SA-SDR: A NOVEL LOSS FUNCTION FOR SEPARATION OF MEETING STYLE DATA

Thilo von Neumann[1]*, Keisuke Kinoshita*[2]*, Christoph Boeddeker*[1]*, Marc Delcroix*[2]*,*
Reinhold Haeb-Umbach[1]

[1]Paderborn University, Germany [2]NTT Corporation, Japan

ABSTRACT

Many state-of-the-art neural network-based source separation systems use the averaged Signal-to-Distortion Ratio (SDR) as a training objective function. The basic SDR is, however, undefined if the network reconstructs the reference signal perfectly or if the reference signal contains silence, e.g., when a two-output separator processes a single-speaker recording. Many modifications to the plain SDR have been proposed that trade-off between making the loss more robust and distorting its value. We propose to switch from a mean over the SDRs of each individual output channel to a global SDR over all output channels at the same time, which we call source-aggregated SDR (SA-SDR). This makes the loss robust against silence and perfect reconstruction as long as at least one reference signal is not silent. We experimentally show that our proposed SA-SDR is more stable and preferable over other well-known modifications when processing meeting-style data that typically contains many silent or single-speaker regions.

***Index Terms* —** Source Separation, Permutation Invariant Training, Signal-to-Distortion Ratio, Loss Function

1. INTRODUCTION

Source separation is an important pre-processing step for many other systems, such as speech recognition or diarization, that often cannot handle recordings of overlapping speech. Advances in the past years using neural network-based source separators have led to impressive results on fully overlapped clean anechoic recordings [1–4]. More realistic and challenging scenarios like meeting-style data recently gained research interest [5–11], where speakers do not fully overlap and a separation system has to handle a varying number of speakers including silence.

Many state-of-the-art separation systems, like the Time-domain Audio Separation Network (TasNet) [1, 2], maximize the Signal-to-Distortion Ratio (SDR) as the objective during training. However, the standard SDR becomes problematic (1) when the system is asked to reconstruct silence, as it is often required in realistic meeting-style conversations when one speaker listens while another utters, and (2) when it reconstructs one reference signal very well. In both cases, the value of the SDR explodes.

Many works address this problem by modifying the SDR for each estimated separated signal [12–17]. To address the instabilities for perfect reconstruction, one can limit the value range of the SDR by introducing a soft maximum [12] or by skewing its curve [16]. Instabilities due to silent targets can be addressed by switching to a log-Mean Squared Error (MSE) variant [17, 15] or by adding small values to the fraction in the SDR [13]. All of these modifications distort the loss value for each separated speech signal.

We propose not to modify the SDR definition for each output channel, but the way it is aggregated across outputs. The common way of aggregation is a simple arithmetic mean over the individual SDRs of each output, — the averaged SDR (A-SDR), e.g., [3, 1, 4]. We propose to transition from these "local" SDRs to a "global" SDR that combines all outputs to one long signal before computing the SDR. This is done by summing the energies of all targets and all error terms — the source-aggregated SDR (SA-SDR).

We found experimentally that the proposed SA-SDR achieves one of the best performances among the presented losses, measured with various metrics, and comes without any hyperparameters to tune. Making the loss robust against silence is important for training on realistic meeting-style data where such a case frequently occurs so that more training data can be used. Limiting the value range of the SDR in general improves the performance of the trained models. We additionally propose to use SA-SDR as a signal-level evaluation metric for meting-style data where the classical SDR cannot be computed. The SA-SDR measures in one metric both how well active and silent sources are estimated.

2. CONVENTIONAL LOSS FUNCTIONS: SDR AND ITS VARIANTS

We consider speech mixtures $\mathbf{y} \in \mathbb{R}^T$ of K speakers. A mixture signal $\mathbf{y} = \sum_{k=1}^{K} \mathbf{s}_k + \mathbf{n}$ is the sum of the speech of individual speakers $\mathbf{s}_k \in \mathbb{R}^T$ and noise $\mathbf{n} \in \mathbb{R}^T$. All signals are represented as vectors of samples with a time length of T.

The process of obtaining estimates $\hat{\mathbf{s}}_k$ for the clean reference signals $\mathbf{s}_k$ from the mixture $\mathbf{y}$ is called source separation. The estimates $\hat{\mathbf{s}}_k$ should reconstruct the clean signals $\mathbf{s}_k$ as closely as possible up to a permutation between estimates and references.

