ONLINE DEEP ATTRACTOR NETWORK FOR REAL-TIME SINGLE-CHANNEL SPEECH SEPARATION

Cong Han, Yi Luo, and Nima Mesgarani

Department of Electrical Engineering, Columbia University, New York, NY

ABSTRACT

Speaker-independent speech separation is a challenging audio processing problem. In recent years, several deep learning algorithms have been proposed to address this problem. The majority of these methods use noncausal implementation which limits their application in real-time scenarios such as in wearable hearing devices and low-latency telecommunication. In this paper, we propose the Online Deep Attractor Network (ODANet), an extension to the Deep Attractor Network (DANet) which is causal and enables real-time speech separation. In contrast with DANet that estimates the global attractor point for each speaker using the entire utterance, ODANet estimates the attractors for each time step and tracks them using a dynamic weighting function with only causal information. This not only solves the speaker tracking problem, but also allows ODANet to generate more stable embeddings across time. Experimental results show that ODANet can achieve a similar separation accuracy as the noncausal DANet in both two speaker and three speaker speech separation problems, which makes it a suitable candidate for applications that require robust real-time speech processing.

Index Terms— Source separation, speaker-independent, attractor network, real-time

1. INTRODUCTION

Speaker-independent monaural speech separation has been a challenging audio processing problem, and recent progress in deep learning has significantly advanced the state of this problem. Two main contributions to the development of this problem are the deep clustering network (DPCL) and the permutation invariant training method (PIT). DPCL [1, 2] maps the time-frequency (T-F) bins to a high-dimensional embedding space such that each T-F bin is represented by an embedding vector. The training objective is set to minimize a given distance metric of embeddings whose T-F bins belong to the same speaker and maximize that of different speaker's embeddings. After training, traditional clustering algorithms can be applied to the embeddings to calculate the source assignments as the estimated T-F masks. In [3, 4], an extension to DPCL, the Deep Attractor Network (DANet), was proposed to incorporate the clustering step into the network, allowing end-to-end training and evaluation. DANet maps each T-F bin of the mixture spectrogram to a high-dimensional embedding space similar to DPCL, and it explicitly forms clusters by calculating the oracle cluster centers based on the oracle embedding assignment (i.e. ideal T-F mask). The oracle cluster centers are called *attractor points*, and the embeddings that correspond to a specific speaker are constrained to be close to the corresponding attractor point. PIT is a general method for any type of objective functions to solve the output permutation problem [5, 6]. It determines the correct output permutation by calculating the lowest value on the selected objective function through all possible output permutations. Variants on the network architecture design and objective function design have proven the effectiveness of this training method [7–10].

On the other hand, the successful separation in many of those systems was contingent upon noncausal configuration, which means they required future information from the utterance. This greatly limits the deployment of such systems in causal or real-time applications such as wearable hearing devices or real-time communication systems. Several approaches have investigated causal system designs [6, 11, 12], but either the performance was not satisfying or the model design was complicated. For example, a source-aware context network was proposed in [13] that models the speaker-independent problem as a speaker-dependent problem with a segment-wise auto-regressive network design, while it requires teacher-forcing during training and more efforts on the auto-regressive architecture design. Another example would be a time-domain separation system, the TasNet [10, 14], that directly performs separation in waveform domain and achieves the state-of-the-art performance. However, for tasks that rely on a time-frequency representation (e.g. end-to-end speech recognition), the time-domain processing might make it harder for joint-training with a backend model.

In this paper, we propose online DANet (ODANet), a causal extension of the previous noncausal DANet that enables causal and real-time separation. ODANet calculates attractor points for the speakers at each frame, and the sequence of attractor points are tracked with a dynamic weighting function motivated by the online clustering methods [15, 16]. This allows us to perform online clustering and estimate the source assignment frame-by-frame. Experiments show that the proposed ODANet has comparable performance with

the noncausal DANet in both two-speaker and three-speaker separation tasks.

