CONVOLUTIVE TRANSFER FUNCTION INVARIANT SDR TRAINING CRITERIA FOR MULTI-CHANNEL REVERBERANT SPEECH SEPARATION

Christoph Boeddeker[1]   *Wangyou Zhang*[2]   *Tomohiro Nakatani*[3]
Keisuke Kinoshita[3]   *Tsubasa Ochiai*[3]   *Marc Delcroix*[3]   *Naoyuki Kamo*[3]
Yanmin Qian[2]   *Reinhold Haeb-Umbach*[1]

[1] Paderborn University, Department of Communications Engineering, Paderborn, Germany
[2] SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[3] NTT Corporation, Japan

ABSTRACT

Time-domain training criteria have proven to be very effective for the separation of single-channel non-reverberant speech mixtures. Likewise, mask-based beamforming has shown impressive performance in multi-channel reverberant speech enhancement and source separation. Here, we propose to combine neural network supported multi-channel source separation with a time-domain training objective function. For the objective we propose to use a convolutive transfer function invariant Signal-to-Distortion Ratio (CI-SDR) based loss. While this is a well-known evaluation metric (BSS Eval), it has not been used as a training objective before. To show the effectiveness, we demonstrate the performance on LibriSpeech based reverberant mixtures. On this task, the proposed system approaches the error rate obtained on single-source non-reverberant input, i.e., LibriSpeech test_clean, with a difference of only 1.2 percentage points, thus outperforming a conventional permutation invariant training based system and alternative objectives like Scale Invariant Signal-to-Distortion Ratio by a large margin.

***Index Terms* —** Multi-channel source separation, acoustic beamforming, complex backpropagation, Signal-to-Distortion Ratio

1. INTRODUCTION

Blind speech separation aims at extracting the individual speech signals present in a mixture. It is considered an important signal enhancement step both in human-to-human communication and for a downstream Automatic Speech Recognition (ASR) system. Having been a topic of extensive research for many years, many solutions have been proposed, such as the Independent Component Analysis (ICA) [1], Independent Vector Analysis (IVA) [2], non-negative matrix factorization [3], spatial mixture model based techniques [4], and deep neural network based separation methods, which are the focus of this contribution.

In a popular variant, the purpose of the neural network (NN) is to estimate time-frequency masks, or, stated differently, a speaker presence probability for each speaker and each time-frequency bin, thus taking advantage of the sparsity and W-disjoint orthogonality [5] of speech signals in this domain. The actual source extraction can be carried out either by applying the mask to a channel of the input signal or by acoustic beamforming. The latter requires multi-channel input, but is known to lead to perceptually more pleasing results exhibiting less artifacts [6]. Furthermore, it takes advantage of spatial information, which can be gleaned from a microphone array, and,

thus, usually leads to better word error rates of a downstream ASR task [7].

While earlier publications, such as permutation invariant training (PIT) [8], deep clustering [9] and variants thereof [10], employed neural network training criteria that were defined in the Short Time Fourier Transform (STFT) domain, more recent publications suggest that loss functions defined in the time-domain, such as the (scale invariant) Signal-to-Distortion Ratio (SDR), generally achieve superior separation performance [11, 12]. In fact, the investigation in [13] showed that the advantage of time-domain loss functions is maintained even if the mask estimation is actually carried out in the frequency domain. However, the combination of time-domain NN training criteria and source extraction by beamforming at training time is widely unexplored, and will be the focus of this work.

In this contribution, we consider source separation as a front-end of a downstream ASR task. One might therefore argue that the front-end should best be trained using an ASR-related criterion, as the latter is closer to the ultimate goal of minimal Word Error Rate (WER). This has been attempted in [14] where a sequence-to-sequence neural ASR system was extended to deal with multi-channel multi-source input. While the results on anechoic speech were promising, their performance on reverberant speech was not yet competitive. For single-source acoustic beamforming the Beamnet architecture [15] has been proposed, which utilizes an ASR-related training objective, and the gradient was backpropagated through the beamformer to the neural mask estimator. It was, however, observed in [16] that joint training of the ASR back-end and the enhancement front-end may have some logistic advantages (no need for parallel clean and distorted training data for the training of the enhancement stage), but that they may not lead to the overall best WER performance. We therefore opted to stick to a signal related training objective for the source separation training in this work, which is actually much simpler to realize, and leave a tighter coupling with the back-end to future work.

