Interspeech 2018
2-6 September 2018, Hyderabad

Single-Channel Dereverberation Using Direct MMSE Optimization and Bidirectional LSTM Networks

Wolfgang Mack, Soumitro Chakrabarty, Fabian-Robert Stöter, Sebastian Braun, Bernd Edler and Emanuël A.P. Habets

International Audio Laboratories Erlangen, 91058 Erlangen, Germany

wolfgang.mack@fau.de
firstname.lastname@audiolabs-erlangen.de

Abstract

Dereverberation is useful in hands-free communication and voice controlled devices for distant speech acquisition. Single-channel dereverberation can be achieved by applying a time-frequency (TF) mask to the short-time Fourier transform (STFT) representation of a reverberant signal. Recent approaches have used deep neural networks (DNNs) to estimate such masks. Previously proposed DNN-based mask estimation methods train a DNN to minimize the mean-squared-error (MSE) between the desired and estimated masks. Recent TF mask estimation methods for signal separation directly minimize instead the MSE between the desired and estimated STFT magnitudes. We apply this direct optimization concept to dereverberation. Moreover, as reverberation exceeds the duration of a single STFT frame, we propose to use a bidirectional long short-term memory (LSTM) network which is able to take the relation between multiple STFT frames into account. We evaluated our method for different reverberation times and source-microphone distances using simulated as well as measured room impulse responses of different rooms. An evaluation of the proposed method and a comparison with a state-of-the-art method demonstrate the superiority of our approach and its robustness to different acoustic conditions.

Index Terms: dereverberation, LSTM

1. Introduction

When capturing acoustic sources at a far distance from the device, the microphone signals often contain a lot of reverberation. In this case, strong reverberation can harm the intelligibility and quality of the signals for communication applications [1], and decrease the performance of automatic speech recognizers [2]. As early reflections are not harmful to or can even contribute to the speech intelligibility [3], it is sufficient for many applications to focus on the reduction of the late reverberation. There have been proposed multi-channel as well as single-channel dereverberation algorithms [4]. Single-channel dereverberation is still a challenging task because no spatial cues can be used but is advantageous due to low hardware requirements.

A widely used approach is single-channel reverberation suppression in the short-time Fourier transform (STFT) domain using a Wiener filter. The Wiener filter requires knowledge of the late reverberation power-spectral-density (PSD), which has to be estimated in advance. Various models for the late reverberation have been developed: i) a noise sequence with an exponentially decaying envelope [5, 6], ii) moving average processes [7, 8], iii) autoregressive processes [9, 10]. In [11], the late reverberation PSD is estimated using a relative-convolutive-transfer-function (RCTF) model, and dereverberation is achieved by a Wiener filter.

Besides those model-based approaches, deep learning techniques have also been proposed recently for dereverberation. Han et al. [12] proposed to use a deep neural network (DNN) to map a reverberant magnitude spectrum to the desired magnitude spectrum. Direct spectral magnitude estimation, however, was shown to perform worse than mask estimation (cf. [13, 14]). A time-frequency (TF) mask for dereverberation can be compared to a STFT domain Wiener filter with the difference that no PSD estimation is required but a DNN estimates the ratio of each TF bin which belongs to the desired signal. The desired signal is then obtained by applying the mask to the spectrum of the reverberant signal.

There are different types of masks. Binary masks [15] assign TF bins completely to the desired or the undesired signal, whereas soft masks [16] assign ratios of each TF bin. There are two types of soft masks: ratio masks (RMs) [16] which apply a real-valued gain to the magnitude spectrum and complex ratio masks (cRMs) [17] which apply a complex-valued gain to the spectrum.

Williamson et al. [18] estimate a cRM for dereverberation, however, do not perform end-to-end optimization. Instead of minimizing the mean-squared-error (MSE) between the desired and the estimated signals, they train their DNN to minimize the error between the estimated and the desired masks.

