INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Investigation of Training Mute-Expressive End-to-End Speech Separation Networks for an Unknown Number of Speakers

Younggwan Kim, Hyungjun Lim, Kiho Yeom, Eunjoo Seo, Hoodong Lee, Stanley Jungkyu Choi, Honglak Lee

Language Lab, LG AI Research, Seoul, Republic of Korea

{glorick.kim, hyungjun.lim, kiho.yeom, eunjoo.seo, hoodong.lee, stanleyjk.choi, honglak}@lgresearch.ai

Abstract

In speech separation, there have been a limited number of prior works for an unknown number of speakers in a speech mixture. To address this situation, one simple solution is to constitute the sufficient number of output channels greater than or equal to the expected number of speakers and ignore invalid outputs containing meaningless signals when the number of speakers is less than the output channels. To detect such invalid outputs, it is an ideal scenario for the meaningless signals to be muted.

In this paper, we investigate several training methods by which separation models can mute the invalid outputs. We first introduce an on-the-fly data mixing scheme adding small random noises to the speech mixtures. As a training criterion, we analyze why the well-known scale-invariant signal-to-noise ratio is not suitable for muting the invalid outputs because of its power amplification problem and also explain why we use the signal-to-noise ratio criterion to avoid the problem.

Index Terms: mute-expressive, speech separation, random silence mixing, scale-invariant signal-to-noise ratio, power amplification problem

1. Introduction

Monaural speech separation is a method for estimating individual speech spoken by different speakers from a single-channel speech mixture [1–5]. End-to-end speech separation (E2E-SS) is one of the main research areas in monaural speech separation, in which separation models are directly trained by waveform speech signals [6–14]. In general, most of E2E-SS researches assume that the number of speakers in the mixture is fixed and known in advance and this assumption is a cause of hindering the broad use of speech separation.

Recently, there have been a limited number of prior works to cope with the condition that the number of speakers in the mixture is unknown. In [15, 16], single speech is separated at a time with sequential forward passes in a recursive way. However, in this method, a certain stopping criterion is required. In [17], multiple models having different number of output channels are trained first. At inference time, the number of speakers is detected to select a proper model having same number of output channels. In [18], the authors use an attractor network to determine the number of speakers, which provides additional information for speech separation. Also, in [19, 20], when the number of speakers is less than the pre-defined number of output channels, the authors suggest a single model-based approach by ignoring the invalid outputs of silent channels containing meaningless signals with certain thresholding methods. This is one of the simple ways to address the unknown number of speakers in the mixture. However, this approach also requires additional procedures to find and apply a proper threshold value.

As mentioned earlier, the prior works for the unknown number of speakers require additional procedures, models, or network modules. In this paper, we investigate several training methods that can allow E2E-SS models to effectively mute the invalid outputs for an unknown number of speakers. This investigation aims to achieve competitive detection performance by enabling straightforward detection and rejection of the invalid outputs by muting them with minimal additional efforts.

For training data, we first introduce a simple data mixing scheme called "random silence mixing" (RSM) randomly replacing speech sources with small random noises to simulate a mismatched condition where the number of speakers is less than the number of output channels. Also, as a training criterion, the scale-invariant signal-to-noise ratio (SI-SNR) has been widely used for speech separation. However, the SI-SNR has a problem gradually amplifying the power of output signals while training, making it unsuitable for muting the invalid outputs. In addition, due to the problem, an additional amplitude normalization process is required. Typically, the normalization process is done by dividing an output signal by its maximum absolute sample value. Unfortunately, if invalid outputs are not properly ignored, the normalization process can hugely amplify small and meaningless signals. To identify the cause of the power amplification problem leading to the aforementioned issues, we analyze the gradient of the SI-SNR. As far as we know, this is the first work to report and analyze the problem. In order to address the issue of power amplification and properly mute the invalid outputs, we use the conventional signal-to-noise ratio (SNR) with the RSM scheme as a power-preserving training criterion. This training combination allows for invalid outputs to be muted, or to remain as small signals even if they are not muted.

