Speaker Attractor Network: Generalizing Speech Separation to Unseen Numbers of Sources

Fei Jiang and Zhiyao Duan, *Member, IEEE*

***Abstract*—Most existing speech separation research focuses on improving the separation performance under consistent source number conditions between training and testing. In real-world applications, however, the source number may be different from that in training sets. In this letter, we address this problem by thoroughly improving the deep attractor network in terms of the network architecture and learning objectives so that it can well generalize to separating an unseen number of sources. Experimental results show that, compared with existing models, the proposed method significantly improves the separation performance when generalizing to an unseen number of speakers, and can separate up to five speakers even the model is only trained on two-speaker mixtures.**

***Index Terms*—Speech separation, unseen numbers of sources, deep clustering, speaker attractor.**

I. Introduction

SPEECH separation is an important task in machine listening with a wide range of applications. Unlike speech enhancement, which only aims to separate a target speaker's voice from the mixture, speech separation needs to separate voices of multiple speakers at the same time. This leads to two significant problems: the *permutation* problem [1] and the *output dimension mismatch* problem. The permutation problem refers to the permutation error of the mapping between separated voices and speaker labels. Two effective approaches have been proposed to address this problem: deep clustering (DC) [1] and permutation invariant training (PIT) [2], [3]. In particular, PIT has been widely adopted in a variety of state-of-the-art speech separation models [4]–[16]. The *output dimension mismatch* problem refers to the mismatch on the number of speakers between training and inference, e.g., training on two-speaker mixtures but testing on three-speaker mixtures or mixtures with a varying number of speakers. The above-mentioned PIT-based methods cannot directly deal with this problem due to their fixed output dimension.

Manuscript received August 17, 2020; revised September 30, 2020; accepted October 1, 2020. Date of publication October 8, 2020; date of current version October 30, 2020. This work was supported in part by the National Science Foundation under Grant 1617107 and Grant 1741472, and in part by the China Scholarship Council (CSC) under Grant 201906030175. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Odette Scharenborg. *(Corresponding author: Zhiyao Duan.)*

Fei Jiang is with the Beijing Institute of Technology, Beijing 100811, China, and also with the University of Rochester, Rochester, NY 14627 USA (e-mail: flyjiang92@gmail.com).

Zhiyao Duan is with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627 USA (e-mail: zhiyao.duan@rochester.edu).

Digital Object Identifier 10.1109/LSP.2020.3029704

Recently, two approaches have been proposed to address the output dimension mismatch problem. One approach is to simply assume a maximum number of sources in the mixture and always output this number of sources during separation [3], [16]–[18]. If a mixture contains less sources than the preset maximum, the model is trained to output either silence [3], [16], [17] or the mixture itself [18] at the redundant output channels, which can then be discarded by evaluating the energy level of the outputs relative to the mixture. Clearly, the choice of this maximum number is critical: being too small limits the applicable scenarios, while being too big may cause extra sources in the separation result. Another approach is to extract speech in a recursive manner [19]–[23], i.e., separating one speaker in each iteration until no speech is left in the residual. In [21], it is shown that this method, after being trained on both two-speaker and three-speaker mixtures, is able to generalize to mixtures with a higher number of sources. However, the iteration termination criteria are not easy to set, and the separation performance decreases in later iterations [21], due to the increasing difficulty of the speakers and the corruptions introduced in earlier iterations.

In addition to these two approaches, theoretically speaking, the deep clustering framework also has the potential to tackle the output dimension mismatch problem. However, all of the previous DC-based models [1], [6], [7], [24], including the deep attractor network (DANet) [25] and its variants [26]–[30], only consider the performance of two-speaker or three-speaker separation. On the one hand, no attempt has been made to separate mixtures with more than three speakers. On the other hand, the source number in the test set is always consistent with that in the training set in the literature [6], [7], [25]–[30]. When generalizing such DC-based models trained on two-speaker mixtures to separating mixtures with three or more speakers, the performance degrades significantly [1], [24].