The Signal-to-Distortion Ratio (SDR) – and variations of it – is a commonly used training objective and evaluation metric for such source separation models. In its basic form, it is defined for a pair of an estimated signal $\hat{\mathbf{s}}$ and a corresponding reference signal $\mathbf{s}$:

$$\mathrm{SDR}(\hat{\mathbf{s}}, \mathbf{s}) = 10 \log_{10} \frac{\|\mathbf{s}\|^2}{\|\hat{\mathbf{s}} - \mathbf{s}\|^2}. \tag{1}$$

The estimation error $\hat{\mathbf{s}} - \mathbf{s}$ is to be minimized at the output of a source separator, so the objective to be minimized becomes the negative SDR, $\mathcal{L}^{(\mathrm{SDR})} = -\mathrm{SDR}$, for each output channel.

We only consider scale-dependent SDRs here, but the same conclusions could be made with scale-invariant losses by re-scaling the target as $\mathbf{s}^{(\text{re-scaled})} = \mathbf{s} \frac{\mathbf{s}^\mathsf{T} \hat{\mathbf{s}}}{\|\mathbf{s}\|^2}$.

The plain SDR is undefined if the target signal is silent ($\mathbf{s} = \mathbf{0}$), when, e.g., a two-output separator is trained to process a single-speaker utterance, or if the reconstruction is perfect ($\hat{\mathbf{s}} = \mathbf{s}$). Even if these edge-cases are not hit, the SDR explodes if the reference is close to $\mathbf{0}$ or the estimation is almost perfect. It is often desired to train with silent references, especially with realistic training data, and perfect reconstruction should never be a problem.

The remainder of this section discusses different modifications to the plain SDR that make it robust. Just preventing the loss value from exploding for silent references often just moves the problem to a later point in training. Since a network can trivially estimate silence, it can easily learn to reconstruct silence (almost) perfectly, so the loss additionally needs to counter perfect reconstruction.

2.1. Soft Maximum

One way to make the SDR robust against perfect reconstruction is to impose a soft maximum with the thresholded SDR (tSDR) [12]:

$$\mathcal{L}^{(\text{tSDR})} = -10 \log_{10} \frac{\|\mathbf{s}\|^2}{\|\mathbf{s} - \hat{\mathbf{s}}\|^2 + \tau \|\mathbf{s}\|^2}, \tag{2}$$

where $\tau = 10^{-\text{SDR}_{\text{max}}/10}$. It can be made robust against silence by adding a small constant $\varepsilon > 0$ to the reference signal [13]:

$$\mathcal{L}^{(\varepsilon\text{-tSDR})} = -10 \log_{10} \frac{\|\mathbf{s}\|^2 + \varepsilon}{\|\mathbf{s} - \hat{\mathbf{s}}\|^2 + \tau(\|\mathbf{s}\|^2 + \varepsilon)}. \tag{3}$$

Note that both τ and ε do not influence the direction of the gradient of $\mathcal{L}^{(\varepsilon\text{-tSDR})}$ but only its scaling and the ratio between different output channels. The ε-tSDR thus gives a smaller weight to the output channels and examples that are well separated.

2.2. Skewing the SDR

Another variation of SDR that tries to combat the numerical instabilities for perfect reconstruction is the skewed SDR [16]. It is originally formulated in a scale-invariant way, but we only consider the scale-dependent variant:

$$\mathcal{L}^{(\text{skewed SDR})} = -10 \log_{10} \frac{\|\mathbf{s}\|^2}{\|\mathbf{s} - \hat{\mathbf{s}}\|^2 + \nu \|\hat{\mathbf{s}}\|^2} \tag{4}$$

The skewing factor $\nu > 0$ controls how much the loss value is skewed for small reconstruction errors. The additional term $\nu\|\hat{\mathbf{s}}\|^2$ pushes the estimation towards $\mathbf{0}$ in the scale-dependent variant.[1] It is unclear if it is well-suited for source separation.

