Optimal scale-invariant signal-to-noise ratio and curriculum learning for monaural multi-speaker speech separation in noisy environment

Chao Ma[1], Dongmei Li[1], Xupeng Jia[1]
[1]Department of Electronic Engineering, Tsinghua University, Beijing, China, 100084
E-mail: {c-ma13, jxp12}@mails.tsinghua.edu.cn, lidmei@tsinghua.edu.cn

***Abstract*— *In* daily listening environments, speech is always distorted by background noise, room reverberation and interference speakers. With the developing of deep learning approaches, much progress has been performed on monaural multi-speaker speech separation. Nevertheless, most studies in this area focus on a simple problem setup of laboratory environment, which background noises and room reverberations are not considered. In this paper, we develop a new objective function named optimal scale-invariant signal-noise ratio (OSI-SNR), which are better than original SI-SNR at any circumstances. In addition, we propose a curriculum learning method based on conv-TasNet to deal with the notable effects of noises and interference speakers. By jointly using the OSI-SNR with curriculum learning method, our algorithm outperforms separation baseline substantially.**

Keywords: conv-TasNet, multi-speaker speech separation, noisy environment, SI-SNR

I. INTRODUCTION

In real world environments, speech is always corrupted by background noise, room reverberation and interference speakers. The presence of such noise, interference and reverberation has a corrupting negative effect on speech intelligibility and speech quality. Many applications, such as speaker identification and automatic speech recognition, become much more challenging in such severe environments, as well as normal hearing and hearing-impaired listeners [1], [2], [3], [4]. Therefore, better enhancement, dereverberation and separation have a significant benefit to not only human listeners but also many speech processing missions.

Over the past few decades, significant efforts have been, and still are being devoted to speech enhancement and speech dereverberation [5], [6], [7], [8]. However, only limited breakthrough has been made in single-channel speaker-independent multi-speaker speech separation task. The most severe difficulties we faced, label permutation problem, are not solved until last ten years.

More recently, several particular approaches have been proposed to deal with the label permutation problem. In [9], [10], Permutation Invariant Training (PIT) and utterance-level PIT choose the speaker arrangement on the basis of the lowest separation error within all possible permutations. In [11], [12], Deep Clustering (DPCL) algorithm achieves label assignment using the clustering methods in a deep embedding space. In [13], Deep Attractor Network (DANet) produces attractors in deep embedding space to achieves label assignment. In [14], a time-domain audio separation network (TasNet) is proposed. In TasNet, traditional short-time Fourier transform (STFT) is replaced with a convolutional encoder-decoder architecture. In [15], the fully-convolutional TasNet (conv-TasNet) is proposed. The use of stacked dilated 1-D convolutional blocks to replace the deep LSTM networks for the separation step not only significantly reduces the model size, but also has a better performance, even surpasses the performance of ideal time-frequency magnitude masks. In [16], a source-aware context network is designed to address the label permutation problem by exploiting temporal dependencies and continuity of the same speech source.

With the astonishing achievements on monaural multi-speaker speech separation, only several works considered the robustness of speech separation algorithms [17], [18], [19], despite of the impressive achievements on clean speech separation. In this paper, a baseline of speech separation in the noisy environment is built.

With the foundation of noisy speech separation dataset and speech separation baseline, we present a novel objective function named optimal scale-invariant signal-to-noise ratio (OSI-SNR）. We not only derive the formula of OSI-SNR, but also show that OSI-SNR is better than original SI-SNR in any circumstances when they are used as loss function. Furthermore, we propose a curriculum learning method applied to noisy speech separation system. We also propose a frequency analysis method to visualize and demonstrate the effect of curriculum learning.

II. ALGORITHM DESCRIPTION

A. *Problem formulation.*

Let $x_i(t)$ and n(t) denote speech from speaker *i* and background noise, respectively. The noisy multi-speaker speech y(t) is modeled by

$$y(t) = \sum_{i=1}^{N} x_i(t) + n(t) \quad (1)$$

The goal of monaural robust speech separation is to estimate the individual speech signals from a given noisy mixture of speech signals and noises. In this work the number of target signals is assumed to be known and set to 2.

