Recursive speech separation for unknown number of speakers

Naoya Takahashi[1], Sudarsanam Parthasaarathy[2], Nabarun Goswami[2], Yuki Mitsufuji[1]

[1]Sony Corporation, Japan
[2]Sony India Software Centre, India

{Naoya.Takahashi, Parthasaarathy.Sudarsanam, Nabarun.Goswami, Yuhki.Mitsufuji}@sony.com

Abstract

In this paper we propose a method of single-channel speaker-independent multi-speaker speech separation for an unknown number of speakers. As opposed to previous works, in which the number of speakers is assumed to be known in advance and speech separation models are specific for the number of speakers, our proposed method can be applied to cases with different numbers of speakers using a single model by recursively separating a speaker. To make the separation model recursively applicable, we propose one-and-rest permutation invariant training (OR-PIT). Evaluation on WSJ0-2mix and WSJ0-3mix datasets show that our proposed method achieves state-of-the-art results for two- and three-speaker mixtures with a single model. Moreover, the same model can separate four-speaker mixture, which was never seen during the training. We further propose the detection of the number of speakers in a mixture during recursive separation and show that this approach can more accurately estimate the number of speakers than detection in advance by using a deep neural network based classifier.

Index Terms: speech separation, deep learning, unknown number of speakers

1. Introduction

Speech communication often occurs in a multi-talker environment. In such a scenario, speech separation is required to selectively process each speaker individually. For example, automatic speech recognition first requires the separation of individual speakers from overlapping speech to successfully transcribe the target speech. Compared with other source separation problems that aim to separate different types of sources such as instrument types in music [1–4], speech separation has been considered very challenging for decades since the statistics of sources are similar or the same in the case of speaker independent speech separation problem. Previously, various approaches including spectral clustering [5] computational auditory scene analysis (CASA) [6], non-negative matrix factorization (NMF) [7–10] were proposed to tackle this problem, yet showed limited success. Recent advances of deep learning based methods including deep clustering (DPCL) [11–14], permutation invariant training (PIT) [15–17], deep attractor network (DANet) [18, 19] dramatically improved the accuracy of separation. Most recently, a time domain method has surpassed the ideal frequency masks performance under a two-speaker condition [17]. However, most of these methods assume that the number of speakers is known in advance. For example, the deep clustering approach requires information of the number of speakers to cluster embeddings and obtain time-frequency (T-F) masks, although a unified model can be used for 2 and 3 speakers mixture [11]. In actual cases, however, the number of speakers is often unknown or varies, making it difficult to robustly estimate the number of speakers in a mixture. In [15, 19],

this problem is partially solved by assuming the maximum number of speakers M in the mixture. The networks are trained to always output M channels regardless of the actual number of speakers N in the input, but when N is smaller than M, $M - N$ channels are enforced to output silent signals. At the test time, the number of speakers is determined by detecting the silent channels. Although the method is shown to work when $M = 3$ [15, 19], it fails when $M < N$.

One way to handle the separation of many speakers is to use visual information to leverage the correlation between speech and mouth movement. In [20], spatio-temporal representations of speakers' faces computed by a neural network trained on the lip reading task are concatenated with an audio signal and a separation network is trained to separate speech sources that correspond to visual information. It is shown to work up to five speakers in [20]. However, such visual information is often not available due to occlusion, frame out or lack of cameras. Thus, speech separation for an unknown number of speakers that operate with audio only is clearly required.

To address this problem, we propose to progressively separate speeches by applying a speech separation network recursively. Instead of separating all speakers in a mixture at once, the proposed model separates only one speaker from a mixture at a time and the residual signal is fed back to the separation model for the recursion to separate the next speaker, as shown in Fig. 1. To this end, we propose one-and-rest permutation invariant training (OR-PIT). The proposed method can handle different numbers of speakers using a single model by controlling the number of iterations. Moreover, the proposed method can separate mixture of multi-speakers whose number is larger than any of that seen during the training time. We further propose a method of robustly determining when to stop the iteration for an unknown number of speakers. With the proposed iteration termination criteria, we can more accurately identify the number of speakers than the number of speaker classifier that accept the mixture as the input, and separate speakers of unknown number.