In this paper, we propose a DNN which estimates a RM from a reverberant magnitude spectrum and is trained to directly optimize the minimum-mean-squared-error (MMSE) between the estimated and the desired magnitude spectra. In recent experiments in signal separation via TF masking, the direct MMSE optimization was preferred (e.g. [13, 14, 19]) over mask optimization. We chose a DNN architecture which can take the temporal relation of multiple STFT frames into account and is similar to commonly used networks (e.g. [20, 21]) for signal separation or enhancement. Our architecture differs from that proposed in [21] only in the DNN output layer in terms of the dimensionality and the activation function due to the different separation scenario.

The paper is structured as follows. In Section 2, we describe the signal model and the dereverberation process with a RM. In Section 3, we propose a DNN architecture and a loss-function to estimate RMs. The data set generation and an overview of the sets is given in Section 4. Finally, we describe our experiments and results in Section 5 and compare our proposed method to [11] and an oracle RM.

2. Problem Formulation

We assume a single speaker in a room captured by a single microphone. We define the recorded reverberant signal as $X(k, n)$ in the short-time Fourier transform (STFT) domain with fre-

10.21437/Interspeech.2018-1296

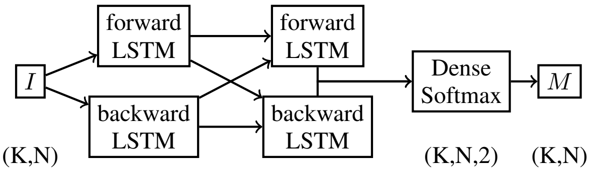

Figure 1: *DNN architecture: forward = in time direction, backward = reverse time direction; 300 Neurons per LSTM*

quency index k and time frame index n. We define the number of frequency bins per time frame of the one-sided discrete STFT spectrum as K and the number of time frames per sample as N. In the STFT domain, the recorded reverberant signal $X(k,n)$ can be decomposed into

$$X(k,n) = X_e(k,n) + X_\ell(k,n), \tag{1}$$

where $X_\mathrm{e}(k,n)$ denotes the direct part plus early reverberation and $X_\ell(k,n)$ the late reverberation. For dereverberation, $X_\mathrm{e}(k,n)$ has to be extracted from $X(k,n)$. In this paper, we propose a RM based algorithm to estimate $X_\mathrm{e}(k,n)$ from $X(k,n)$.

We want to estimate the positive, real-valued RM, denoted by M, which minimizes the mean-squared-error (MSE),

$$\text{MSE} = \sum_{n=0}^{N-1} \sum_{k=0}^{K-1} (|\widehat{X}_\mathrm{e}(k,n)| - |X_\mathrm{e}(k,n)|)^2, \tag{2}$$

where the estimation of $X_\mathrm{e}(k,n)$ is

$$\widehat{X}_\mathrm{e}(k,n) = M(k,n) \cdot X(k,n). \tag{3}$$

Note that the phase of $\widehat{X}_\mathrm{e}$ is equal to the phase of X. The minimization of (2) results in the ideal-ratio-mask (IRM) [16],

$$\text{IRM}(k,n) = \frac{|X_\mathrm{e}(k,n)|}{|X(k,n)|}. \tag{4}$$

The time-domain signal of $\widehat{X}_\mathrm{e}$ is obtained by computing the inverse STFT. Hence, the goal is to obtain an accurate estimate of M. In the following, our approach for mask estimation is presented.

3. Proposed Method

We propose to estimate M with a DNN trained to minimize the MSE between $|X_\mathrm{e}|$ and $|\widehat{X}_\mathrm{e}|$ as in (2). We refer to the procedure as direct MMSE optimization.