2. ONLINE DEEP ATTRACTOR NETWORK

2.1. Deep attractor network recap

We first review the DANet model in [3, 4] since it forms the foundation of the proposed ODANet. DANet treats the T-F mask estimation problem as a clustering problem in a high-dimensional embedding space, where the source assignments are used as the estimated T-F masks. DANet first maps the T-F bins in the mixture T-F representation to a K-dimensional embedding using a neural network

$$\mathbf{V} = \mathcal{H}(\mathbf{X}) \tag{1}$$

where $\mathbf{V} \in \mathbb{R}^{K \times TF}$ is the matrix containing all the embeddings and $\mathcal{H}(\cdot)$ is the mapping function defined by the network. C attractors, $\mathbf{A} \in \mathbb{R}^{C \times K}$, are formed by calculating the weighted average of the embeddings given the oracle speaker assignment $\mathbf{Y} \in \mathbb{R}^{C \times TF}$ in the mixture

$$\mathbf{a}_i = \frac{\mathbf{y}_i \mathbf{V}^\top}{\sum_{f,t} \mathbf{y}_i}, i = 1, \ldots, C \tag{2}$$

where $\mathbf{a}_i \in \mathbb{R}^{1 \times K}$ is the attractor of source i and $\mathbf{y}_i \in \mathbb{R}^{1 \times TF}$ is the oracle assignment of source i . The oracle speaker assignment $\mathbf{Y}$ can be any type of real-valued ideal T-F masks calculated on the mixture, where in [3] the ideal binary mask (IBM) was used. The mask is then calculated by transforming the distance between the embeddings and the attractors into a probability distribution that sum to 1

$$\mathbf{M} = Softmax(\mathbf{AV}) \tag{3}$$

where the Softmax function is applied column-wise.

For a purely end-to-end mask estimation without the need for oracle speaker assignemnt, the anchored DANet (ADANet) was proposed in [4] which introduced N trainable *anchors* in the embeddings space as the initialization of the attractors. During the training of the network, the position of the anchor points is jointly optimized to maximize the separability of the mixture sounds. After the training is performed, the anchor points are fixed. To separate a mixture that contains C speakers during the test phase, we first choose all possible C combinations of the N anchor points, resulting in $\binom{N}{C}$ subsets of the N anchors. The C actual attractors for a particular mixture are the ones in the subset that minimize in-set similarity between the attractors (i.e., maximizing the in-set distance between the chosen attractor points). Then masks are estimated through equation 3.

2.2. Online deep attractor network

The proposed ODANet estimates the attractors at each frame $\mathbf{A}_t \in \mathbb{R}^{C \times K}, t = 1, \ldots, T$ with stacked deep uni-directional LSTM layers followed by a fully-connected layer, and track them through a weighting mechanism across history information. The per-frame T-F masks can then be calculated with $\mathbf{A}_t$ and the current embeddings $\mathbf{V}_t \in \mathbb{R}^{K \times F}$ through equation 3. Figure 1 shows the general architecture of ODANet.

The estimation of attractors varies for the first frame and the following frames. To estimate the attractors $\mathbf{A}_1$ at the first frame, we apply ADANet described in section 2.1 on a single-frame input $\mathbf{x}_1 \in \mathbb{R}^{1 \times F}$. The corresponding masks $\mathbf{m}_1 \in \mathbb{R}^{1 \times F}$ are then calculated in the same way as equation 3 with the embeddings of the first frame $\mathbf{V}_1$. Starting from the second frame, $\mathbf{A}_{t-1}$ is used as anchors in ADANet. By applying the ADANet procedure on $\mathbf{V}_t$ and $\mathbf{A}_{t-1}$, the candidate attractors $\hat{\mathbf{A}}_t$ can be calculated by

$$\mathbf{Y}_t = Softmax(\mathbf{A}_{t-1}\mathbf{V}_t) \tag{4}$$

$$\hat{\mathbf{a}}_{t,i} = \frac{\mathbf{y}_{t,i}\mathbf{V}_t^\top}{\sum_f \mathbf{y}_{t,i}}, i = 1, \ldots, C \tag{5}$$

However, the embeddings at current frame $\mathbf{V}_t$ can have different distribution than those of previous frames $\mathbf{V}_{j,j<t}$, due to either the instability of the learning process of the network or the drastic change in the mixture contents (e.g. speaker silence or transition). This might push $\hat{\mathbf{A}}_t$ too far away from $\mathbf{A}_{t-1}$ and thus cause unstable speaker tracking and mask estimation in upcoming frames. To achieve a smoother transition between attractors as well as a more stable speaker tracking, we design a weighting mechanism across history information to update the current attractors. We introduce two types of update mechanisms, the context-based weighting and dynamic weighting, for attractor update at the current frame.