B. *baselines settings.*

The baselines are based on conv-TasNet [15]. The systems consist of three modules: an encoder module, a separation module and a decoder module.

The first baseline system is exactly same as conv-TasNet. The noisy mixture y(t) is input to the 1-D convolutional encoder module and embedded to a spectrum space. At this paper, we will call this embedded space matrix as spectrum, because we considered and demonstrated later that this embedded space matrix is quite similar to traditional spectrum. The temporal convolutional network (TCN) separation module estimates the masks based on the encoder output.

The dilate factors in the separation module increase exponentially, which guarantee an enough reception field to take advantage of the long-range dependencies of the speech signal. The output of the separation module multiplied with the output of encoder is passed to the decoder module and transferred to clean separated speech signal.

Since the output of the network are the waveforms of the estimated clean signals, here the scale-invariant source-to-noise ratio is used and Permutation invariant training (PIT) is applied during training to settle the permutation problem. Consequently, the loss function of baseline is:

$$L_{PIT} = \min_{\pi \in P} \sum_c -SISNR(x_c(t), \widehat{x_{\pi(c)}}(t)) \quad (2)$$

Where P is the set of all possible permutations over the set of sources {1, …, C}, $x_c(t)$ denotes the recovery separated speech, $\widehat{x_{\pi(c)}}(t)$ denotes the original clean speech. The definition and improvement of SI-SNR will be explained in next section.

C. *optimal SI-SNR.*

The use of Scale-Invariant Source-to-noise ratio (SI-SNR) is a remarkable improvement against SNR [21]. The definition of frequently used SI-SNR is given as:

$$s_{target} = \frac{\langle \hat{s}, s \rangle s}{\|s\|^2} \quad (3)$$

$$e_{noise} = \hat{s} - s_{target} \quad (4)$$

$$SI - SNR := 10log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \quad (5)$$

Where s represents the original speech signal and $\hat{s}$ represents the reconstructed speech signal. This method adjusts original speech signal to a proper scale, and calculates an adjusted SNR.

It is obvious that the length of s_{target} is not relevant of original signal s. We can calculate that the equal definition of s_{target} is as below:

$$s_{target} = |\hat{s}| cos\theta \vec{s} \quad (6)$$

Where $\vec{s}$ is the unit vector at same direction of s, θ is the angle between s_{target} and s.

While the length of s_{target} and length of s are proportional, the SNR after adjusted is not relevant of length of s_{target} or s, but only relevant of angle θ. The equal definition of SI-SNR is as below:

$$\text{SI} - \text{SNR} = 10log_{10} \frac{1}{\tan^2 \theta} \quad (7)$$

As [21] explained, we get this formula by finding a point in s which is closest to $\hat{s}$, i.e. $\alpha s \perp (\alpha s - \hat{s})$. The orange brace in Fig.1. illustrates this definition of SI-SNR. However, there is no strong reason to do so. The source vector and noise vector don't have to be orthogonal. If we adjust s to be closest to $\hat{s}$, that will lead to a different result, which is also a scale-invariant result.

In other words, the definition of scale-invariant is not unique. A calculating method can be called a scale-invariant SNR as long as the scale of s_{target} is not relevant to the scale of original speech signal.

Table.1 shows different definitions of SI-SNR. But which one is better? If we can compute a maximum SI-SNR with a fixed s and $\hat{s}$, then this calculating method is easier to optimize and less likely to fall into a local best while training.

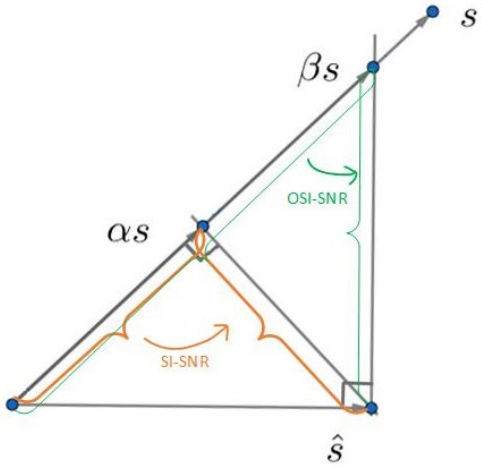

Fig. 1: Illustration of the definitions of SI-SNR and OSI-SNR

Therefore, let's find the maximum SI-SNR.

