*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Single-Channel Blind Direct-to-Reverberation Ratio Estimation Using Masking

*Wolfgang Mack, Shuwen Deng, Emanuël A. P. Habets*

International Audio Laboratories Erlangen, Germany
a joint institution of the University of Erlangen-Nuremberg and Fraunhofer IIS.
{wolfgang, emanuel}.{mack, habets}@audiolabs-erlangen.de
shuwen.deng@fau.de

## Abstract

Acoustic parameters, like the direct-to-reverberation ratio (DRR), can be used in audio processing algorithms to perform, e.g., dereverberation or in audio augmented reality. Often, the DRR is not available and has to be estimated blindly from recorded audio signals. State-of-the-art DRR estimation is achieved by deep neural networks (DNNs), which directly map a feature representation of the acquired signals to the DRR. Motivated by the equality of the signal-to-reverberation ratio and the (channel-based) DRR under certain conditions, we formulate single-channel DRR estimation as an extraction task of two signal components from the recorded audio. The DRR can be obtained by inserting the estimated signals in the definition of the DRR. The extraction is performed using time-frequency masks. The masks are estimated by a DNN trained end-to-end to minimize the mean-squared error between the estimated and the oracle DRR. We conduct experiments with different preprocessing and mask estimation schemes. The proposed method outperforms state-of-the-art single- and multi-channel methods on the ACE challenge data corpus.

**Index Terms**: acoustic parameter, direct-to-reverberation ratio (DRR) estimation, time-frequency mask, deep learning, ACE challenge

## 1. Introduction

The reverberation time $T_{60}$ and the direct-to-reverberation ratio (DRR) are essential for many audio processing algorithms. The DRR is defined as the energy ratio between the direct and reverberant parts of a room impulse response (RIR) [1]. It is used, for example, in speech dereverberation [2–4], or source distance estimation [5, 6]. In practice, the DRR often has to be estimated blindly from captured audio, as the RIR is not available.

In 2015, the acoustic characterization of environments (ACE) challenge provided a data corpus to evaluate and compare the performance of $T_{60}$ and DRR estimation algorithms [7]. The estimation of the DRR, thereby, showed to be challenging. Multi-channel recordings are generally exploited for DRR estimation due to the spatial information several microphones provide [8–13]. The direct and reverberation components are usually estimated separately, and the DRR is based on their ratio. Hioka et al. [8] proposed to use two beamformers to estimate the power spectral density (PSD) of direct sound and reverberation, respectively. This algorithm shows the best results in the ACE challenge. In [9], two spatial correlation matrices, one for the direct and the other for reverberant sound, are utilized to estimate the corresponding power spectra. Chen et al. [11] proposed an algorithm based on a theoretical relationship between particle velocities and sound pressure, and the DRR. However, these methods heavily depend on a priori information about the direction-of-arrival (DOA) of the sound source w.r.t. the microphone array. In [14, 15], the authors proposed single- and multi-channel DRR estimation algorithms based on variants of the speech-to-reverberation modulation energy ratio (SRMR) metric, which can be mapped to the DRR, linearly.

For single-channel DRR estimation, state-of-the-art results are obtained using deep neural networks (DNNs), which directly map a feature representation of the input to the DRR. In [16], the authors proposed to use a recurrent neural network (RNN) to learn a nonlinear mapping from 134 frame-based features extracted from the captured speech to the DRR. Xiong et al. [17–19] proposed to estimate the DRR and the $T_{60}$ via a multi-layer perceptron which maps a 2D Gabor feature representation to the $\mathrm{T}_{60}$ and the DRR. The authors showed that jointly estimating the DRR and the $T_{60}$ with a single model yields better results than separate estimation [18]. However, DRR estimation remains challenging, as obtained Pearson correlation coefficients are 0.6 or lower.

