic speech recognition (ASR) since speech recordings are generally contaminated by interference, background noise, and reverberation. On the one hand, blind source separation (BSS) such as independent component analysis (ICA) [1], independent vector analysis (IVA) [2, 3], independent low-rank matrix analysis (ILRMA) [4], dereverberation techniques such as weighted prediction error (WPE) [5], and their joint optimization such as ILRMA-T [6] have been an area of intense research. On the other hand, supervised learning of deep neural networks (DNN) has been proposed for the single-channel source separation [7–9]. It has also been extended to the multi-channel setup, where the time-frequency (TF) masks or separated signals estimated by the single-channel network are used to obtain spatial statistics for beamforming [10–12]. Such linear filtering techniques have empirically been shown to be better as a front-end to ASR than non-linear separation approaches [13].

We focus on linear dereverberation and separation techniques. Among them, the cascade connection of dereverberation and separation such as WPE [5] and neural beamforming [10–12] or their joint otimization [14] have shown good performance in both speech metrics and WER. However, most of them have two drawbacks. First, they are sensitive to domain mismatch, because they highly depend on the front-end mask estimation network. Second, the number of sources to be separated is fixed. Most methods prepare an output layer for the number of sources, which is practically inconvenient because the number of sources is usually unknown in advance.

*This work was done during an internship at LINE Corporation.

BSS with a neural source model [15–17] is one of the methods that work regardless of the number of sources. In [15, 16], a pre-trained network is utilized as the source model. End-to-end training of the source model of IVA was also proposed in [17], where stable training was enabled thanks to the inverse-free spatial model update for IVA, i.e., iterative source steering (ISS) [18]. However, [17] tackled only separation, not dereverberation. Recently, an extension of ISS to joint dereverberation and separation, time-decorrelation ISS (T-ISS) [19] has been proposed. Still, in [19], the conventional non-negative low-rank model [4, 6], which has no trainable parameters, was used as source model, limiting the performance.

In this work, we propose to replace the non-trainable source model in the original T-ISS with a trainable neural network. We train this neural source model in an end-to-end manner with a permutation invariant loss [12] on the time-domain separation output signals. Although end-to-end training of iterative methods requires memory proportional to the number of iterations, we introduce a memory-efficient gradient computation technique, *demixing matrix checkpointing* (DMC), a derivative of checkpointing technique [20] which makes the memory cost nearly constant regardless of the number of iterations. Our proposed method is robust to domain mismatch because the trained model can be applied whenever there at least as many microphones as sources, regardless of their number. While the neural source model extracts a single source from a single channel input, the parameter-free spatial model update of T-ISS handles the difference in the number of microphones or sources.

The key contributions are summarized as follows. 1) This is the first work to train a neural source model for T-ISS. We show the superiority of our proposed method to T-ISS with conventional models [19] and cascade connection of dereverberation [5] and separation [12, 17], regardless of the number of speakers. 2) We demonstrate that the proposed method trained using synthetic mixtures with no overlap-free data also works well on the low-overlap recordings of LibriCSS [21] without any modification. Our proposed method outperformed a strong Conformer-based MVDR beamforming approach [22].

2. Joint Dereverberation and Separation

Assuming N sources are captured by M microphones, the microphone input signal in the short-time Fourier transform (STFT) domain is modeled auto-regressively, namely,

$$\boldsymbol{x}_{f,t} = \boldsymbol{A}_f \boldsymbol{s}_{f,t} + \sum_{l=D+1}^{D+L} \boldsymbol{Z}_{f,l} \boldsymbol{x}_{f,t-l}, \tag{1}$$

where $\boldsymbol{s}_{f,t}$ is the clean sources vector, $\boldsymbol{A}_f \in \mathbb{C}^{M\times N}$ the mixing matrix, and $\boldsymbol{Z}_{f,l}$ produces the reverberation component from past samples. L is the tap length and D is the delay to separate the direct signal and reverberation. $f = 1, \ldots, F$ and

