INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Which Ones Are Speaking? Speaker-inferred Model for Multi-talker Speech Separation

Jing Shi[1,2], *Jiaming Xu*[1,*], *Bo Xu*[1,2,3]

[1]Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China
[2]University of Chinese Academy of Sciences, China
[3]Center for Excellence in Brain Science and Intelligence Technology, CAS, China
{shijing2014,jiaming.xu,xubo}@ia.ac.cn

Abstract

Recent deep learning methods have gained noteworthy success in the multi-talker mixed speech separation task, which is also famous known as the Cocktail Party Problem. However, most existing models are well-designed towards some predefined conditions, which make them unable to handle the complex auditory scene automatically, such as a variable and unknown number of speakers in the mixture. In this paper, we propose a speaker-inferred model, based on the flexible and efficient Seq2Seq generation model, to accurately infer the possible speakers and the speech channel of each. Our model is totally end-to-end with several different modules to emphasize and better utilize the information from speakers. Without a priori knowledge about the number of speakers or any additional curriculum training strategy or man-made rules, our method gets comparable performance with those strong baselines.

Index Terms: Speech Separation, Cocktail Party Problem, Generation Model, Attention, Deep Learning

1. Introduction

With the rapid development of electronic devices and artificial intelligence technologies, human-computer speech interaction has become increasingly prominent in recent years. However, the performance of these technologies in open complex environments, such as in cocktail parties, is far from satisfactory. It is still a challenging task to develop a computational auditory system with strong adaptivity and robustness at present.

As one of the most important parts in front-end auditory processing system, speech separation greatly affects or even restricts the performance of whole pathways. Various speech-based technologies and tasks could benefit from speech separation methods, such as automatic meeting transcription, automatic captioning for audio/video recordings, and multi-party human-computer interactions [1]. Therefore, the focus on speech separation in complex auditory environment is worthy and necessary. Since the first description of the so-called Cocktail Party Problem in 1953 [2], many approaches have been proposed. Before the wide popularity of deep learning methods, there were signal processing based methods, rule-based methods and decomposition-based methods [3, 4]. However, none of these could offer a satisfying solution for this complicated task.

In the past few years, deep learning methods have shown great effectiveness and adaptability in speech processing. There are many great works in speaker recognition [5, 6], speech enhancement and denoising [7] and the like. For the task of multi-talker single-channel speech separation, many works have also been presented and shown remarkable progress. Unfortunately, most existing representative works are still well-designed to fit into the ideal setting, making them difficult to conduct the complex and dynamic circumstance. To be specific, there are two common problems when processing a piece of mixed speech: the permutation problem and the output dimension mismatch problem [8]. To prevent the confusion brought by different permutation, Yu et al. [1] proposed a Permutation Invariant Training (PIT) technique to pool over all possible permutations for N mixed sources ($N!$ permutations), and minimize the source reconstruction error no matter how labels are ordered. Although this technique works well when facing a limited number of speech sources, it still suffers from the output dimension mismatch problem because it assumes a fixed number of sources and also suffers from computation efficiency [8]. So as to solve both permutation and output dimension problems, Hershey et al. [9] proposed a Deep Clustering (DPCL) method which first maps the time-frequency units into an embedding space, and afterward creates several groups of time-frequency units by utilizing a clustering algorithm, such as k-means. Following DPCL, Chen et al. [8] proposed a Deep Attractor Network (DANet) which first structures k cluster centers in the embedding space to act as the attractor points. After that, DANet pulls together the time-frequency units according to these attractor pointers. Although DPCL and DANet are sufficiently flexible to handle speech separation of different number of sources in the mixture without retraining, both of them require the cluster number during evaluation. Additionally, some artificial designs have to be adopted to ensure performance such as the weights to shield silent region or emphasize salient parts. Although all the abovementioned works have been improved [10, 11, 12] to enhance performance, the difficulties when facing variable and unknown number of speakers have not been eliminated.