Considering the definition of the DRR, it can be obtained based on the (prewhitened) direct and (prewhitened) reverberant signal components (see Section 3) [4]. Consequently, we propose to formulate DRR estimation as an extraction task. Time-frequency masking techniques are widely used to extract desired signal components from a mixture. Typically, such masks are estimated via a DNN from an input feature representation of the captured audio. Subsequently, the masks are applied to the short-time Fourier transform (STFT) representation of the input to extract the desired signal components. Masking techniques have also been applied for dereverberation [20, 21], where the direct and reverberant signal components have been estimated [20]. In this paper, we propose to use masks similar to [20] to obtain direct and reverberant signal components. Subsequently, these components are used for DRR computation. In contrast to [20], we propose to train the mask-DNN end-to-end with the mean-squared-error between the estimated and the oracle DRR.

The remainder of this paper is structured as follows. In Section 2, we introduce a signal model for DRR estimation. Subsequently, in Section 3, we present our proposed methods. The data sets are described in Section 4 followed by a thorough evaluation and comparison of our proposed methods to state-of-the-art in Section 5.

## 2. Problem Formulation

We assume a single microphone capturing a time-domain mixture $y[t]$ consisting of reverberant speech $x[t]$, and noise $v[t]$,

$$y[t] = x[t] + v[t], \tag{1}$$

http://dx.doi.org/10.21437/Interspeech.2020-2171

with the discrete time-index $t$. The speech $x[t]$ is obtained by convolving a source signal $s[t]$ with a RIR $h[t]$, i.e.,

$$x[t] = s[t] * h[t] = s[t] * (h_d[t] + h_r[t]), \tag{2}$$

where $h[t]$ can be divided in a direct $h_d[t]$ and a reverberant component $h_r[t]$. Similarly, we can divide $x[t]$ in its direct $x_d[t] = s[t] * h_d[t]$ and reverberant $x_r[t] = s[t] * h_r[t]$ components. Our objective is to extract from $y[t]$ the direct-to-reverberation ratio (DRR), an acoustic parameter which specifies the ratio of the direct to the reverberant energy in dB. The DRR, thereby, can be obtained from $h[t]$,

$$\begin{aligned}\mathrm{DRR} &= 10\log_{10}\left(\frac{\sum_t h_d^2[t]}{\sum_t h_r^2[t]}\right)\\ &= 10\log_{10}\left(\frac{\sum_{t=t_d-t_0}^{t_d+t_0} h^2[t]}{\sum_{t=0}^{t_d-t_0} h^2[t] + \sum_{t=t_d+t_0}^{\infty} h^2[t]}\right)\mathrm{dB},\end{aligned} \tag{3}$$

where $t_d$ and $t_0$ specify the central sample index and the temporal spread of $h_d$, respectively. From (3), it is clear that the DRR is a purely channel-based parameter which is independent of $s[t]$ and $v[t]$.

Under certain conditions, the DRR is equivalent to the signal-to-reverberation ratio (SRR),

$$\mathrm{SRR} = 10\log_{10}\left(\frac{\sum_t x_d^2[t]}{\sum_t x_r^2[t]}\right)\mathrm{dB}, \tag{4}$$

which is dependent on $s[t]$. For further analysis of the equivalence between the DRR and the SRR, we define the STFT representations of $s$, $y$, $x_d$, $x_r$, $v$ as $S[n,k]$, $Y[n,k]$, $X_d[n,k]$, $X_r[n,k]$, $V[n,k]$, respectively, where $n$ is the time-frame and $k$ the frequency index. In [4], the authors showed the DRR equals the SRR if $|S|$ is frequency independent. This can be verified via Parseval's theorem, with SRR $= \sum_k |S[k]|^2 |H_d[k]|^2 / \sum_k |S[k]|^2 |H_r[k]|^2$. When $|S[k]| = |S|$, then $|S[k]|^2$ can be taken out of the sum in the numerator and denominator in Parseval's theorem and cancels out yielding the same result as the DRR. Additionally, the authors investigated spectral prewhitening methods to enable DRR estimation from the prewhitened versions of $X_d$ and $X_r$. The definition of $X_d$, $X_r$ in STFT domain and the computation of the prewhitening, thereby, showed to be crucial and challenging.

## 3. Proposed Method

Motivated by the findings about the DRR and the SRR in [4], we formulate single-channel DRR estimation as a data-driven signal extraction task. In the next section, we describe the DRR estimation procedure, followed by details of the DNN architecture and training.