Copyright © 2022 ISCA

10.21437/Interspeech.2022-271

$t = 1, \ldots, T$ are the frequency bin and the time frame index in STFT-domain. We can rewrite (1) as

$$\boldsymbol{x}_{f,t} = \boldsymbol{A}_f \boldsymbol{s}_{f,t} + \bar{\boldsymbol{Z}}_f \bar{\boldsymbol{x}}_{f,t}, \tag{2}$$

where $\bar{\boldsymbol{Z}}_f = [\boldsymbol{Z}_{f,D+1}, \ldots, \boldsymbol{Z}_{f,D+L}] \in \mathbb{C}^{M \times ML}$ and $\bar{\boldsymbol{x}}_{f,t} = [\boldsymbol{x}_{f,t-D-1}^\top, \ldots, \boldsymbol{x}_{f,t-D-L}^\top]^\top \in \mathbb{C}^{ML}$. Supposing we know $\boldsymbol{W}_f = \boldsymbol{A}_f^{-1}$, i.e., the demixing matrix, and $\bar{\boldsymbol{Z}}_f$, we recover

$$\boldsymbol{s}_{f,t} = \boldsymbol{W}_f (\boldsymbol{x}_{f,t} - \bar{\boldsymbol{Z}}_f \bar{\boldsymbol{x}}_{f,t}). \tag{3}$$

Here we assume the determined case, i.e., $M = N$. In the following, $\boldsymbol{I}$ and $\boldsymbol{e}_n$ denote the identity matrix and the nth canonical basis vector, * the complex conjugate, and $^\top$ and $^\mathsf{H}$ the transpose and Hermitian transpose of vectors or matrices.

Since conventional dereverberation methods such as WPE [5] assume the existence of only one source, the cascade connection of dereverberation and separation is not optimal. To circumvent this problem, a *joint* dereverberation and separation framework was introduced as part of ILRMA-T [6], where the estimated clean sources vector $\boldsymbol{y}_{f,t}$ is obtained with a unified filter $\boldsymbol{P}_f = \boldsymbol{W}_f[\boldsymbol{I}, -\bar{\boldsymbol{Z}}_f] \in \mathbb{C}^{M \times M(L+1)}$ as $\boldsymbol{y}_{f,t} = \boldsymbol{P}_f \tilde{\boldsymbol{x}}_{f,t}$ where $\tilde{\boldsymbol{x}}_{f,t} = [\boldsymbol{x}_{f,t}^\top, \bar{\boldsymbol{x}}_{f,t}^\top]^\top \in \mathbb{C}^{M(L+1)}$. Since both $\boldsymbol{y}_{f,t}$ and $\boldsymbol{P}_f$ are unknown, we solve it by maximum likelihood with a generative model of the sources. ILRMA-T assumes that each TF bin of the nth source, $y_{n,f,t}$, belongs to a complex Gaussian distribution, $p(y_{n,f,t}) = \frac{1}{\pi r_{n,f,t}} \exp\left(-\frac{|y_{n,f,t}|^2}{r_{n,f,t}}\right)$, where $r_{n,f,t}$ is the time-varying variance of the nth source and modeled as low-rank non-negative [4]. Assuming each source is independent, we obtain the following negative log-likelihood [6],

$$\mathcal{J} = \sum_{f,t} \left[-2 \log |\det \boldsymbol{W}_f| + \frac{1}{T} \sum_n \frac{|\boldsymbol{p}_{n,f}^\mathsf{H} \tilde{\boldsymbol{x}}_{f,t}|^2}{r_{n,f,t}} \right], \tag{4}$$