To overcome these disadvantages, we think speaker identities are worthy to be concerned. Actually, for speech separation task with only audio stream, speaker information really matters. First, from our own observation and some evidence from other work [13], we find it is very difficult to perform speech separation with speeches from the same person using only characteristic speech frequencies contained in the audio. This fact means that the speaker information contains some clues to promote speech separation. Second, the accurate prediction about the speakers in mixture speech could directly provide the number of speakers there, enabling to handle dynamic and flexible circumstances, such as a variable number of mixed speakers which is not given in advance. In addition, for application level, users usually want to get the information about not only the separated speeches but also which ones speak what. Based on these considerations, we think the speech separation model integrated

* Corresponding author.

http://dx.doi.org/10.21437/Interspeech.2019-1591

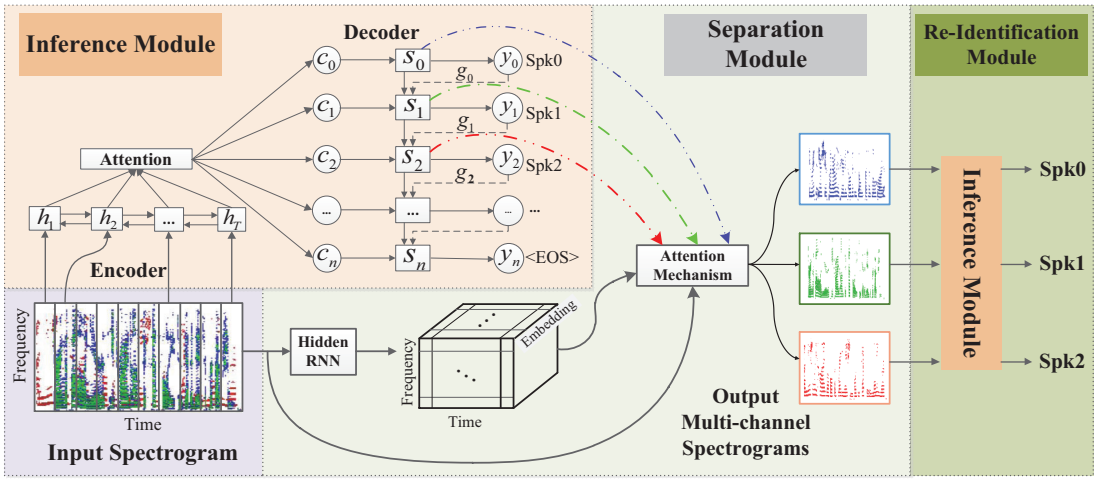

Figure 1: *Illustration of our whole speaker-inferred model. The automatic inference module (top left part) and attention based separation module (center part) are organically combined in our proposed framework to make better use of the speakers' information. Optionally, a Re-Identification module (right part) based on the same network from inference module could be used in training phase to re-infer the speaker for each separated spectrogram respectively.*

with speaker information is helpful.

In this paper, we present a speaker-inferred model to build an end-to-end framework for single-channel speech separation with an alterable and unknown number of speakers. The inference module serves to provide the predicted speakers and also the corresponding hidden embeddings of each to guide the separation module to extract corresponding speech channel. Based on the encoder-decoder generation network, possible candidates are predicted automatically. Compared with existing representative methods like DPCL, DANet, PIT, and others, this work progresses a step further to fit the complex real scene, which does not give the number of speakers in advance. Without any artificial or man-made setting like a threshold for detecting salient region used in [8, 10], our work achieves good performance. Code, models, and video will be available on: https://github.com/shincling/TDAAv2/tree/master/TDAAv2_2019.

2. Related work

Recently, single-channel speech separation task gains much attention. Among those kinds of works, extensive exploration has been done with deep learning methods to make significant progress. For most deep learning based models, there exists a neural network to extract the hidden representation of the input mixture. After getting the hidden representation, a pre-set structure with a fixed number of branches will be used to output several separated channels. As mentioned in the former section, PIT based methods [1, 12] measure similarity between outputted and ground-truth spectrograms to determines the best label assignment by comparing separation errors of all possible orders. In the lasted research, some further works based on PIT have been proposed. For example, grid LSTM [14] has been used to improve the network structure, adversarial training was introduced with a complicated netowrk [15], and the time-domain audio separation network (TasNet) [16, 17] gained very impressive results. However, these works are still built on a strategy to handle the predefined conditions as much as possible. In other words, these designs may overfit the mixture with a fixed number of speakers. Similarly, DPCL [9, 10] and DANet [8, 11] need to know the number of channels to complete the pipeline.