Another advantage of the proposed method is that it tends to separate first a speaker that is easy to separate and sequentially tackle those that are harder to separate. Thus the first separation usually has the highest quality and the quality gradually decreases with increasing number of iterations. This is a preferable property since one can design a system that focuses on separating some of the clearest speakers, that is, a few speakers who are close to the microphone. Recently, similar recursive separation approaches are proposed [21, 22]. However, their works focus on the case where the number of speakers are same or less than training time, and evaluated only up to 2 speaker mixture. Moreover, [22] requires speaker ID during the training time. On the other hand, we show that our approach works even for 4 speaker mixture, which is greater than the number of speakers in the mixtures used for training.
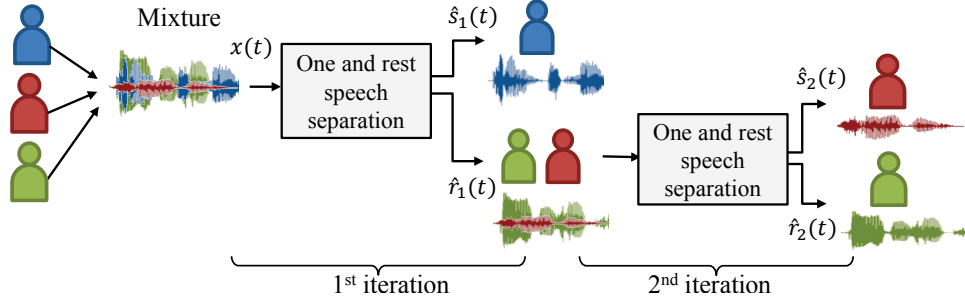
Our contributions are fourfold:

Figure 1: *Illustration of recursive speech separation when $N = 3$. The speech separation model is trained to separate one speaker from remaining speakers with OR-PIT, and is recursively applied to the second output.*

1. We propose a recursive speech separation method for separating a mixture of different numbers of speakers with a single model, even for mixtures which have more number of speakers than the mixtures used for training. To train the recursive separation model, we propose OR-PIT.

2. We further propose a robust and efficient recursion stopping method that enables to operate the recursive speech separation model for an unknown number of speakers.

3. Experimental results showed that our proposed method achieves state-of-the-art results on WSJ0-2mix and WSJ0-3mix datasets using a single model. Moreover, the proposed model can work surprisingly well for a four-speaker mixture, which was never encountered during the training.

4. We further showed that our proposed approach can more accurately detect the number of speakers in a mixture than the naive approach of directly classifying the number of speakers.

2. Recursive speech separation

In this section we first introduce the proposed recursive single-channel speech separation without prior knowledge of the number of speakers. Then we describe the training method for the recursive speech separator, followed by the loss function and the recursion stopping criterion.

2.1. Recursive speech separation

Time domain single channel speech separation entails estimation of N speaker sources $s_1(t), s_2(t), ... s_N(t)$ from a mixture signal $x(t)$, where $x(t) = \sum_{i=1}^{N} s_i(t)$. In our work we consider the number of speakers N to be unknown. At the jth recursion step, the recursive speech separator separates one speaker source $\hat{s}^j(t)$ and the mixture of residual speaker sources $\hat{r}^j(t)$ from $\hat{r}^{j-1}(t)$ as

$$\hat{s}^j(t), \hat{r}^j(t) = F(\hat{r}^{j-1}(t)), \qquad (1)$$

where $F()$ denotes the recursive speech separator modeled as a neural network. We define $\hat{r}^0(t) = x(t)$. The residual signal estimated at each step is input to F recursively to obtain subsequent speaker sources; thus, $\hat{r}^j(t)$ consists of $N - j$ speakers. The procedure is illustrated in Fig. 1. The criterion for deciding the number of recursion steps required to estimate all N speaker sources is described in 2.3.

