

SDR – HALF-BAKED OR WELL DONE?

Jonathan Le Roux¹, Scott Wisdom², Hakan Erdogan³, John R. Hershey²

¹Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

²Google AI Perception, Cambridge, MA

³Microsoft Research, Redmond, WA

ABSTRACT

In speech enhancement and source separation, signal-to-noise ratio is a ubiquitous objective measure of denoising/separation quality. A decade ago, the `BSS_eval` toolkit was developed to give researchers worldwide a way to evaluate the quality of their algorithms in a simple, fair, and hopefully insightful way: it attempted to account for channel variations, and to not only evaluate the total distortion in the estimated signal but also split it in terms of various factors such as remaining interference, newly added artifacts, and channel errors. In recent years, hundreds of papers have been relying on this toolkit to evaluate their proposed methods and compare them to previous works, often arguing that differences on the order of 0.1 dB proved the effectiveness of a method over others. We argue here that the signal-to-distortion ratio (SDR) implemented in the `BSS_eval` toolkit has generally been improperly used and abused, especially in the case of single-channel separation, resulting in misleading results. We propose to use a slightly modified definition, resulting in a simpler, more robust measure, called scale-invariant SDR (SI-SDR). We present various examples of critical failure of the original SDR that SI-SDR overcomes.

Index Terms— speech enhancement, source separation, signal-to-noise-ratio, objective measure

1. INTRODUCTION

Source separation and speech enhancement have been an intense focus of research in the signal processing community for several decades, and interest has gotten even stronger with the recent advent of powerful new techniques based on deep learning [1–11]. An important area of research has focused on single-channel methods, which can denoise speech or separate one or more sources from a mixture recorded using a single microphone. Many new methods are proposed, and their relevance is generally justified by their outperforming some previous method according to some objective measure.

While the merits of various objective measures such as PESQ [12], Loizou’s composite measure [13], PEMO-Q [14], PEASS [15], or STOI [16], could be debated and compared, we are concerned here with an issue with the way the widely relied upon `BSS_eval` toolbox [17] has been used. We focus here on the single-channel setting. The `BSS_eval` toolbox reports objective measures related to the signal-to-noise ratio (SNR), attempting to account for channel variations, and to report a decomposition of the overall error, referred to as signal-to-distortion ratio (SDR), into components indicating the type of error: source image to spatial distortion ratio (ISR), signal to interference ratio (SIR), and signal to artifacts ratio (SAR). In version 3.0, `BSS_eval` featured two main functions, `bss_eval_images` and `bss_eval_sources`.

- `bss_eval_sources` completely forgives channel errors that can be accounted for by a time-invariant 512-tap filter, modifying the reference to best fit each estimate. This includes very strong modifications of the signal, including low-pass or high-pass filters. Thus, obliterating some frequencies of a signal by setting them to 0 could absurdly still result in near infinite SDR.
- `bss_eval_images` reports channel errors (including gain errors) as errors in the ISR measure, but its SDR is nothing else than vanilla SNR. While not as fatal as the modification of the reference in `bss_eval_sources`, `bss_eval_images` suffers from some issues. First, it does not even allow for a global rescaling factor, which may occur when one tries to avoid clipping in the reconstructed signal. Second, as does SNR, it takes the scaling of the estimate at face value, a loophole that algorithms could (potentially unwittingly) exploit, as explained in section 2.2.

An earlier version (2.1) of the toolbox does provide, among other functions, a decomposition which only allows a constant gain via the function `bss_decomp_gain`. Performance criteria such as SDR can then be computed from this decomposition, but most papers on single-channel separation appear to be using `bss_eval_sources`. The `BSS_eval` website¹ actually displays a warning about which version should be used. Version 3.0 “*is recommended for mixtures of reverberated or diffuse sources (aka convolutive mixtures), due to longer decomposition filters enabling better correlation with subjective ratings. It [is] also recommended for instantaneous mixtures when the results are to be compared with SiSEC.*” On the other hand, version 2.1 “*is practically restricted to instantaneous mixtures of point sources. It is recommended for such mixtures, except when the results are to be compared with SiSEC.*” It appears that this warning has not been understood, and most papers use Version 3.0 without further consideration. The desire to compare results to (early editions of) SiSEC should also not be a justification for using a flawed measure. The same issues apply to an early Python version of `BSS_eval`, `bss_eval`² [18]. Recently, `BSS_eval` v4 was released as a Python implementation³ [19]: the authors of Version 4 acknowledged the issue with the original `bss_eval_sources`, and recommended using `bss_eval_images` instead. This however does not address the scaling issue.