In this contribution, we focus on reverberant scenarios. A time-domain loss can be extremely sensitive to changes in the input waveform, e.g., caused by the different reflection pattern when selecting another channel of the microphone array as input to the mask estimation network [17]. Although inaudible, those changes can have a drastic impact on the loss, and unfavorably influence the performance. To overcome this problem, we propose a training objective that is invariant to such errors, i.e., Convolutive transfer function Invariant Signal-to-Distortion Ratio (CI-SDR) loss, inspired by the BSS Eval SDR [18] measure. We give experimental evidence that

it is superior to the well-known Scale Invariant Signal-to-Distortion Ratio (SI-SDR) if combined with an Minimum Variance Distortionless Response (MVDR)-based beamformer. On a reverberant source separation task compiled from LibriSpeech data [19], which was developed during the JHU JSALT 2020 workshop, the proposed system achieves a word error rate which is only 1.2 percentage points higher than on the non-reverberant oracle source signals. The CI-SDR PyTorch code is publicly available on GitHub[1].

The paper is organized as follows. In the next section we introduce the signal model underlying our investigations. In Section 3 the system architecture is introduced. To learn the neural mask estimator with a time-domain training criterion, the loss has to be backpropagated through the beamformer coefficient computation. We use a beamformer implementation that requires computing the dominant eigenvector to estimate the steering vectors. We propose using the power estimation to simplify such a computation during training. Section 4 discusses the training objectives, which are experimentally evaluated in Section 5. The paper closes with some conclusions drawn in Section 6.

2. SIGNAL MODEL

Assuming I concurrent speakers and an array of M microphones, the vector of signals at the microphones, $\mathbf{y}_\ell = [y_{\ell,1}, \ldots, y_{\ell,M}]^\top$, at sampling time ℓ can be written as follows

$$\mathbf{y}_\ell = \sum_{i=1}^{I} \mathbf{x}_{\ell,i} + \mathbf{n}_\ell, \tag{1}$$

where

$$\mathbf{x}_{\ell,i} = \mathbf{d}_{\ell,i} + \mathbf{r}_{\ell,i} = \sum_{\tau=0}^{L_\tau^{\text{early}}-1} \mathbf{a}_{\tau,i} s_{\ell-\tau,i} + \sum_{\tau=L_\tau^{\text{early}}}^{L_\tau-1} \mathbf{a}_{\tau,i} s_{\ell-\tau,i}. \tag{2}$$

Here, $\mathbf{x}_{\ell,i}$ is the image of the ith source at the microphones and $\mathbf{n}_\ell$ the noise vector. Further, $\mathbf{a}_{\tau,i}$ is the room impulse response (RIR) vector from the ith source to the microphone array at time lag τ. It can be decomposed in an early and a late part, resulting in an "early" part of the image at the microphones, $\mathbf{d}_{\ell,i}$, and a late contribution $\mathbf{r}_{\ell,i}$, where the former contains the direct signal and early reflections (typically up to 50 ms of the RIR) and the latter the late reverberation.

In the STFT domain this model can be approximated as follows

$$\mathbf{y}_{t,f} = \sum_{i=1}^{I} \mathbf{d}_{t,f,i} + \sum_{i=1}^{I} \mathbf{r}_{t,f,i} + \mathbf{n}_{t,f}, \tag{3}$$

where t and f denote the time frame index and the frequency bin index, respectively. The early and late arriving speech are given by

$$\mathbf{d}_{t,f,i} = \sum_{\lambda=0}^{\Delta-1} \mathbf{a}_{\lambda,f,i} s_{t-\lambda,f,i} \approx \mathbf{v}_{f,i} s_{t,f,i} = \tilde{\mathbf{v}}_{f,i,r} d_{t,f,i,r}, \tag{4}$$

$$\mathbf{r}_{t,f,i} = \sum_{\lambda=\Delta}^{L_\lambda-1} \mathbf{a}_{\lambda,f,i} s_{t-\lambda,f,i}, \tag{5}$$

where Δ is set to correspond to approximately 50 ms.