where $\boldsymbol{p}_{n,f}^\mathsf{H}$ is the nth row vector of $\boldsymbol{P}_f$. For the sake of generality, we hereafter refer to the reciprocal of the time-varying variance as the function of $\boldsymbol{y}$, i.e., $u_{f,t}(\boldsymbol{Y}_n) = 1/r_{n,f,t}$, where $(\boldsymbol{Y}_n)_{f,t} = (\boldsymbol{y}_{f,t})_n$. The unified filter $\boldsymbol{P}_f$ can be estimated by iteratively minimizing (4). To avoid large matrix inversions in the update rule derived from iterative projection algorithm [6,23], the inverse-free rank-1 update rule, T-ISS has been proposed [18, 19]. For $1 \leq n \leq M$, the update is done as,

$$\boldsymbol{P}_f \leftarrow \boldsymbol{P}_f - \boldsymbol{v}_{n,f} \boldsymbol{p}_{n,f}^\mathsf{H}, \tag{5}$$

where $\boldsymbol{v}_{n,f} = [v_{1,n,f}, \ldots, v_{M,n,f}]^\top$ is chosen to minimize (4), which yields,

$$v_{m,n,f} = \begin{cases} 1 - \left(\frac{1}{T} \sum_t u_{f,t}(\boldsymbol{Y}_n) |y_{n,f,t}|^2\right)^{-\frac{1}{2}}, & \text{if } m = n, \\ \frac{\sum_t u_{f,t}(\boldsymbol{Y}_m) y_{m,f,t} y_{n,f,t}^*}{\sum_t u_{f,t}(\boldsymbol{Y}_m) |y_{n,f,t}|^2}, & \text{otherwise.} \end{cases} \tag{6}$$

For $n > M$, the update is $\boldsymbol{P}_f \leftarrow \boldsymbol{P}_f - \boldsymbol{v}_{n,f} \boldsymbol{e}_{n,f}^\top$ and $v_{m,n,f}$ is,

$$v_{m,n,f} = \frac{\sum_t u_{f,t}(\boldsymbol{Y}_m) y_{m,f,t} \tilde{x}_{n,f,t}^*}{\sum_t u_{f,t}(\boldsymbol{Y}_m) |\tilde{x}_{n,f,t}|^2}, \tag{7}$$

where $\tilde{x}_{n,f,t}$ is the nth component of $\tilde{\boldsymbol{x}}_{f,t}$.

3. Proposed Source Model Learning

Conventionally, the source mask $u_{f,t}(\boldsymbol{Y}_n)$ is derived from a probabilistic model of speech such as low-rank non-negative model [6]. Instead, we propose to replace it with a neural network and learn its weights by gradient descent in the independence-based joint dereverberation and separation framework of T-ISS [19]. The overview of our proposed method

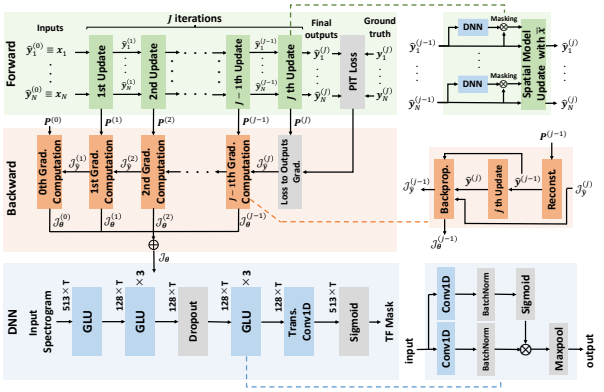


Figure 1: *Overview of our proposed method. We train a source model DNN (blue) directly through J iterations with a memory-efficient gradient computation. Unified filter $\boldsymbol{P}^{(j)}$ is stored in the forward pass (green) and used to reconstruct the separated signals in the backward pass (orange). The gradient to update the DNN parameter $\mathcal{J}_{\boldsymbol{\theta}}$ is simply obtained by summing up the gradient at each iteration.*

is shown in Fig. 1. We iterate estimating $u_{f,t}(\boldsymbol{Y}_n)$ and updating spatial model with (5) for multiple times, and compute loss between the ground truth clean signals and the final separation outputs of multiple iterations (top of Fig. 1). To reduce memory cost to train through multiple iterations, we introduce a memory-efficient gradient computation technique, DMC, as an alternative to backpropagation (BP) (middle of Fig. 1). Since the neural source model extracts a single source from a single channel input and the spatial model update of T-ISS has no trainable parameters, our proposed method can be applied to mixtures with any number of microphones or sources within an (over-) determined case and is robust to domain mismatch.