These problems shed doubt on many results, including some in our own older papers, especially in cases where algorithms differ by a few tenths of a dB in SDR. This paper is intended both to illustrate and propagate this message more broadly, and also to encourage the use, for single-channel separation evaluation, of simpler, scale-aware, versions of SDR: scale-invariant SDR (SI-SDR) and scale-dependent SDR (SD-SDR). We also propose a definition

¹http://bass-db.gforge.inria.fr/bss_eval/

²http://github.com/craffel/mir_eval/

³<https://sigsep.github.io/sigsep-mus-eval/museval.metrics.html>

we obtain:

$$\text{SD-SDR} = 10 \log_{10} \left(\frac{\|\alpha s\|^2}{\|s - \hat{s}\|^2} \right) = \text{SNR} + 10 \log_{10} \alpha^2 \quad (6)$$

Going back to the example in Section 2.2, SI-SDR is independent of the rescaling of x , while SD-SDR for $\hat{s} = \mu x$ is equal to

$$10 \log_{10} \left(\frac{\|\mu s\|^2}{\|s - \mu x\|^2} \right) = 10 \log_{10} \left(\frac{\mu^2 \|s\|^2}{\|(1 - \mu)s - \mu n\|^2} \right) \quad (7)$$

$$= 10 \log_{10} \left(\frac{\mu^2}{(1 - \mu)^2 + \mu^2} \right), \quad (8)$$

which does peak at $\mu = 1$. While this measure properly accounts for down-scaling errors where $\mu < 1$, it only decreases to -3 dB for large up-scaling factors $\mu \gg 1$. For those applications where both down-scaling and up-scaling are critical, one could consider the minimum of SNR and SD-SDR as a relevant measure.

2.4. SI-SIR and SI-SAR

In the original `BSS_eval` toolkit, the split of SDR into SIR and SAR is done in a mathematically non intuitive way: in the original paper, the SAR is defined as the “sources to artifacts ratio,” not the “source to artifacts ratio,” where “sources” refers to all sources, including the noise. That is, if the estimate contains more noise, yet everything else stays the same, then the SAR actually goes up. There is also no simple relationship between SDR, SIR, and SAR.

Similarly to `BSS_eval`, we can further decompose e_{res} as $e_{\text{res}} = e_{\text{interf}} + e_{\text{artif}}$, where e_{interf} is defined as the orthogonal projection of e_{res} onto the subspace spanned by both s and n . But differently from `BSS_eval`, we define the scale-invariant signal to interference ratio (SI-SIR) and the scale-invariant signal to artifacts ratio (SI-SAR) as follows:

$$\text{SI-SIR} = 10 \log_{10} \left(\frac{\|e_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2} \right), \quad (9)$$

$$\text{SI-SAR} = 10 \log_{10} \left(\frac{\|e_{\text{target}}\|^2}{\|e_{\text{artif}}\|^2} \right). \quad (10)$$

These definitions have the advantage over those of `BSS_eval` that they verify

$$10^{-\text{SI-SDR}/10} = 10^{-\text{SI-SIR}/10} + 10^{-\text{SI-SAR}/10}, \quad (11)$$

because the orthogonal decomposition leads to $\|e_{\text{res}}\|^2 = \|e_{\text{interf}}\|^2 + \|e_{\text{artif}}\|^2$. There is thus a direct relationship between the three measures. Scale-dependent versions can be defined similarly.