3.1. DNN Input-Output Representations and Architecture

With our approach, we process the entire STFT sample of a reverberant signal with a total dimension $N \times K$. One single element of the DNN-input, denoted as I, is

$$I(k,n) = \log_{10}(|X(k,n)| + \epsilon), \tag{5}$$

where a constant $\epsilon \in R^+$ is added to avoid zeros in the logarithm. It is mapped via the DNN to the estimated output mask M of equal shape. Figure 1 depicts the DNN architecture we used. A similar architecture has been proposed by Hershey et al. [21] for binary mask estimation. The DNN consists of two bidirectional long short-term memory (LSTM) [22] layers. LSTMs are a special kind of recurrent DNNs with an internal memory. We chose a recurrent architecture to be able to process time-sequences of variable length. Furthermore, the internal memory of an LSTM allows information to flow through time without significant modification and solves the well-known vanishing gradient problem [23] of normal recurrent DNNs. Bidirectional LSTMs consist of an LSTM where information is passed forward in time and another one where information is passed backward in time. In dereverberation, the temporal context is especially important as X_ℓ and X_e are both time-shifted, filtered versions of the same source signal. We chose the bidirectional LSTM architecture in order to cover the temporal context in the best possible way.

The DNN output is a dense layer with dimension (K,N,2). The activation function is a softmax over the last dimension which introduces the bound $0 \leq M(k,n) \leqslant 1$. The DNN was only trained for M, however, the DNN output yields additionally to M a mask for X_ℓ due to the softmax activation. This can be useful for applications where the reverberant signal is of importance.

Note that the number of time frames N can be variable during training and testing because we chose a recurrent DNN architecture, whereas the number of frequencies K has to be fixed.

3.2. Direct MMSE Optimization

A typically used loss function for mask estimation reduces the error between M and a desired mask (e.g. [18, 21]). In the case of RMs, the loss, denoted as J, is given by

$$J = \sum_{n=0}^{N-1} \sum_{k=0}^{K-1} (\text{IRM}(k,n) - M(k,n))^2. \tag{6}$$

Optimization according to this loss function has two disadvantages. Firstly, the IRM is ill-defined. For $X(k,n) = 0$ and $X_\mathrm{e}(k,n) \neq 0$, the IRM given by (4) tends to infinity. A solution is to add a small constant to the denominator in (4) if $X(k,n)$ approaches zero or to compress the mask as in [18]. Secondly, and more severely, the error in (6) only correlates with the ratio of $|X(k,n)|$ and $|X_\mathrm{e}(k,n)|$ but not with their actual magnitudes. Hence, low-amplitude TF bins can have a significant impact on the loss.

DNN-based signal separation contributions via mask estimation propose to directly minimize the reconstruction MSE (e.g. [13, 14, 25]). Directly optimizing the reconstructed magnitude signal with an adapted loss, given by (2), solves both problems introduced in (6). As the IRM is no longer part of the loss function, the numerical instability is resolved. In addition, the loss is expressed in terms of the signal magnitudes instead of their ratios which increases the impact of energy-rich TF bins and reduces the impact of energy-poor TF bins. Please note, that the DNN output is still M, only the loss changed.

The softmax output activation of our approach is, in the case of destructive interference of $X_\mathrm{e}(k,n)$ and $X_\ell(k,n)$, not compatible with the goal of estimating the IRM. Destructive interference describes the case for which

$$|X_\mathrm{e}(k,n) + X_\ell(k,n)| < |X_\mathrm{e}(k,n)| < |X_\mathrm{e}(k,n)| + |X_\ell(k,n)|, \tag{7}$$

i.e., where the magnitude of the addition of both signals is smaller than the addition of their magnitudes. To investigate this case, we reformulate the denominator of the IRM given by (4) with (1) as

$$\text{IRM}(k,n) = \frac{|X_\mathrm{e}(k,n)|}{|X_\mathrm{e}(k,n) + X_\ell(k,n)|}. \tag{8}$$