3.1. Network Architecture

The source model network architecture is shown at the bottom of Fig. 1. It consists of multiple gated linear units (GLU) [24] and a transpose convolution layer, all with kernel size three. It has around 2.2M trainable parameters in total. The input is down-sampled in the first block and then fed into six GLU blocks with a dropout layer with a probability of 0.5 between the third and the fourth. Finally, the transpose convolution layer up-samples the intermediate feature to the same size as the original input, and a TF mask is output thorough the sigmoid activation. The mask may be understood as hiding the target source in the current source spectrogram estimate.

3.2. Loss Function

To train the neural source model, we use convolutive transfer function invariant [25] signal-to-distortion ratio (CI-SDR) loss, which was shown to be effective for multi-channel enhancement [12]. In CI-SDR, a K tap filter $\boldsymbol{\alpha} \in \mathbb{R}^K$, where K is usually set to 512, is used to make the SDR short-impulse-response-invariant. Let the time-domain estimated signal and the ground truth be $\hat{\boldsymbol{s}}$ and $\boldsymbol{s} \in \mathbb{R}^I$, respectively. Further define a matrix containing K shifts of $\boldsymbol{s}$ in its columns, i.e., $\bar{\boldsymbol{S}} = [\bar{\boldsymbol{s}}_1, \ldots, \bar{\boldsymbol{s}}_I]^\top \in \mathbb{R}^{I \times K}$, with ith row $\bar{\boldsymbol{s}}_i = [s_{i-1}, \ldots, s_{i-K}]^\top \in \mathbb{R}^K$, where $i = 1, \ldots, I$ is the time index. Then, let $\boldsymbol{\alpha} = (\bar{\boldsymbol{S}}^\top \bar{\boldsymbol{S}})^{-1} \bar{\boldsymbol{S}}^\top \hat{\boldsymbol{s}}$, the CI-SDR loss is,

$$\mathcal{L}_{\text{CI-SDR}} = -10 \log_{10} \left(\frac{||\bar{\boldsymbol{S}} \boldsymbol{\alpha}||^2}{||\bar{\boldsymbol{S}} \boldsymbol{\alpha} - \hat{\boldsymbol{s}}||^2} \right). \tag{8}$$

3.3. Demixing Matrix Checkpointing

BP starts from the final loss and works its way backward through the computational graph computing the gradient along the way. Thus, BP needs all the intermediate results and memory cost becomes a bottleneck for iterative framework. However, by storing the demixing matrices during the forward pass, we can recreate the intermediate separation results. Based on this idea, we introduce a memory-efficient gradient computation technique, demixing matrix checkpointing (DMC). The backward pass of DMC is shown in Fig.1. DMC proceeds by computing the contribution of each iteration to the gradient in reverse order. At jth iteration, we first reconstruct the separated signals of the previous iteration $\hat{\boldsymbol{y}}^{(j-1)}$ using $\boldsymbol{P}^{(j-1)}$ stored during the forward pass. Then, we re-compute $\hat{\boldsymbol{y}}^{(j)}$ in the same way as the forward pass. The gradient of the model parameter $\boldsymbol{\theta}$ at jth iteration, $\mathcal{J}_{\boldsymbol{\theta}}^{(j-1)}$, is computed by BP. Since the DNN parameters are shared in all the iterations, the gradient to update the DNN parameters is obtained by summing up $\mathcal{J}_{\boldsymbol{\theta}}^{(j)}$ for all j.