In this letter, we propose a new speech separation model named speaker attractor network (SANet) to improve the separation performance on mixtures with an unseen number of sources. It can be viewed as a thoroughly improved version of DANet [25] along several aspects. The key idea is to learn time-frequency (T-F) embeddings that show clustering effects among speakers in the same mixture and consistent positioning for the same speaker across mixtures. Specifically, we propose to combine three training objectives under the DC framework: 1) good reconstruction of sources, 2) compact distribution of T-F embeddings of each speaker in a mixture, and 3) good speaker discrimination among speaker attractors (i.e., average T-F embeddings for different speakers) across mixtures. In addition, inspired by the two-stage TasNet [15], we replace the magnitude spectrogram input in the original DC and DANet framework with a pre-trained 1-d

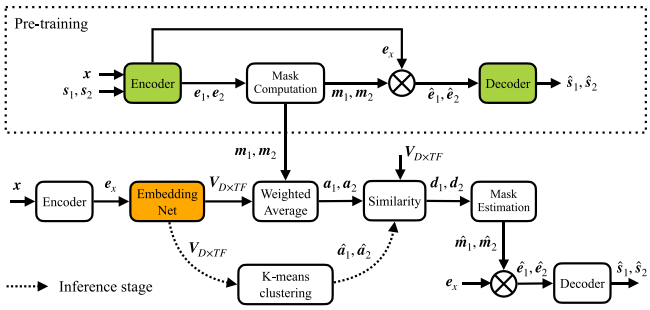

Fig. 1. The diagram of SANet. It adopts a two-stage training strategy: pre-training the encoder-decoder and then solely training the embedding network. Colored blocks have trainable parameters, while white blocks do not.

convolutional encoder-decoder that takes the raw waveform as input, which significantly improves the overall separation performance. Experimental results show that the proposed SANet significantly improves the separation performance on mixtures with an unseen number of sources over state-of-the-art methods, including DC-based methods and recursive separation methods.

The main contributions of this letter are threefold. First, to our best knowledge, this is the first work to improve the generalization ability of DC framework in terms of separating an unseen number of sources. Second, a new set of embedding learning objectives is proposed for the DC framework. Third, for the first time, the speech separation model is generalized to cope with up to five sources even the model is only trained on two-speaker mixtures, showing better generalization ability than the state of the art (e.g., recursive separation method [21]).

II. Speaker Attractor Network

As shown in Fig. 1, the proposed SANet consists of a pre-trained encoder-decoder and an embedding network, followed by clustering and mask estimation modules. The encoder transforms the speech waveform into a latent space, which can be viewed as a T-F representation. The embedding network then takes the mixture T-F representation as input, and outputs embedding vectors for each T-F unit. Then attractors of each source are computed by mask-weighted average of the embeddings in the training phase and approximated by k-means centroids of the embeddings in the test phase. By comparing the similarity between embeddings and attractors, masks for the sources can be estimated. The masks are then multiplied with the mixture T-F representation to compute the source T-F representations, which are passed to the decoder to reconstruct the source waveforms. It is noted that the encoder-decoder is frozen during the training of the embedding network.

A. Speech Encoder-Decoder

We use a 1-d convolutional layer followed by a ReLU activation function as the speech encoder in the time domain, and a 1-d transposed convolutional layer without any nonlinear activation as the decoder. The kernel size, stride, and number of channels are 16, 8, 128, respectively. Compared to the magnitude spectrogram domain, encoding and decoding in the time domain avoids the phase reconstruction issue, and is adopted by many recent source separation methods [8]–[15].

We first train this encoder-decoder as in [15]. Let $\boldsymbol{x}$ be the additive mixture of C clean sources, $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_C$, i.e., $\boldsymbol{x} = \sum_{i=1}^{C} \boldsymbol{s}_i$. The T-F representations of the mixture and the sources calculated by the encoder are $\boldsymbol{e}_x, \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_C \in \mathbb{R}^{TF}$, respectively. Then the ideal mask $\boldsymbol{m}_i \in \mathbb{R}^{TF}$ for the i-th source can be defined as

$$\boldsymbol{m}_i = \boldsymbol{e}_i \oslash \sum_{j=1}^{C} \boldsymbol{e}_j, \tag{1}$$

where $\oslash$ denotes element-wise division. The estimated T-F representation of the i-th source can be obtained by $\hat{\boldsymbol{e}}_i = \boldsymbol{m}_i \odot \boldsymbol{e}_x$, where $\odot$ is element-wise multiplication. The reconstructed source $\hat{\boldsymbol{s}}_i$ is finally obtained through decoding $\hat{\boldsymbol{e}}_i$ by the decoder. We adopt the scale-invariant signal-to-distortion ratio (SI-SDR) [31] between the clean source $\boldsymbol{s}$ and the estimated source $\hat{\boldsymbol{s}}$ as the training objective of this encoder-decoder.

