Beam-Guided TasNet: An Iterative Speech Separation Framework with Multi-Channel Output

Hangting Chen^{1,2}, *Yang Yi*^{1,2}, *Dang Feng*^{1,2} *and Pengyuan Zhang*^{1,2}

^{1}Key Laboratory of Speech Acoustics & Content Understanding, Institute of Acoustics, CAS, China
^{2}University of Chinese Academy of Sciences, Beijing, China

{chenhangting,yangyi,dangfeng,zhangpengyuan}@hccl.ioa.ac.cn

Appendix

A. Frame-by-frame processing

For online frame-by-frame processing, the permutation solver calculates metrics based on the received signal to conduct source reorder in a frame-by-frame method. In our practice, the distance measurement methods, such as Euclidean norm and correlation, can achieve similar performance. Here we use SNR to reorder the sources, which corresponds to Euclidean norm. The causal permutation solver obtains the order $\hat{\pi}_{c,t}$, which can be expressed as,

$$\hat{\pi}_{c,t} = \underset{\pi_{c,t}}{\mathrm{argmax}} \sum_{s=1}^{S} \mathrm{SNR}\left(\hat{x}_{s,1}[0:n_t], \hat{x}_{\pi_{c,t}(s),c}[0:n_t]\right), \tag{1}$$

where n_t denotes the number of received samples until frame t. The SCMs are updated as the followings:

$$\hat{\mathbf{\Phi}}^{\mathrm{Target}_s}_{t,f} = \frac{t-1}{t}\hat{\mathbf{\Phi}}^{\mathrm{Target}_s}_{t-1,f} + \frac{1}{t}\hat{\mathbf{Z}}_{s,t,f}\hat{\mathbf{Z}}^{\mathrm{H}}_{s,t,f}, \tag{2}$$

$$\begin{aligned}\hat{\mathbf{\Phi}}^{\mathrm{Interfer}_s}_{t,f} =& \frac{t-1}{t}\hat{\mathbf{\Phi}}^{\mathrm{Interfer}_s}_{t-1,f} \\ &+ \frac{1}{t}(\mathbf{Y}_{t,f} - \hat{\mathbf{Z}}_{s,t,f})(\mathbf{Y}_{t,f} - \hat{\mathbf{Z}}_{s,t,f})^{\mathrm{H}},\end{aligned} \tag{3}$$

where $\hat{\mathbf{Z}}_{s,t,f}$ is reordered by $\hat{\pi}_{c,t}$.

B. Theoretical explanation

Different from Beam-TasNet, the proposed iterative scheme focus on finding a distribution $p(y_c; x_{s,c})$ parameterized by $x_{s,c}$, which maximizes the probability of generating the observed data. According to [1], the loglikelihood $\log p(y_c; x_{s,c})$ can be decomposed into 2 terms using latent variable $\mathbf{\Phi}$:

$$\begin{aligned}\log p\left(y_c; x_{s,c}\right) =& \mathrm{KL}\left[q(\mathbf{\Phi}) \| p\left(\mathbf{\Phi} \mid y_c; x_{s,c}\right)\right] \\ &+ E_{q(\mathbf{\Phi})}\left[\log \frac{p\left(y_c, \mathbf{\Phi}; x_{s,c}\right)}{q(\mathbf{\Phi})}\right]\end{aligned} \tag{4}$$

where $\mathbf{\Phi}$ is the spatial correlation matrix $\hat{\mathbf{\Phi}}^{\mathrm{Target}_s}_{f}$ and $\hat{\mathbf{\Phi}}^{\mathrm{Interfer}_s}_{f}$. We can use MC-Conv-TasNet to estimate signals and then obtain $\hat{\mathbf{\Phi}}$ with the given y_c and estimated parameters $\hat{x}_{s,c}$, which corresponds to $p(\mathbf{\Phi}|y_c; x_{s,c})$. Since the neural network directly generates estimation, we can view the distribution as an impulse function. Then by setting $q(\Phi) = p(\Phi|y_c; x_{s,c})$, maximizing the second item is equal to $\hat{x}_c = \mathrm{argmax}_{x_{s,c}}\, p\left(y_c, \hat{\mathbf{\Phi}}; x_{s,c}\right)$. With Bayes' rule, the optimal $\hat{x}_c$ is equal to $argmax_{x_c} p(x_c|y_c, \hat{\mathbf{\Phi}})$, which can be viewed as MVDR beamforming. Different from classic statistical models [2, 3], TasNet does not guarantee the estimated $\hat{\mathbf{\Phi}}$ closer to the oracle one. Thus, the proposed method may exhibit performance degradation in iterations.