As we proposed in Section 1, a framework taking advantages of the speaker information may avoid this dilemma. To our delight, the community begins to pay attention to the speaker information in cocktail party problem and the corresponding attention mechanism. For instance, speaker identification loss is added to the final loss function to reduce separation and permutation error in [18]. Xu et al. [19] modeled a network with memory module storing the information of each speaker. Moreover, the work done in [20] proposed a source-aware network to use pure speech from each speaker to direct the continuous separation. However, these models need to get additional information on the aim speakers in advance, and none of them gave a solution to directly process the input mixtures.

The work proposed here focuses on the speech separation task, especially the automatic inference about the possible channels. In last year, Shi el at. [21] first raised the Top-Down Auditory Attention Model (TDAA) to fix this problem. Our work takes the advantage of TDAA model, which consists of a bottom-up inference module and top-down attention module. The former one in our work is built with a Sequence-to-Sequence model to accurately predict the candidate speakers. The detail of our framework will be explained in the next section.

3. Speaker-inferred model

Our network can be divided into two organically combined parts: automatic inference module and attention based separation module. An optional module named Re-Identification could also be added in the training phase to further utilize the same network from inference module. The whole framework is illustrated in Figure 1.

3.1. Automatic inference module

Assume the total number of speakers in the training dataset is N. The purpose of this automatic inference module is to predict the possible speakers Spk_ϵ, $\epsilon = 1, .., N$, given one piece of mixed spectrogram χ (transformed from the original mixture speech x by Short-Time Fourier Transform). As the design of the bottom-up inference module from TDAA, this module could be viewed as a task of Multi-Label Classification (MLC). MLC is an important yet challenging task in many areas. In our design, it carries the duty of accurately estimating not only the true number of speakers but also identifying them. In this paper, inspired by the great success of the sequence-to-sequence model in natural language processing related tasks, such as machine

translation [22, 23] and abstractive summarization [24] , we use a sequence generation model from [25] to infer multiple possible speakers in the input mixed speech. The proposed sequence generation model consists of an encoder and a decoder with the attention mechanism.

The whole process of the generation model could be concluded as follows.

Overview A given spectrogram χ, containing T frames and F frequencies in total, is viewed as a sequence of frames. χ is first encoded to the hidden states, which are aggregated to a context vector c_t by the attention mechanism at time-step t in the decoder phase. The decoder takes the context vector c_t, the last hidden state s_{t-1} of the decoder and the embedding vector $g(y_{t-1})$ as the inputs to produce the hidden state s_t at time-step t. Here y_{t-1} is the predicted probability distribution over the label space L [1] at time-step $t-1$. The function g takes y_{t-1} as input and produces the embedding vector which is then passed to the next step of the decoder. Finally, the softmax layer is used to output the probability distribution for each speaker y_t. The whole process stops as soon as the <EOS> label is predicted at one step.

Encoder Due to the strong sequential dependencies of speech signal, we use a bidirectional LSTM [26] to read the spectrogram χ from both directions and compute the hidden states h_i for each frame, which embodies the information of the speech centered around the i-th frame.

Attention Actually, for mixed speech, every speaker gets his specific salient area, most of which is usually not overlapped with others'. This phenomenon could be observed from the mixed and separated spectrogram [8]. For our work, this provides favorable support for the use of attention mechanism over different frames when predicting each speaker at one decoder step. Specially, the attention mechanism assigns the weight α_{t_i} to the i-th frame at time-step t as $\alpha_{t_i} = \mathrm{Softmax}(v_E^T tanh(W_E s_t + U_E h_i))$, where W_E, U_E, v_E are weight parameters and s_t is the current hidden state of the decoder at time-step t. The final context vector c_t which is passed to the decoder at time-step t is calculated as $c_t = \sum_{i=1}^{T} \alpha_{t_i} h_i$.

