On Training Speech Separation Models With Various Numbers of Speakers

Hyeonseung Kim, *Student Member, IEEE*, and Jong Won Shin, *Senior Member, IEEE*

***Abstract*—Many monaural speech separation models assume that the exact number of speakers is known in advance, which is not applicable to many real-world scenarios. To deal with an unknown number of speakers, previous approaches either iteratively separate one speech at a time, or employ a more relaxed assumption that the maximum number of speakers is known a priori and set the number of outputs accordingly. When the number of speakers in the mixture is smaller than the number of outputs in the latter case, the extra outputs that are not mapped onto signals in the input mixture are trained to produce predefined target signals such as the silence or the input mixture. In this letter, we propose to ignore the extra outputs in training instead of evaluating the cost with a certain target for separation models with a fixed number of output channels. We also introduce a method to select valid output signals. Experimental results showed that assigning any type of predefined targets degraded separation performance compared with ignoring the extra outputs.**

***Index Terms*—Speaker counting, speech separation, unknown number of speakers.**

I. Introduction

SPEECH separation is the task to separate individual speeches from a mixture. Compared to other source separation problems, speech separation has been considered to be difficult because the signals to be separated have similar characteristics, especially when the speakers are not known in advance [1], [2], [3], [4], [5]. However, recent advances in deep learning have dramatically improved speech separation for the open-speaker condition under an environment without noise and reverberation [6], [7], [8], [9], [10], [11], [12], [13], [14], [15]. With the success in this limited setting, many studies have been proposed to separate speeches in more realistic environments such as under noisy and reverberant conditions [16], [17], [18], [19], [20], [21], with an unknown number of speakers [22], [23], [24], [25], [26], [27], [28], [29], or from a continuous audio stream [31], [32], [33], [34].

One strategy to separate the unknown number of speeches is to extract one speech at a time from the input or residual mixture [22], [23], [24], [25], [28]. The extraction stops when the residual mixture does not contain any speech signal or after a certain number of iterations given by the speaker counting. However, computational delay of this iterative separation increases linearly as the number of speakers increases and the quality of the signals extracted later may not be as good as those extracted earlier [25]. An alternative strategy is to select a suitable model among multiple models trained assuming different numbers of speakers [26]. At first, the input mixture is separated with the model for the assumed largest number of speakers. If one of the separated outputs has too low power, the input is processed with the model with one less number of speakers. This method requires longer process time for smaller number of speakers and the model size increases with a maximum number of speakers as it requires a model for each number of speakers.

Another strategy is to set the number of output channels of the separation network as many as the expected maximum number of speakers and determine if each output is valid after separation [9], [27], [28], [29], [30]. When an input mixture contains fewer speakers than the assumed maximum number of speakers, some of the outputs would be regarded as separated speeches, while the rest of the outputs which we call invalid outputs are discarded. To train models using training data with various numbers of speakers, previous approaches [9], [27], [28], [29], [30] assigned certain target signals to the invalid outputs. In [9], [28], [29], [30], silent signals were used as the targets for the invalid outputs in training, and the valid outputs were identified in the separation phase by comparing the power of each output with a predetermined threshold. Although this approach was simple and effective, the adoption of a silent target signal restricts the use of losses such as the scale-invariant signal-to-noise ratio (SI-SNR) [37], and the fixed threshold may make it difficult to deal with the variability in the input speech power [27]. To overcome these weaknesses, the input mixture was used as the target in [27]. However, guiding the invalid outputs to become similar to the input mixture may hinder finding better valid outputs.

In this letter, we propose to ignore the extra outputs in training instead of evaluating any type of cost with a specific target for models with a fixed number of output channels. We also propose an algorithm to identify valid output signals when the extra outputs are not guided to have a specific form. We have also tested another approach to assign one of the individual speeches that is closest to the output as the target for each invalid channel. We applied these strategies and those in [27], [28], [29], and [30] to the speech separation system based on the Dual-Path RNN (DPRNN) [14] to evaluate the separation performances. Experimental results showed that the proposed strategy to ignore the invalid outputs in training outperformed other approaches in terms of the SI-SNR.