### 3.1. Signal-Based DRR Estimation

We propose to estimate two signals, $X_1$ and $X_2$, from $Y$ such that, when inserted into the DRR definition, we obtain an estimate of the DRR in dB, i.e.,

$$\widehat{\mathrm{DRR}} = 10\log_{10}\left(\frac{\sum_{n,k}|X_1[n,k]|^2}{\sum_{n,k}|X_2[n,k]|^2}\right). \tag{5}$$

The signals $X_1$ and $X_2$, thereby, are obtained using time-frequency masks. These masks are estimated with a DNN and subsequently applied element-wise to $Y$ to compute $X_1$ and $X_2$. The use of time-frequency masks for this problem is motivated by [20, 21] where the authors achieved state-of-the-art results for the extraction of direct and reverberant signal components from $Y$.

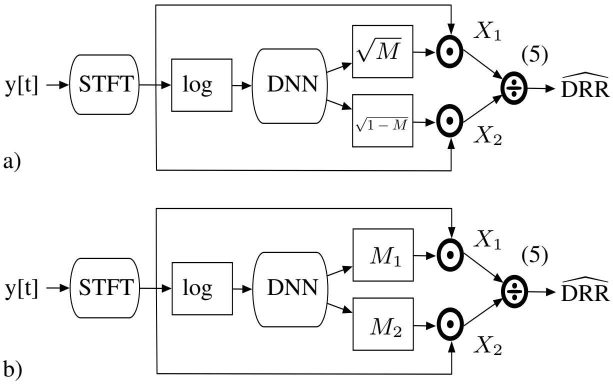

Figure 1: *Overview of the proposed methods for DRR estimation with a single mask in a) and two masks in b). In both cases estimates $X_1$ and $X_2$ are obtained. The DRR is estimated by inserting these estimates in (5).*

#### 3.1.1. Single Mask

First, we assume that the DNN estimates a single mask $M[n,k] \in [0,1]$. An overview of the procedure is given in Figure 1a. As the DRR is a power-based parameter, we propose

$$\begin{aligned}X_1[n,k] &= Y[n,k]\cdot\sqrt{M[n,k]},\\ X_2[n,k] &= Y[n,k]\cdot\sqrt{(1-M[n,k])},\end{aligned} \tag{6}$$

where the overall signal power is preserved as $(|Y[n,k]|\cdot\sqrt{1-M[n,k]})^2 + (|Y[n,k]|\cdot\sqrt{M[n,k]})^2 = |Y[n,k]|^2$. In a frequency band, the relation of the estimated and the residual masks are fixed in (6).

In [4], the authors proposed spectral prewhitening to estimate the DRR from signals, i.e., a frequency band dependent gain $G[k]$. With a single estimated mask with power preservation, such a gain cannot be frequency dependent. In Section 5, we compare the performance of a DNN trained with a spectrally prewhitened $Y[k]\cdot G[k]$ to another DNN trained with the non-prewhitened $Y[k]$. Please note that when prewhitening is applied, the complete pipeline from DNN input to mask application uses the prewhitened $G[k]\cdot Y[k]$ instead of $Y[k]$.

#### 3.1.2. Dual Mask

To allow the DNN to learn the prewhitening, we propose to estimate two independent masks instead of one mask with the DNN, such that

$$\begin{aligned}X_1[n,k] &= Y[n,k]\cdot M_1[n,k],\\ X_2[n,k] &= Y[n,k]\cdot M_2[n,k],\end{aligned} \tag{7}$$

where $M_1$ and $M_2$ are the estimated masks. As $M_1$ and $M_2$ are estimated without the constraint that $M_1^2 + M_2^2 = 1$ as for a single mask, the DNN can learn the prewhitening factor $G[k]$. An overview of the procedure is given in Figure 1b.