4. Experiments

Experimental comparisons were conducted to verify the effectiveness of our proposed method. The baselines were as follows. **WPE + DNN-MVDR** [5, 12]: Cascade connection of WPE and neural MVDR beamforming. Here, we used the BLSTM-based network from [12] rather than the one shown in Fig. 1. **WPE + DNN-IVA** [5, 17]: Cascade connection of WPE and IVA with a neural source model. **ILRMA-T-ISS** [19]: T-ISS with low-rank non-negative source model [4]. It had no trainable parameters and performs joint dereverberation and separation blindly.

All the algorithms were implemented in Pytorch [26]. Allocated memory and computational time were measured with the nvidia-smi command and the profiler of Pytorch. Experiments were conducted on a Linux workstation with an Intel® Xeon® Gold 6230 CPU 2.10 GHz with 8 cores, an NVIDIA® Tesla® V100 graphical processing unit (GPU), and 64 GB of RAM.

4.1. Datasets and Experimental Setup

WSJ-mix: We used the synthetic reverberant noisy mixtures introduced in [17]. It consisted of speech from the WSJ0 [27] and WSJ1 [28] datasets, and noise from the CHIME3 dataset [29] sampled in 16 kHz. The reverberation times were chosen from 200 ms to 600 ms. The relative power of sources was chosen 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 0.85 hours of mixtures, respectively. We created two, three, and four channels mixtures with two, three and four speakers, respectively. Only the two-speaker mixtures were used for training, and two, three and four-speaker mixtures were used for validation and test. Note that it contained no overlap-free samples, i.e., all the mixtures consisted of multiple speeches. To evaluate word error rate (WER), we trained an ASR system with clean anechoic signals using the wsj/asr1 recipe from the ESPNet framework [30]. WER for the clean, anechoic test set was 9.25 %. During training, the batch size was eight and the input signal length was seven seconds. The network parameters were optimized using the Adam optimizer [31] with the learning rate of 1.0×10^{-4}. For STFT, we used the Hann window with the size of 1024, and hop size was 256. The number of iterations of T-ISS and IVA were set to 20 in training and 50, 75 and 100 for two, three and four-speaker mixtures in test. The number of iterations for WPE was 3 both in training and test. The delay was set to 1 in T-ISS and 3 in WPE because such setting gave the best performance. Both used $L = 5$ taps. We applied projection back [32] to the final output.

Table 1: *Average SDR in decibels, STOI, PESQ, SRMR and WER of separated signals from WSJ-mix test set. The number of sources equals that of channels. Training was done using only 2ch. data. The proposed method is indicated by a star ($\star$).*

Ch.	Algo.	SDR$^{\uparrow}$	STOI$^{\uparrow}$	PESQ$^{\uparrow}$	SRMR$^{\uparrow}$	WER$^{\downarrow}$
2	Unprocessed	-0.4	0.728	1.21	5.13	111.4%
	WPE+DNN-MVDR	10.0	0.875	1.63	6.66	54.2%
	WPE+DNN-IVA	9.8	0.880	1.60	6.57	53.2%
	ILRMA-T-ISS	7.9	0.860	1.55	6.45	49.3%
	DNN-T-ISS*	**12.3**	**0.907**	**1.78**	**6.87**	**29.2%**
3	Unprocessed	-3.6	0.640	1.14	4.58	145.9%
	WPE+DNN-IVA	7.9	0.850	1.47	6.41	72.3%
	ILRMA-T-ISS	4.4	0.799	1.37	6.16	77.9%
	DNN-T-ISS*	**9.7**	**0.879**	**1.62**	**6.65**	**43.8%**
4	Unprocessed	-5.4	0.584	1.12	4.12	162.8%
	WPE+DNN-IVA	6.6	0.828	1.40	6.25	87.0%
	ILRMA-T-ISS	2.0	0.755	1.28	5.81	98.1%
	DNN-T-ISS*	**7.7**	**0.849**	**1.50**	**6.40**	**57.3%**