Manuscript received 29 June 2023; revised 17 August 2023; accepted 25 August 2023. Date of publication 31 August 2023; date of current version 8 September 2023. This work was supported by the MSIT (Ministry of Science and ICT), Korea, through the ITRC (Information Technology Research Center) supported Program under Grant IITP-2021-2021-0-01835, supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniele Giacobello. *(Corresponding author: Jong Won Shin.)*

The authors are with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea (e-mail: kimhs355@gm.gist.ac.kr; jwshin@gist.ac.kr).

Digital Object Identifier 10.1109/LSP.2023.3310881

1070-9908 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Kwangju Institute of Science and Technology. Downloaded on September 12,2023 at 06:39:27 UTC from IEEE Xplore. Restrictions apply.

II. Conventional Target Assignment Strategies for the Invalid Outputs of Separation Models

Most of separation networks using learnable encoder-decoder structures have a fixed number of output channels. Therefore, when the number of speakers in the mixture is smaller than the number of channels in the model, it is not straightforward to assign target signals in training to the output channels that are not mapped onto one of the mixed signals. We introduce some recently proposed methods to configure target signals for invalid outputs in this section and propose alternative approaches in Section III.

A. Silent Target for Invalid Outputs

In [28], two methods to separate various numbers of speakers were proposed. One is to separate one source at a time and the other is to separate multiple sources simultaneously. For the latter one, a silent signal is used as the target for each extra output channel which is not mapped to one of the signals in the input mixture. Specifically, when there are C output channels but only M signals are present in the input mixture, the permutation invariant training (PIT) [9] is applied with M valid targets and $(C - M)$ silent target signals. The negative SI-SNR [37] loss is defined as

$$\mathcal{L}^{\mathrm{S}}(\mathbf{s}, \hat{\mathbf{s}}) = -10 \log_{10} \frac{\|\gamma \mathbf{s}\|_2^2}{\|\hat{\mathbf{s}} - \gamma \mathbf{s}\|_2^2}, \tag{1}$$

where $\mathbf{s}$ is the time domain signal vector, $\hat{\mathbf{s}}$ is the estimate of it, and γ is the optimal scaling factor. Although it is widely used as a loss function for speech separation task, it cannot be used with silent target signals since the scaling factor $\gamma = \langle \mathbf{s}, \hat{\mathbf{s}} \rangle / \|\mathbf{s}\|_2^2$ cannot be defined on them. Instead, [28] proposed T-L1PMSE loss, which is the logarithm of one plus mean squared error in the time domain defined as

$$\mathcal{L}^{\mathrm{T}}(\mathbf{s}, \hat{\mathbf{s}}) = 10 \log_{10} \left(1 + \|\mathbf{s} - \hat{\mathbf{s}}\|_2^2\right). \tag{2}$$

The loss with M mixed signals and C outputs then becomes

$$\mathcal{L}^{\mathrm{1PMSE}} = \frac{1}{C} \left(\sum_{m=1}^{M} \mathcal{L}^{\mathrm{T}}(\mathbf{s}_m, \hat{\mathbf{s}}_m) + \sum_{n=M+1}^{C} \mathcal{L}^{\mathrm{T}}(\mathbf{0}, \hat{\mathbf{s}}_n) \right), \tag{3}$$

where $\mathbf{s}_m$ and $\hat{\mathbf{s}}_m$ are the m-th valid target and corresponding estimated signals found by the PIT, $\hat{\mathbf{s}}_n$ is the output signal that is not chosen by the PIT, and $\mathbf{0}$ is a zero vector.