2.2. One-and-Rest PIT

According to Eq. (1) the separation model F is to be trained to separate one speaker at a time and be recursively applica-

ble. However, the choice of one speaker is ambiguous, e.g., there are N valid choices of one target speaker $s_i(t)$ and corresponding residual signal $r_i = \sum_{n \neq i} s_n(t)$. The training with a random or constant choice of the target speaker fails since we do not assume any prior on the order of sources, e.g., we do not assume s_1 to be female and s_2 to be male or so on, and the model becomes confused how to choose the speaker during the test time. To address this problem, we propose novel training method called OR-PIT. Inspired by uPIT [15], OR-PIT computes the error l between the network output and the target for N possible splits of one and rest assignment, $s_i(t), r_i(t)$. The assignment that yields the lowest loss is used for the training objective L to optimize the network,

$$L = \min_i l(\hat{s}(t), s_i(t)) + \frac{1}{N - 1} l(\hat{r}(t), \sum_{n \neq i} s_n(t)). \qquad (2)$$

We omit j for simplicity. The error of the residual signal output channel is divided by the number of speaker sources in the residual mixture to balance it with the error of the single-speaker output channel. For the error, in this work, we use the scale-invariant signal-to-noise ratio (SI-SNR)[1], which has successfully been used in speech separation in the literature [11,16,17]. SI-SNR is formulated as:

$$\begin{cases} s_{target} := \frac{\langle \hat{s}, s \rangle s}{\|s\|^2} \\ e_{noise} := \hat{s} - s_{target} \\ l_{\text{SI-SNR}}(\hat{s}, s) := 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \end{cases} \qquad (3)$$

where $\hat{s}$ and s are the mean normalized estimates and targets, respectively. The mean normalization of the sources ensures the scale invariance property of the loss function.

In the case when the input to the network is a two speaker mixture, OR-PIT is equivalent to the conventional uPIT [15]. However, when the input is a mixture of more than two speakers, the permutations are computed by taking combinations of one speaker source and the sum of other speaker sources. Thus, the number of permutations in our case is N rather than $N!$, which is the case in uPIT. Another key difference from uPIT is that the sum of rest speaker sources (residual, $\hat{r}_j(t)$) is always trained to be on the second output channel. The purpose of OR-PIT is to ensure that the best combination of one speaker source and residual speaker sources are separated during the training. This allows the model to be used recursively in the second output channel until the stopping criterion is met.

One notable advantage of the proposed method is that it is not required to predefine the maximum number of speakers

---

[1]Also denoted as SI-SDR in [12,23].

and can be applied to an arbitrary number of sources even those never seen during the training. We verify this in Sec. 3.2.

2.3. Iteration termination criteria

As the proposed method recursively separates one speaker from a mixture at a time, we obtain J speaker sources by J recursion steps, where $J \leqq N$. If we wish all speaker sources to be separated, the number of iteration steps should be equal to the number of speakers, i.e., $J = N$. ($J - 1 = N$ is also possible since the residual signal at the $J - 1$th recursion step contains a single speaker.) A naive approach is to estimate the number of speakers N using a neural network [24]. We argue that estimating the number of speakers directly from the mixture is relatively difficult and propose to leverage the recursive speech separation model. We propose a simple deep neural network based binary classifier that accepts the residual outputs $\hat{r}^j$, ($j \geqq 1$) and predicts whether the signal is speech or not at each recursion step j. If $\hat{r}^j$ is predicted as speech, we proceed to the next recursion step. Otherwise, we stop the recursion and estimate N as j. Note that the energy based approach in [15, 19] cannot be applied to our approach since we use SI-SNR as a training objective and separating a single speaker input does not guarantee to produce a silent signal in one of the outputs, $\hat{s}^j(t), \hat{r}^j(t)$.

3. Experiments

3.1. Network training

We trained our model for the speech separation task using the Wall Street Journal data set (WSJ0). The model was trained concurrently with 2-speaker and 3-speaker mixture inputs. Following [11], the input mixtures were generated by randomly selecting utterances of different speakers from WSJ0 and mixing them at random SNR between -2.5 dB and 2.5 dB. The mixture was resampled to 8 kHz to reduce computations.