Our goal is to extract the early part of the source image $d_{t,f,i,r}$ at a reference microphone r for each source i. By doing so, we
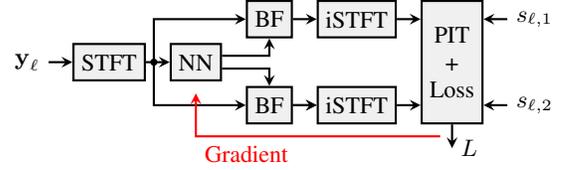
[1]https://github.com/fgnt/ci_sdr

Fig. 1: System overview. The gradient is backpropagated from the time-domain loss through the inverse STFT and beamforming to the NN parameters.

not only aim at removing the competing speakers $i' \neq i$ and noise, but also wish to remove the late reverberation, which is known to significantly degrade recognition performance. For acoustic beamforming the multiplicative transfer function approximation is applied in eq. (4), which also introduces the relative transfer function (RTF) vector $\tilde{\mathbf{v}}_{f,i,r} = \mathbf{v}_{f,i}/v_{f,i,r}$, where $v_{f,i,r}$ is the rth component of the steering vector $\mathbf{v}_{f,i}$.

3. ENHANCEMENT SYSTEM ARCHITECTURE

The block diagram of the enhancement system is depicted in Fig. 1. The multi-channel input $\mathbf{y}_\ell$ is transformed to the STFT domain, where I beamformers (BF) are used to extract the source signals from the mixture. The masks for the beamformer coefficient computation are estimated by a NN, whose input is one of the microphone channels. The loss function for the network training is computed in the time-domain on the beamformed signals and then backpropagated to the NN. Note that a source permutation problem occurs in the loss computation because the network output order may not match the order of the target signals. To address this, we use the PIT loss [8]: We compute the loss for each permutation of the network output and search for its minimum. Then, we take this minimal loss to compute the gradients for backpropagation.

3.1. Mask estimation

The input to the NN is the log magnitude of the STFT (with size 1024 and shift 256 at 16 kHz) of the microphone signal at a reference channel r: $\log(1 + |y_{f,t,r}|)$. The constant is added to avoid $\log(0)$.

The NN consists of 3 Bidirectional Long Short-Term Memory (BLSTM) layers with 600 hidden units in each direction, followed by two feed forward layers. The first feed forward layer keeps the feature size and the second expands it to $512 \cdot 3 \cdot I$. That is, the network outputs three masks per speaker: $m_{\mathbf{d},t,f,i}$, $m_{\mathbf{n},t,f,i}$, and $m_{\tilde{\mathbf{n}},t,f,i}$. They are used for the estimation of

- The spatial covariance matrix of the target speaker $\mathbf{R}_{\mathbf{d},f,i}$,
- The spatial covariance matrix of the distortions (competing speakers plus noise) $\mathbf{R}_{\mathbf{n},f,i}$ of target speaker i, to be used in the MVDR beamformer computation (see Section 3.2), and
- The spatial covariance matrix of the distortions $\mathbf{R}_{\tilde{\mathbf{n}},f,i}$ to be used in the dominant eigenvector estimation (see Section 3.3).

The covariance matrices are computed as follows:

$$\mathbf{R}_{\boldsymbol{\nu},f,i} = \frac{1}{T} \sum_t (\varepsilon + m_{\boldsymbol{\nu},t,f,i}) \mathbf{y}_{t,f} \mathbf{y}_{t,f}^{\mathrm{H}}, \tag{6}$$

where $\boldsymbol{\nu} \in \{\mathbf{d}, \mathbf{n}, \tilde{\mathbf{n}}\}$, and where $\varepsilon = 0.01$ is a small value to improve the training convergence.