$$SI - SNR(\lambda) = 10log_{10}(\frac{\|\lambda s\|^2}{\|\lambda s - \hat{s}\|^2}) \quad (8)$$

$$\lambda = \underset{\lambda}{\text{argmax}}\, SI - SNR(\lambda) \quad (9)$$

Where λ indicates the scale adjust factor. The derivative of this SI-SNR is calculated below.

$$F(\lambda) = \frac{\|\lambda s\|^2}{\|\lambda s - \hat{s}\|^2} \quad (10)$$

$$F'(\lambda) = \frac{2ks^2(\hat{s}^2 - ks\hat{s})}{|ks - \hat{s}|^4} = 0 \quad (11)$$

$$\lambda = \frac{|\hat{s}|^2}{\langle s, \hat{s} \rangle} \quad (12)$$

This maximum SI-SNR will be called optimal SI-SNR (OSI-SNR) in this paper. The performance of OSI-SNR will be demonstrated in chapter 4.

Table. 1: The definitions and equal definitions of SI-SNR and OSI-SNR.

SI-SNR	OSI-SNR
$s_{target} = \frac{\langle s, \hat{s} \rangle s}{\|s\|^2}$	$s_{target} = \frac{\|\hat{s}\|^2 s}{\langle s, \hat{s} \rangle}$
$s_{target} = \|\hat{s}\| cos\theta \vec{s}$	$s_{target} = \frac{\|\hat{s}\| \vec{s}}{cos\theta}$
$\text{SNR} = 10log_{10} \frac{1}{\tan^2 \theta}$	$\text{SNR} = 10log_{10} \frac{1}{\sin^2 \theta}$

Table.1 shows definitions of different SI-SNRs. SI-SNR is the commonly used SI-SNR. If we compare the SI-SNR and OSI-SNR, we will find they are opposite in many ways. The green brace in Fig.1 also demonstrates the differences between SI-SNR and OSI-SNR.

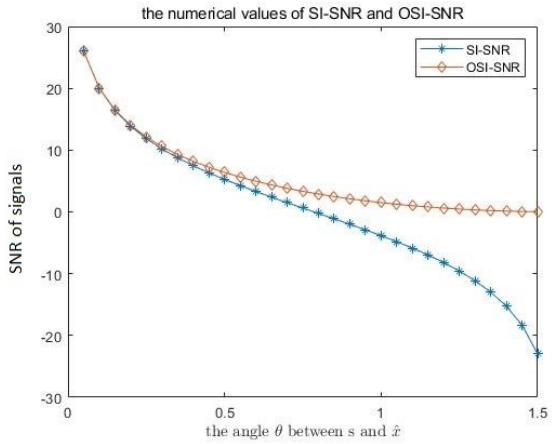

Fig. 2: The illustration of SI-SNR and OSI-SNR.

Fig.2 shows the numerical values of different SI-SNRs. The boundaries of different SI-SNRs is easy to observe from Fig.2 and calculate from the equal definitions in Table.1. When θ ranges from 0 to $\frac{\pi}{2}$, the value of SI-SNR ranges from $-\infty$ to $+\infty$, while the values of OSI-SNR range from 0 to $+\infty$. The meaning and explanation of fig.2 will be shown in section of experimental results.

Here we made a simple summary of OSI-SNR.

The original idea behind SI-SNR is using the scale-invariant-ness of this metric to get a stable gradient to quantify the processed speech. But the source vector and noise vector don't have to be orthogonal as original defined. There are all kinds of different SI-SNRs, which we only mentioned two of them in this paper. The key idea of this paper is that if a $SI-SNR(\theta,\lambda_{max})$ is larger than any other $SI-SNR(\theta,\lambda)$, then this $SI-SNR(\theta,\lambda_{max})$ is the ridge line of metric in high dimension space when we maximize SI-SNR by optimize θ. As we derived before and demonstrate later, OSI-SNR will be easier to optimize and faster to converge, and not likely to fall into a local optimum.