The remainder of this paper organized as follows: In Section 2, we first briefly describe the pipeline of speech separation. Afterward, we introduce the RSM scheme, analyze how SI-SNR causes the power amplification problem, and explain why we replace the SI-SNR with the SNR. Section 3, we demonstrate the power amplification phenomenon that interferes with muting the invalid outputs through actual waveforms, spectrograms, and a prediction-to-target signal power ratio curve in dB scale. We also compare separation performances between the SI-SNR and the SNR with the RSM applied to both. In addition, invalid output detection rates are reported in this section. Finally, in Section 4, we summarize this paper.

2. Methodology

2.1. Speech Separation Pipeline

The general goal of speech separation is to estimate individual speech sources $\boldsymbol{s}_i \in \mathbb{R}^{T\times 1}$, where $i \in [1, ..., S]$, from a speech

10.21437/Interspeech.2023-1158

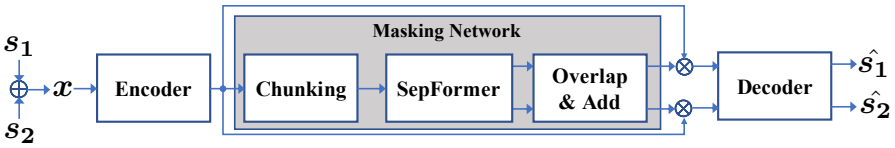


Figure 1: *Block diagram of SepFormer-based speech separation.*

mixture which is defined as $\boldsymbol{x} = \sum_{i=1}^{S} \alpha_i \boldsymbol{s}_i$, where α_i and S denote a randomly assigned scaling factor for mixing and the number of speakers.

There have been various approaches for E2E-SS and masking-based speech separation is one of the most reliable methods which generate mask information for each output channel. In this paper, we use a masking-based speech separation model with SepFormer as shown in Figure 1 [21].

2.2. Random Silence Mixing

For training, the speech mixture is typically made up of the sum of the speech sources, where the number of speakers S is equal to the number of output channels C. However, with this speech mixture, the E2E-SS models are not able to learn how to mute the invalid outputs when $S < C$. To simulate the mismatched condition, in [20], the authors use predefined two- and three-speaker mixture training sets with white Gaussian noise having relatively low power compared to the speech sources. For procedural simplification, in this paper, we introduce an on-the-fly data mixing scheme called *random silence mixing* (RSM) instead of preparing separate training sets. The RSM only utilizes the speech mixtures that meets the condition $S = C$ and randomly substitutes the speech sources with small white Gaussian noises. By the RSM, the modified speech mixture is defined as:

$$\bar{\boldsymbol{x}} = \sum_{i=1}^{C} \alpha_i \bar{\boldsymbol{s}}_i, \ \bar{\boldsymbol{s}}_i = \begin{cases} \boldsymbol{s}_i, & \text{if } \eta_i \leq \rho \\ \beta \boldsymbol{\epsilon}, & \text{otherwise} \end{cases} \tag{1}$$

where η_i is a sampled value drawn from a uniform distribution $\mathcal{U}(0,1)$, ρ denotes a fixed probability for assigning real speech sources, β is a fixed scaling factor for silence, and $\boldsymbol{\epsilon}$ is a $1 \times T$ vector whose values are randomly drawn from a Gaussian distribution $\mathcal{N}(0,1)$. Note that we assign at least one real speech source to $\bar{\boldsymbol{x}}$ and there is no need to calculate relative power for silence. With $\bar{\boldsymbol{x}}$, we can instantly augment training speech mixtures from 1- to C-speakers. In the following, the source index i is omitted for brevity.

2.3. Power Amplification Problem of SI-SNR

The SI-SNR has been widely used for a learning criterion to estimate the various types of target signals including speech [7,8,22]. However, it has a problem that gradually amplifies the power of the output speech much greater than the target speech. Therefore, the output speech by the SI-SNR requires further amplitude normalization causing the series of problems as mentioned in Section 1. Also, this problem interferes with muting the invalid outputs in speech separation. In this section, we analyze how the SI-SNR criterion causes the power amplification problem during training according to its gradient.