Clearly, one could use $\mathbf{R}_{\mathbf{n},f,i} = \mathbf{R}_{\tilde{\mathbf{n}},f,i}$, but we decided to make this distinction because it showed slightly better performance.

3.2. MVDR beamformer

For source extraction we employ the well-known MVDR beam-former [20]:

$$\mathbf{w}_{f,i,r} = \frac{\mathbf{R}_{\mathbf{n},f,i}^{-1} \tilde{\mathbf{v}}_{f,i,r}}{\tilde{\mathbf{v}}_{f,i,r}^{\mathrm{H}} \mathbf{R}_{\mathbf{n},f,i}^{-1} \tilde{\mathbf{v}}_{f,i,r}}, \tag{7}$$

$$\hat{d}_{t,f,i,r} = \mathbf{w}_{f,i,r}^{\mathrm{H}} \mathbf{y}_{t,f}, \tag{8}$$

where $\mathbf{w}_{f,i,r}$ is a vector containing the beamformer coefficients and $\hat{d}_{t,f,i,r}$ is the estimate of the desired signal at the reference channel r.

3.3. Relative transfer function estimation

For the RTF estimation we follow [21]:

$$\mathbf{v}_{f,i} = \mathbf{R}_{\tilde{\mathbf{n}},f,i} \, \mathrm{MaxEig} \left\{ \mathbf{R}_{\tilde{\mathbf{n}},f,i}^{-1} \mathbf{R}_{\mathbf{d},f,i} \right\}, \tag{9}$$

$$\tilde{\mathbf{v}}_{f,i,r} = \mathbf{v}_{f,i}/v_{f,i,r}, \tag{10}$$

where $\mathrm{MaxEig}\{\cdot\}$ extracts the eigenvector corresponding to the largest eigenvalue. However, this approach has a theoretical and practical drawback: First, the gradient is complicated [22]. For this work we used PyTorch and at the time of writing, the support for complex numbers was not yet finished. Therefore, each complex operation was mapped to real operations[2]. However, for the eigenvalue decomposition we couldn't find such a mapping that worked. See [22] for a discussion of some of the issues.

As an alternative we employed the power iteration algorithm to obtain the eigenvector corresponding to the dominant eigenvalue [23]. The algorithm to estimate the RTF is then:

1: $\boldsymbol{\Phi}_{f,i} = \mathbf{R}_{\tilde{\mathbf{n}},f,i}^{-1} \mathbf{R}_{\mathbf{d},f,i}$
2: $\mathbf{v}_{f,i} \leftarrow \mathbf{u}_r$ # $\mathbf{u}_r$ ... one hot vector, i.e., rth component is 1
3: **for** $\eta \in (1, \ldots, \eta_{\max})$ **do**
4:     $\mathbf{v}_{f,i} \leftarrow \boldsymbol{\Phi}_{f,i} \mathbf{v}_{f,i}$
5: $\mathbf{v}_{f,i} \leftarrow \mathbf{R}_{\tilde{\mathbf{n}},f,i} \mathbf{v}_{f,i}$
6: $\tilde{\mathbf{v}}_{f,i,r} = \mathbf{v}_{f,i}/v_{f,i,r}$

where line 2 to line 4 is the power iteration algorithm.

As can be seen, passing the gradient through the power iteration involves rather elementary operations, such as the derivative of a matrix inverse. For those complex-valued gradients, the mapping to real operations works just fine.

4. TRAINING OBJECTIVES

In the following description of loss functions we ignore the source permutation problem in the notation for better readability.

4.1. Frequency-Dependent SDR

As a reference, we first consider a loss function that is defined in the STFT domain, the mean squared error between the complex-valued target signal, $d_{t,f,i,r}$, and its estimate obtained at the beamformer output $\hat{d}_{t,f,i,r}$:

$$L^{\mathrm{F-SDR}} = \frac{10}{I} \sum_i \log_{10} \left(\frac{\sum_{t,f} |d_{t,f,i,r} - \hat{d}_{t,f,i,r}|^2}{\sum_{t,f} |d_{t,f,i,r}|^2} \right), \tag{11}$$

where the denominator, which is a constant w.r.t. network parameters, is introduced for better interpretability of the loss as a

frequency-dependent signal-to-distortion ratio[3]. Note that this loss is equivalent to the phase sensitive loss [8], except for the log operation and a scale.