B. Embedding Network

The embedding network takes the mixture T-F representation $\boldsymbol{e}_x$ as input, and outputs an embedding matrix $\boldsymbol{V} \in \mathbb{R}^{D \times TF}$, where each column corresponds to each T-F unit of $\boldsymbol{e}_x$. We adopt a temporal convolutional network (TCN) as the embedding network in our model. Its configuration is the same as the TCN used in Conv-TasNet [9], except that the last 1×1 convolutional layer is an embedding layer with $D \times TF$ output channels rather than a mask regression layer with $C \times TF$ output channels. We do not use any activation function in the embedding layer, and we normalize the embeddings $\boldsymbol{v}_i \in \mathbb{R}^D$ to have unit norm, i.e., $\|\boldsymbol{v}_i\|_2 = 1$. Note that this is different from DANet [25]–[30], which applies a tanh activation to the embedding layer and the embeddings are not normalized.

In the training phase, the attractor of the i-th source, lying on the unit sphere in $\mathbb{R}^D$, is computed as

$$\boldsymbol{a}_i = \frac{\boldsymbol{V} \cdot (\boldsymbol{w} \odot \boldsymbol{m}_i)}{\|\boldsymbol{V} \cdot (\boldsymbol{w} \odot \boldsymbol{m}_i)\|_2}, \tag{2}$$

where $\boldsymbol{w} = \boldsymbol{e}_x / \|\boldsymbol{e}_x\|_1$ is the weighting factor for each T-F unit, and $\boldsymbol{m}_i$ is the ideal mask obtained by Eq. (1). In the test phase, the attractors are approximated by the spherical k-means clustering [32] centers of the embeddings. With the attractors, we then compute the similarities between embeddings and attractors using cosine similarity as

$$\boldsymbol{d}_i = \boldsymbol{V}^T \boldsymbol{a}_i \in \mathbb{R}^{TF}, \tag{3}$$

and the mask for the i-th source can be estimated by

$$\hat{\boldsymbol{m}}_i = \exp(\alpha \boldsymbol{d}_i) \oslash \sum_{j=1}^{C} \exp(\alpha \boldsymbol{d}_j), \tag{4}$$

where α is a scale factor empirically set to 10 to control the hardness of the mask assignment. With the estimated masks we can estimate the T-F representation of each source, and finally reconstruct the source using the decoder.

Because the spherical k-means in our attractor estimation also uses cosine similarity, it is consistent with the similarity measure for embedding assignment in Eq. (3). In contrast, DANet uses dot product in embedding-attractor similarity calculation but uses Euclidean distance for attractor calculation through

k-means clustering, making the final embedding assignment suboptimal [27]. Also note that the mask computation in Eq. (4) is different from that in DANet, where either sigmoid function or softmax without the scale factor is used to obtain the estimated mask $\hat{\boldsymbol{m}}_i$. We do not use the former because the cosine similarities between T-F bins and non-target attractors in SANet are all generally larger than 0. We do not use the latter because $\boldsymbol{d}_i \in [-1, 1]$ in Eq. (3) would lead to a very limited mask range.

C. Model Training Objectives

An ideal embedding space of the deep clustering framework should meet: embeddings of the same source are close while embeddings of different sources are far apart. Both the deep clustering loss and DANet are proposed for this purpose. However, this still cannot be guaranteed when the source number in a mixture differs from that in training data. We propose to extend the embedding learning objective to

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{spk} + \lambda_2 \mathcal{L}_{com}, \tag{5}$$

which consists of the *reconstruction loss*, the *speaker loss* and the *compactness loss*, with λ_1 and λ_2 controlling the weights. The loss terms are explained in the following.