2.3. log-MSE

A simple way to avoid the instability for silent targets is to ignore the numerator $\|\mathbf{s}\|^2$. This leads to the log-MSE loss [17]:

$$\mathcal{L}^{(\text{log-MSE})} = \log_{10} \|\mathbf{s} - \hat{\mathbf{s}}\|^2. \tag{5}$$

It has the same gradients as $\mathcal{L}^{(\text{SDR})}$ but scaled differently. As discussed earlier, the loss additionally has to be made robust against perfect reconstruction, e.g., by adding a constant to the argument of the logarithm [14]:

$$\mathcal{L}^{(\text{log1p-MSE})} = \log_{10}(\|\mathbf{s} - \hat{\mathbf{s}}\|^2 + 1). \tag{6}$$

The log-MSE loss has the disadvantage that its value depends on the scaling of the signals and thus varies more and is more difficult to interpret than the SDR. Especially when the best model is selected based on the development loss, a sub-optimal model might be selected. Similar modifications are possible as for the SDR-based variants, such as adding a soft minimum similar to Eq. (2) [15].

[1] This effect is not present in the scale-invariant variant, but we only consider scale-dependent losses here.

2.4. Extra Loss for Silence

A different way to handle problematic inputs is to identify them and use an alternative loss where the SDR is not applicable. One example for this is using the mixture signal $\mathbf{y}$ instead of the target in a thresholded loss where the target is silent but the mixture is not [15], here as a variant of the log-tMSE:

$$\mathcal{L}_0^{(\text{log-tMSE})} = 10 \log_{10}(\|\hat{\mathbf{s}}\|^2 + \tau \|\mathbf{y}\|^2). \tag{7}$$

This loss is only applied where the target is silent, i.e., $\mathbf{s} = \mathbf{0}$. Applying different losses to different outputs can create discontinuities in the gradients. Besides that, the decision which outputs are silent is not always trivial, e.g., for very short segments of speech.

3. AGGREGATING SDR ACROSS OUTPUTS

The modifications discussed so far all modify the SDR for each individual output. But, a source separator has multiple outputs and the losses for different outputs have to be combined. Most source separation techniques that use an SDR-based loss average the loss over the output channels, e.g., [1, 2, 5, 12, 13, 17]. For the standard SDR, this can be written as

$$\mathcal{L}^{(\text{A-SDR})} = \frac{1}{10} \sum_{k=1}^{K} \mathcal{L}(\mathbf{s}_k, \hat{\mathbf{s}}_k) = -\frac{10}{K} \sum_{k=1}^{K} \log_{10} \frac{\|\mathbf{s}_k\|^2}{\|\mathbf{s}_k - \hat{\mathbf{s}}_k\|^2}. \tag{8}$$

Extensions to all other single-channel losses described in Section 2 are straightforward. We call this conventional way of combining the single-channel SDRs the averaged SDR (A-SDR). It suffers from the aforementioned problems with the standard SDR: It becomes unstable if any output channel has perfect reconstruction or a silent reference signal.

We propose to stabilize the loss by, instead of computing the arithmetic mean, summing the energies of the targets and distortions:

$$\mathcal{L}^{(\text{SA-SDR})} = -10 \log_{10} \frac{\sum_{k=1}^{K} \|\mathbf{s}_k\|^2}{\sum_{k=1}^{K} \|\mathbf{s}_k - \hat{\mathbf{s}}_k\|^2}. \tag{9}$$

This is equivalent to concatenating all output channels to compute a global SDR and we call it source-aggregated SDR (SA-SDR). It is stable as long as at least one reference is not perfectly reconstructed and at least one is not completely silent. The case of complete silence, i.e., all reference signals are zero, is not considered here since separation is trivial in that case and silence can easily be detected.

Both A-SDR and SA-SDR are aggregations over the SDRs of the individual output channels and thus bounded by them, i.e.,

$$\min_k \text{SDR}(\hat{\mathbf{s}}_k, \mathbf{s}_k) \leq \text{A-SDR} \leq \max_k \text{SDR}(\hat{\mathbf{s}}_k, \mathbf{s}_k), \tag{10}$$

$$\min_k \text{SDR}(\hat{\mathbf{s}}_k, \mathbf{s}_k) \leq \text{SA-SDR} \leq \max_k \text{SDR}(\hat{\mathbf{s}}_k, \mathbf{s}_k). \tag{11}$$

From this follows that for a special case where the SDRs of all individual output channels are equal ($\text{SDR}(\hat{\mathbf{s}}_1, \mathbf{s}_1) = \text{SDR}(\hat{\mathbf{s}}_2, \mathbf{s}_2) = \ldots$), A-SDR and SA-SDR are also equal.