2.2.1. Context-based weighting

To achieve a middle ground between equation 2 with global frames and equation 5 with only the current frame, we can approximate $\mathbf{A}_t$ with past information by a weighted moving average process as

$$\mathbf{a}_{t,i} = \frac{\sum_{j=t-\tau}^{t} \mathbf{y}_{j,i}\mathbf{V}_j^\top}{\sum_{j=t-\tau}^{t} \sum_f \mathbf{y}_{j,i}} \tag{6}$$

$$= \frac{\sum_{j=t-\tau}^{t-1} \mathbf{y}_{j,i}\mathbf{V}_j^\top + \mathbf{y}_{t,i}\mathbf{V}_t^\top}{\sum_{j=t-\tau}^{t} \sum_f \mathbf{y}_{j,i}} \tag{7}$$

$$\approx \frac{\mathbf{a}_{t-1,i} \sum_{j=t-\tau}^{t-1} \sum_f \mathbf{y}_{j,i} + \hat{\mathbf{a}}_{t,i} \sum_f \mathbf{y}_{t,i}}{\sum_{j=t-\tau}^{t} \sum_f \mathbf{y}_{j,i}} \tag{8}$$

$$:= (1 - \alpha_{t,i})\mathbf{a}_{t-1,i} + \alpha_{t,i}\hat{\mathbf{a}}_{t,i} \tag{9}$$

where $\tau \in [0, t-1]$ is the length of the context window, and $\alpha_{t,i}$ is the update coefficient defined as

$$\alpha_{t,i} = \frac{\sum_f \mathbf{y}_{t,i}}{\sum_{j=t-\tau}^{t} \sum_f \mathbf{y}_{j,i}} \tag{10}$$

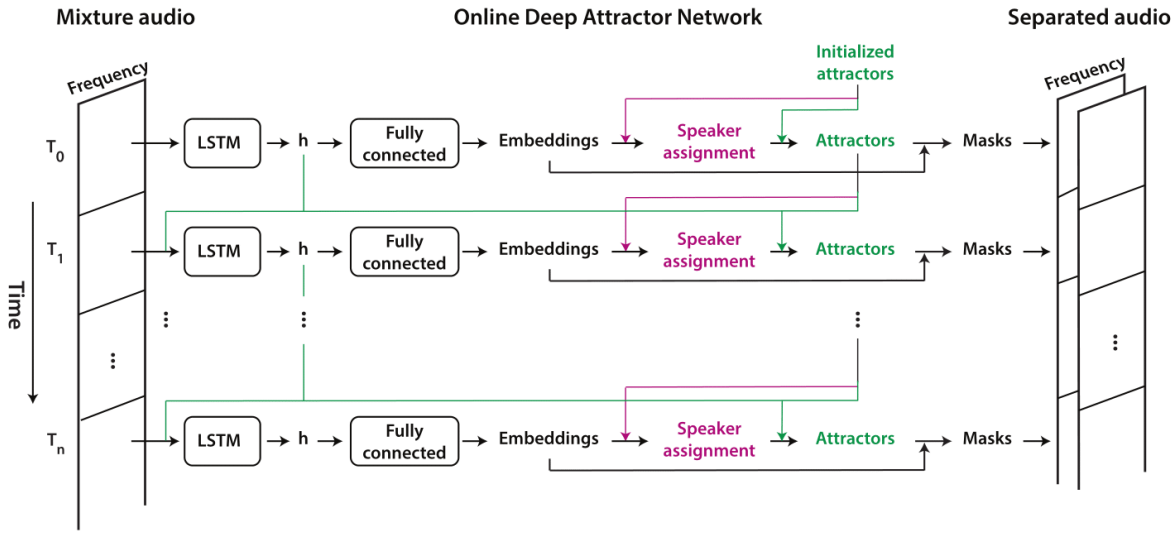

Fig. 1. The architecture of ODANet. In the first frame, attractors are estimated with the help of anchor points. In subsequent frames, the attractor points are updated by dynamically merging the previous and current attractors estimates.

This is similar to an exponential weighted average with the weights $\boldsymbol{\alpha}_t \in \mathbb{R}^{C\times 1}$ controlled by the relative importance of speaker assignments in the context window.

2.2.2. *Dynamic weighting*

In context-based weighting, the current frame and each previous frame are of equivalent importance to determine update rate, which is suboptimal in online system where a new frame could be more salient than previous ones or not. Motivated by the gating mechanism which is proven effective in many recent deep learning frameworks [17], we design an extra gating function in equation 10 for a dynamic control of the update coefficient $\boldsymbol{\alpha}_{t,i} \in \mathbb{R}^{1\times K}$.