### 3.2. DNN Architecture and Training

We define

$$I[n,k] = \log_{10}\left(|Y[n,k]| + \epsilon\right), \tag{8}$$

as DNN input, where $\epsilon \in \mathbf{R}^+$ is a small constant to avoid zeros in the log. The DNN architecture consists of 2 bidirectional long short-term memory layers [22] (BLSTMs) with 600 neurons each followed by a feed-forward layer of shape $(600, 257)$ for one and $(600, 2 \cdot 257)$ for two masks with sigmoid activation to ensure that $M[n,k] \in [0,1]$. The DNN architecture is similar to the one employed in [20], where it was used to extract $X_d$ and $X_r$. Here, we extract $X_1$ and $X_2$, instead.

We propose to train the DNN end-to-end for the DRR with the loss function

$$J = \left(\mathrm{DRR} - \widehat{\mathrm{DRR}}\right)^2, \tag{9}$$

such that we do not need to define $X_1$ and $X_2$ for training, which is required in [20]. Please note that consequently, $X_1$ and $X_2$ may differ from the (prewhitened) $X_d$ and $X_r$ as they are defined by the DNN during training. In addition, as the DRR is a purely channel-based parameter, the end-to-end training approach allows training with measured signals given the DRR is known or can be obtained from measured RIRs.

All DNNs were trained for 100 epochs using the Adam optimizer [23] with a learning rate of 1e-3 and a dropout layer of 0.5 after the first BLSTM layer. The batch-size was 256. For evaluation, we selected the model with the lowest validation loss.

## 4. Data Sets

We generated two training and validation sets using simulated RIRs, noise, and speech from LIBRI [24]. The test set is from the ACE challenge (ACE Eval), which allows direct comparison to state-of-the-art methods [7] (measured RIRs). Each training set consists of 36000, each validation set of 3600, and the test set of 27000 files. Each file has a duration of four seconds and a sampling frequency of 16 kHz.

### 4.1. Data Set Generation

In the following, we describe the process of generating a single file in the training and validation set. We convolve speech from the respective set of LIBRI with the simulated RIRs introduced in Section 4.3. Subsequently, noise is added with a signal-to-noise ratio (SNR) $\in [0, 10, 20]$ dB. The noise is simulated similar to the noise in the ACE challenge and is of type ambient, fan, or babble. For the noise simulation, we adopted the procedure of [25].

The ACE Eval corpus consists of 6 different microphone configurations with 4500 samples, each. The configurations are *Single* (1 mic.), *Chromebook* (2 mic.), *Mobile* (3 mic.), *Crucif* (5 mic.), *Lin8Ch* (8 mic.), and *EM32* (32 mic.). For single-channel evaluation, only the first microphone was used. We refer to the merge of all first microphones of all configurations as seventh configuration *All*.

The following processing is applied to training, validation, and test files. The files are normalized in time-domain such that $\max(|y[t]|) = 1$ and cut/zero-padded to a length of four seconds. Please note that the normalization showed to be crucial. After the normalization, $y$ is transformed in STFT domain with a hop-size of 10 ms and a Hann window of 32 ms.

### 4.2. Prewhitening

In Section 3.1, we described that extraction-based methods require prewhitening to estimate the DRR. For that, we compute the prewhitening factor $G$ from the speech files in the LIBRI training set by

$$\frac{1}{G[k]} = \sqrt{\frac{1}{FN}\sum_{f=1}^{F}\sum_{n=1}^{N}|S_f[n,k]|^2}, \tag{10}$$

where $N$ is the total number of time-frames per file, $f$ is the file index and $F$ the total number of files in the training set. We conduct experiments with and without prewhitening, yielding a training, validation, and test set with and without prewhitening. Subsequently, prewhitening is marked by the subindex $\circ_{\mathrm{pw}}$ with the DNN name.

### 4.3. Room Impulse Response Generation

The training and validation RIRs are generated using the source-image-method [26, 27]. The simulation parameters include seven rooms as specified by the ACE challenge [7]. We considered different source-microphone distances $\in \{0.4\,\mathrm{m}, 0.5\,\mathrm{m}, 0.7\,\mathrm{m}, 1.1\,\mathrm{m}, 1.3\,\mathrm{m}, 1.5\,\mathrm{m}, 1.7\,\mathrm{m}, 2\,\mathrm{m}, 3\,\mathrm{m}\}$ for which 10 source-microphone positions were sampled per room. This configuration yields $7 \cdot 9 \cdot 10 = 630$ source-microphone position pairs. For the rooms, we consider a reverberation time ranging from 0.3 s to 0.7 s, with an increment of 0.1 s, to cover a similar DRR range as present in the ACE challenge.