3.1. Energy of the Reference Signals

The A-SDR weights each output channel equally, independent of its energy level. This is often not desired: When a reference signal contains only a short segment of speech (i.e., low energy), it gets weighted the same as a longer speech signal in another output channel. This gives the samples in the short speech fragment an extraordinarily large weight. The SA-SDR is less sensitive to these outliers as it implicitly weights the output channels by their energy and focuses less on low-energy signals.

3.2. Energy of the Distortions

Having a single well-separated output signal $\hat{\mathbf{s}}_l$ is enough to push the A-SDR to extremely good values even if other outputs are separated poorly. The A-SDR thus focuses the already well separated outputs while the SA-SDR minimizes the total distortions by focusing the poorly separated outputs. This can be seen from the gradients.

The gradient of the l-th output $\hat{\mathbf{s}}_l$ of A-SDR depends only on the l-th output signal

$$\nabla_{\hat{\mathbf{s}}_l} \mathcal{L}^{\text{(A-SDR)}} = \frac{20}{K \ln 10} \frac{\hat{\mathbf{s}}_l - \mathbf{s}_l}{\|\hat{\mathbf{s}}_l - \mathbf{s}_l\|^2}, \tag{12}$$

while the gradients of SA-SDR depend on all output signals:

$$\nabla_{\hat{\mathbf{s}}_l} \mathcal{L}^{\text{(SA-SDR)}} = \frac{20}{\ln 10} \frac{\hat{\mathbf{s}}_l - \mathbf{s}_l}{\sum_k \|\hat{\mathbf{s}}_k - \mathbf{s}_k\|^2}. \tag{13}$$

One would expect the gradients of the output with worse quality to be larger, i.e., $\left\|\nabla_{\hat{\mathbf{s}}_k} \mathcal{L}^{\text{(A-SDR)}}\right\| > \left\|\nabla_{\hat{\mathbf{s}}_l} \mathcal{L}^{\text{(A-SDR)}}\right\|$ if $\|\hat{\mathbf{s}}_k - \mathbf{s}_k\|^2 > \|\hat{\mathbf{s}}_l - \mathbf{s}_l\|^2$. But the opposite is true for A-SDR:

$$\frac{\left\|\nabla_{\hat{\mathbf{s}}_k} \mathcal{L}^{\text{(A-SDR)}}\right\|}{\left\|\nabla_{\hat{\mathbf{s}}_l} \mathcal{L}^{\text{(A-SDR)}}\right\|} = \frac{\|\hat{\mathbf{s}}_l - \mathbf{s}_l\|}{\|\hat{\mathbf{s}}_k - \mathbf{s}_k\|} < 1 \text{ if } l \text{ is better separated.} \tag{14}$$

The SA-SDR has the expected behavior:

$$\frac{\left\|\nabla_{\hat{\mathbf{s}}_k} \mathcal{L}^{\text{(SA-SDR)}}\right\|}{\left\|\nabla_{\hat{\mathbf{s}}_l} \mathcal{L}^{\text{(SA-SDR)}}\right\|} = \frac{\|\hat{\mathbf{s}}_k - \mathbf{s}_k\|}{\|\hat{\mathbf{s}}_l - \mathbf{s}_l\|} > 1 \text{ if } l \text{ is better separated.} \tag{15}$$

The SA-SDR is not only an elegant way to make the SDR robust against silent targets and perfect reconstruction in common use-cases where some speakers make a pause, it also leads to a better balance between the output channels.

4. EXPERIMENTS

4.1. Data

We evaluate the different loss functions on fully overlapped mixtures from the WSJ0-2mix database [18] and on artificially generated meetings [13] based on WSJ [19]. Each meeting is about 120 s long, contains 5-8 speakers, an overlap ratio between 0.2 and 0.4 and is corrupted by white microphone noise of 20 dB to 30 dB. Following the ideas of Continuous Speech Separation (CSS) [5], there are never more than two speakers overlapping at the same time.

We randomly cut 2 s long segments from the meetings for training. This segment size was shown to work well on this data in [13].

4.2. Model Training

We use a Dual-Path Recurrent Neural Network (DPRNN) [2] with two outputs and the default configuration from [2] for experiments on fully overlapped data, i.e., six blocks, a feature size of 64 and a window size and shift of 100 and 50, respectively. To speed up our experiments on meeting-style data, we use a shallower model with only three blocks. We train all models with Permutation Invariant Training (PIT) for the same number of iterations with the same batch size. We pick the best checkpoint for evaluation based on the loss on the development set.