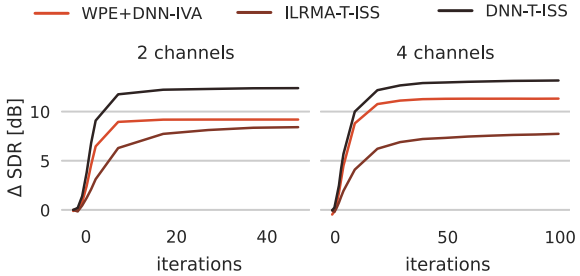


Figure 2: *Convergence of SDR improvement as a function of the number of the iterations on WSJ-mix test set.*

LibriCSS: We used LibriCSS dataset [21] to evaluate the performance of the models trained with WSJ-mix on out-of-domain real-world recorded speeches. Utterances taken from LibriSpeech test clean set were played back from loudspeakers placed in a room, and recorded by a seven channel circular microphones with radius of 4.25 cm. The distances from the loudspeakers and the microphones ranged from 33 cm to 409 cm. Separation performance were evaluated by WERs of separated signals using the ASR systems provided in [21]. LibriCSS contained data whose average overlap ratio ranges from 0 to 40%, 0S, 0L, 10, 20, 30 and 40% overlaps, where 0S/0L were overlap-free recordings with short/long inter-utterance silence. We conducted *utterance-wise* evaluation using two, three and seven channels of microphones with 50, 30 and 15 iterations, respectively. Because basically one or two speech sources were in an observation, using three or more channels corresponded to the over-determined condition. When over-determined case, two separated signals with the highest power were evaluated.

4.2. Results on WSJ-mix dataset

Table 1 shows the evaluation results on the WSJ-mix test set. The evaluation metrics were SDR, the short-time objective intelligibility (STOI) [33], the perceptual evaluation of speech quality (PESQ) [34], the speech-to-reverberation modulation energy ratio (SRMR) [35], and WER. Since DNN-MVDR estimated two signals, it was evaluated with only two-speaker mixtures. For two-speaker mixtures, our proposed method, DNN-

Table 2: *WERs evaluated with LibriCSS dataset. Training is done with two-speaker mixtures from WSJ-mix dataet. 0S/0L are 0% overlap case with short/long inter-utterance silence.*

Ch.	Algo.	WER per overlap ratio in %						
		0S	0L	10.0	20.0	30.0	40.0	Avg
	Unprocess. [21]	11.8	11.7	18.8	27.2	35.6	43.3	26.4
7	Chen *et al.* [22]	7.2	7.5	9.6	11.3	13.7	15.1	11.2
	Wang *et al.* [36]	5.8	5.8	5.9	6.5	7.7	8.3	6.8
2	WPE+DNN-MVDR	8.8	8.9	12.9	18.6	23.6	29.0	18.0
	WPE+DNN-IVA	7.9	8.1	12.3	17.2	21.8	27.1	16.7
	ILRMA-T-ISS	7.5	7.7	12.2	17.1	22.2	26.5	16.6
	DNN-T-ISS*	**7.3**	**7.5**	**10.6**	**14.3**	**18.0**	**21.2**	**13.9**
3	WPE+DNN-MVDR	10.0	10.0	11.9	14.8	18.6	21.4	15.1
	WPE+DNN-IVA	8.0	8.0	9.9	13.0	15.8	18.0	12.7
	ILRMA-T-ISS	6.6	6.6	9.2	13.3	17.0	20.2	12.9
	DNN-T-ISS*	**6.5**	**6.3**	**7.6**	**10.8**	**13.5**	**13.9**	**10.2**
7	WPE+DNN-MVDR	11.6	12.5	13.6	16.4	19.5	21.0	16.2
	WPE+DNN-IVA	8.7	8.9	10.5	12.8	15.9	17.4	12.8
	ILRMA-T-ISS	**6.6**	6.9	**8.5**	12.9	15.4	18.1	12.0
	DNN-T-ISS*	6.9	**6.6**	8.9	**10.9**	**13.6**	**14.1**	**10.6**