D. curriculum learning.

a) frequency analysis.

The monaural multi-speaker speech separation system used in this work is based on conv-TasNet [15]. The time-domain raw waveform was input to a 1-D encoder and become time-frequency representations. A TCN processes the representations to complete the separation function and output two masks. After multiplied with original speech representations, the 1-D decoder recover the both single clean speech from the recovered single speech representations.

The encoder is consisting of a finite impulse response filter bank and a nonlinear activate function, which is ReLu function in this work. This structure of encoder is easy to analyze and visualize.

The technique of analyze the frequency of encoder is demonstrated below.

The size of kernels in the encoder, which is also the frame length of speech signal, is set to 20. This is a pretty small number for frame length.

So, we zero-padding the kernels to a large number, like 800 in this work. (512 or 1024 are convenient as well.) After 800-points DFT, we take the left 400 points and calculate the absolute value of it. That's how we get the response of each kernel against each frequency bin.

With the same technique, we can visualize the decoder of conv-TasNet as well, since the decoder is a deconvolution layer, and can be considered as transposition of encoder.

b) curriculum learning method

One of the key ideas of curriculum learning [20] is let the network learn easy datasets before hard datasets just like human learning everything. This can force the model to get a better local optimum and fasten the speed of training.

In this work, the method of curriculum learning is to train the model with the clean dataset first as a pretraining step if we want to train a noisy speech separation system.

Fig. 3 demonstrate visualization of encoders of three systems.

The top two sub-figures of fig.3 show encoder of clean speech separation system. We can see that this encoder does complete time-to-frequency transformation. The distribution of filters' frequency matches human cochlear very well. Just like gammatone filterbanks or Mel-frequency filterbanks.

The middle two sub-figures of fig.3 show encoder of noisy speech separation system trained without curriculum training, i.e. trained with random initial. We can't say that the filter response must be cochlear-like distribution. But absolutely not like sub-figure middle-right. Many filters have approximately the same frequency response, which is a waste of computation and network memory. This indicates the noisy speech separation system are hard to converge and easy to fall into a local optimum.

The bottom two sub-figures of fig.3 show encoder of noisy speech separation system trained with curriculum learning, i.e. trained with model initialed with clean model. This time the results perform well and are easy to understand. With curriculum learning, the model is initialed with clean model and already got a better circumstance, so the model is much likely to fall into a better local optimum or even the global optimum relatively.

By the way, we also tried to transfer the encoder and decoder from clean speech separation system directly. It is useful but not as good as curriculum learning. The model after transfer still need a fine-tune step, which makes it almost the same as method of curriculum learning.

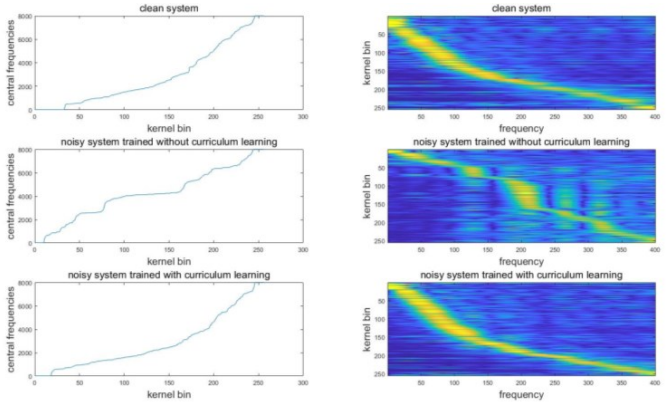

Fig. 3: the central frequency distribution and frequency response of filters in the encoder of clean system (the top two sub-figures), noisy system trained without curriculum learning (the middle two sub-figures) and noisy system trained with curriculum learning (the bottom two sub-figures).