C. The effect of STFT window size on MVDR

Table 1: *The performance of different window size for oracle IRMs.*

Window size (ms)	Causal	SDR$_\uparrow$ (dB) $\hat{z}_{s,1}$	SDR$_\uparrow$ (dB) $\hat{x}_{s,1}$	WER$_\downarrow$ (%) $\hat{z}_{s,1}$	WER$_\downarrow$ (%) $\hat{x}_{s,1}$
512	✗	11.0	14.5	28.1	15.8
32	✗	**12.9**	**17.6**	**12.4**	**12.8**
512	✓	11.0	10.6	28.1	20.9
32	✓	**12.9**	**14.0**	**12.4**	**13.6**

The STFT settings affect the performance of oracle IRMs. A longer window size and stride will lead to worse SDRs as the phase plays a more important role. A window size of 512ms results in an SDR of 11.0dB (w/o MVDR) and 14.7dB (w/ MVDR), similar to the Beam-TasNet paper, while a window size of 32ms results in an SDR of 12.9dB (w/o MVDR) and 17.6dB (w/ MVDR).

D. Unmatched noisy condition

To evaluate the proposed framework under the noisy condition, we simulated the noisy training and evaluation sets by mixing WSJ0-2MIX dataset with real recorded noise from the REVERB challenge [4]. The training set contains noise recorded in a small room with an SNR range from 10 dB to 20 dB. The evaluation set contains noise recorded in a medium and a large room with an SNR range from 0 dB to 10 dB.

The experiment result in Table 4 indicates that the proposed framework can deal with the noisy condition under unmatched noise settings. Compared with the baseline Beam-TasNet, our method achieved an SDR improvement of 2.5 dB and a WER reduction of 3.3%.

E. Multi-speaker condition

We deployed the proposed methods on 2- and 3-speaker conditions with a non-causal model using the 2- and 3-speaker spatialized WSJ0-2MIX and WSJ0-3MIX datasets. We used A2PIT [5] for training, which can be integrated with the proposed Beam-guided TasNet by introducing multiple outputs.

The experimental results are listed in Table 3. We have found that the proposed Beam-guided TasNet could outperform

This work is partially supported by the National Natural Science Foundation of China (Nos.62071461).

*Pengyuan Zhang is the corresponding author.

Table 2: *The performance of non-causal models under the unmatched noisy condition.*

Model	Iteration number	SDR$_{\uparrow}$ (dB)		WER$_{\downarrow}$ (%)	
		$\hat{z}_{s,1}$	$\hat{x}_{s,1}$	$\hat{z}_{s,1}$	$\hat{x}_{s,1}$
Beam-TasNet	-	10.5	14.0	31.1	19.6
1-Stage	-	9.7	13.6	33.7	20.3
2-Stage	1	15.2	15.1	17.7	18.1
2-Stage	2	16.2	15.5	16.5	17.7
2-Stage	4	**16.5**	15.6	**16.3**	17.7
Oracle IRM	-	12.3	15.4	12.6	17.2
Oracle signal	-	∞	**18.2**	**11.7**	16.5