Decoder The hidden state s_t of the decoder at time-step t is computed as follows:

$$s_t = \mathrm{LSTM}(s_{t-1}, [g(y_{t-1}); c_{t-1}]), \tag{1}$$

where $[g(y_{t-1}); c_{t-1}]$ means the concatenation of the vectors $g(y_{t-1})$ and c_{t-1}. $g(y_{t-1})$ is the embedding of the label which has the highest probability under the distribution y_{t-1}. y_{t-1} is the probability distribution over the label space L at time-step $t-1$ and is computed as follows:

$$y_t = \mathrm{Softmax}(W_{D1} f(W_{D2} s_t + W_{D3} c_t)), \tag{2}$$

where W_{D1}, W_{D2}, W_{D3} are all weight parameters and f is a nonlinear activation function. At the training phase, the loss function in the inference module is the cross-entropy loss function. We employ the beam search algorithm [27] to find the top-ranked prediction path at inference phase.

3.2. Attention based speech separation module

For the mixture χ, the separation module first adopts a Bidirectional Long-Short Term Memory network to map the input into hidden states $H_{\tau,m}$, where the m equals the number of hidden units in this layer. Following that, a feed-forward layer embeds the hidden states to form an embedding matrix whose every time-frequency unit $\theta_{\tau,f} \in \mathbb{R}^d$ implicitly represents features in the hidden states of the unit (τ, f) from the mixture spectrogram.

For each candidate ϵ, the corresponding hidden vector E_ϵ is extracted from every step in the inference module. Different actions have been trialled to serve as the E_ϵ. Usually, we directly take the final hidden states s_t in the decoder as E_ϵ, and we could take these hidden states after the Schmidt orthogonalization to increase the diversity over each step. Moreover, the global embedding (GE) raised in [25] after the prediction at each decoder step is another quite powerful choice. After that, the attention mechanism takes $\theta \in \mathbb{R}^{T \times F \times d}$ and E_ϵ to compute $mask_\epsilon \in \mathbb{R}^{T \times F}$ as estimation of the Ideal Ratio Mask (IRM) for channel ϵ in the mixture. Towards each unit $\theta_{\tau,f} \in \mathbb{R}^d$ and speaker E_ϵ, $\mathrm{Sigmoid}(W_{S1} tanh(W_{S2}\theta_{\tau,f} + W_{S3} E_\epsilon))$ attention mechanism is used, where W_{S1}, W_{S2} and W_{S3} are all learned parameters.

Table 1: *The performance of the inference module. The number after the model name means the number of speakers used in the mixed training dataset. The metrics, HL, P, R, F1, SR2, SR3 denote hamming loss, micro-precision, micro-recall, micro-F1, success rate in 2 mixed test dataset and 3 mixed test data respectively. All the results in this table are in percentage. Except the hamming loss, the higher the other metrics are, the better the model performs. GE here means the hidden vectors using global embedding in the decoder.*

Model	Valid Dataset				Test Dataset	
	HL	P	R	F1	SR2	SR3
TDAA 2	0.59	71.0	**99.1**	82.6	2.8	-
TDAA 2&3	0.41	83.6	95.2	88.9	19.7	7.5
TDAA+Recu	-	-	-	-	87.2	48.3
Our 2	0.36	91.4	91.4	91.4	**100**	-
Our 2+GE	**0.17**	**95.5**	95.5	**95.5**	**100**	-
Our 2&3	0.23	93.8	94.4	94.1	94.8	**85.6**

3.3. Re-Identification Module

In our model, one important characteristic of the inference module lies in the ability using the unified model to predict variable number of candidate speakers, even when there is no speaker or just one speaker in given spectrogram. Based on this characteristic, we could use the same network from inference module to re-identify each outputted speech channel in training phase, without any additional parameter. Through this additional module, the seq2seq model could be trained adequately, along with the whole framework. From the viewpoint of loss function, the loss in this module could be seen as one kind of normalization towards the Mean-Square-Error (MSE) loss in separation module. From the performance in Section 4.3, promotion could be observed with the proposed Re-identification.