Table 1: *Overview of the data sets: the first row defines the RIR data sets; measured RIRs in Bar-Ilan; simulated RIRs in the others; A and B represent rooms; $\mathcal{D}$ and $\mathcal{T}$ represent the T_{60} and source-microphone distance sets of the RIRs of each room; the third row contains information about the number of different RIRs per set*

	Train-A1	Train-A2	Test-A1	Test-A2	Test-B	Bar-Ilan [24]
T_{60} [s]	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_1$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_B$
Dist. [m]	$\mathcal{D}_1$	$\mathcal{D}_1$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_2$	$\mathcal{D}_B$
Number of RIRs	525	210	525	420	420	156

In the case that (7) holds, then the element $\text{IRM}(k,n)$ performs a magnitude amplification because $\text{IRM}(k,n) \geq 1$ according to (8). If an estimated mask is not bounded to one, estimation errors in M can lead to an amplification of noise or to musical tones when applied to $X(k,n)$ according to (3). We set the output layer activation function of our DNN to a softmax to introduce an upper bound of one to the mask values $M(k,n)$ and mitigate the risk of noise amplification and musical artifacts.

We define the mask with the upper performance bound of our approach in (2) as the oracle mask,

$$M_o(k,n) = \begin{cases} \text{IRM}(k,n), & \text{if IRM}(k,n) \leq 1 \\ 1, & \text{otherwise} \end{cases}. \quad (9)$$

We compare our results in the evaluation with M_o to show the best performance our approach can achieve.

4. Data Sets

DNNs need labeled training data to learn a task. We generated data with separated X_e, X_ℓ and X by simulating the reverberation process with artificial room impulse responses (RIRs) for the training sets and with measured and artificial RIRs for the test sets.

4.1. Room Impulse Responses and Acoustic Parameters

We generated RIRs with the RIR-generator by Habets [26]. It is an implementation of the image-method by Allen and Berkley [27]. We divided each RIR in two parts, one for the generation of X_e and one for X_ℓ. The transition index between the two respective RIRs was set 50 ms after the direct part in the RIR. The RIR index of the direct part was estimated with the energy-decay-curve (EDC) [28] of the respective RIR. We assumed the direct part at the EDC index which marked a decay of 0.01 dB in energy.

The computed RIRs are organized in sets shown in Table 1. For the sets, we consider two simulated rooms, room A with the dimensions 6 m×7.5 m×2.4 m and room B with the dimensions 9 m×4 m×3 m. The measured RIRs are from the Multichannel Impulse Response Database from Bar-Ilan university [24].

Furthermore, we define source-microphone distance (SMD) sets $\mathcal{D}$ and reverberation time sets $\mathcal{T}$, which we assign in Table 1 to the RIR sets:
$\mathcal{T}_1 = \{0.3\text{ s}, 0.5\text{ s}, 0.7\text{ s}, 1\text{ s}, 1.5\text{ s}\}$, $\mathcal{T}_2 = \{0.8\text{ s}\}$,
$\mathcal{T}_3 = \{0.2\text{ s}, 0.3\text{ s}, ..., 1.5\text{ s}\}$, $\mathcal{T}_B = \{0.36\text{ s}, 0.61\text{ s}\}$.
$\mathcal{D}_1 = \{0.5\text{ m}, 0.7\text{ m}, 1\text{ m}, 1.5\text{ m}, 2\text{ m}, 3\text{ m}, 4\text{ m}\}$,
$\mathcal{D}_2 = \{0.6\text{ m}, 2.5\text{ m}, 4.5\text{ m}\}$, $\mathcal{D}_B = \{1\text{ m}, 2\text{ m}\}$.
The RIR sets are used to provide separate training and test sets for the DNN and to investigate the impact of different acoustic parameters on the dereverberation performance in Section 5. In each simulated data set, for every possible combination of elements in $\mathcal{D}$ and $\mathcal{T}$, 15 different RIRs were generated for Train-A1, 30 for Train-A2, and 10 each for Test-A1, Test-A2, and Test-B. Source and microphone positions were, with respect to their distance, randomly selected for each $\mathcal{T}$-$\mathcal{D}$ pair in the training sets. In the test sets, the source-microphone positions were fixed over the T_{60} variation to exclude the position influence in the evaluation over the T_{60}. The minimum distance of a source or a microphone to a wall was 0.5 m.