As the network architecture, we adopted TASNet [17], which is a recently proposed time domain speech separation network and produced state-of-the-art results on WSJ-2mix and WSJ-3mix datasets. We used the best performing configuration described in [17]. We replaced the softmax non-linearity used to generate the masks with a ReLU non-linearity in our architecture as we found that it worked better in our recursive model. We chose a time domain approach as the phase reconstruction is shown to be important for source separation to improve the performance [12,25] and operating directly on time domain signals is possible. However, our proposed method is also applicable to the T-F domain approach.

While training, we forced the first network output channel to always have one speaker and the second channel to collect all the remaining speakers in the mixture input. The model was trained using the OR-PIT with the SI-SNR loss function explained in Sec. 2.2. The network was initially trained for 100 epochs with 2- and 3-speaker mixture inputs. The initial learning rate was set to $1e^{-3}$. The Adam optimizer was used with a weight decay of $1e^{-5}$. The input mixtures were 4 seconds long with 50% overlap between two successive frames. In the decoder, the overlapping segments were added together to generate the final reconstructions as in [17].

To further improve the performance of recursion, the model was fine tuned on a 2-speaker mixture obtained from a separation of the first iteration of 3-speaker mixture, instead of clean 2-speaker mixture. The loss was accumulated on both the first iteration (clean 3-speaker separation) and the second iteration (residual 2-speaker separation), and back-propagated.

Table 1: *SI-SNR improvement (dB) for 2- and 3-speaker separations before and after fine tuning*

Model	WSJ0-2mix	WSJ0-3mix
Before fine tune	15.0	12.2
After fine tune	14.8	12.6

Although the fine tuning slightly decreased the performance of 2-speaker separation, it more significantly improved the performance of the 3-speaker separation. The SI-SNR improvement of our model before and after fine tuning is shown in Table 1.

3.2. Comparison with other approaches

We compared the proposed method with other state-of-the-art methods [11, 15, 17–19] on WSJ0-2mix and WSJ0-3mix datasets [26]. The baseline methods are categorized into three groups, namely the *2-speaker model* which is trained for the 2-speaker separation task, the *3-speaker model* which is trained for 3-speaker separation task, and the *2&3-speaker model* which is trained so that it can be applied to both 2- and 3-speaker separation tasks with a unified model. Note that our model is not grouped as *2&3-speaker model* since it can be applied to arbitrary number of speakers even though it was trained on two and three speaker mixture. On the other hand, [18, 19], *2&3-speaker model* is designed to handle an unknown number of speakers which is smaller than the maximum number of speakers it is trained for. To confirm whether our proposed method can handle the number of speakers that was never seen during the training time, we also evaluated the proposed method on a newly created four-speaker mixture. The 4-speaker evaluation set (WSJ0-4mix) was created from the WSJ0-3mix by adding one more speaker source to each of the 3-speaker mixture and mixing them at random SNRs between -3 and 3dB. We also include the ideal binary mask as the baseline.

The SI-SNR improvement (SI-SNRi), signal-to-distortion ratio improvement (SDRi) [27] and perceptual evaluation of speech quality score (PESQ) [28] are shown in Table 2. In the cases where the number of speakers in the mixture is different from that of the model target, we marked them as Not Applicable (N/A). We assume an oracle iteration termination for OR-PIT. The termination method is discussed in Sec. 3.3. As shown in the table, the proposed method achieved the best results on SI-SNRi and SDRi on both WSJ0-2mix and 3mix datasets. Even when compared with the models specifically trained for 2- or 3-speaker separation, the proposed method outperforms most baselines with a single model. It is worth noting that our model uses the same network architecture as Conv-TasNet-gLN [17] except the nonlinearlity of the last layer. Comparison with these models suggests the effectiveness of the proposed recursive separation method and generalization capability to the number of input speakers. This effectiveness is further supported by the evaluation on four-speaker mixture. Even though the proposed method never encountered the four-speaker mixture during the training, it separated four speaker surprisingly well with three recursions.