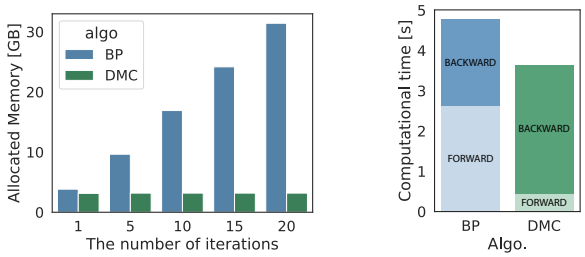
T-ISS (marked with a $\star$), outperformed all the baselines on all the metrics. The significant performance gap between ILRMA-T-ISS and DNN-T-ISS shows the effectiveness of learning the source model. Also, our proposed method greatly outperformed the cascade approaches, which confirms the effectiveness of joint optimization of dereverberation and separation. For three and four-speaker mixtures, our proposed method also gave the best performance. In particular, our proposed method led to a significant improvement in WER, over 20%, 30% and 40% reduction for two, three and four-speakers mixtures compared to the next best method, respectively. This confirms that our proposed method can straightforwardly be applied to mixtures with different number of sources than during training.

Fig.2 shows the convergence of SDR improvement as a function of the number of iterations on two and four-speaker mixtures. Note that iterations of WPE is not included in Fig.2. For both case, the neural source model led to faster convergence than the conventional source model. In particular, in two-speaker case, DNN-T-ISS converged in 20 iterations, whereas ILRMA-T-ISS took approximately 40 iterations.

4.3. Results on LibriCSS dataset

Table 2 shows the evaluation results on the LibriCSS. Focusing on the separation with two channels, i.e., the same condition as training, our proposed method achieved the best performance. For all overlap ratios, WPE+DNN-MVDR performed poorly compared to other baselines. This may be due to the configuration of the DNN-MVDR such that it outputs two masks at the same time and *over-separates* the sources. In contrast, our proposed method, which learns a source prior for a *single* source, worked well, although training set did not contain overlap-free data. This result gave a new insight that our proposed method can also be straightforwardly applied to low-overlap mixtures including overlap-free cases. Compared to two channels case, using more channels, i.e., over-determined separation, significantly boosted the performance. Using three channels led to as much as 3.7% reduction of WER in our proposed method. Although using all seven channels did not improve the performance over three channels case due to a slight over-separation, it still achieved much better performance than when using only two channels.

Our proposed method with 2.2M parameters is also compared with several prior works, Conformer-based MVDR beamforming with 58.7M parameters [22], and deep convolutional



(a) Allocated memory (b) Comp. time

Figure 3: *(a) Allocated memory of BP and DMC. (b) Computational time for forward and backward pass, where batch size is 8 and the number of iterations is 20.*

network-based MVDR beamforming followed by another enhancement network with around 13.8M parameters [36]. Note that the training set in [36] were simulated assuming the knowledge of the target domain, e.g., the microphone position was the same as LibriCSS and the average overlap ratio was relatively low, about 33%. Compared to the Conformer-based approach [22], which was trained with more than twice as much data as ours, the proposed method achieved better performance for all overlap ratios with much smaller number of parameters. Although the proposed method did not reach the state-of-the-art [36], it showed promising performance without any knowledge of the target domain and with fewer parameters and microphones.

4.4. Effectiveness of DMC

Fig. 3 shows the memory cost as a function of the number of iterations (left), and runtimes of the forward and backward passes per batch of 8 samples (right). Under the setting of our experiment, DMC reduced the memory cost from 31 GB to only 3 GB. In addition, surprisingly, even though the backward of DMC is obviously slower, DMC was found to be faster than BP in the total of the forward and the backward passes. Although we set the training parameters that allowed for learning without DMC, there are several advantages using DMC. To begin with, it enables training with limited resources without decreasing the number of iterations or input signal length. In preliminary experiments, we found out that reducing them makes training unstable. In addition, the batch size can be enlarged, which could also contribute to stable training. We investigate the performance with larger batch size in future work.