1) Speech Reconstruction: Unlike DANet which optimizes the reconstruction of the magnitude spectrogram, we directly optimize the reconstruction of the source waveform. We use the negative SI-SDR between estimated sources and the clean sources as the reconstruction loss:

$$\mathcal{L}_{rec} = -\text{SI-SDR}(\hat{\boldsymbol{s}}, \boldsymbol{s}). \tag{6}$$

2) Speaker Discriminative Attractors: In most previous DC-based models, the attractors or cluster centroids of different speakers in a mixture are trained to be far apart. They are, however, not speaker-discriminative across mixtures, as the positioning of attractors of the same speaker is not consistent across mixtures. This inconsistent positioning makes it difficult to generalize the speech separation model to an unseen number of speakers. In [30], the attractors are trained to be speaker-discriminative across mixtures for the first time, making the generalization possible. However, cross-entropy loss is used and the model is prone to overfitting to speakers in the training set, and the model cannot generalize to an unseen number of speakers. In this letter, we use metric learning instead: We define two attractors of the same speaker in different mixtures as a positive pair, and two attractors of different speakers as a negative pair, and adopt the circle loss [33] defined as

$$\mathcal{L}_{spk} = \log \left[1 + \sum_{i=1}^{K_p} \sum_{j=1}^{K_n} \exp(\gamma(\alpha_n^j (s_n^j - \Delta_n) - \alpha_p^i (s_p^i - \Delta_p))) \right], \tag{7}$$

where K_p and K_n are the number of positive pairs and negative pairs in one mini-batch, respectively, s_p and s_n are their corresponding cosine similarity, and Δ_n and Δ_p are their corresponding margins, α_p^j and α_n^i are adaptive weighting factors that vary with s_p and s_n, and γ is a scale factor. More details can be found in [33].

3) Compact Embeddings: Only using $\mathcal{L}_{rec}$ and $\mathcal{L}_{spk}$ is not sufficient for our model. The reconstruction loss $\mathcal{L}_{rec}$ is only related to the estimated mask $\hat{\boldsymbol{m}}$, which is derived from Eq. (4), based on the relative differences of embedding-attractor similarities $\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_C$. This relative similarity difference does not necessarily lead to a compact distribution of embedding vectors of the same speaker nor a robust estimate of attractors during the test phase. The speaker loss $\mathcal{L}_{spk}$ only ensures that the centroids of the embeddings of different speakers are far apart, but no constraints are imposed upon the variance of embeddings of each speaker, i.e., embeddings of different speakers may overlap. Therefore, compact embedding distributions for each speaker is very important for ensuring its embeddings are close to each other. To this end, we propose the following compactness loss:

$$\mathcal{L}_{com} = -\sum_{j=1}^{TF} w_j m_k^j \boldsymbol{v}_j^T \boldsymbol{a}_k, \tag{8}$$

where w_j is the same weight as in Eq. (2). m_k^j is the j-th T-F unit's mask for the k-th speaker, where $k = \arg\max_{i \in \{1,2,\ldots,C\}} |e_i^j|$, i.e, the dominating speaker for the j-th T-F unit. In other words, we force the T-F embedding to be close to only the attractor of the dominant speaker instead of all speakers, which is experimentally found to perform better.

III. Experiments

A. Experimental Setup

The LibriMix [34], derived from the LibriSpeech corpus [35], is used to evaluate the performance of SANet.[1] It consists of two main subsets, Libri2Mix and Libri3Mix, which are two-speaker and three-speaker mixtures, respectively. Each of the two subsets contains two training sets (`train-100`, `train-360`), one validation set (`dev`), and one test set (`test`). It also contains a Sparse3Mix dataset, a sparsely overlapping versions of Libri3Mix test set, which is used to simulate more realistic, conversation-like scenarios. We use the `train-100` sets for training and the `dev` sets for validation in the following experiments. As for the test sets, we use the `test` sets in Libri2Mix and Libri3Mix, and Sparse3Mix, Sparse4mix, Sparse5mix. Sparse4mix and Sparse5mix are generated following the same recipe of generating Sparse3Mix [34]. Regarding overlap ratios, we find the definition used in [34] less intuitive, and propose to calculate it regarding different overlapping source numbers as L_n/L_{total}, where L_n is the total length of n-speaker overlapping clips in the mixture and L_{total} is the total length of the mixture. With this definition, the average overlap ratios of the three sparse LibriMix sets are: Sparse3Mix - 40% (2-speaker) and 6% (3); Sparse4Mix - 22% (2), 24% (3), and 8% (4); Sparse5Mix - 18% (2), 13% (3), 17% (4), and 3% (5). We select the 8 kHz *min* mode of all these data sets. The training sets contain 251 different speakers in total, while the test sets contain 40 unseen speakers.