4. Experiments

4.1. Dataset and Setup

We evaluated our work on speech mixtures from the Wall Street Journal (WSJ0) corpus. For the condition of two-speaker, the WSJ0-2mix dataset was introduced in [9]. In this dataset, 20 k, 5 k and 3 k mixtures of utterances were created for training, validation and test. For three-speaker experiments, similar methods were adopted. In the validation set, TDAA could be used to evaluate the source separation performance on known talkers, which is the so-called Closed Conditions (CC) in [9, 10]. As a contrast, Open Condition (OC) will provide unknown speakers. To quantitatively evaluate results, we report the overall perfor-

[1] $L = N + 2$, with additional <BOS> (Begin-of-Sequence) and <EOS> (End-of-Sequence).

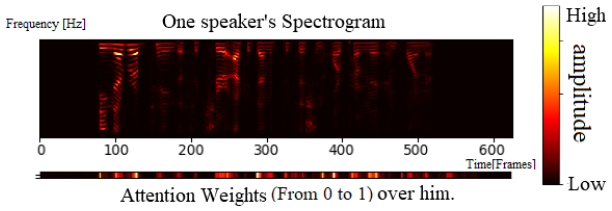

Figure 2: *Example of one speaker's spectrogram from mixed speech and corresponding attention weights α_{t_i} towards him.*

mance via the Signal-to-Distortion Ratio (SDR) [28] metric.

In our experiments, all data are resampled to 8 kHz. The magnitude spectra is used as the input feature, computed from Short-Time Fourier Transform (STFT) with 32 ms window length, 8 ms hop size and the sine window.

For the architecture of networks, in seq2seq model, the hidden sizes of the encoder and decoder are 256 and 512, respectively. The number of BiLSTM layers of encoder and decoder is 2. The label embedding size is 512. The separation network adopts 3-layer BiLSTM with 300 hidden units and a following feed-forward layer to embed each unit to dimension $d = 50$.

4.2. Performance in inference module

First of all, we evaluate the performance of the inference module, which uses the Seq2Seq generation network to output the possible candidates speakers. Similar experiments were done in the TDAA model. Table 1 shows the results of our work and the compared TDAA model with different settings. Here we use 4 metrics in the 2-mixed valid dataset (CC) and 2 metrics in the test dataset (OC) to describe the performance. For the CC, hamming loss, micro-precision, micro-recall, and micro-F1 are used. For the OC, due to the not-overlapped unknown speakers, we use the success rate to correctly predict the exact number of speakers to measure the inference module, which is the guarantee to the following separation performance.

From the results we could see that, the inference module in our work considerably outperforms the TDAA model. For the same number of speakers, our model with different settings could reach the 100% success rate to output the right number while TDAA model gets some inaccurate results which may degrade the exact separation after the inference module.

4.3. Speech separation performance

Table 2 summarizes the SDR performance in the WSJ0-2mix dataset of different methods. Here we list 4 different settings in our model. As illustrated in Figure 1, the basic model adopts the last hidden states from the decoder to serve as the query vector E_ϵ to extract the speech channel, while the "Sch" means the hidden vectors normalized by Schmidt orthogonalization. The "GE" means taking the global embedding after decoder's prediction at each step as E_ϵ. The "Re-ID" means the adoption of Re-Identification module in Section 3.3. In common with the TDAA, "top-1" means to extract the mask for the first candidate and the $1 - mask$ for the other channel.

From the result we observed that, our speaker-inferred model of relatively few parameters gets competitive with those baselines, especially in closed condition. It should be noticed that our work is totally end-to-end, without any additional curriculum training strategy or man-made rules to select the silent or salient region in [8, 10]. And most important, the information about the number of speakers in mixtures to be conducted is not provided, the module could infer it automatically. In addition, our model is optimized with the simple IRM, compared with the Wiener-Filter like Mask (WFM) used in DANet [8] and

Table 2: *SDR improvements (dB) with different separation methods on the WSJ0-2mix dataset.*