$$\begin{aligned}\boldsymbol{\alpha}_{t,i} =&(\mathbf{g}_{t,i} \cdot \sum_{f} \mathbf{y}_{t,i}) \oslash \\ &(\mathbf{f}_{t,i} \cdot \sum_{j=t-\tau}^{t-1} \sum_{f} \mathbf{y}_{j,i} + \mathbf{g}_{t,i} \cdot \sum_{f} \mathbf{y}_{t,i})\end{aligned} \tag{11}$$

where $\oslash$ means element-wise division and $\mathbf{f}_{t,i} \in \mathbb{R}^{1\times K}$ and $\mathbf{g}_{t,i} \in \mathbb{R}^{1\times K}$ are gating vectors calculated as

$$\begin{cases}\mathbf{f}_{t,i} = \sigma(\mathbf{h}_{t-1}\mathbf{W}_f + \mathbf{x}_t\mathbf{U}_f + \mathbf{a}_{t-1,i}\mathbf{J}_f + \mathbf{b}_f) \\ \mathbf{g}_{t,i} = \sigma(\mathbf{h}_{t-1}\mathbf{W}_g + \mathbf{x}_t\mathbf{U}_g + \mathbf{a}_{t-1,i}\mathbf{J}_g + \mathbf{b}_g)\end{cases} \tag{12}$$

where $\sigma(\cdot)$ is the Sigmoid activation function, $\mathbf{W}_f, \mathbf{W}_g \in \mathbb{R}^{K\times H}$, $\mathbf{U}_f, \mathbf{U}_g \in \mathbb{R}^{K\times F}$, $\mathbf{J}_f, \mathbf{J}_g \in \mathbb{R}^{K\times K}$ and $\mathbf{b}_f, \mathbf{b}_g \in \mathbb{R}^{K\times 1}$ are trainable parameters, and $\mathbf{h}_{t-1} \in \mathbb{R}^{1\times H}$ is the output at frame $t-1$ of the last LSTM layer.

3. EXPERIMENTS AND DISCUSSIONS

3.1. Data

We evaluated our system on two-speaker and three-speaker speech separation problem using the WSJ0-2mix and WSJ0-3mix datasets, which are used in many recent literatures [2, 4, 6, 18]. All audio data are downsampled to 8 kHz before processing. The input feature is the log magnitude spectrogram computed by STFT with 32 ms window length, 8 ms hop size and the square root of Hanning window. Wiener-filter like mask (WFM) [4, 19] is used as the target mask during training phase.

3.2. Evaluation metrics

We report the scale-invariant signal-to-noise ratio (SI-SNR) [2,4,20], signal-to-distortion ratio (SDR) [21], and perceptual evaluation of speech quality score (PESQ) [22] as objective measures of speech separation accuracy.

Table 1. SDR improvement (dB) on WSJ0-2mix with different choice of context window size τ.

τ	context-based weighting	dynamic weighting
0	8.4	8.4
1	8.6	8.7
10	8.8	9.0
20	8.9	9.2
50	9.0	9.3
t-1	9.0	9.4

3.3. Network architecture

All networks contain 4 uni-directional LSTM layers with 600 hidden units in each layer. The embedding dimension is set to 20 for a fair comparison with previous systems [2, 4]. Following [4], the number of anchors is set to be 6. Batch size is set to 128. Adam [23] is used as the optimizer with the initial learning rate of $1e^{-4}$ and then halved if no best validation model is found in 3 epochs. The total number of epochs is set to 150, and early stopping is applied if no best model on the validation set is found for 10 consecutive epochs. All models are initialized using a pretrained DANet-LSTM model.

Table 2. Model comparison on WSJ0-2mix dataset. SI-SNR and SDR improvements (dB) and PESQ scores are reported.

Method	Causal	SI-SNRi	SDRi	PESQ
DPCL+ [2]	X	9.4	-	-
DPCL++ [2]	X	10.3	-	-
uPIT-BLSTM [6]	X	-	7.4	-
ADANet-LSTM	X	9.1	9.5	2.73
uPIT-LSTM [6]	✓	-	7.0	-
Source-aware Context [13]	✓	-	9.5	-
ODANet	✓	9.0	9.4	2.70

Table 3. Model comparison on WSJ0-3mix dataset. SI-SNR and SDR improvements (dB) and PESQ scores are reported.

Method	Causal	SI-SNRi	SDRi	PESQ
DPCL++ [2]	X	7.1	-	-
uPIT-BLSTM [6]	X	-	7.4	-
ADANet-LSTM	X	7.0	7.4	2.13
ODANet	✓	6.7	7.2	2.03

3.4. Results

We first evaluate the effect of different context window size τ on the performance of the entire utterance. Table 1 compares models with different values of τ. We can see that the larger context window leads to constantly better performance, while with full window size (i.e. $\tau = t - 1$), the model achieves the best performance. A possible explanation might be that small context window may cause unstable speaker tracking and separation in short periods where the one or more speakers are silent, while longer context window is more robust.