In [29], the silence signal was also used as a target for invalid outputs with another loss function, which is the soft-thresholded SNR loss given by

$$\mathcal{L}^{\mathrm{tSNR}} = \frac{1}{C} \left(\sum_{m=1}^{M} \mathcal{L}^{\mathrm{act}}(\mathbf{s}_m, \hat{\mathbf{s}}_m) + \sum_{n=M+1}^{C} \mathcal{L}^{0}(\mathbf{x}, \hat{\mathbf{s}}_n) \right), \tag{4}$$

in which

$$\mathcal{L}^{\mathrm{act}}(\mathbf{s}, \hat{\mathbf{s}}) = 10 \log_{10} \left(\|\mathbf{s} - \hat{\mathbf{s}}\|_2^2 + \tau \|\mathbf{s}\|_2^2 \right), \tag{5}$$

$$\mathcal{L}^{0}(\mathbf{x}, \hat{\mathbf{s}}) = 10 \log_{10} \left(\|\hat{\mathbf{s}}\|_2^2 + \tau \|\mathbf{x}\|_2^2 \right), \tag{6}$$

and $\mathbf{x}$ is the input mixture signal.

Because using different losses for different outputs can be unstable, the source-aggregated source-to-distortion ratio (SA-SDR) loss was proposed in [30]. Instead of averaging losses for all sources, SA-SDR sums up the energies for the targets and distortion, which results in

$$\mathcal{L}^{\text{SA-SDR}} = -10 \log_{10} \frac{\sum_{m=1}^{M} \|\mathbf{s}_m\|_2^2}{\sum_{m=1}^{M} \|\mathbf{s}_m - \hat{\mathbf{s}}_m\|_2^2 + \sum_{n=M+1}^{C} \|\hat{\mathbf{s}}_n\|_2^2}. \tag{7}$$

B. Mixture Target for Invalid Outputs

Since the negative SI-SNR loss cannot be adopted and it is tricky to set the power threshold when silent signals are used as the targets for invalid outputs, the input mixture signal was used as the targets for invalid outputs in [27]. The authors called the loss for the invalid outputs the auxiliary autoencoding loss as the input and the target are the same. Since reproducing the input mixture is a relatively easy task compared with separating individual sources, the loss for the invalid outputs may dominate the final loss function if the negative SI-SNR loss is applied to all outputs. Therefore, the negative α-skewed SI-SNR loss was used as the auxiliary autoencoding loss [27], which is defined as

$$\mathcal{L}^{\alpha\text{-S}}(\mathbf{s}, \hat{\mathbf{s}}) = -10 \log_{10} \left(\frac{c(\mathbf{s}, \hat{\mathbf{s}})^2}{1 + \alpha - c(\mathbf{s}, \hat{\mathbf{s}})^2} \right), \tag{8}$$

in which $c(\mathbf{a}, \mathbf{b}) = (\mathbf{a} \cdot \mathbf{b}) / (\|\mathbf{a}\| \|\mathbf{b}\|)$ is the cosine similarity between $\mathbf{a}$ and $\mathbf{b}$ and α is a perturbation factor which was set to 0.3 in [27] and our experiments. The loss function becomes

$$\mathcal{L}^{\mathrm{A2PIT}} = \frac{1}{C} \left(\sum_{m=1}^{M} \mathcal{L}^{\mathrm{S}}(\mathbf{s}_m, \hat{\mathbf{s}}_m) + \sum_{n=M+1}^{C} \mathcal{L}^{\alpha\text{-S}}(\mathbf{x}, \hat{\mathbf{s}}_n) \right). \tag{9}$$

III. Proposed Strategies for the Invalid Outputs of Separation Models

A. Ignoring Losses for Invalid Outputs

At the inference phase, the outputs that are decided to be invalid will be disregarded eventually. As we are not interested in the invalid output signals, enforcing them to be similar to specific target signals may disrupt speech separation rather than regularizing it. In this regard, we propose to compute the loss function in training with only M outputs that matches the individual target signals best, which we call Choose the Best and Ignore the Rest (CBIR) strategy. The loss function is therefore just the average of losses for the valid outputs:

$$\mathcal{L}^{\mathrm{CBIR}} = \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}^{\mathrm{S}}(\mathbf{s}_m, \hat{\mathbf{s}}_m). \tag{10}$$