To generate one RIR, we randomly sample from the $\mathrm{T}_{60}$ range and the source-microphone position pairs. For each reverberant speech file, a new RIR is generated. We calculate the oracle DRR from the RIR as in (3). Since the corresponding equalization filters in [28] are not available, the index $t_d$ is obtained by selecting the absolute peak position of the RIR, and $t_0$ is set to 128 as in [28].

## 5. Performance Evaluation

We evaluate three DNNs on the ACE Eval set for DRR estimation and report the bias, the mean-squared-error (MSE), and the Pearson correlation coefficient ($\rho$) as proposed by the ACE challenge. We trained two DNNs with (5), (6), (9), one with and the other without prewhitening. These single mask models are denoted by SiM and $\mathrm{SiM}_{\mathrm{pw}}$, respectively. The third model is trained without prewhitening for two masks with (5), (7), (9) without prewhitening and is denoted by DuM.

### 5.1. Evaluation of Different Microphone Configurations

We compare our proposed methods on all files in ACE Eval in Table 1 in the microphone configuration *All.* The proposed DuM performs best in terms of MSE compared to SiM and $\mathrm{SiM}_{\mathrm{pw}}$. In our experiments, the MSE contained a lot of outliers, which makes a model comparison based on the MSE hard (as shown in Figure 2). In terms of the correlation coefficient $\rho$, all proposed models perform on par. Contrary to our expectations, neither prewhitening nor the use of two masks helped to improve $\rho$. For that reason and due to space constraints, we subsequently only report the results of DuM as it achieved the lowest MSE and bias in Table 1.

The ACE challenge provides performance results of different state-of-the-art algorithms for different microphone configurations in ACE Eval. We report these results for recapitulation and the performance of DuM in Table 1. Our proposed

| Method | Mic. Conf. | Bias | MSE | $\rho$ |
|---|---|---|---|---|
| DuM | *All* | **-0.6** | **9.0** | 0.62 |
| SiM | *All* | -1.02 | 9.7 | 0.61 |
| SiM$_{\text{pw}}$ | *All* | -1.29 | 10.2 | **0.63** |
| DuM | *Single* | **-0.18** | **8.4** | **0.71** |
| ROPE [19] | *Single* | - | - | 0.56 |
| jROPE-IV [18] | *Single* | - | - | 0.62 |
| NIRAv2 [16] | *Single* | -1.85 | 14.8 | 0.56 |
| DuM | *Mobile* | **-0.24** | **6.4** | **0.69** |
| PSD* [8] | *Mobile* | 1.07 | 8.1 | 0.58 |
| DuM | *Chr.b.* | **-1.52** | **13.7** | 0.30 |
| DENBE* [13] | *Chr.b.* | -4.25 | 34.1 | **0.31** |
| DuM | *Crucif* | **-0.19** | **8.0** | **0.73** |
| NOSRMR* [14] | *Crucif* | -4.1 | 31.1 | 0.08 |
| DuM | *EM32* | **0.18** | **7.3** | **0.50** |
| Part. vel.* [11] | *EM32* | -2.38 | 10.4 | 0.45 |
| DuM | *Lin8Ch* | -1.67 | 10.3 | 0.61 |

Table 1: *Experimental results for blind DRR estimation on ACE Eval of our proposed and state-of-the-art methods. The results of the state-of-the-art methods are extracted from the respective papers. Note that [18, 19] estimated the early to reverberation ratio not the DRR and show that $\rho$ can nevertheless be compared. The MSE and bias values are in dB and multi-channel algorithms are marked with *.*