The Bar-Ilan data set was measured in a single, varechoic room with a linear array of 8 microphones. For each T_{60} element in $\mathcal{T}_B$, there are three linear array variants characterizing the inter-microphone distance [cm]: $[3,3,3,8,3,3,3]$, $[4,4,4,8,4,4,4]$, $[8,8,8,8,8,8,8]$. The source was placed in a half-circle around the array center with the distances in $\mathcal{D}_B$. The half-circle-resolution is 15 degree. For each of these scenarios, we randomly selected a measured RIR of one of the 8 microphones.

4.2. Speaker Data Sets and Parameters

The training and validation speakers were selected from the Libri Free Speech Corpus [29], from the training and validation set respectively. The testing speakers were all from the TIMIT [30] test set. We generated one artificially reverberated training set with Train-A1 and another one with Train-A2 with 18000 training and 2000 validation samples each. The sample duration is five seconds. All signals were resampled to a sampling frequency of 8 kHz. The STFT parameters were: 10 ms hop-size, 25 ms frame-length, and Hann window. We used [11] as reference method with the proposed parameters. The RCTF length was set to 13 and the number of frames modelling the early reverberation to 2 for the simulated RIRs and to 3 for the measured RIRs [24].

5. Performance Evaluation

Our implementation was done in Python with Keras 1.2.2 [31] and Theano [32]. We set the dropout [33] to 0.5, recurrent dropout [34] to 0.2, the clipnorm to 200, the optimizer to *rmsprop*, the batch size to 128, the learning rate to 0.001, $K = 129$ and $N = 500$. The dropout parameters and the clipnorm are equal to those in [35], whereas the batchsize and the input spectrum size parameters were selected to fit in the GPU memory.

We evaluated the two DNNs trained with the RIR sets Train-A1 and Train-A2 from Table 1 by the improvement in terms of Cepstral Distance (CD) [36], in terms of PESQ [37] and in terms of frequency-weighted-segmental-signal-to-reverberation-ratio (fwSegSRR) [38]. Furthermore, we compared our results to those obtained using the method in [11].

5.1. Impact of the Reverberation Time and the Distance on the Dereverberation-Performance

In this subsection, we evaluate the influence of the T_{60} of a room and the SMDs. Figure 2 depicts the influence of the T_{60} of a room and the SMD.

A low T_{60} is equivalent to a low amount of reverberation in the signal. Hence, the recorded signal is not degraded as strongly as with a high T_{60}. As a result, the ΔCD of all algorithms is lower as depicted in Figure 2a. For a T_{60} of 0.2 s,

Table 2: *Results dereverberation (mean-values) of DNN-Train-A1, RCTF [11], and the oracle mask*

	ΔCD			ΔPESQ			ΔfwSegSRR (dB)		
	Oracle	RCTF	DNN-Train-A1	Oracle	RCTF	DNN-Train-A1	Oracle	RCTF	DNN-Train-A1
Train-A1	2.68	0.42	0.90	1.33	0.25	0.49	12.14	2.04	4.29
Test-A1	2.68	0.43	0.89	1.34	0.26	0.49	12.15	2.07	4.28
Test-A2	2.83	0.40	0.93	1.36	0.23	0.48	12.43	1.87	4.40
Test-B	2.78	0.39	0.90	1.34	0.23	0.48	12.15	1.73	4.18
Bar-Ilan [24]	1.42	0.13	0.37	1.01	0.31	0.40	8.79	0.41	1.27