We use a stitching approach [5, 13] to evaluate our model on the 120 s long meetings. The input signal is segmented into overlapping segments of 2.4 s length, each segment is processed by the separator, and the separated signals from adjacent segments are aligned to minimize the mean squared error between the overlapping signal parts. The stitcher uses a future and history context of 1 s each.

We choose $\nu = 0.3$ for the skewed losses, and set $\text{SDR}_{\text{max}} = 30\,\text{dB}$ and $\varepsilon = 10^{-6}$ for the thresholded losses (prefixed with "t").

4.3. Metrics

4.3.1. Word Error Rate (WER)

To obtain a Word Error Rate (WER), we use a speech recognizer from the ESPnet toolkit [20] trained on clean WSJ data. It achieves a WER of 5.6 % on the clean eval92 set of WSJ.

We do not compute the WER for the full meetings because of two reasons: The speech recognizer poorly generalizes to long signals and the alignment of estimated transcriptions with the ground truth is difficult. We therefore cut the separated signals at the ground truth utterance boundaries and compute the average WER over these utterances, choosing the output channel with the lower WER. This explicitly ignores regions in the output that should be silent.

4.3.2. Signal-to-distortion Ratio (SDR)

As a signal-level metric, we compute the BSSEval-SDR [21, 22]. Similar to WER, it is not meaningful to compute the BSSEval-SDR over a whole output signal for meeting-style data because each output channel can contain more than one utterance and processing one output channel would follow the source-aggregated idea while the channels are averaged, i.e., the aggregation would be a mixture of source-aggregation and averaging. BSSEval-SDR usually uses averaging, hence we use the same processing as for WER: We cut utterances from the separated signals and compute the BSSEval-SDR for each utterance independently. For WSJ0-2mix, we compute the BSSEval-SDR over the full signals using the *min* sub-set.

4.3.3. Attenuation Ratio for Silence

To judge how well the systems can suppress speech where the output should be silent, we compute an attenuation ratio

$$\text{attenuation-ratio} = 10 \log_{10} \frac{\left\|\mathbf{y}^{\text{(sil)}}\right\|^2}{\left\|\hat{\mathbf{s}}^{\text{(sil)}}\right\|^2}, \tag{16}$$

where $\mathbf{y}^{\text{(sil)}}$ and $\hat{\mathbf{s}}^{\text{(sil)}}$ are the signal parts that should be silent in the mixture and separated streams, respectively. When the evaluated system favors a suppression, e.g. as the skewed SDR, the value may be overoptimistic for those systems.

4.3.4. Voice Activity Error Rate (VAER)

We use webrtcvad[2] to obtain hypotheses for speech activity from the separated signals. From these, we compute a Voice Activity Error Rate (VAER) by comparing the estimated speech activity with the ground truth overlap-free voice activity labels using pyannote [23]. This metric has the advantage compared to the WER, SDR and attenuation ratio that it judges the quality of the full output streams including silence. It, however, only judges how well the system can discriminate where speakers are active and not the separation quality in general.

4.3.5. SA-SDR with Graph-PIT

As a signal-level metric that measures the overall quality of a system output, we propose to use the SA-SDR. We use the ideas from Graph-PIT[3] [13, 24] to construct reference signals for the output streams for meeting-style data from the reference utterance signals because the placement of utterances on output channels is irrelevant. A-SDR is not well applicable here for the same reasons as BSSEval-SDR while SA-SDR does not depend on the placement of utterances. The optimal assignment of utterances to output channels for the reference signal is much more efficient to compute for SA-SDR than for A-SDR [24].

[2] https://github.com/wiseman/py-webrtcvad

[3] https://github.com/fgnt/graph_pit

Table 1. Comparison of the separation performance of A-SDR and SA-SDR on WSJ0-2mix. Separation performance is evaluated with BSS-eval SDR [22].

Loss	BSSEval SDR	A-SDR	SA-SDR
no separation	0.2	0.0	0.0
A-SDR [1, 2]	17.8	17.5	17.8
A-tSDR [12]	17.8	17.5	17.8
SA-SDR	**18.0**	**17.7**	**18.0**
SA-tSDR	17.7	17.5	17.8

Table 2. Comparison of the separation performance of SDR variants on meeting-style data. Averaged losses are prefixed with "A-" and source-aggregated losses with "SA-". Best numbers are **bold** and best numbers among conventional averaged SDRs are underlined.