B. Best Matched Targets for Invalid Outputs

Unlike two previous approaches introduced in Section II, the proposed CBIR strategy does not constrain the invalid outputs. However, since the output channels invalid for the current input are valid output channels for other input signals including mixtures of C signals, the invalid output signal may be speech-like. We observe that each invalid output signal for the CBIR strategy resembles one of the speeches in the mixture. Based on this observation, we propose another strategy to assign one of the speech signals closest to the given output as a target even for the invalid outputs, which we call the best matching target (BMT)

Authorized licensed use limited to: Kwangju Institute of Science and Technology. Downloaded on September 12,2023 at 06:39:27 UTC from IEEE Xplore. Restrictions apply.

strategy. The loss function for this strategy is defined as

$$\mathcal{L}^{\text{BMT}} = \frac{1}{C}\left(\sum_{m=1}^{M}\mathcal{L}^{\text{S}}(\mathbf{s}_m, \hat{\mathbf{s}}_m) + \sum_{n=M+1}^{C}\mathcal{L}^{\text{S}}(\mathbf{s}_n, \hat{\mathbf{s}}_n)\right), \quad (11)$$

where $\mathbf{s}_n$ is one of the individual speech signals in the mixture which minimizes loss for the invalid output $\hat{\mathbf{s}}_n$.

C. Identifying Valid Output Channels

In the inference phase, the validity of each output channel should be identified. In the method with a silent target for invalid outputs introduced in Section II-A, the power of the output signal is used to determine the validity of the signal, while the SI-SNR between the mixture input and the given output is used to discriminate valid channels from invalid ones in the method with a mixture target explained in Section II-B. In contrast to these approaches which utilize the specific targets for the invalid outputs to check validity of each output signal, it is not clear how to identify the channel validity for the CBIR strategy. We have observed that each invalid output signal for the CBIR method resembled one of the signals in the mixture, possibly because an invalid output channel for the current input becomes a valid output channel for other input mixtures and thus was trained to produce speech-like signals. As for the BMT strategy, the invalid output signals are forced to become similar to one of the signals in the mixture. Therefore, we can identify valid output channels based on the similarity among output signals for both the CBIR and the BMT strategies.

The procedure to determine the number of speakers in the output mixture and the validity of output signals for the CBIR and the BMT methods are as follows. First, we determine if only one speaker is active by checking whether the magnitude of the cosine similarity between the input and the output is higher than a threshold η_1 for all output signals. Once the input turns out to be a mixture of multiple speeches, we initially assume all output channels are active and evaluate the absolute values of the cosine similarities for all pairs of output signals. At each stage with k remained output signals, we decide the input contains k speakers if the maximum of cosine similarities between the remained signals is less than a threshold η_{k-1}. If not, one of the output signals that produce the maximum absolute cosine similarity is classified as an invalid output and the cosine similarities related to that output are removed. Between the two most similar signals, we discard the output channel which was less frequently selected as a valid output for the validation set to select the output channel that is more likely to produce well-separated sources for the unseen samples. This procedure is repeated until the maximum similarity is less than the threshold or k becomes 2. The thresholds η_k were determined using the validation set. This method also worked for noisy mixtures to an extent.

IV. Experiments

A. Experimental Configurations

Although WSJ0-2mix and 3mix datasets [6] have been widely used for the training and evaluation of speech separation model, some utterances have preceding silence, which makes some of the trimmed target signals silent signals. These silent target signals may disrupt the comparison of the target assignment strategies and source counting methods. In this regard, we made a fully overlapped version of WSJ0-2mix, 3mix, and 4mix datasets by removing the silent regions at the beginning and the end of the source signals to train and test our implemented models. Like the original dataset, the fully overlapped version of the dataset used `si_tr_s` subset of WSJ0 corpus to construct the training and validation sets and `si_dt_05` and `si_et_05` to form the test dataset. The relative SNRs for 2 and 3mix datasets were also set in the same way as the original dataset. When producing 4mix data, the levels of the first two signals were adjusted by $+r$ and $-r$ dB in which r was a random number between 0 and 2.5. Those for the rest of two signals were modified by $+k + r'$ and $+k - r'$ dB in which k and r' were randomly chosen in ranges $[-2.5, 2.5]$ and $[0, 2.5\text{-}|k|]$, respectively. For training, we generated 20,000 mixture samples for each number of speakers with the length of 4 s every epoch to increase the diversity in the training data. The validation and test sets consist of 5,000 and 3,000 mixture samples, respectively. The sampling rate was 8 kHz.