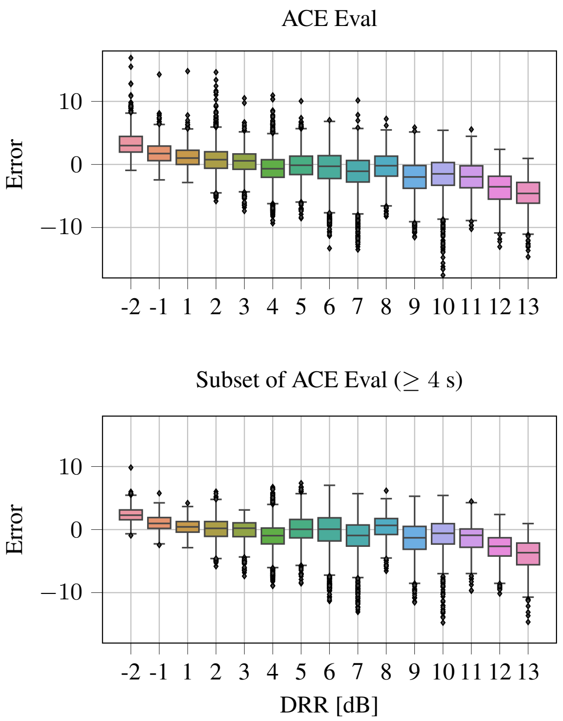

Figure 2: *Estimation error of DuM on ACE Eval and a subset of ACE Eval, where only files larger/equal four seconds were evaluated (16200 files). Error specifies the difference between the estimated (dB) minus the oracle DRR (dB).*

method outperforms single-channel state-of-the-art in terms of MSE, bias, and $\rho$ in the ACE microphone configuration *Single*. The difference in $\rho$ to the second-best method, jROPE-IV, is 0.09. In contrast to our approach, jROPE-IV was trained for $T_{60}$ and a metric similar to the DRR (see caption Table 1) estimation, simultaneously.

When compared to multi-channel DRR estimators, our single-channel method outperforms the baselines in terms of MSE, bias, and $\rho$, except in *Chromebook*. There both methods performed poorly with $\rho \approx 0.3$. Note that the performance of the proposed and the baseline methods strongly depends on the microphone configuration. For the proposed method, the range of $\rho$ is between 0.3 and 0.71 for different microphone configurations. We assume this to be caused by very different configuration-dependent recording circumstances. Please also note that we always used the first microphone in each configuration for our evaluations and that the DRR may be microphone dependent.

### 5.2. Evaluation Using the Oracle DRR

To further investigate the performance of DuM, we show the estimation error over the oracle DRR in the upper plot of Figure 2 for ACE Eval *All*. The mean values of the error have slight offsets at the extrema of the oracle DRR range. For high oracle DRRs, DuM underestimates the DRR, whereas, for low oracle DRRs it tends to overestimate. For oracle DRRs from approximately 1 to 11 dB, the performance is comparable, and the bias is close to zero. The opposite bias at the extrema is expected as the training DRR range is similar to the test DRR range. The DNN was not trained for DRRs surpassing the test range. Consequently, assuming uncertainty, the DNN tends to map inside the training DRR range but not outside, which causes the bias.

We like to draw attention to the large number of outliers in the upper plot of Figure 2. These outliers distort the MSE results such that it is hard to compare methods based on the MSE. To investigate whether the outliers are caused by insufficient temporal context for the DNN, we show a similar boxplot for files of four seconds or longer from ACE Eval in the lower plot of Figure 2. As expected, the number of outliers reduces as more temporal context is given.

## 6. Conclusion

We proposed to estimate the DRR by two DNN-based masking techniques for signal extraction with and without prewhitening. The effect of different masking techniques and prewhitening on the Pearson correlation coefficient was minor, and the MSE and bias were slightly better for the dual masking technique. Our proposed data-driven single-channel methods outperform single- and multi-channel state-of-the-art methods on the widely used and well known ACE Eval set in terms of MSE, bias, and performed better or on par in terms of Pearson correlation coefficient. Furthermore, the recurrent DNN structure allows DRR estimation based on variable length inputs.

## 7. Acknowledgements

The authors would like to thank the Erlangen Regional Computing Center (RRZE) for providing computing resources and support.

# 8. References

[1] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. New York, NY, USA: Springer, 2010.

[2] K. Lebart, J.-M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[3] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, 2010.