That being said, we feel compelled to note that, whether it is still relevant to split SDR into SIR and SAR is a matter of debate: machine-learning based methods tend to perform a highly non-stationary type of processing, and using a global projection on the whole signal may thus not be guaranteed to provide the proper insight.

3. EXAMPLES OF EXTREME FAILURE CASES

We present some failure modes of SDR that SI-SDR overcomes.

3.1. Optimizing a filter to minimize SI-SDR

For this example, we optimize an STFT-domain, time-invariant filter to minimize SI-SDR. We will show that despite SI-SDR being minimized by the filter, SDR performance remains relatively high since it is allowed to apply filtering to the reference signal.

Optimization of the filter that minimizes SI-SDR is implemented in Keras with a Tensorflow backend, where the trainable weights are an F -dimensional vector \mathbf{w} . A sigmoid nonlinearity is applied to this vector to ensure the filter has values between 0 and 1, and the final

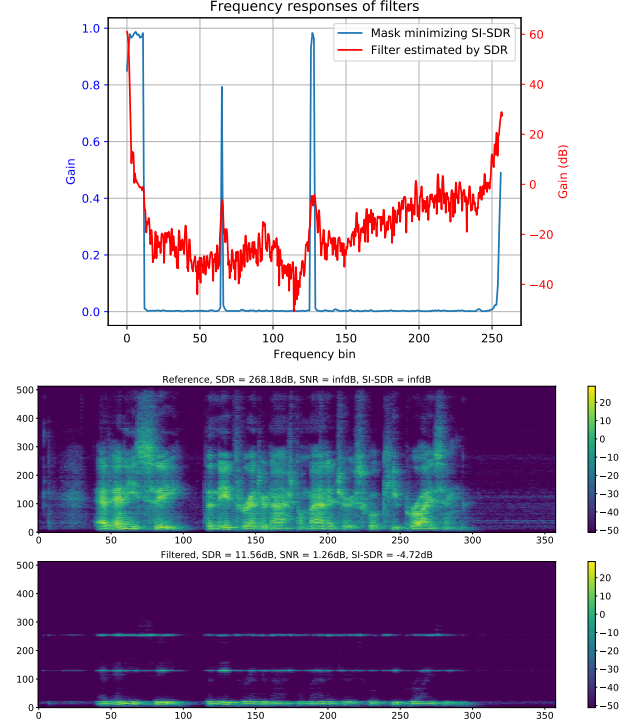


Fig. 2. Top: filter applied to a clean speech signal that minimizes SI-SDR (blue) and magnitude response of the FIR filter estimated by SDR (red). Bottom: spectrograms of a clean speech signal (top) and the same signal processed by the optimized filter in blue above.

filter \mathbf{m} is obtained by renormalizing $\mathbf{v} = \text{sigm}(\mathbf{w})$ to have unit ℓ_∞ -norm: $\mathbf{m} = \mathbf{v} / \|\mathbf{v}\|_\infty$. The filter is optimized on a single speech example using gradient descent, where the loss function being minimized is SI-SDR. Application of the masking filter is implemented end-to-end, where gradients are backpropagated through an inverse STFT layer.

An example of a learned filter and resulting spectrograms for a single male utterance from CHiME2 is shown in Fig. 2. To minimize SI-SDR, the filter learns to remove most of the signal’s spectrum, only passing a couple of narrow bands. This filter achieves -4.7 dB SI-SDR, removing much of the speech content. However, despite this destructive filtering, we have the paradoxical result that the SDR of this signal is still high at 11.6 dB, since `BSS_eval` is able to find a filter to be applied to the reference signal that removes similar frequency regions. This filter is shown in red in the top part of Fig. 2, somewhat matching the filter minimizing SI-SDR in blue.

3.2. Progressive deletion of frequency bins

The previous example illustrated that SDR can yield high scores despite large regions of a signal’s spectrum being deleted. Now we examine how various metrics perform when frequency bins are progressively deleted from a signal.