III. EXPERIMENTAL SETTINGS

A. *datasets.*

The 2-speaker speech separation dataset we evaluated our system on is based on wsj0-2mix [11], which contains 30 hours of training data, 10 hours of validation data and 5 hours of evaluation data. When generating mixtures, the way of randomly choosing speakers and utterances, the SNR adjustment between two speakers and other settings are exactly same as wsj0-2mix. However, we add some noise after mixing the utterances. The noise file is from NoiseX [25] noise sets. The chosen noise type is babble, destroyer engine, destroyer ops and factory1. The SNR between clean mixture and noise is normally distributed in -5dB to 5dB, i.e. for most noisy speech, $20 * log_{10} \frac{|x_1(t)+x_2(t)|}{|n(t)|} \approx 0$.

We also compared and evaluated SI-SNR and OSI-SNR in clean speech separation systems. When we test them in clean speech separation system, the wsj0-2mix dataset is used.

B. *network settings.*

The networks are trained for 100 epochs on 4-seconds long segments. Adam optimizer [22] is used. The learning rate is initialed to 1e-3 and halved if the accuracy of validation set is not improved in three epochs. The hyperparameters of the network are same as conv-TasNet [15].

C. *criteria for evaluation.*

We use signal-to-distortion ratio improvement (SDRi) as objective measures of separation accuracy. The scale-invariant signal-to-noise ratio improvement (SI-SNRi) and the optimal scale-invariant signal-to-noise ratio improvement (OSI-SNRi) is also compared afterwards. In addition, we also evaluated the quality of the separated speech using both the perceptual evaluation of subjective quality (PESQ [23]) and the short-time objective intelligibility (STOI [24]). The PESQ scores is between [-0.5, 4.5], while the STOI scores range from 0 to 1. Higher values in PESQ and STOI are reflection of better speech quality.

D. EVALUATION RESULTS

In this study, five objective metrics mentioned before are employed to evaluate the performance of separation systems.

Table 2: *performance comparison.*

	PESQ	STOI	SDRi	SI-SNRi	OSI-SNRi
Baseline	2.215	0.800	9.985	11.757	5.744
OSI-SNR	2.249	0.811	10.010	12.230	6.253
Curriculum learning	2.235	0.805	10.103	12.047	5.959
OSI-SNR + curriculum learning	**2.291**	**0.817**	**10.590**	**12.627**	**6.494**

Table.2 shows performance of different systems trained with techniques proposed in this paper.

From the results demonstrated before, we can see that the use of OSI-SNR and curriculum learning can improve the performance of noisy speech separation system separately and independently. By combine these two methods together, the performance is improved substantially. The contribution of OSI-SNR and curriculum learning are approximately 70% and 30%, respectively.

It must be clear that this paper only confirmed that OSI-SNR is doubtlessly better than SI-SNR when they are used as the training targets. It would be better if we continue use SI-SNRi as criteria of evaluation.

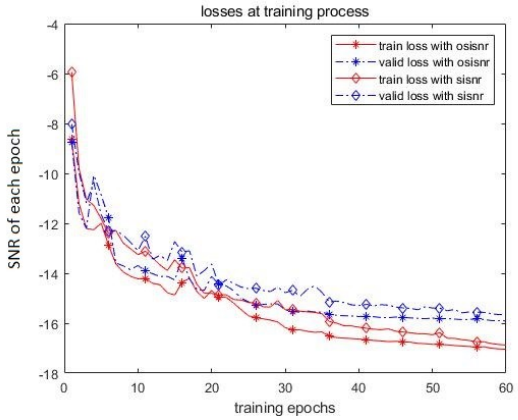

Fig. 4: *losses comparison between different SI-SNRs while trained in noise-free separation system*

Fig.2 shows the curve that SI-SNR and OSI-SNR change with difference of angles θ between original signal and reconstructed signal. They are quite the same when θ is extremely low. That's why the original SI-SNR also have a remarkable performance when you deal with the noise-free monaural speech separation task. Still, we trained the clean speech separation system with SI-SNR and OSI-SNR, separately. Like Fig.5 shows, although they all converge to same point and have a same performance eventually, training process using OSI-SNR converges more rapidly. Both the training losses and validation losses decrease faster. The lines with * markers are usually below the lines with diamond markers.