TABLE I
The SI-SNRi (dB) Scores for Various Training Strategies on the Fully Overlapped Version of WSJ0-Mix. The Performances on the Original WSJ0-Mix are Shown in Parentheses

Training		2 speakers	3 speakers	4 speakers
Matched		17.39 (17.72)	14.64 (14.91)	12.42
A2PIT [27]		16.57 (16.85)	14.93 (15.10)	12.72
Silence	1PMSE [28]	16.51 (16.81)	14.96 (15.14)	12.65
	tSNR [29]	16.58 (16.93)	14.95 (15.22)	12.66
	SA-SDR [30]	16.92 (17.21)	15.01 (15.17)	12.52
BMT		16.89 (17.16)	14.98 (15.23)	12.80
CBIR		17.09 (17.42)	15.30 (15.53)	13.08

We used the DPRNN [14] with 4 outputs as a backbone model to compare target assignment strategies to deal with various number of speakers. For the encoder and decoder, we set the window length and stride as 16 (2 ms) and 8 (1 ms), respectively, and the encoded feature dimension was 64. The chunk size for dual-path modeling was set to 90 frames. All the other parameters such as the bottleneck dimension, the number of DPRNN blocks, and the number of hidden units in each BLSTM layer were configured to be the same as those in [14]. The systems with 2, 3, and 4 outputs trained and tested with matched numbers of speakers were also prepared as performance benchmarks. We used Adam optimizer with initial learning rate to 0.001 and applied the cosine annealing with warm restart scheduler [40] implemented in Pytorch [41] where the number of epochs till the first restart was 4 and the number of epochs between each restart was increased by a factor of 2. The training stopped before the 5th restart, resulting in a total of 124 epochs.

B. Experimental Results

To assess the speech separation performance when the estimated number of speakers $\hat{M}$ may not be the same as the true number of speakers M, we have evaluated penalized-SI-SNRi (P-SI-SNRi) proposed in [35], given by

$$\text{P-SI-SNRi} = \frac{\sum_{m=1}^{\min(M,\hat{M})}\text{SI-SNRi}(\bar{\mathbf{s}}_m, \hat{\bar{\mathbf{s}}}_m) + |M-\hat{M}|\mathcal{P}_{\text{ref}}}{\max(M,\hat{M})}, \quad (12)$$

Authorized licensed use limited to: Kwangju Institute of Science and Technology. Downloaded on September 12,2023 at 06:39:27 UTC from IEEE Xplore. Restrictions apply.

TABLE II
SPEAKER COUNTING ACCURACIES AND PENALIZED SI-SNRI SCORES. THE ROW INDICES REPRESENT THE ACTUAL NUMBER OF SPEAKERS AND THE COLUMN INDICES DENOTE THE ESTIMATED NUMBER OF SPEAKERS

	1	2	3	4	P-SI-SNRi (dB)
1	100	0	0	0	-
2	0.30	98.30	1.40	0	16.28
3	0	0	98.33	1.67	14.75
4	0	0	0.37	99.63	12.70
Average	99.07				14.58

(a) A2PIT [27]

	1	2	3	4	P-SI-SNRi (dB)
1	99.79	0.22	0	0	-
2	0	99.90	0.10	0	16.91
3	0	0.07	99.50	0.43	14.96
4	0	0	1.07	98.93	12.49
Average	99.53				14.79

(b) SA-SDR [30]