The SI-SNR-based learning criterion is defined as:

$$\mathcal{L}_{\text{SI-SNR}} = -10 \log_{10} \frac{||\tilde{\boldsymbol{s}}||^2}{||\tilde{\boldsymbol{e}}||^2} \tag{2}$$

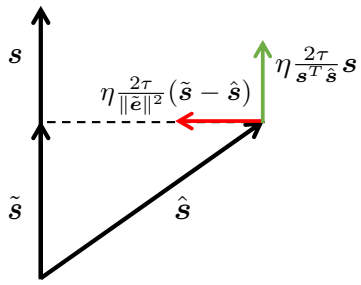


Figure 2: *Illustration of error signals at a network output layer of SI-SNR in descent direction.*

where $\tilde{\boldsymbol{s}} = \frac{\boldsymbol{s}^T \hat{\boldsymbol{s}}}{||\boldsymbol{s}||^2} \boldsymbol{s}$ is a scaled version of target speech and $\tilde{\boldsymbol{e}} = \tilde{\boldsymbol{s}} - \hat{\boldsymbol{s}}$ is a error vector between the scaled target speech and the predicted output speech of the separation model. As shown in (2), the SI-SNR criterion scales the amplitude of the target speech and it makes the output speech power unpredictable. As described earlier, the power of the output speech tends to be much greater than the target speech and this phenomenon can be explained by analyzing the gradient of (2). At first, we define the gradient descent-based parameter update rule as follows:

$$\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} - \eta \frac{\partial \mathcal{L}_{\text{SI-SNR}}}{\partial \hat{\boldsymbol{s}}} \frac{\partial \hat{\boldsymbol{s}}}{\partial \boldsymbol{\Theta}} \tag{3}$$

where $\boldsymbol{\Theta}$ denotes the parameter set of the speech separation model. Generally, $\frac{\partial \mathcal{L}_{\text{SI-SNR}}}{\partial \hat{\boldsymbol{s}}}$ is called the error signal at the output layer, which is back-propagated to lower layers to update the model parameters. The error signal can be divided into two terms as follows:

$$\frac{\partial \mathcal{L}_{\text{SI-SNR}}}{\partial \hat{\boldsymbol{s}}} = \frac{\partial 10 \log_{10} ||\tilde{\boldsymbol{e}}||^2}{\partial \hat{\boldsymbol{s}}} - \frac{\partial 10 \log_{10} ||\tilde{\boldsymbol{s}}||^2}{\partial \hat{\boldsymbol{s}}}. \tag{4}$$

The first term on the right-hand side of (4) generates gradient to decrease the distance between $\tilde{\boldsymbol{s}}$ and $\hat{\boldsymbol{s}}$ as shown below:

$$\frac{\partial 10 \log_{10} ||\tilde{\boldsymbol{e}}||^2}{\partial \hat{\boldsymbol{s}}} = \frac{\tau}{||\tilde{\boldsymbol{e}}||^2} \frac{\partial ||\tilde{\boldsymbol{e}}||^2}{\partial \hat{\boldsymbol{s}}} \tag{5}$$