We add white noise at 15 dB SNR to the same speech signal used in Section 3.1. Then time-invariant STFT-domain masking is used to remove varying proportions of frequency bins, where the mask is bandpass with a center frequency at the location of median spectral energy of the speech signal averaged across STFT frames. We measure four metrics: SDR, SNR, SI-SDR, and SD-SDR. The re-

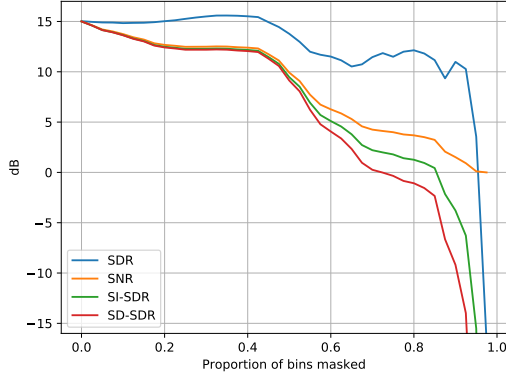


Fig. 3. Various metrics plotted versus proportion of frequency bins deleted for a speech signal plus white noise at 15dB SNR.

sults are shown in Fig. 3. Despite more and more frequency bins being deleted, SDR (blue) remains between 10 dB and 15 dB, until nearly all frequencies are removed. In fact, SDR even *increases* for a masking proportion of 0.4. In contrast, the other metrics more appropriately measure signal degradation since they monotonically decrease.

An important practical scenario in which such behavior would be fatal is that of bandwidth extension: it is not possible to properly assess the baseline performance, where upper frequency bins are silent, using SDR.

3.3. Varying band-stop filter gain for speech corrupted with band-pass noise

In this example, we consider adding bandpass noise to a speech signal, then applying a mask that filters the noisy signal in this band with varying gains, as a crude representation of a speech enhancement task. We mix the speech signal with a bandpass noise signal, where the local SNR within the band is 0 dB, and the band is 1600 Hz wide (20% of the total bandwidth for a sampling frequency of 16 kHz), centered at the maximum average spectral magnitude across STFT frames of the speech signal. In this case, the optimal time-invariant Wiener filter should be bandstop, with a gain of 1 outside the band and a gain of about 0.5 within the band, since the speech and noise have approximately equal power, and the Wiener filter is $P_{\text{speech}} / (P_{\text{speech}} + P_{\text{noise}})$.

We consider the performance of such filters when varying the band-stop gain from 0 to 1 in steps of 0.025, again for SDR, SNR, SI-SDR, and SD-SDR. The results are shown in Fig. 4. Notice that SNR, SI-SDR have a peak around a gain of 0.5 as expected. However, SDR monotonically increases as gain decreases. This is an undesirable behavior, as SDR becomes more and more optimistic about signal quality as more of the signal’s spectrum is suppressed, because it is all too happy to see the noisy part of the spectrum being suppressed and modify the reference to focus only on the remaining regions. SD-SDR peaks slightly above 0.5, because it penalizes the down-scaling of the speech signal within the noisy band.

4. COMPARISON ON A SPEECH SEPARATION TASK

Both SI-SDR and BSS_{eval} ’s SDR have recently been used by various studies [6–9, 11, 21–23, 25, 26] in the context of single-channel speaker-independent speech separation on the wsj0-2mix dataset [6], some of these studies reporting both figures [21–23, 25]. We gather

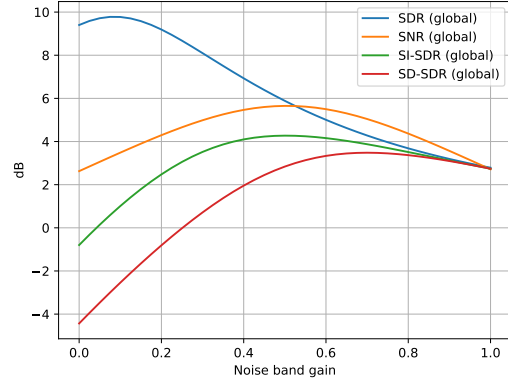


Fig. 4. Various metrics plotted versus bandstop filter gain for a speech signal plus bandpass white noise with 0dB SNR in the band.