5. Conclusion

We proposed an independence-based joint dereverberation and separation method with a neural source model. The source model was trained in an end-to-end manner directly with permutation invariant loss at the final output in the framework of T-ISS. We also introduced DMC to reduce the training memory cost of the iterative approach. We analyzed our proposed method from a variety of perspectives using in-domain synthetic mixtures and out-of-domain low-overlap real-world recordings. In experiments, we showed that our proposed method worked better than the conventional model and the cascade connection of dereverberation and separation. High performance was achieved regardless of the number of speakers or channels, including in the single-speaker case. Furthermore, we demonstrated that our proposed method trained with synthetic mixtures also performed well on real-world recordings without any modification.

6. References

[1] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[2] T. Kim, T. Eltoft, and T.-W. Lee, "Independent vector analysis: An extension of ica to multivariate components," in *International conference on independent component analysis and signal separation*. Springer, 2006, pp. 165–172.

[3] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2006, pp. 601–608.

[4] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1626–1641, 2016.

[5] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[6] R. Ikeshita, N. Ito, T. Nakatani, and H. Sawada, "A unifying framework for blind source separation based on a joint diagonalizability constraint," in *Proc. IEEE EUSIPCO*, 2019, pp. 1–5.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE ICASSP*, 2016, pp. 31–35.

[8] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE ICASSP*, 2017, pp. 241–245.

[9] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE ICASSP*, 2018, pp. 696–700.

[10] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE ICASSP*, 2016, pp. 196–200.

[11] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *Proc. ICML*, 2017, p. 2632–2641.

[12] C. Boeddeker et al., "Convolutive transfer function invariant SDR training criteria for multi-channel reverberant speech separation," in *Proc. IEEE ICASSP*, 2021, pp. 8428–8432.

[13] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2020.

[14] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Blind and neural network-guided convolutional beamformer for joint denoising, dereverberation, and source separation," in *Proc. IEEE ICASSP*, 2021, pp. 6129–6133.

[15] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[16] N. Makishima et al., "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 10, pp. 1601–1615, 2019.

[17] R. Scheibler and M. Togami, "Surrogate source model learning for determined source separation," in *Proc. IEEE ICASSP*, 2021, pp. 176–180.

[18] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. IEEE ICASSP*, 2020, pp. 236–240.

[19] T. Nakashima, R. Scheibler, M. Togami, and N. Ono, "Joint dereverberation and separation with iterative source steering," in *Proc. IEEE ICASSP*, 2021, pp. 216–220.

[20] J. Martens and I. Sutskever, "Training deep and recurrent networks with hessian-free optimization," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 479–535.

[21] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *Proc. IEEE ICASSP*, 2020, pp. 7284–7288.

[22] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *Proc. IEEE ICASSP*, 2021, pp. 5749–5753.

[23] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. IEEE WASPAA*, 2011, pp. 189–192.

[24] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, Aug 2017, pp. 933–941.

[25] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jun. 2006.

[26] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[27] J. S. Garofolo *et al.*, *CSR-I (WSJ0) Complete LDC93S6A*, Linguistic Data Consortium, Philadelphia, 1993, web Download.

[28] Linguistic Data Consortium, and NIST Multimodal Information Group, *CSR-II (WSJ1) Complete LDC94S13A*, Linguistic Data Consortium, Philadelphia, 1994, web Download.

[29] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE ASRU*, 2015, pp. 504–511.

[30] S. Watanabe et al., "ESPnet: End-to-end speech processing toolkit," in *Proc. ISCA INTERSPEECH*, 2018, pp. 2207–2211.

[31] P. K. Diederik and B. Jimmy, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[32] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1, pp. 1–24, 2001.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[34] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE ICASSP*, vol. 2, 2001, pp. 749—752.

[35] T. H. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[36] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2001–2014, 2021.