4.2. SDR

Casting the loss of eq. (11) to the time-domain, we obtain the time-domain SDR loss,

$$L^{\mathrm{SDR}} = \frac{10}{I} \sum_i \log_{10} \left(\frac{\sum_\ell |d_{\ell,i,r} - \hat{d}_{\ell,i,r}|^2}{\sum_\ell |d_{\ell,i,r}|^2} \right), \tag{12}$$

which is the loss proposed in [13].

4.3. Scale Invariant SDR

In [13] it is also shown that this loss is closely related to the SI-SDR loss used in [11]:

$$L^{\mathrm{SI-SDR}} = \frac{10}{I} \sum_i \log_{10} \left(\frac{\sum_\ell |d_{\ell,i,r} \hat{a}_i - \hat{d}_{\ell,i,r}|^2}{\sum_\ell |d_{\ell,i,r} \hat{a}_i|^2} \right), \tag{13}$$

where the scaling term $\hat{a}_i$ is introduced to compensate for a potential scaling error between target and estimate. It is computed as

$$\hat{a}_i = \underset{a_i}{\mathrm{argmin}} \left\{ \sum_\ell \left| d_{\ell,i,r} a_i - \hat{d}_{\ell,i,r} \right|^2 \right\}. \tag{14}$$

4.4. Convolutive transfer function Invariant SDR

Separation networks trained with the SI-SDR loss have been reported to deliver very good separation results [11, 12]. However, those observations have been mostly made in single-channel scenarios. When considering a multi-channel reverberant setup, it was shown in [17] that SI-SDR produces strange artifacts. For example, if the SI-SDR is calculated between one channel as estimate and another channel as target, i.e., they have a different RIR, there is no audible difference, but the SI-SDR indicates a huge difference [17].

This observation led to the opinion that the invariance to a short impulse response that is given in the original SDR measure of the BSS Eval toolbox [18], and that was criticized in [24], is actually beneficial in reverberant scenarios. We therefore propose the following training objective, which we call Convolutive transfer function Invariant Signal-to-Distortion Ratio (CI-SDR):

$$L^{\mathrm{CI-SDR}} = \frac{10}{I} \sum_i \log_{10} \left(\frac{\sum_\ell |\sum_\tau s_{\ell-\tau,i} \hat{a}_{\tau,i} - \hat{d}_{\ell,i}|^2}{\sum_\ell |\sum_\tau s_{\ell-\tau,i} \hat{a}_{\tau,i}|^2} \right), \tag{15}$$

$$\hat{a}_{\tau,i} = \underset{a_{\tau,i}}{\mathrm{argmin}} \left\{ \sum_\ell \left| \sum_\tau s_{\ell-\tau,i} a_{\tau,i} - \hat{d}_{\ell,i} \right|^2 \right\}. \tag{16}$$

where the estimation of $\hat{a}_{\tau,i}$ uses a solution of the Wiener-Hopf equation. Note that, unlike in SI-SDR, $\hat{a}_{\tau,i}$ is a finite impulse response filter with 512 coefficients[4]. As far as we know, this is the first time, that CI-SDR is used as a training objective. Both [17, 24] tried to interpret them as metric, here we want to compare them as NN training objectives.

In [24], the authors argued that the CI-SDR measure of the BSS Eval toolbox does not punish all kinds of distortions, e.g., the suppression of some frequencies. Instead, they proposed to use SI-SDR

---

[2] https://github.com/kamo-naoyuki/pytorch_complex

[3] Since NNs want to minimize the loss function, the loss function is the negative SDR.

[4] To be precise, we reimplemented "BSS Eval v3" with gradient support.

Table 1: Scores for no enhancement, beamforming with oracle masks (Wiener Like Mask (WLM) [28]) and Eq. (9), and directly using oracle signals as ASR input.