	1	2	3	4	P-SI-SNRi (dB)
1	99.41	0.59	0	0	-
2	0.07	99.83	0.10	0	16.85
3	0	0.10	99.10	0.80	14.89
4	0	0	0.63	99.37	12.77
Average	99.43				14.84

(c) BMT

	1	2	3	4	P-SI-SNRi (dB)
1	99.89	0.05	0	0.05	-
2	0.03	99.70	0.13	0.13	16.83
3	0	0.13	97.77	2.10	14.98
4	0	0	2.80	97.20	12.85
Average	98.64				14.89

(d) CBIR

where $\mathcal{P}_{\text{ref}}$ was set as -30 dB, and $\bar{\mathbf{s}}_m$ and $\hat{\bar{\mathbf{s}}}_m$ are separated signals which is determined as valid by the thresholding algorithm and corresponding target signals found by the PIT, respectively. P-SI-SNRi is the same as the SI-SNRi when the number of speakers and valid channels are correctly identified, and penalizes each counting error by $\mathcal{P}_{\text{ref}}$.

Firstly, we have compared the speech separation performance for various target assignment strategies, selecting the M output channels that produced the best performance to exclude the effect of the source counting and channel validity estimation. The P-SI-SNRi scores, which are the same as SI-SNRi scores in this experiment, are summarized in Table I. The performances for models trained and tested with a single number of speakers (Matched) are also presented for performance benchmarks. The performance with the original WSJ0-2mix and -3mix datasets are provided in the parentheses to make the comparison with other papers easier. Existing target assignment strategies did not show significant differences in separation performance. Though the SA-SDR [30] method showed the second best performance for 2- and 3-mix data, it had the lowest performance for the 4-mix data. The proposed BMT method showed little difference with SA-SDR method for 2- and 3-mix data, and second-best performance for 4-mix data. The proposed CBIR strategy exhibited the best performance for 2, 3, and 4 speakers with a significant difference compared to all other training strategies. The p-values comparing the SI-SNRi scores for the CBIR and the conventional approach that exhibited the best performance, SA-SDR, were 4.3e-4, 1.2e-9, and 2.6e-25 for 2-, 3-, and 4-mixes, respectively. For the 4-mix data, p-value between CBIR and A2PIT was 4.7e-16. Therefore, we may conclude that evaluating loss only for the output channels that matches the source speeches best and ignoring the rest of the channels was effective in training the networks with a fixed number of output channels. It is interesting that the performances for 3 and 4 speakers using the networks trained with various number of speakers were better than those for the models trained with a matched number of speakers. It may be because the fully overlapped version of the 3 or 4 mixtures also has intervals with less number of speakers corresponding to the speech pauses of one or more speakers. Another possible reason would be the training with simpler mixtures might help the training of more complex mixtures.

Next, we have evaluated the speaker counting accuracies and the the separation performance using the thresholding method. The results of speaker counting and corresponding P-SI-SNRi scores are given in Table II. We only report the performance of SA-SDR loss for the silence target methods, as it showed the best overall separation performance among them. For each sub-table, the row indices represent the actual numbers of speakers and the column indices denote the predicted ones. Although models were trained only with 2-, 3-, and 4-mix data, it was possible to classify a single speech as observed in [28]. The proposed CBIR method showed slightly lower counting accuracy compared to existing methods, possibly because there were no proper way to discriminate the valid and invalid channel. Nevertheless, the average P-SI-SNRi scores of the CBIR method is higher than existing two methods, indicating that separation performance is good enough to compensate for the penalty of mis-counting.

V. CONCLUSION

In this letter, we have proposed a strategy to ignore the invalid output signals when computing the loss function to train the separation network, which we call Choose the Best and Ignore the Rest (CBIR) strategy and a method to assign one of the target signals that matches the given output best, which is called Best Matching Target (BMT) strategy. Experimental results showed that the proposed CBIR and BMT methods outperformed conventional strategies to assign mixture or silent signals as targets for invalid output channels in terms of separation performance.

REFERENCES

[1] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space," in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, pp. 89–92.