[4] P. A. Naylor, N. D. Gaubitch, and E. A. P. Habets, "Signal-based performance evaluation of dereverberation algorithms," *J. of Electr. and Comp. Eng.*, vol. 2010, p. 1, 2010.

[5] Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1793–1805, 2010.

[6] A. Brendel and W. Kellermann, "Learning-based acoustic source-microphone distance estimation using the coherent-to-diffuse power ratio," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Sig. Proc. (ICASSP)*, 2018, pp. 61–65.

[7] A. H. M. J. Eaton, N. D. Gaubitch and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1681–1693, 2016.

[8] Y. Hioka and K. Niwa, "PSD estimation in beamspace for estimating direct-to-reverberant ratio from a reverberant speech signal," in *Proc. ACE Challenge Workshop*, 2015.

[9] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using d/r spatial correlation matrix model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2374–2384, 2011.

[10] Y. Hioka, K. Furuya, K. Niwa, and Y. Haneda, "Estimation of direct-to-reverberation energy ratio based on isotropic and homogeneous propagation model," in *Proc. of the Intl. Workshop on Acoust. Sig. Enh. (IWAENC)*, 2012, pp. 1–4.

[11] H. Chen, P. N. Samarasinghe, T. D. Abhayapala, and W. Zhang, "Estimation of the direct-to-reverberant energy ratio using a spherical microphone array," in *Proc. ACE Challenge Workshop*, 2015.

[12] J. Eaton, A. H. Moore, P. A. Naylor, and J. Skoglund, "Direct-to-reverberant ratio estimation using a null-steered beamformer," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Sig. Proc. (ICASSP)*, 2015, pp. 46–50.

[13] J. Eaton and P. A. Naylor, "Direct-to-reverberant ratio estimation on the ACE corpus using a two-channel beamformer," in *Proc. ACE Challenge Workshop*, 2015.

[14] M. Senoussaoui, J. F. Santos, and T. H. Falk, "SRMR variants for improved blind room acoustics characterization," in *Proc. ACE Challenge Workshop*, 2015.

[15] S. Braun, J. F. Santos, E. A. P. Habets, and T. Falk, "Dual-channel modulation energy metric for direct-to-reverberation ratio estimation," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Sig. Proc. (ICASSP)*, 2018, pp. 206–210.

[16] P. P. Parada, D. Sharma, T. van Waterschoot, and P. A. Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ACE challenge," in *Proc. ACE Challenge Workshop*, 2015.

[17] F. Xiong, S. Goetze, and B. T. Meyer, "Joint estimation of reverberation time and direct-to-reverberation ratio from speech using auditory-inspired features," in *Proc. ACE Challenge Workshop*, 2015.

[18] F. Xiong, S. Goetze, B. Kollmeier, and B. T. Meyer, "Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 255–267, 2018.

[19] F. Xiong, S. Goetze, B. Kollmeier, and B. Meyer, "Exploring auditory-inspired acoustic features for room acoustic parameter estimation from monaural speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1809–1820, 2018.

[20] W. Mack, S. Chakrabarty, F.-R. Stöter, S. Braun, B. Edler, and E. A. P. Habets, "Single-channel dereverberation using direct MMSE optimization and bidirectional LSTM networks." in *Interspeech*, 2018.

[21] D. S. Williamson and D. Wang, "Speech dereverberation and denoising using complex ratio masks," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Sig. Proc. (ICASSP)*, 2017, pp. 5590–5594.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Intl. Conf. on Learn. Repr. (ICLR)*, 2015.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Sig. Proc. (ICASSP)*, 2015, pp. 5206–5210.

[25] H. Gamper and I. J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *Proc. of the Intl. Workshop on Acoust. Sig. Enh. (IWAENC)*, 2018, pp. 136–140.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust.. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[27] E. A. P. Habets, *Room impulse response generator. [Online]*, 2008, available: http://github.com/ehabets/RIR-Generator.

[28] S. Mosayyebpour, H. Sheikhzadeh, T. A. Gulliver, and M. Esmaeili, "Single-microphone lp residual skewness-based inverse filtering of the room impulse response," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1617–1632, 2012.