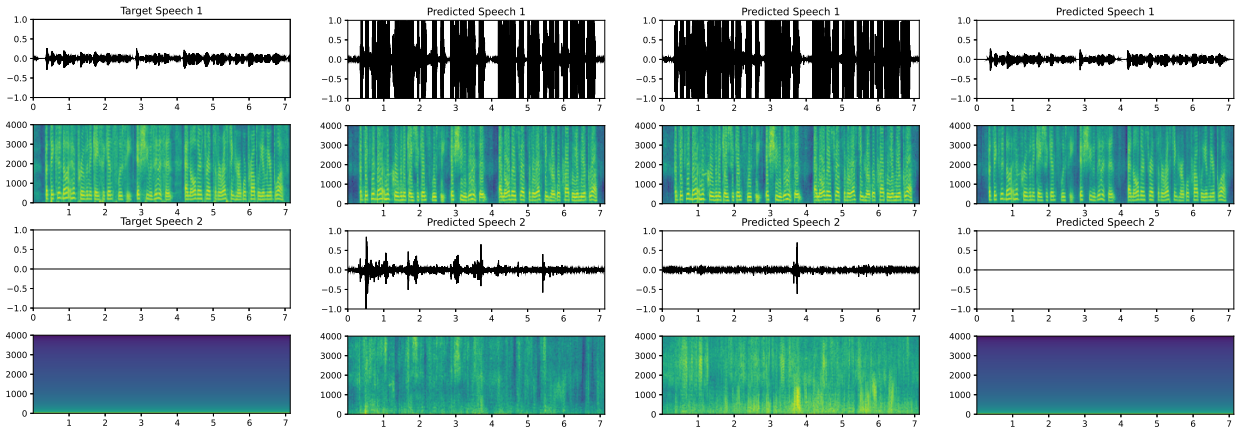
where τ denotes $\frac{10}{\ln 10}$ and $\frac{\partial ||\tilde{\boldsymbol{e}}||^2}{\partial \hat{\boldsymbol{s}}}$ is calculated as:

$$\frac{\partial ||\tilde{\boldsymbol{e}}||^2}{\partial \hat{\boldsymbol{s}}} = \frac{\partial \left(\frac{\boldsymbol{s}^T \hat{\boldsymbol{s}}}{||\boldsymbol{s}||^2}\right)^2 \boldsymbol{s}^T \boldsymbol{s} - 2 \frac{\boldsymbol{s}^T \hat{\boldsymbol{s}}}{||\boldsymbol{s}||^2} \boldsymbol{s}^T \hat{\boldsymbol{s}} + \hat{\boldsymbol{s}}^T \hat{\boldsymbol{s}}}{\partial \hat{\boldsymbol{s}}} \tag{6}$$

$$= 2 \frac{\boldsymbol{s}^T \hat{\boldsymbol{s}}}{||\boldsymbol{s}||^2} \boldsymbol{s} - 4 \frac{\boldsymbol{s}^T \hat{\boldsymbol{s}}}{||\boldsymbol{s}||^2} \boldsymbol{s} + 2 \hat{\boldsymbol{s}} \tag{7}$$

$$= -2 \frac{\boldsymbol{s}^T \hat{\boldsymbol{s}}}{||\boldsymbol{s}||^2} \boldsymbol{s} + 2 \hat{\boldsymbol{s}} = -2 \tilde{\boldsymbol{s}} + 2 \hat{\boldsymbol{s}}. \tag{8}$$

In the manner of gradient descent, (8) just tries to minimize the distance between $\tilde{\boldsymbol{s}}$ and $\hat{\boldsymbol{s}}$. However, the second term on the



(a) Target speech (b) SI-SNR (c) SI-SNR + RSM (d) SNR + RSM

Figure 3: *Comparisons on output waveforms and spectrograms.*

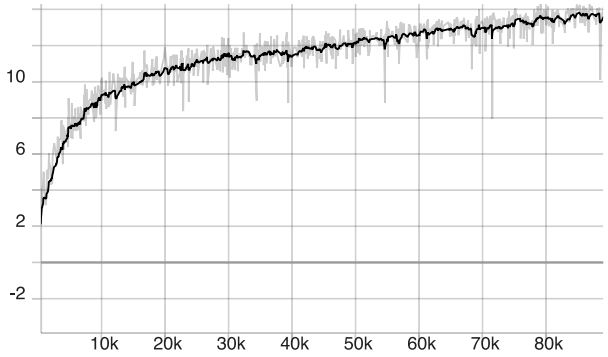


Figure 4: *Output-to-target power ratio curve of the SI-SNR in dB scale while training. Transparent and thick line indicate raw and smoothed curve with a factor of 0.9 for a better observation.*