Loss	#spk train	Metrics				
		WER	atten. ratio	BSSEval SDR	VAER	SA-SDR
no separation	—	48.1	0.0	7.3	65.6	0.0
A-SDR [1]	2	13.5	25.5	19.1	12.6	13.8
A-log-MSE [17]	2	13.1	18.3	19.5	13.2	14.8
A-log1p-MSE [14]	1+2	13.5	25.3	19.6	9.9	16.8
A-skewed-SDR [16]	2	15.6	24.7	18.7	12.5	10.1
A-tSDR [12]	2	13.6	21.1	18.8	14.0	13.3
A-ε-tSDR [13]	1+2	12.8	25.9	19.6	11.8	15.5
A-log-tMSE+$\mathcal{L}_0$ [15]	1+2	12.8	26.4	19.6	10.7	14.5
SA-SDR	1+2	12.5	30.3	19.8	9.7	16.1
SA-log-MSE	1+2	13.3	**31.5**	19.3	11.6	14.7
SA-log1p-MSE	1+2	15.1	25.1	18.7	11.4	15.7
SA-skewed-SDR	1+2	15.1	28.9	18.6	12.6	10.6
SA-tSDR	1+2	**12.2**	30.8	**19.9**	**8.2**	**17.9**
SA-ε-tSDR	1+2	12.8	27.5	19.6	9.1	16.3

4.4. A-SDR and SA-SDR on Fully Overlapped Data

To show that the SA-SDR is well suited for general source separation purposes, we first compare it with the conventional A-SDR on the common task to separate fully overlapped speech from the WSJ0-2mix database in Table 1. The model trained with SA-SDR performs slightly better than the model trained with A-SDR, while the variants with a threshold, A-tSDR and SA-tSDR, show a comparable performance. The value of SA-SDR as a metric is close to BSSEval-SDR and A-SDR for this data. It is hence an alternative metric for source separation in clean anechoic scenarios.

4.5. A-SDR and SA-SDR on Meeting-Style data

We compare the performance of models trained with different loss variants for meeting-like data in Table 2. Training is performed on 2 s long segments randomly cut from the meeting-style data. These segments can contain any number of speakers so we discard any segments with more than two speakers (that our two-output separator cannot handle) and less than one speaker. Some losses cannot handle single-speaker segments, i.e., when one reference signal is silent. For these losses, we additionally discard all single-speaker segments during training (roughly 50 %). The number of speakers seen during training is indicated by the "#spk train" column in Table 2. Two-speaker segments always contain some speech in each reference signal but are not necessarily fully overlapped.

4.5.1. Averaged SDR Losses

The upper half of Table 2 compares the different conventional averaged losses (prefixed with "A-"). The A-log-MSE and A-SDR, as expected, show similar numbers where the A-log-MSE is slightly worse, probably due to model selection discussed in Section 2.3. Modifying the log-MSE so that it can handle single-speaker training segments improves the VAER significantly while the WER and BSSEval SDR are unchanged. This is expected as additional single-speaker training segments likely improve the silence estimation that is judged by VAER and SA-SDR while WER and BSSEval-SDR only judge speech regions. A similar effect can be observed when switching from A-tSDR to A-ε-tSDR. A-ε-tSDR and log-tMSE+$\mathcal{L}_0$ achieve the best performance in speech regions where A-ε-tSDR is preferable because it does not require switching the loss function depending on the energy of the reference signals. The best overall performance among the averaged loss variants, including silence evaluation, can be achieved with A-log1p-MSE.

The scale-dependent skewed SDR does not seem to be well-suited for training a source separation system. We can observe that the loss pushes the outputs towards silence: The attenuation ratio and VAER are relatively good while all other metrics show a poor performance.

4.5.2. Source-Aggregated SDR Losses

Comparing the averaged losses with the source-aggregated losses, we can observe a consistent improvement for most loss variants. Many modifications of the standard SDR from the upper half of Table 2 improve the performance when averaged because they allow training on single-speaker segments. They, however, distort the loss values, trading-off between more realistic data and an undistorted loss. For the source-aggregated losses that always allow single-speaker segments, they lose the benefit of better training data and most of them degrade the performance (compare "log-1p" and "ε-tSDR" losses). The SA-log1p-MSE variant, for example, is now slightly worse than the SA-log-MSE because of the constant 1.