Table 1. Comparison of improvements in SI-SDR and SDR for various speech separation systems on the wsj0-2mix dataset test set [6].

Approaches	SI-SDR [dB]	SDR [dB]
Deep Clustering [6, 7]	10.8	-
Deep Attractor Networks [22, 25]	10.4	10.8
PIT [8, 9]	-	10.0
TasNet [26]	10.2	10.5
Chimera++ Networks [11]	11.2	11.7
+ MISI-5 [11]	11.5	12.0
WA [21]	11.8	12.3
WA-MISI-5 [21]	12.6	13.1
Conv-TasNet-gLN [23]	14.6	15.0
Oracle Masks:		
Magnitude Ratio Mask	12.7	13.2
+ MISI-5	13.7	14.3
Ideal Binary Mask	13.5	14.0
+ MISI-5	13.4	13.8
PSM	16.4	16.9
+ MISI-5	18.3	18.8
Ideal Amplitude Mask	12.8	13.2
+ MISI-5	26.6	27.1

in Table 1 various SI-SDR and BSS_{eval} SDR improvements (in dB) on the test set of the wsj0-2mix dataset mainly from [11], to which we add the recent state-of-the-art score of [23]. The difference between the SI-SDR and the SDR scores for the algorithms considered are around 0.5 dB, but vary from 0.3 dB to 0.6 dB. Note furthermore that the algorithms considered here all result in signals that can be considered of good perceptual quality: much more varied results could be obtained with algorithms that give worse results. If the targets and interferences in the dataset were more stationary, such as in some speech enhancement scenarios, it is also likely there could be loopholes for SDR to exploit, where a drastic distortion that can be well approximated by a short FIR filter happens to lead to similar results on the mixture and the reference signals.

5. CONCLUSION

We discussed issues that pertain to the way BSS_{eval} ’s SDR measure has been used, in particular in single-channel scenarios, and presented a simpler scale-invariant alternative called SI-SDR. We also showed multiple failure cases for SDR that SI-SDR overcomes.

Acknowledgements: The authors would like to thank Dr. Shinji Watanabe (JHU) and Dr. Antoine Liutkus and Fabian Stötter (Inria) for fruitful discussions.

6. REFERENCES

- [1] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. ISCA Interspeech*, 2013.
- [2] F. J. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. GlobalSIP Machine Learning Applications in Speech Processing Symposium*, 2014.
- [3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, 2014.
- [4] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Apr. 2015.
- [5] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA)*, 2015.
- [6] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2016.
- [7] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. ISCA Interspeech*, Sep. 2016.
- [8] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2017.
- [9] 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, 2017.
- [10] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," in *arXiv preprint arXiv:1708.07524*, 2017.
- [11] 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)*, Apr. 2018.
- [12] 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. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.
- [13] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.
- [14] R. Huber and B. Kollmeier, "Pemo-q – a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, 2006.
- [15] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, 2011.
- [16] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.
- [17] 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, Jul. 2006.
- [18] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "mir_eval: A transparent implementation of common mir metrics," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2014.
- [19] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Proc. International Conference on Latent Variable Analysis and Signal Separation (LVA)*, 2018.
- [20] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.
- [21] Z.-Q. Wang, J. Le Roux, D. Wang, and J. R. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. ISCA Interspeech*, Sep. 2018.
- [22] Z. Chen, Y. Luo, and N. Mesgarani, "Deep Attractor Network for Single-Microphone Speaker Separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.
- [23] Y. Luo and N. Mesgarani, "TasNet: Surpassing ideal time-frequency masking for speech separation," *arXiv preprint arXiv:1809.07454*, Sep. 2018.
- [24] S. Venkataramani, R. Higa, and P. Smaragdis, "Performance based cost functions for end-to-end speech separation," *arXiv preprint arXiv:1806.00511*, 2018.
- [25] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.
- [26] Y. Luo and N. Mesgarani, "TasNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," in *arXiv preprint arXiv:1711.00541*, 2017.