right-hand side of (4) works in a different way when we observe its gradient which is given by

$$-\frac{\partial 10\log_{10}||\tilde{\boldsymbol{s}}||^2}{\partial\hat{\boldsymbol{s}}} = -\frac{\partial 10\log_{10}||\frac{\boldsymbol{s}^T\hat{\boldsymbol{s}}}{||\boldsymbol{s}||^2}\boldsymbol{s}||^2}{\partial\hat{\boldsymbol{s}}} \tag{9}$$

$$= -\frac{\partial 10\log_{10}\left[\left(\frac{1}{||\boldsymbol{s}||^2}\boldsymbol{s}^T\hat{\boldsymbol{s}}\right)^2||\boldsymbol{s}||^2\right]}{\partial\hat{\boldsymbol{s}}}. \tag{10}$$

By omitting the term independent of $\hat{\boldsymbol{s}}$, (10) can be simplified into

$$-\frac{\partial 10\log_{10}(\boldsymbol{s}^T\hat{\boldsymbol{s}})^2}{\partial\hat{\boldsymbol{s}}} = \frac{-2\tau}{\boldsymbol{s}^T\hat{\boldsymbol{s}}}\boldsymbol{s}. \tag{11}$$

As shown in (11), the error signal obtained from $-10\log_{10}||\tilde{\boldsymbol{s}}||^2$ has no proper relation with $\hat{\boldsymbol{s}}$ except for its scale $\frac{-2\tau}{\boldsymbol{s}^T\hat{\boldsymbol{s}}}$. For more detailed explanation, the error signals of (4) in descent direction is shown in Figure 2. As can be seen in Figure 2, the error signal of $10\log_{10}||\tilde{\boldsymbol{e}}||^2$ makes $\hat{\boldsymbol{s}}$ get close to $\tilde{\boldsymbol{s}}$. Contrarily, the error signal of $-10\log_{10}||\tilde{\boldsymbol{s}}||$ continuously amplify the $\hat{\boldsymbol{s}}$ in the direction of $\boldsymbol{s}$. This power amplification problem disrupts muting the invalid outputs and we demonstrate this experimentally.

2.4. Power-Preserving Learning Criterion

To avoid the power amplification problem, we consider the simple signal-to-noise ratio (SNR) learning criterion given by

$$\mathcal{L}_{\text{SNR}} = -10\log_{10}\frac{||\boldsymbol{s}||^2}{||\boldsymbol{e}||^2} \tag{12}$$

where $\boldsymbol{e} = \boldsymbol{s} - \hat{\boldsymbol{s}}$. As shown in (12), the SNR does not modify the power of the target signal $\boldsymbol{s}$ and only makes $\hat{\boldsymbol{s}}$ get close to $\boldsymbol{s}$. Thus, to train a mute-expressive speech separation model, we apply the SNR with the RSM scheme.

3. Experiments

3.1. Experimental Setup

We perform experiments on one-, two-, and three-speaker speech separation on LibriMix, an open source dataset for clean and noisy speech separation [23]. Training mixture is constructed by `train-clean-360` data in Librispeech [24]. 50,800 and 33,900 utterances are used for training our separation models with two- and three-channels. For testing, 2,000 utterances are used for one-speaker and 3,000 utterances are used for each two- and three-speaker mixture, respectively. We use 8kHz wave data. For detailed information on data configuration, please refer to [23].

All experiments are conducted on the SpeechBrain toolkit [25]. We use pre-trained SepFormer-based E2E-SS models also provided in SpeechBrain, which were trained by WSJ0-2 and -3mix [26]. The SepFormer models have 2 blocks of dual path module with 8 layers for each intra- and inter-transformer. A kernel size of 16 and 50% stride are used for 1-D CNN-based encoders and transposed 1-D CNN-based decoders. For masking operation, ReLU activation is used. We use an Adam optimizer and initial learning rate is set to 1e-4 which is further controlled by the `ReduceLROnPlateau` scheduler with a patient factor of 4 and a decaying factor of 0.5. We utilize the utterance-level permutation invariant training (u-PIT) scheme [20, 27]. For the RSM, ρ and β are set to 0.7 and 1e-7, respectively.