3.3. Identification of number of speakers

Since our model can perform speech separation of an unknown number of speakers in input by recursion, it is very important to know when to terminate the recursion. We evaluated the proposed iteration termination criteria described in Sec.2.3. We trained the Alexnet model [29] for the task of binary classifica-

Table 2: *Performance comparison of models trained on WSJ0 data sets. SiSNRi(dB), SDRi(dB) and PESQ of models on WSJ0-2mix, WSJ0-3mix and WSJ0-4mix are compared. ('N/A' - Not applicable, '-' - Data not available)*

Method		WSJ0-2mix			WSJ0-3mix			WSJ0-4mix		
		SI-SNRi	SDRi	PESQ	SI-SNRi	SDRi	PESQ	SI-SNRi	SDRi	PESQ
2 speaker model	DPCL++ [11]	10.8	-	-	N/A	N/A	N/A	N/A	N/A	N/A
	uPIT-BLSTM-ST [15]	-	10.0	-						
	DANet [18]	10.5	-	2.64						
	ADANet [19]	10.4	10.8	2.82						
	Conv-TasNet-gLN [17]	14.6	**15.0**	**3.25**						
3 speaker model	DPCL++ [11]	N/A	N/A	N/A	7.1	-	-	N/A	N/A	N/A
	uPIT-BLSTM-ST [15]				-	7.7	-			
	DANet [18]				8.6	8.9	1.92			
	ADANet [19]				9.1	9.4	2.16			
	Conv-TasNet-gLN [17]				11.6	12.0	2.5			
2&3 speaker model	DPCL++ [11]	10.5	-	-	7.1	-	-	N/A	N/A	N/A
	uPIT-BLSTM-ST [15]	-	10.1	-	-	7.8	-			
	ADANet [19]	10.4	-	-	8.5	-	-			
OR-PIT (Proposed)		**14.8**	**15.0**	3.12	**12.6**	**12.9**	**2.60**	**10.2**	**10.6**	**2.26**
Oracle mask (Ideal binary mask)		13.0	13.5	3.33	13.2	13.6	2.91	11.8	12.0	2.42

Table 3: *Test accuracy of speech or noise binary classifier and multi-class count speakers classifier*

Model	Accuracy
Binary classifier	**95.7%**
Multi-class classifier	77.9 %

Figure 2: *SI-SNR of dominant-speaker separation on various numbers of interference speakers.*

tion of speech or noise on the residual inputs coming out from the second channel. As a baseline, we also trained the Alexnet model for a multiclass classification task of counting the number of speakers in the input mixture. The baseline model can be used with, for example, DPCL [11] to decide the number of clusters. Please note that the capability of the multiclass classifier is limited by the maximum number of speakers in a mixture input in the training set. On the other hand, the binary classifier is independent of the number of speakers in the mixture. For both models, input segments of 10 seconds long mel spectrogram with window length 1024, 50% overlap and downsized to 128 mel bands were used as input features. The test set consisted of 3000 samples each of clean 1, 2 and 3 speakers from the WSJ0 evaluation sets, WSJ0-2mix and WSJ0-3mix, respectively. For the binary classifier, when the network predicts all the iterations except the last as speech, it is considered as a successful classification and vice versa. As shown in Table 3, the binary classifier more accurately detects the number of speakers than the multiclass classifier. This clearly indicates the effectiveness of leveraging the recursive speech separation model for the detection of the number of speakers.

3.4. Dominant speech separation

A notable property of the proposed method is that it tends to separate the most dominant (easiest) speaker first and successively tackles the separation of less dominant (harder) speakers. This is useful when we consider extracting a few speakers close to the microphone in a crowd since we can often assume that the conversation takes place within a small area and one of speakers can hold or attach a microphone. As a special case, we consider extracting the most dominant speaker from a mixture of a large number of speakers. To simulate a speech in a crowd recorded
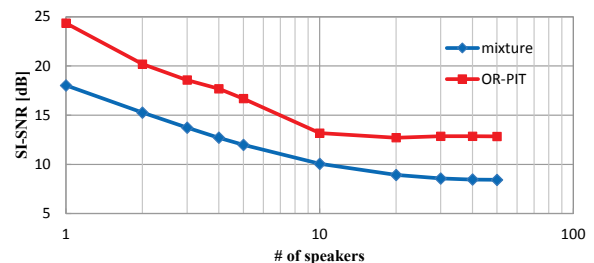
by a microphone attached to the target speaker, we created an evaluation dataset by adding N interference speakers to the target speaker. We vary N from 1 to 50 and created an evaluation set of 500 samples for each of the cases. The first interference speaker level is scaled to be 18dB less than target speaker and every Nth speaker was scaled to have 0.5 dB lower than the $N-1$th interference speaker. The same model in Sec.3.2 was used and the model was never trained or fine-tuned for this task. , i.e., the model never saw a mixture of more than three speakers during the training. Fig. 2 shows the SI-SNR of the proposed method and the mixture as a baseline. It is shown that the proposed method consistently improved SI-SNR and achieved high SI-SNR even for a mixture of more than 10 speakers. It indicates that the proposed method can be robustly applied for a dominant-speaker separation from a mixture of a large number of speakers.