Table 3: *The performance on the 2-/3-speaker dataset using non-causal models.*

Speaker number	Model	Iteration number	SDR$_{\uparrow}$ (dB)		WER$_{\downarrow}$ (%)	
			$\hat{z}_{s,1}$	$\hat{x}_{s,1}$	$\hat{z}_{s,1}$	$\hat{x}_{s,1}$
2	Beam-TasNet	-	11.8	16.7	25.3	14.0
	1-Stage	-	11.0	16.1	28.4	14.6
	2-Stage	1	18.4	19.1	14.2	12.4
	2-Stage	2	20.0	19.8	13.2	12.1
	2-Stage	4	**20.9**	20.3	13.1	**11.9**
	Oracle IRM	-	12.9	17.6	12.4	12.8
	Oracle signal	-	∞	**23.5**	**11.7**	11.9
3	Beam-TasNet	-	7.3	11.4	48.0	23.8
	1-Stage	-	6.4	10.6	52.6	25.8
	2-Stage	1	12.4	13.7	22.5	17.2
	2-Stage	2	14.6	14.8	17.5	15.5
	2-Stage	4	**15.8**	15.5	15.9	**14.5**
	Oracle IRM	-	9.8	14.8	12.2	14.8
	Oracle signal	-	∞	**22.1**	**11.7**	12.3

the Beam-TasNet consistently under the 2- and 3-speaker condition.

F. Learning anechoic signals

Previous experiments use models to learn single-speaker reverberant signals. Here we set the learning target to single-speaker anechoic signals to perform both dereverberation and separation tasks.

Table 4: *The performance of non-causal models on spatialized WSJ0-2MIX. The learning target and the reference signal for SDR calculation is single-speaker anechoic signals.*

Model	Iteration number	SDR$_{\uparrow}$ (dB)		WER$_{\downarrow}$ (%)	
		$\hat{z}_{s,1}$	$\hat{x}_{s,1}$	$\hat{z}_{s,1}$	$\hat{x}_{s,1}$
Beam-TasNet	-	10.8	14.6	29.8	15.2
1-Stage	-	9.4	14.0	38.9	17.1
2-Stage	1	14.5	16.4	17.9	13.6
2-Stage	2	16.5	17.1	14.9	12.8
2-Stage	4	17.1	**17.3**	14.2	**12.4**
Oracle IRM	-	11.4	12.0	11.1	15.5
Oracle signal	-	∞	**21.1**	**10.2**	11.4

The experiment results in Table 5 exhibit that the Beam-guided TasNet achieves an SDR of 17.3dB and a WER of 12.4%, far exceeding Beam-TasNet.

Table 5: *The performance of non-causal models on LibriCSS.*

Model	Iteration number	WER$_{\downarrow}$ (%)					
		0S	0L	OV10	OV20	OV30	OV40
Unprocessed	-	11.8	11.7	18.8	27.2	35.6	43.3
DPT-FSNET	-	7.1	7.3	7.6	8.9	10.8	11.3
1-Stage	-	7.3	7.3	7.8	8.9	10.6	11.1
2-Stage	1	7.1	7.1	7.1	8.0	9.2	9.7
2-Stage	2	**7.0**	**7.1**	**6.9**	7.9	**8.8**	9.3
2-Stage	4	**7.0**	**7.1**	7.0	**7.7**	**8.8**	**9.0**

G. Experiments on LibriCSS with frequency-domain model

LibriCSS is a real-recorded dataset. The ASR engine uses the original hybrid model [6]. We validate the iterative framework on a frequency-domain model, named DPT-FSNET [7]. After iterations, we achieve a WER of 9.0% on OV40 subset, 3.0% lower than DPT-FSNETs.

1. References

[1] C. M. Bishop, *Pattern recognition and machine learning, 5th Edition: 10.1 Variational Inference*, ser. Information science and statistics. Springer, 2007.

[2] N. Ito, S. Araki, and T. Nakatani, “Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing,” in *24th European Signal Processing Conference, EUSIPCO 2016, Budapest, Hungary, August 29 - September 2, 2016.* IEEE, 2016, pp. 1153–1157.

[3] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, “Online MVDR beamformer based on complex gaussian mixture model with spatial prior for noise robust ASR,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 4, pp. 780–793, 2017.

[4] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, “A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research,” *EURASIP J. Adv. Signal Process.*, vol. 2016, p. 7, 2016.

[5] N. Kanda, S. Horiguchi, R. Takashima, Y. Fujita, K. Nagamatsu, and S. Watanabe, “Auxiliary interference speaker loss for target-speaker speech recognition,” in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 236–240.

[6] 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 *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020.* IEEE, 2020, pp. 7284–7288. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9053426

[7] F. Dang, H. Chen, and P. Zhang, “Dpt-fsnet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement,” in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022.* IEEE, 2022.