Table 1: *Comparison of SI-SNR / SI-SNRi performance on clean speech condition.*

	Number of speakers in a mixture		
	1-speaker	2-speaker	3-speaker
2-channel			
SI-SNR	8.32 / -88.91	19.97 / 19.97	–
3-channel			
SI-SNR	22.12 / -75.11	15.25 / 15.25	**14.87 / 18.27**
+ RSM	50.38 / -46.85	18.27 / 18.27	14.71 / 18.11
SNR	20.56 / -76.67	15.47 / 15.47	14.18 / 17.57
+ RSM	**59.27 / -37.96**	**19.43 / 19.43**	14.69 / 18.09

Table 2: *Comparison of SI-SNR / SI-SNRi performance on noisy speech condition.*

	Number of speakers in a mixture		
	1-speaker	2-speaker	3-speaker
2-channel			
SI-SNR	10.86 / 7.92	13.09 / 15.09	–
3-channel			
SI-SNR	11.88 / 8.94	11.15 / 13.14	**10.65 / 15.06**
+ RSM	13.88 / 10.95	12.37 / 14.36	10.42 / 14.83
SNR	10.95 / 8.01	10.48 / 12.47	10.10 / 14.51
+ RSM	**14.71 / 11.77**	**13.10 / 15.09**	10.55 / 14.96

3.2. Results

3.2.1. Power Amplification Problem

We first demonstrate the power amplification problem of the SI-SNR learning criterion and show how the SNR with the RSM can maintain the power of the output speech and mute the invalid outputs. In Figure 3, we compare output waveforms and spectrograms to show the effectiveness of our training methodology. For inference, note that no small noises are added to a silence target signal and only a small constant of 1e-9 is added for proper spectrogram visualization. As shown in Figure 3 (b) and (c), it can be seen that the outputs are significantly amplified, which require an additional normalization process. Also, it is noticeable that unexpected signals are generated on the channels which need to be muted. Only with the SNR and the RSM together, the power of output speech is close to the target speech and the invalid output is muted. Figure 4 shows output-to-target power ratio in dB scale. As described in Section 2.3, the power of the output speech based on the SI-SNR is rapidly growing in the early stage of training and continuously increased.

3.2.2. Speech Separation Results

In this section, we compare the speech separation performances on clean and noisy speech mixtures in terms of the SI-SNR and the SI-SNR improvement (SI-SNRi) [7, 8]. As shown in Table 1 and 2, it is obvious that the SNR with the RSM shows better performance than the SI-SNR with the RSM when the number of speakers and the number of channels are mismatched. In case of matched conditions, it is also observable that the SNR with the RSM shows marginal performance losses compared to the SI-SNR. Interestingly, in spite of separating single-speaker mixture, the E2E-SS models can cause severe signal distortion without the RSM as indicated by its low SI-SNR. By applying the RSM, we can see that it is possible to fully cover varying number of speakers by the stand-alone E2E-SS model.

Table 3: *Confusion matrix of invalid output detection rates (%) for clean speech mixture. The numbers in parentheses indicate the number of utterances.*

Prediction	Number of invalid outputs (oracle)		
	0	1	2
0	100.0 (3000)	1.8 (53)	0.0 (0)
1	0.0 (0)	98.2 (2947)	1.7 (34)
2	0.0 (0)	0.0 (0)	98.3 (1966)

Table 4: *Confusion matrix of invalid output detection rates (%) for noisy speech mixture. The numbers in parentheses indicate the number of utterances.*

Prediction	Number of invalid outputs (oracle)		
	0	1	2
0	99.9 (2999)	4.3 (129)	0.0 (0)
1	0.1 (1)	95.7 (2871)	2.4 (48)
2	0.0 (0)	0.0 (0)	97.6 (1952)