The encoder-decoder is pre-trained as illustrated in Section II-A. It is frozen while training the embedding network. Both stages are trained on 2-second long segments. The embedding dimension is 32. The learning rate is initialized to 0.001 and then halved if no best validation model is found in 6 consecutive

[1] Code available at https://github.com/fjiang9/sanet

TABLE I
SI-SDR IMPROVEMENT (dB) ON DIFFERENT TEST SETS. MODELS IN THE UPPER BLOCK ARE TRAINED ON 2-SPEAKER MIXTURES, WHILE THOSE IN THE LOWER BLOCK ARE TRAINED ON 2- & 3-SPEAKER MIXTURES

Training Set	Method	Test Set					Aver. on Unseen
		Libri2Mix	Libri3Mix	Sparse3Mix	Sparse4Mix	Sparse5Mix	
Libri2Mix	DPCL [1]	8.8	1.1	2.6	0.2	-1.3	0.7
	DANet [25]	8.4	2.0	3.4	2.0	1.2	2.2
	Conv-TasNet (Light) [9]	**13.5**	-	-	-	-	-
	Conv-DANet	12.4	3.4	4.6	2.4	1.6	3.0
	SANet (Proposed)	12.0	**7.3**	**12.6**	**10.1**	**8.2**	**9.6**
	SANet (w/o $\mathcal{L}_{spk}$)	12.8	4.0	7.0	4.9	3.8	4.9
	SANet (w/o $\mathcal{L}_{com}$)	0.4	0.9	2.7	2.3	2.2	2.0
Libri2Mix & Libri3Mix	Conv-TasNet-OR-PIT [21]	**13.6**	**12.4**	14.7	12.0	9.8	10.9
	SANet (Proposed)	12.9	11.3	**15.3**	**13.0**	**10.8**	**11.9**

epochs. The total number of training epochs is 200, and the training is early stopped if no improvement on the validation set is observed for 20 epochs.

B. Results

We compare the proposed SANet with two DC-based models, DPCL [1] and DANet [25], and Conv-TasNet [9]. For a fair comparison, we implement a light Conv-TasNet using the same number of channels in the encoder-decoder as SANet. We also build a baseline model named Conv-DANet using the same similarity computation as the original DANet [25], but replace its loss function with $\mathcal{L}_{rec}$ (Eq. (6)), encoder-decoder and embedding network architecture as those in the proposed SANet, to reduce the architectural effects in the comparison. In addition, we further train two models, SANet (w/o $\mathcal{L}_{spk}$) and SANet (w/o $\mathcal{L}_{com}$), to perform an ablation study regarding the training objectives of SANet. All of these models are trained solely on the Libri2Mix. Weights between different loss functions are tuned using the validation sets in Libri2Mix and Libri3Mix. We assume the source number in the mixture is known during testing. We evaluate the separation performance by the SI-SDR improvement from the unprocessed mixture.

The separation results on different test sets are shown in the top block of Table I. Several interesting observations can be made. First, compared to the four baselines, the proposed SANet achieves significantly better results on mixtures with three or more (i.e., unseen numbers of) speakers, even though it slightly underperforms Conv-DANet and Conv-TasNet on two-speaker mixtures. This shows the superior power of generalization to unseen numbers of speakers. In particular, the improvement from the best Conv-DANet is 6.6 dB on average. A main reason for this improvement is better discrimination of attractors of different speakers. Fig. 2 visualizes the estimated attractors of 3000 Libri2Mix test mixtures in Conv-DANet and SANet using t-SNE [36]. We can see that the estimated attractors of different speakers in SANet are much better grouped.

Second, we can see that for SANet, without $\mathcal{L}_{spk}$, the two-speaker separation result is slightly better but the result on mixtures with unseen numbers of speakers degrades significantly. Without $\mathcal{L}_{com}$, the model almost fails to separate mixtures in the test phase. This supports our analyses in Section II-C, and shows the effectiveness of combining the three training objectives. Third, among the three baselines, Conv-DANet improves significantly from DANet across all settings. This suggests the effectiveness of the two-stage training strategy and network architecture in our model.

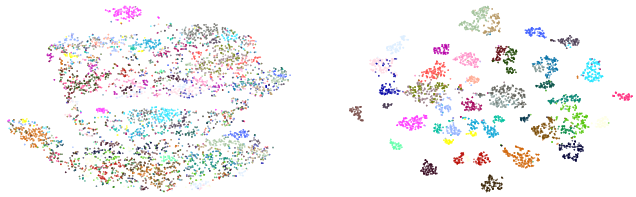

Fig. 2. Estimated attractors (k-means centroids) of test mixtures visualized by t-SNE. Each color represents a speaker. Left: Conv-DANet. Right: SANet.