Signal	Mask	Enh.	PESQ	BSS Eval SDR	STOI	WER
Obs. $y_{\ell,r}$	—	—	1.22	-0.48	0.715	96.4
Obs. $\mathbf{y}_{t,f}$	WLM	MVDR(eig)	2.26	16.37	0.911	3.8
Early $\mathbf{d}_{\ell,i}$	—	—	2.76	19.26	0.920	3.2
Source $s_{\ell,i}$	—	—	4.64	289.48	1.000	3.0

in scenarios of anechoic single channel mixture recordings. Nevertheless, the comparison of SI-SDR with CI-SDR in more realistic scenarios without reverberation did not show a clear advantage of one metric over the other [24].

Note that we here consider a different scenario, multi-channel reverberant recordings. We argue that the above issue of completely suppressing some frequencies cannot occur because of the regularizing effect of the MVDR beamformer. Its distortionless constraint ensures that unwanted solutions that may still drive the SDR to large values are not allowed.

Another difference between eq. (15) and (13): SI-SDR needs a target signal that is aligned with the input. So, a reverberated signal is used as target. Here, we follow [25], who used just the early part (i.e., 50 ms) of the RIR to generate the target. In eq. (15) we use the source signal $s_{\ell,i}$ convolved with the best (in the Mean Square Error (MSE) sense) RIR, of the length 32 ms, as target. We conjecture that this drives the beamformer to have also a dereverberating effect.

5. EXPERIMENTS

5.1. Dataset

For the experiments we used a simulated dataset that was developed during the JSALT 2020 workshop at JHU. This dataset uses the clean utterances from LibriSpeech [19] at a sample rate of 16 kHz and generates 600 h, 3.6 h and 3.6 h of training, development and testing data, respectively. Each mixture contains 2 utterances which either have full or partial overlap. The reverberation time T_{60} and Signal-to-Noise Ratio (SNR) (spherical noise) is uniformly sampled from 0.15 s to 0.6 s and 10 dB to 20 dB, respectively. The RIRs are generated by image method [26, 27] for a circular array with 7 microphones and minimum angle between the speakers is $5°$.

5.2. Results

For judging the trained systems, we compare 4 metrics: Perceptual Evaluation of Speech Quality (PESQ) [29], BSS Eval SDR [18], Short Time Objective Intelligibility (STOI) [30] and Word Error Rate (WER). For the speech enhancement metrics, PESQ, BSS Eval SDR and STOI, the source signal $s_{\ell,i}$ is used as reference. The BSS Eval SDR has to be taken with care, because the CI-SDR objective tries to optimize this score. For the WER calculation we used a pretrained system from ESPnet [31], a transformer-based ASR system [32]. Note that the ASR system was only trained on clean utterances and was not adapted to any enhancement artifacts. Table 1 shows results of some reference systems. Without any enhancement the WER is close to 100 %. Using oracle masks and beamforming, the WER can be drastically reduced. The final two rows show the performance for the source signal $s_{\ell,i}$ and the early signal $d_{\ell,i,r}$ as input to the ASR system, respectively.

Table 2: Separation performance of enhancement (Enh.) systems that use masking or MVDR beamforming. In *MVDR* the RTF estimation is done with 3 power iterations, while *MVDR(eig)* uses the eigenvalue decomposition. The training loss is varied from frequency SDR over time-domain SDR to Convolutive transfer function Invariant SDR.

Train		Test	PESQ	BSS Eval SDR	STOI	WER
Enh.	Loss	Enh.				
Masking	F-SDR	Masking	1.49	8.37	0.787	45.6
Masking	F-SDR	MVDR(eig)	1.94	13.63	0.884	8.2
MVDR	F-SDR	MVDR	1.99	15.38	0.893	7.9
MVDR	SDR	MVDR	1.98	15.08	0.893	6.7
MVDR	SI-SDR	MVDR	2.01	15.58	0.895	6.9
MVDR	CI-SDR	MVDR	2.46	20.40	0.930	4.4
MVDR	CI-SDR	MVDR(eig)	2.50	20.61	0.930	4.2

In table 2 we compare the performance of different loss functions. As a reference, the first row displays the performance of a classical PIT system. The mask estimator is trained with a frequency loss, eq. (11), and source extraction is done by masking. In the second row the enhancement at test time is changed to a beamformer.