Model	of params.	2mix CC SDR	2mix OC SDR
DPCL [9]	-	5.9	5.8
DPCL+ [8]	-	-	9.1
DANet [8]	9.1M	-	9.6
DANet‡ [8]	9.1M	-	10.5
DPCL++ [10]	10.6M	-	9.4
DPCL++‡ [10]	16.9M	-	10.8
PIT-DNN [1]	-	5.2	5.2
PIT-CNN [1]	-	7.6	7.6
uPIT-BLSTM [12]	46.4M	9.4	9.4
uPIT-BLSTM-ST [12]	94.6M	10.0	10.0
TasNet-LSTM [17]	32.0M	-	**11.1**
TDAA-basic [21]	9.5M	8.5	4.1
TDAA top-1 [21]	9.5M	9.1	7.5
Our	18.3M	9.2	8.7
Our + Sch	18.3M	9.8	9.3
Our + Re-ID	18.3M	10.0	9.2
Our + GE	18.4M	10.7	8.0
Our + GE + top-1	18.4M	**11.0**	10.1
IRM		12.3	12.5

Ideal Phase Sensitive Mask (IPSM) used in uPIT [12]. Further attempt towards more complicated and phase sensitive masks may lead to better performance with our model.

Compared with the original TDAA model, with the same separation module, our inference module gains promotion by using the hidden vectors from the inference part to facilitate the following separation module, making the whole structure much more powerful for the variable numbers of mixed speakers. In our work, the Re-Identification gets an obvious promotion based on basic structure. For the hidden embedding E_ϵ, the Schmidt orthogonalization brings certain advantage to diversify the prediction in the decoder, which decreased the possibility of predicting similar speakers towards one channel. Similarly, the global embedding method is quite powerful, but it is also likely to overfit to the known speakers. We hypothesize that a large corpus with much more speakers would be more suitable for our work to maximum capacity.

In addition, we visualized one example in Figure 2 about the real spectrogram of one speaker in the mixture and the corresponding attention weights α_{t_i} towards him from the encoder in the inference module. As we expect, the attention weights not only present obvious beginning and end of the aim speech but also show significant consistency with the spectrogram.

5. Conclusions

In this paper, we propose a speaker-inferred model to perform the speech separation task. Through the Seq2Seq model, our work utilizes the information from speakers and accurately predicts the candidates speakers in mixed speech while taking the dynamic hidden states as stimuli to extract speech channel for each one. Our model is trained end-to-end and of relatively small size. Without any additional curriculum training strategy or man-made rules to select the silent or the salient region, our work gets similar or better results with the existing baselines.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (61602479), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB32070000) and the Beijing Brain Science Project (Z181100001518006).

7. References

[1] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 241–245.

[2] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[3] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH)*, 2006.

[4] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[5] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv: Computation and Language*, 2016.

[6] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 4835–4839.

[7] M. Kolbaek, Z. H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[8] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 246–250.

[9] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[10] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[11] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[12] M. Kolbaek, D. Yu, Z. H. Tan, J. Jensen, M. Kolbaek, D. Yu, Z. H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[13] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *Acm Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, 2018.

[14] C. Xu, R. Wei, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 6–10.

[15] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "Cbldnn-based speaker-independent speech separation via generative adversarial training," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 711–715.

[16] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network." in *Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH)*, 2018, pp. 342–346.

[17] ——, "Tasnet:time-domain audio separation network for real-time, single-channel speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 696–700.

[18] L. Drude, T. von Neumann, and R. Haeb-Umbach, "Deep attractor networks for speaker re-identification and blind source separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2018, pp. 11–15.

[19] J. Xu, J. Shi, G. Liu, X. Chen, and B. Xu, "Modeling attention and memory for auditory selection in a cocktail party environment," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 2564–2571.

[20] Z. Li, Y. Song, L. Dai, and L. McLoughlin, "Source-aware context network for single-channel multi-speaker speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2018, pp. 681–685.

[21] 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 *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Computer Science*, 2014.

[23] K. Cho, B. Van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *Computer Science*, 2014.

[24] Y. Zhang, Y. Wang, J. Liao, and W. Xiao, "A hierarchical attention seq2seq model with copynet for text summarization," in *International Conference on Robots and Intelligent System*, 2018, pp. 316–320.

[25] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "Sgm: Sequence generation model for multi-label classification," 2018.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] S. Wiseman and A. M. Rush, "Sequence-to-sequence learning as beam-search optimization," 2016.

[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.