4. Conclusion

We proposed a novel recursive speech separation approach that deal with different numbers of speakers cases using a single model. Experimental results show that our proposed method achieves state-of-the-art results on two and three speaker mixture with the same model and even worked on a four-speaker mixture even though the model has never seen the four-speakers mixture during the training.

5. References

[1] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel music separation with deep neural networks," in *Proc. EUSIPCO*, 2015.

[2] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving Music Source Separation Based On Deep Networks Through Data Augmentation And Network Blending," in *Proc. ICASSP*, 2017, pp. 261–265.

[3] N. Takahashi and Y. Mitsufuji, "Multi-scale Multi-band DenseNets for Audio Source Separation," in *Proc. WASPAA*, 2017, pp. 261–265.

[4] N. Takahashi, N. Goswami, and Y. Mitsufuji, "Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. IWAENC*, 2018.

[5] F. R. Bach and M. I. Jordan, "Learning Spectral Clustering, with Application to Speech Separation," *The Journal of Machine Learning Research*, vol. 7, pp. 1963–2001, Jan 2006.

[6] K. Hu and D. Wang, "An Unsupervised Approach to Cochannel Speech Separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 21, p. 122131, Jan 2013.

[7] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. Interspeech*, 2006.

[8] T. Virtanen and A. T. Cemgil, "Mixtures of gamma priors for non-negative matrix factorization based speech separation," in *Proc. Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 646–653.

[9] G. J. Mysore and P. Smaragdis, "A non-negative approach to language informed speech separation," in *Proc. Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2012, pp. 356–363.

[10] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," in *Proc. ICASSP*, 2014, pp. 3749–3753.

[11] 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.

[12] Z.-Q. Wang, J. L. Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, 2018.

[13] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. ICASSP*, 2018.

[14] S. Wang, G. Naithani, and T. Virtanen, "Low-latency deep clustering for speech separation," in *Proc. ICASSP*, 2019.

[15] M. Kolbk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, p. 19011913, 2017.

[16] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*, 2018.

[17] Y. Luo and N. Mesgarani, "Tasnet: Surpassing ideal time-frequency masking for speech separation," *CoRR*, 2018. [Online]. Available: http://arxiv.org/abs/1809.07454v2

[18] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. ICASSP*, 2017.

[19] 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, April 2018.

[20] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[21] M. D. T. N. Keisuke Kinoshita, Lukas Drude, "Listening to each speaker one by one with recurrent selective hearing networks," in *Proc. ICASSP*, 2018.

[22] G. L. B. X. Jing Shi, Jiaming Xu, "Listen, think and listen again: Capturing top-down auditory attention for speaker-independent speech separation," in *Proc. IJCAI*, 2018.

[23] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?" *CoRR*, vol. abs/1811.02508, 2018. [Online]. Available: http://arxiv.org/abs/1811.02508

[24] F.-R. Stoeter, S. Chakrabarty, B. Edler, and E. A. P. Habets, "Classification vs. regression in supervised learning for single channel speaker count estimation," in *Proc. ICASSP*, 2018.

[25] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "Phasenet: Discretized phase modeling with deep neural networks for audio source separation," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[26] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech and Language Processing*, no. 4, pp. 1462–1469, 2006.

[28] 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 *Proc. ICASSP*, 2001, p. 749752.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.