[2] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[3] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Comput. Speech Lang.*, vol. 24, no. 1, pp. 1–15, Jan. 2010.

[4] Y. Wang and D. Wang, "Towards scaling up classification based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[5] J. Le Roux, F. J. Weninger, and J. R. Hershey, "Sparse NMF - Half-baked or well done?" MERL, Cambridge, MA, USA, Tech. Rep. TR2015-023, Mar. 2015.

[6] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 31–35.

Authorized licensed use limited to: Kwangju Institute of Science and Technology. Downloaded on September 12,2023 at 06:39:27 UTC from IEEE Xplore. Restrictions apply.

[7] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Int. Conf. Spoken Lang. Process.*, 2016, pp. 545–549.

[8] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," in *Proc. Int. Conf. Spoken Lang. Process.*, 2017, pp. 2456–2460.

[9] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[10] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 4, pp. 787–796, Apr. 2018.

[11] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 696–700.

[12] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[13] Y. Liu and D. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.

[14] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 46–50.

[15] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -RF: Efficient networks for universal audio source separation," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process.*, 2020, pp. 1–6.

[16] Y. Luo, C. Han, N. Mesgarani, E. Ceolini, and S. Liu, "FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2019, pp. 260–267.

[17] J. Zhang, C. Zorilă, R. Doddipatla, and J. Barker, "On end-to-end multi-channel time domain speech separation in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6389–6393.

[18] R. Gu et al., "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7319–7323.

[19] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6394–6398.

[20] D. Wang, Z. Chen, and T. Yoshioka, "Neural speech separation using spatially distributed microphones," in *Proc. Int. Conf. Spoken Lang. Process.*, 2020, pp. 339–343.

[21] C. Fan, J. Tao, B. Liu, J. Yi, and Z. Wen, "Gated recurrent fusion of spatial and spectral features for multi-channel speech separation with deep embedding representations," in *Proc. Int. Conf. Spoken Lang. Process.*, 2020, pp. 3321–3325.

[22] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5064–5068.

[23] J. Shi, J. Xu, G. Liu, and B. Xu, "Listen, think and listen again: Capturing top-down auditory attention for speaker-independent speech separation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 4353–4360.

[24] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," in *Proc. Int. Conf. Spoken Lang. Process.*, 2019, pp. 1348–1352.

[25] Z. Jin, X. Hao, and X. Su, "Coarse-to-fine recursive speech separation for unknown number of speakers," 2022, *arXiv:2203.16054*.

[26] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 7164–7175.

[27] Y. Luo and N. Mesgarani, "Separating varying numbers of sources with auxiliary autoencoding loss," in *Proc. Int. Conf. Spoken Lang. Process.*, 2020, pp. 2622–2626.

[28] T. V. Neumann et al., "Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR," in *Proc. Int. Conf. Spoken Lang. Process.*, 2020, pp. 3097–3101.

[29] S. Wisdom et al., "What's all the fuss about free universal sound separation data?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 186–190.

[30] T. V. Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "SA-SDR: A novel loss function for separation of meeting style data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 6022–6026.

[31] Z. Chen et al., "Continuous speech separation: Dataset and analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7284–7288.

[32] S. Chen et al., "Continuous speech separation with conformer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 5749–5753.

[33] S. Chen et al., "Don't shoot butterfly with rifles: Multi-channel continuous speech separation with early exit transformer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6139–6143.

[34] 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.

[35] J. Zhu, R. A. Yeh, and M. H. Johnson, "Multi-decoder DPRNN: Source separation for variable number of speakers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 3420–3424.

[36] S. E. Chazan, L. Wolf, E. Nachmani, and Y. Adi, "Single channel voice separation for unknown number of speakers under reverberant and noisy settings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 3730–3734.

[37] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[38] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 47–54.

[39] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 933–941.

[40] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Representations*, 2017.

[41] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

Authorized licensed use limited to: Kwangju Institute of Science and Technology. Downloaded on September 12,2023 at 06:39:27 UTC from IEEE Xplore. Restrictions apply.