We further compare SANet with one-and-rest permutation invariant training (OR-PIT) [21], which is a state-of-the-art recursive separation methodology that can be applied to any separation models to generalize to separating unseen numbers of speakers. We use the same Conv-TasNet mentioned before as the separation network. Because OR-PIT needs to be trained on both two-speaker and three-speaker mixtures, we also train another SANet on two- and three-speaker mixtures. The separation results are shown in the lower block of Table I. Two observations can be made: 1) Comparing with OR-PIT, the proposed SANet can better generalize to separating unseen numbers of speakers, while maintaining comparable performance on two- and three-speaker separation. 2) SANet also improves 0.6 dB from OR-PIT on Sparse3Mix, suggesting better robustness to instantaneous speaker number variations within a mixture.

IV. CONCLUSION

We proposed a new speech separation model that can generalize to separating mixtures with unseen numbers of speakers. The proposed method thoroughly improves the deep attractor network in terms of the network architecture and embedding learning objectives, and it results in outstanding performance compared to the state of the art when separating mixtures with unseen numbers of speakers. As the source number in the mixture is assumed to be known during the inference phase in this letter, we leave the speaker counting under our framework for future work. Future work also includes generalizing the model to separating mixtures in noisy conditions.

REFERENCES

[1] J. R. Hershey, Z. Chen, J. Le 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.

[2] D. Yu, M. Kolbak, Z. 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.*, 2017, pp. 241–245.

[3] M. Kolbak, 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.

[4] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "CBLDNN-based speaker-independent speech separation via generative adversarial training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 711–715.

[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.*, 2018, pp. 6–10.

[6] Z. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 686–690.

[7] Z.-Q. Wang, J. Le Roux, D. Wang, and J. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, 2018, pp. 2708–2712.

[8] 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.

[9] 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.

[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.*, 2020, pp. 46–50.

[11] Z. Shi, H. Lin, L. Liu, R. Liu, S. Hayakawa, and J. Han, "End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *Proc. Interspeech*, 2019, pp. 4614–4618.

[12] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6359–6363.

[13] D. Ditter and T. Gerkmann, "A multi-phase gammatone filterbank for speech separation via TasNet," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 36–40.

[14] B. Kadioglu, M. Horgan, X. Liu, J. Pons, D. Darcy, and V. Kumar, "An empirical study of Conv-TasNet," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7264–7268.

[15] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 31–35.

[16] 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.

[17] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 2623–2634.

[18] Y. Luo and N. Mesgarani, "Separating varying numbers of sources with auxiliary autoencoding loss," 2020, *arXiv:2003.12326*.

[19] 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. 28th Int. Joint Conf. Artif. Intell.*, 2018, pp. 4353–4360.

[20] 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.

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

[22] T. von Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, and R. Haeb-Umbach, "All-neural online source separation, counting, and diarization for meeting analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 91–95.

[23] T. von Neumann *et al.*, "Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR," 2020, *arXiv:2006.02786*.

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

[25] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 246–250.

[26] 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.

[27] Y. Luo and N. Mesgarani, "Augmented time-frequency mask estimation in cluster-based source separation algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 710–714.

[28] G.-P. Yang, C.-I. Tuan, H.-Y. Lee, and L. shan Lee, "Improved speech separation with time-and-frequency cross-domain joint embedding and clustering," in *Proc. Interspeech*, 2019, pp. 1363–1367.

[29] Y. Xiao and H. Zhang, "Improved source counting and separation for monaural mixture," 2020, *arXiv:2004.00175*.

[30] L. Drude, T. von Neumann, and R. Haeb-Umbach, "Deep attractor networks for speaker re-identification and blind source separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 11–15.

[31] 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.

[32] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Mach. Learn.*, vol. 42, no. 1, pp. 143–175, 2001.

[33] Y. Sun *et al.*, "Circle loss: A unified perspective of pair similarity optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6397–6406.

[34] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation," 2020, *arXiv:2005.11262*.

[35] 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.*, 2015, pp. 5206–5210.

[36] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.