The other systems in this table use a beamformer for source extraction, both at training and test time. It can be observed that the system trained with the CI-SDR criterion clearly outperforms the other systems.

If the three power iterations for dominant eigenvector estimation on the test data are replaced by an eigenvector decomposition the WER can be further slightly reduced to 4.2 %. This is only 1.2 percentage points worse than the WER on single-source non-reverberant input, shown in table 1. Comparing this system with an oracle mask (second row of Table 1), our system outperforms it in terms of the speech enhancement metrics and comes close in terms of WER performance: the WER of the oracle mask-based system is only 0.4 percentage points better.

6. CONCLUSIONS

This paper proposes to use a Convolutive transfer function Invariant Signal-to-Distortion Ratio (CI-SDR), i.e., BSS Eval SDR, as the training criterion of a NN supported multi-channel beamforming based source separation system. The effectiveness is shown on an artificially mixed reverberant speech database. It outperforms a classical PIT system, irrespective of whether source extraction is done by masking or by beamforming at test time in that system, but also outperforms systems trained on alternative time-domain objectives. Furthermore, the system is compared with an oracle mask definition, where it outperforms the oracle mask in the speech enhancement metrics and approaches it in terms of WER performance. The final WER is 4.2 %, while the WER of 3.0 % on the non-reverberant single speaker source signals is only 1.2 % better. To achieve this no fine tuning of the ASR system was necessary.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] P. Comon, "Independent component analysis, a new concept?," *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[2] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 70–79, 2006.

[3] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[4] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Normalized observation vector clustering approach for sparse source separation," in *2006 14th European Signal Processing Conference*. IEEE, 2006, pp. 1–5.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[6] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time–frequency masks from spatial features," *Speech communication*, vol. 68, pp. 97–106, 2015.

[7] 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*, vol. 13, no. 4, pp. 815–826, 2019.

[8] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

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

[10] Z. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *ICASSP*. IEEE, 2018, pp. 1–5.

[11] Y. Luo and N. Mesgarani, "TasNet: Time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP*. IEEE, 2018, pp. 696–700.

[12] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 825–838, 2020.

[13] J. Heitkaemper, D. Jakobeit, C. Boeddeker, L. Drude, and R. Haeb-Umbach, "Demystifying TasNet: A dissecting approach," in *ICASSP*. IEEE, 2020, pp. 6359–6363.

[14] X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, "MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition," in *ASRU*. IEEE, 2019, pp. 237–244.

[15] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *ICASSP*. IEEE, 2017, pp. 5325–5329.

[16] J. Heymann, M. Bacchiani, and T. N. Sainath, "Performance of mask based statistical beamforming in a smart home scenario," in *ICASSP*. IEEE, 2018, pp. 6722–6726.

[17] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv preprint arXiv:1910.13934*, 2019.

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

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[20] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[21] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *ICASSP*. IEEE, 2017, pp. 681–685.

[22] C. Boeddeker, P. Hanebrink, L. Drude, J. Heymann, and R. Haeb-Umbach, "Optimizing neural-network supported acoustic beamforming by algorithmic differentiation," in *ICASSP*. IEEE, 2017, pp. 171–175.

[23] R. Mises and H. Pollaczek-Geiringer, "Praktische Verfahren der Gleichungsauflösung.," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 9, no. 1, pp. 58–77, 1929.

[24] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *ICASSP*. IEEE, 2019, pp. 626–630.

[25] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech & Language*, vol. 46, pp. 374–385, 2017.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[27] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.

[28] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *ICASSP*. IEEE, 2015, pp. 708–712.

[29] 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 *ICASSP*. IEEE, 2001, vol. 2, pp. 749–752.

[30] 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.

[31] 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 *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[32] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, et al., "A comparative study on transformer vs RNN in speech applications," in *ASRU*. IEEE, 2019, pp. 449–456.