The overall best performance on meeting-style data can be achieved with the SA-tSDR. It is more elegant and easier to use than the A-ε-tSDR or the A-log-tMSE+$\mathcal{L}_0$ because it has fewer hyperparameters to tune. Very close performance can be achieved in all metrics by the SA-SDR loss that has no hyperparameters.

We observed that the averaged loss variants often become unstable late in training, probably because they focus the best separated outputs. The source-aggregated loss variants that better balance different outputs did not become unstable in our experiments.

5. CONCLUSIONS

We compared different SDR-based objective functions for source separation that allow training a neural-network-based separator on more realistic data, including silent reference signals for single outputs. We found that stabilizing losses for perfect reconstruction and allowing silent targets, which result in training data closer to the evaluation data, often improves the performance. We proposed a novel way of combining the SDRs computed on individual output channels that elegantly addresses the problems of conventional SDR-based losses and improves the performance of trained models.

6. ACKNOWLEDGEMENTS

Computational resources were provided by the Paderborn Center for Parallel Computing. C. Boeddeker was supported by DFG under project no. 448568305.

7. REFERENCES

[1] Y. Luo and N. Mesgarani, "TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 696–700.

[2] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 46–50.

[3] 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, Oct. 2017.

[4] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention Is All You Need In Speech Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 21–25.

[5] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous Speech Separation: Dataset and Analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7284–7288.

[6] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous Speech Separation with Conformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 5749–5753.

[7] T. v. Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, and R. Haeb-Umbach, "All-neural Online Source Separation, Counting, and Diarization for Meeting Analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 91–95.

[8] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A Pre-announcement," in *Machine Learning for Multimodal Interaction*, S. Renals and S. Bengio, Eds., Berlin, Heidelberg, 2006, Lecture Notes in Computer Science, pp. 28–39, Springer.

[9] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[10] M. Van Segbroeck, A. Zaid, K. Kutsenko, C. Huerta, T. Nguyen, X. Luo, B. Hoffmeister, J. Trmal, M. Omologo, and R. Maas, "DiPCo — Dinner Party Corpus," in *Interspeech*. Oct. 2020, pp. 434–436, ISCA.

[11] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, D. Snyder, A. Shanmugam Subramanian, J. Trmal, B. Ben Yair, C. Boeddeker, Z. Ni, S. Horiguchi, N. Kanda, T. Yoshioka, and N. Ryant, "CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *6th International Workshop on Speech Processing in Everyday Environments (CHiME 2020)*, 2020.

[12] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised Speech Separation Using Mixtures of Mixtures," in *Advances in Neural Information Processing Systems*. 2020, vol. 33, pp. 3846–3857, Curran Associates, Inc.

[13] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "Graph-PIT: Generalized permutation invariant training for continuous separation of arbitrary numbers of speakers," in *Interspeech*. 2021, ISCA.

[14] T. von Neumann, C. Boeddeker, L. Drude, K. Kinoshita, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "Multi-Talker ASR for an Unknown Number of Sources: Joint Training of Source Counting, Separation and ASR," in *Interspeech*. Oct. 2020, pp. 3097–3101, ISCA.

[15] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the Fuss about Free Universal Sound Separation Data?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 186–190.

[16] Y. Luo and N. Mesgarani, "Separating Varying Numbers of Sources with Auxiliary Autoencoding Loss," in *Interspeech*. Oct. 2020, pp. 2622–2626, ISCA.

[17] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A Dissecting Approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6359–6363.

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

[19] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete," *Linguistic Data Consortium*, 2007.

[20] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Interspeech*. Sept. 2018, pp. 2207–2211, ISCA.

[21] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[22] C. Raffel, B. Mcfee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. W. Ellis, C. C. Raffel, B. Mcfee, and E. J. Humphrey, "mir_eval: A Transparent Implementation of Common MIR Metrics," in *15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[23] H. Bredin, "pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Interspeech*, Aug. 2017.

[24] T. von Neumann, C. Boeddeker, K. Kinoshita, M. Delcroix, and R. Haeb-Umbach, "Speeding Up Permutation Invariant Training for Source Separation," in *Speech Communication; 14th ITG-Symposium*, Sept. 2020.