3.2.3. Invalid Output Detection

In this section, we show experimental results about how accurately the SNR with the RSM can mute the invalid outputs to filter out them. In this experiment, a very simple condition is used to predict the number of invalid outputs based on whether the power of each output speech is exactly zero or not. As shown in Table 3 and 4, despite the simple condition, the detection results are quite accurate. These results show that the invalid outputs can be properly filtered out when the numbers of speakers and channels are mismatched without considering any specific thresholds. While a fair comparison is difficult, the author of [19] reported detection results using an SI-SNR based threshold and compared to this, our results show overall improvements in terms of detection rates without the need for the process of determining and applying a threshold. In addition, there are two types of errors on the invalid output detection. The first type is that the separation model falsely detect the invalid outputs as valid and in this case, since the falsely detected outputs has significantly low energies, it can be regarded as a minor problem. However, in the opposite case, the problem is critical because it means that we might lose real speech information. Surprisingly, unlike the results of [19], there is only one error for the latter case as shown in Table 4. In a single-speaker mixture case (the oracle number of invalid outputs = 2), it is also noticeable that our experimental results show high accuracy although the detection almost always fails in [19].

4. Conclusion

In this paper, we investigate the training methodology for mute-expressive E2E-SS. We introduce the on-the-fly data mixing scheme, analyze the power amplification problem of the SI-SNR, and apply the SNR that makes the power of output speech close to the target speech. With the SNR and the RSM, experimental results show that the SepFormer-based E2E-SS can properly mute the invalid outputs and the separation performance can be improved when the numbers of speakers and channels are mismatched while incurring marginal losses for the matched condition. In our future work, we will further investigate the performance of our training methodology with various E2E-SS models having over 3 channels.

5. References

[1] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2014, pp. 1562–1566.

[2] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2014, pp. 3734–3738.

[3] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, p. 1849–1858, 2014.

[4] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, p. 483–492, 2016.

[5] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 6–10.

[6] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for endto-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[7] 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. (ICASSP)*, 2018, pp. 696–700.

[8] ——, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, p. 1256–1266, 2019.

[9] M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Filterbank design for end-to-end speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 6364–6368.

[10] 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. (ICASSP)*, 2020, pp. 46–50.

[11] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech-separation," in *Proc. Interspeech*, 2020, pp. 2642–2646.

[12] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, p. 2840–2849, 2021.

[13] Z. Zhang, B. He, and Z. Zhang, "TransMask: A compact and fast speech separation model based on transformer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 5764–5768.

[14] C. Li, Y. Luo, C. Han, J. Li, T. Yoshioka, T. Zhou, M. Delcroix, K. Kinoshita, C. Boeddeker, Y. Qian, S. Watanabe, and Z. Chen, "Dual-path RNN for long recording speech separation," in *Proc. of IEEE Spoken Lang. Tech. Workshop (SLT)*, 2021, pp. 865–872.

[15] N. Takahashi, S. Parthasaarathy, N. Goswami, and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," in *Proc. Interspeech*, 2019, pp. 1348–1352.

[16] 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. (ICASSP)*, 2018, pp. 5064–5068.

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

[18] S. R. Chetupalli and E. Habets, "Speech separation for an unknown number of speakers using transformers with encoder-decoder attractors," in *Proc. Interspeech*, 2022, pp. 5393–5397.

[19] Y.Luo, *End-to-end speech separation with neural networks*. Ph.D. disseratation, Columbia University, 2021.

[20] M. Kolbæk, Z.-H. T. D. Yu, 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, p. 1901–1913, 2017.

[21] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2021, pp. 21–25.

[22] 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. (ICASSP)*, 2019, pp. 626–630.

[23] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speechseparation," *arXiv preprint arXiv:2005.11262*, 2020.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2015, pp. 5606–5610.

[25] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, and J. Z. *et al.*, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[26] 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. (ICASSP)*, 2016, pp. 31–35.

[27] 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 Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2017, pp. 241–245.