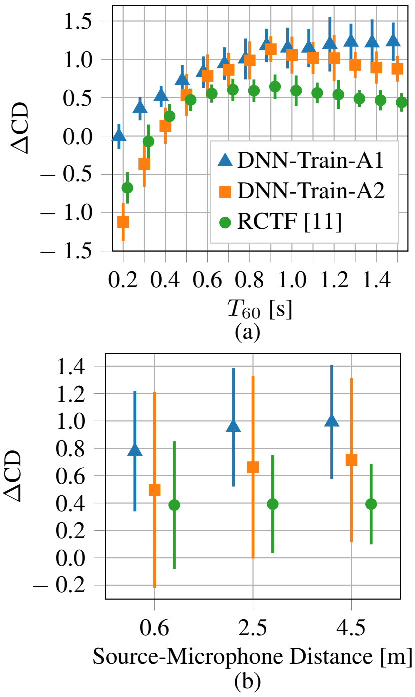

Figure 2: *ΔCD for Test-B over the reverberation time (a) and the source-microphone distance (b); mean-values and standard deviation shown by the error-bars*

there is even a further degradation of the signal in terms of CD. Only the DNN trained with Train-A1 does not further degrade the signal on average. For a higher reverberation time, the improvement of DNN-Train-A1 rises until it saturates at a T_{60} of approximately 0.8 s, whereas the performance of the others decreases. DNN-Train-A1 performs best because it was trained with a variety of T_{60}s. DNN-Train-A2 was trained with a single T_{60} of 0.8 s. Near 0.8 s in Figure 2a, DNN-Train-A2 performs similar to DNN-Train-A1. The further apart of the single training T_{60} of DNN-Train-A2, the higher is the improvement gap between both networks. Interestingly, DNN-Train-A2 also performs worse for T_{60}s lower than 0.8 s. Hence, training a network only with a maximum T_{60} is not expedient. Nevertheless, both DNNs adapted to T_{60}s with which they were not trained. Furthermore, DNN-Train-A1 outperforms [11] for all T_{60}s, regardless whether they were part of the training set or not.

Figure 2b shows the ΔCD over the SMD with distances which were not in the training set. Both DNNs show an increased performance for a higher SMD, whereas [11] is relatively stable in terms of the mean-improvement value. Only the variance is reduced for a higher distance. By training DNN-Train-A1 with more T_{60}s than DNN-Train-A2, the variance of the SMD results was reduced. Furthermore, as none of the depicted SMDs was in the training set, the networks adapt well to unseen SMDs. The results of the PESQ and the fwSegSRR show a similar trend.

5.2. Results of Different Simulated and Measured Rooms

Experiments with respect to the room and measured RIRs are covered in this subsection. We focus here on DNN-Train-A1 and compare it to [11] and to the oracle mask defined in (9). The results are shown in Table 2.

The oracle mask yields the maximum improvement our proposed algorithm can achieve. There is a significant gap in all metrics compared to the oracle results. However, our approach achieved superior results compared to [11] in all metrics.

We tested whether the DNN adapted to the source-microphone positions in Train-A1. Therefore, we evaluated all RIRs in Train-A1 and computed Test-A1 with the same parameters but different room positions. The results depicted in Table 2 show no significant performance difference indicating no overfitting of DNN-Train-A1 to the room positions in Train-A1.

Test-A2 was generated with a different set of T_{60}s and distances. Because of the impact of those parameters as shown in Figure 2, we cannot directly compare the results of Test-A1 or Train-A1 with Test-A2. Test-B has the same T_{60} and SMD parameters as Test-A2 but a different room geometry. The improvement results do not show a major difference which means that the DNN did not overfit to those parameters. Hence, the test performance of DNN-Train-A1 was independent of whether the room, the T_{60}, or the SMD were parameters of the training set. This shows the robustness of our approach.

The results of the measured RIRs of Bar-Ilan show a significantly decreased performance for all methods under test including the oracle mask. We assume this to be caused by the different T_{60} and SMD parameters of the measured RIRs. Both are in areas where the proposed algorithm showed limited results. A normalization of the results with the respective oracle mask results shows a similar performance of our algorithm given simulated and tested RIRs although the DNNs were only trained with simulated RIRs.