When considering each frame, the energy of one speaker is often larger than other speakers in some period and the number of speakers even varies, e.g., one speaker is in silence, which may cause the attractors tracking speakers with low energy to shift severely and even to fail to track if the window size is too small. A larger window size increases tracking accuracy, although sacrifices the performance of speaker switching. In this dataset, the speakers in each utterance are fixed, so full window size ($\tau = t - 1$) forces attractors converged, which greatly reduces speaker tracking failure cases. Moreover, applying the gate mechanism again improves performance. This proves the effectiveness of the gates on controlling the update of attractors across time. Figure 2 visualizes the embedding space and the attractor points at different frames during separation for the dynamic weighting model.

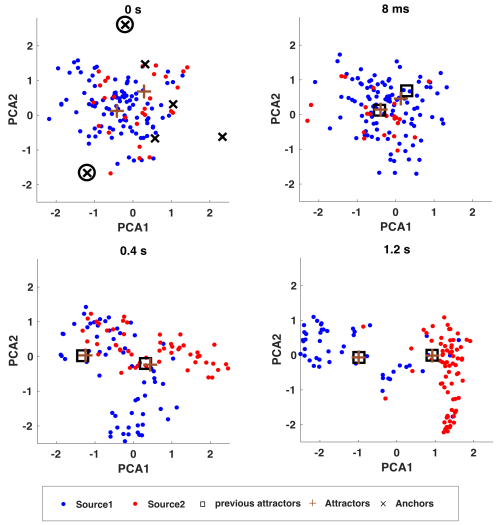

Fig. 2. Embedding space visualization at different frames in a 2-second mixture example. In the first frame, several anchor points serve as the initialization for attractors. From the second frame, the attractors are being updated smoothly. As time goes by, the attractor positions are stable as the embeddings are well-separated.

We then compared our system with previous systems on both two-speaker and three-speaker tasks. Table 2 and 3 provides the results for ODANet with other causal or noncausal systems on WSJ0-2mix and WSJ0-3mix datasets. We observe that in the two-speaker separation task, the causal ODANet significantly outperforms the causal uPIT and have on par performance with a noncausal ADANet with LSTM layers and global embedding clustering. Among T-F mask based methods, ODANet and Source-aware Context network both achieves the best performance, but ODANet is easily applied in the three-speaker separation task. In the three-speaker separation task, ODANet also achieves comparable performance with several noncausal systems, further proving its potential.

4. ACKNOWLEDGEMENT

This work was funded by a grant from the National institute of health, National Science Foundation (NSF-Career) and MIT-Lincoln Lab.

5. REFERENCES

[1] 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.

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

[3] 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.

[4] 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. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2018.2795749

[5] 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.

[6] M. Kolbæk, 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.

[7] C. Xu, X. Xiao, 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*. IEEE, 2018.

[8] X. Chang, Y. Qian, and D. Yu, "Adaptive permutation invariant training with auxiliary information for monaural multi-talker speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5974–5978.

[9] L. Chen, M. Yu, Y. Qian, D. Su, and D. Yu, "Permutation invariant training of generative adversarial network for monaural speech separation," *Proc. Interspeech 2018*, pp. 302–306, 2018.

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

[11] M. Sunohara, C. Haruta, and N. Ono, "Low-latency real-time blind source separation for hearing aids based on time-domain implementation of online independent vector analysis with truncation of non-causal components," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 216–220.

[12] Y. Wang, D. Wang, and K. Hu, "Real-time method for implementing deep neural network based speech separation," Mar. 2 2017, uS Patent App. 14/536,114.

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

[14] Y. Luo and N. Mesgarani, "Real-time single-channel dereverberation and separation with time-domain audio separation network," in *Proc. Interspeech*, 2018, pp. 342–346.

[15] L. Bottou and Y. Bengio, "Convergence properties of the k-means algorithms," in *Advances in neural information processing systems*, 1995, pp. 585–592.

[16] E. Liberty, R. Sriharsha, and M. Sviridenko, "An algorithm for online k-means clustering," in *2016 Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*. SIAM, 2016, pp. 81–89.

[17] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *CoRR*, vol. abs/1612.08083, 2016. [Online]. Available: http://arxiv.org/abs/1612.08083

[18] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[19] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 708–712.

[20] Y. Luo and N. Mesgarani, "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*. IEEE, 2018.

[21] 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.

[22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 749–752.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.