As Fig.2 demonstrated, when θ becomes bigger, the difference between SI-SNR and OSI-SNR also grows. So, the adoption of OSI-SNR becomes important when we deal with noisy monaural speech separation task. The more noise we face, the more performance improvement we will observe when using OSI-SNR instead of SI-SNR as a training objective function.

E. CONCLUSION

In this paper, we have developed a new objective function for speech separation systems, which is very easy to adopt in speech processing systems. A curriculum learning method is proposed to improve training process when tackle the noisy problem. Systematic evaluation demonstrated that our OSI-SNR combined with curriculum learning improves separation performance substantially in terms of all metrics, SDRi, SI-SNRi, OSI-SNRi, PESQ and STOI.

ACKNOWLEDGMENT

This work is supported by Huawei Innovation Research Program.

REFERENCES

[1] E. L. J. George, S. T. Goverts, J. M. Festen, and T. Houtgast, "Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners," Journal of Speech, Language, and Hearing Research, vol. 53, pp. 1429–1439, 2010.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noiserobust automatic speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, pp. 745–777, 2014.
[3] D. Gelbart and N. Morgan, “Double the trouble: handling noise and reverberation in far-field automatic speech recognition.,” in INTERSPEECH, 2002.
[4] K. A. Al-Karawi, A. H. Al-Noori, F. F. Li, and T. Ritchings, “Automatic speaker recognition system in adverse conditions - implication of noise and reverberation on system performance,” International Journal of Information and Electronics Engineering, vol. 5, pp. 423–427, 2015
[5] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, “Discriminatively trained recurrent neural networks for single-channel speech separation,” in IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2014, pp. 577–581.
[6] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1562–1566.
[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, pp. 7–19, 2015.
[8] Y. Zhao, D. L. Wang, I. Merks, and T. Zhang, “DNN-based enhancement of noisy and reverberant speech,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 6525–6529.
[9] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speakerindependent multi-talker speech separation,” in IEEE Int. Conf. on Acoust., Speech and Signal Process. IEEE, 2017, pp. 241–245.
[10] 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 Trans. on Audio, Speech, and Lang. Process., vol. 25, no. 10, pp. 1901–1913, 2017.
[11] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in IEEE Int. Conf. on Acoust., Speech and Signal Process. IEEE, 2016, pp. 31–35.
[12] Y. Isik, J. Le Roux, Z. Chen, S.Watanabe, and J. R. Hershey, “Single-channel multi-speaker separation using deep clustering,” INTERSPEECH, pp. 545–549, 2016.
[13] Z. Chen, Y. Luo, and N. Mesgarani, “Deep attractor network for single-microphone speaker separation,” in IEEE Int. Conf. on Acoust., Speech and Signal Process. IEEE, 2017, pp. 246–250.
[14] Y. Luo and N. Mesgarani, "TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," *2018 IEEE* International *Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, 2018, pp. 696-700.
[15] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256-1266, Aug. 2019.
[16] Z. Li, Y. Song, L. Dai and I. McLoughlin, "Source-Aware Context Network for Single-Channel Multi-Speaker Speech Separation," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 2018, pp. 681-685.
[17] G. Wichern, J. Antognini, M. Flynn, et al. "WHAM!: Extending Speech Separation to Noisy Environments," in INTERSPEECH, pp. 1368-1372. 2019.
[18] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Joint separation and denoising of noisy multi-talker speech using recurrent neural networks and permutation invariant training,” in Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Sep. 2017, pp. 1–6.
[19] L. Drude and R. Haeb-Umbach, “Integration of neural networks and probabilistic spatial models for acoustic blind source separation,”IEEE Journal of Selected Topics in Signal Processing, 2019.
[20] Bengio, Yoshua, et al. “Curriculum Learning.” Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 41–48.
[21] J. L. Roux, S. Wisdom, H. Erdogan and J. R. Hershey, "SDR – Half-baked or Well Done?," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 626-630.
[22] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
[23] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2001, vol. 2, pp. 749–752.
[24] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125–2136, 2011.
[25] A. Varga and H. J. M. Steeneken. Assessment for automatic speech recognition: Ii.noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun., 12(3):247–251, 1993.