6. Conclusion

We presented a single-channel dereverberation approach with a DNN for TF mask estimation which directly optimizes the MMSE between the estimated and the desired magnitude spectra. This approach ensures that the training-loss is related to the signal-amplitudes and, therefore, optimizes directly for the signal magnitude reconstruction. An evaluation showed the superior performance of our algorithm compared to [11]. Furthermore, we investigated the influence of some selected room acoustic properties on the performance. Especially the reverberation time showed a significant impact on the performance. Our approach adapted to measured and simulated room impulse responses, source-microphone distances and reverberation times. Moreover, we showed the significant gap to the oracle mask which motivates further investigations. Mask estimation methods compensating for destructive interference with mask-values greater than one could further improve the results, however, at the risk of noise amplification and musical tones.

7. References

[1] A. K. Nábělek and D. Mason, "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," *Journal of Speech and Hearing Research*, vol. 24, pp. 375–383, 1981.

[2] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[3] H. Haas, "The influence of a single echo on the audibility of speech," *Journal Audio Eng. Soc.*, vol. 20, no. 2, pp. 146–159, 1972.

[4] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. London, UK: Springer, 2010.

[5] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech de-reverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[6] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–774, Sep. 2009.

[7] J. Erkelens and R. Heusdens, "Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1746–1765, Sep. 2010.

[8] B. Schwartz, S. Gannot, and E. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 2, pp. 394–406, 2015.

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and J. Biing-Hwang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[10] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 231–246, Feb. 2009.

[11] S. Braun, B. Schwartz, S. Gannot, and E. A. P. Habets, "Late reverberation PSD estimation for single-channel dereverberation using relative convolutive transfer functions," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Xi'an, China, Sep. 2016.

[12] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[13] D. Yu, M. Kolbk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 241–245.

[14] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[15] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Kluwer Academic, 2005, pp. 181–197.

[16] C. Hummersone, T. Stokes, and T. Brookes, *Blind Source Separation*. Springer, 2014, ch. On the Ideal Ratio Mask as the Goal of Computational Auditory Scene Analysis, pp. 349–368.

[17] F. Mayer, D. S. Williamson, P. Mowlaee, and D. Wang, "Impact of phase estimation on single-channel speech separation based on time-frequency masking," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4668–4679, 2017.

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

[19] F. Weninger, J. R. Hershey, J. L. Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *IEEE Glob. Conf. on Sig. and Inf. Proc.*, Dec. 2014, pp. 577–581.

[20] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. of the 12th Int. Conf. on Lat.Var. An. and Sig. Sep.*, ser. LVA/ICA. New York, USA: Springer-Verlag, 2015, pp. 91–99.

[21] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 31–35.

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

[23] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[24] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Antibes - Juans les Pins, France, Sep. 2014, pp. 313–317.

[25] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 246–250.

[26] E. A. P. Habets. (2008, May) Room impulse response (RIR) generator. [Online]. Available: https://github.com/ehabets/RIR-Generator

[27] 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, Apr. 1979.

[28] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, 1965.

[29] 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 Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210.

[30] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, Feb. 1993.

[31] F. Chollet *et al.*, "Keras 1.2.2," 2015.

[32] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2670313

[34] Y. Gal and Z. Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Proc. Neural Information Processing Conf*, ser. NIPS'16. USA: Curran Associates Inc., 2016, pp. 1027–1035. [Online]. Available: http://dl.acm.org/citation.cfm?id=3157096.3157211

[35] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech Conf.*, Sep. 2016, pp. 545–549.

[36] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low bit-rate speech coding systems," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 2, pp. 262–273, 1988.

[37] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2001, pp. 749–752.

[38] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 1978, pp. 586–590.