Single-channel late reverberation power spectral density estimation using denoising autoencoders

Ina Kodrasi, Hervé Bourlard

Idiap Research Institute, Speech and Audio Processing Group, Martigny, Switzerland

{ina.kodrasi, herve.bourlard}@idiap.ch

Abstract

In order to suppress the late reverberation in the spectral domain, many single-channel dereverberation techniques rely on an estimate of the late reverberation power spectral density (PSD). In this paper, we propose a novel approach to late reverberation PSD estimation using a denoising autoencoder (DA), which is trained to learn a mapping from the microphone signal PSD to the late reverberation PSD. Simulation results show that the proposed approach yields a high PSD estimation accuracy and generalizes well to unseen data. Furthermore, simulation results show that the proposed DA-based PSD estimate yields a higher PSD estimation accuracy and a similar dereverberation performance than a state-of-the-art statistical PSD estimate, which additionally also requires knowledge of the reverberation time.

Index Terms: late reverberation PSD, denoising autoencoder, dereverberation

1. Introduction

In hands-free communication, the received microphone signal typically contains not only the desired speech signal but also delayed and attenuated copies of the desired speech signal due to reverberation. While early reverberation may be desirable [1, 2], severe reverberation yields a degradation in speech quality and intelligibility [3, 4]. With the continuously growing demand for high quality hands-free communication, in the last decades many single-channel and multi-channel dereverberation techniques have been proposed [5]. Although multi-channel techniques have become increasingly popular, several applications rule out multi-channel solutions due to, e.g., hardware limitations, and hence, effective single-channel dereverberation techniques remain necessary. Many single-channel dereverberation techniques aim at suppressing the late reverberation in the spectral domain using an estimate of the late reverberation power spectral density (PSD) [6–11]. The effectiveness of such techniques depends on the accuracy of the late reverberation PSD estimate.

Existing single-channel late reverberation PSD estimators can be broadly classified into two classes, i.e., statistical estimators [7–9] and model-based estimators [10, 11]. Statistical estimators are based on the assumption that the room impulse response (RIR) can be represented by a zero-mean Gaussian random sequence multiplied by an exponentially decaying function. The late reverberation PSD is then estimated using knowledge of the reverberation time [7, 8] or also of the direct-to-reverberation ratio [9]. Model-based estimators rely on a convolutive transfer function (CTF) model of the RIR in the short-time Fourier transform domain (STFT) [10, 11]. In order to estimate the late reverberation PSD, the CTF coefficients are either estimated taking inter-frame correlations into account [10] or using a Kalman filter [11]. In [11] it is shown that model-based PSD estimators yield a similar estimation accuracy as the statistical PSD estimators in [7–9].

In this paper, we propose a third class of single-channel late reverberation PSD estimators based on denoising autoencoders (DAs) [12, 13]. In the context of dereverberation, DAs have already been used for generating robust dereverberated features for speech recognition [14, 15] as well as for enhancing reverberant speech [16–18]. In [16–18], a DA has been used to learn a spectral mapping from the magnitude spectrogram of reverberant speech to the magnitude spectrogram of clean speech. In [18] it is shown that by incorporating information of the reverberation time during the training stage, the dereverberation performance can be further improved. In the present approach, instead of estimating the clean speech magnitude spectrogram from the reverberant speech magnitude spectrogram as in [16–18], we propose to use a DA to estimate the late reverberation PSD from the microphone signal PSD. The estimated late reverberation PSD can then be used in a spectral enhancement technique such as the Wiener filter in order to achieve dereverberation. Hence, a DA is used to estimate the signal statistics, while speech enhancement is still performed using traditional signal processing techniques. This allows for a controlled evaluation of the possible benefits of combining machine learning techniques with traditional speech enhancement techniques. In addition, such an approach gives the user the flexibility to select the most advantageous spectral enhancement technique to use depending on the application. Our proposed approach differs from [16–18] not only in estimating the late reverberation PSD instead of the clean speech magnitude spectrogram, but also in the used DA architecture.

Simulation results show the effectiveness of the proposed approach, with the DA-based late reverberation PSD estimate yielding a higher PSD estimation accuracy and a similar dereverberation performance than the state-of-the-art statistical estimate in [7] (which additionally requires knowledge of the reverberation time).

2. Speech Dereverberation

We consider a reverberant acoustic system with a single speech source and a single microphone. The microphone signal $y(n)$ at time index n is given by

$$y(n) = \underbrace{\sum_{p=1}^{L_e} h_n(p)s(n-p)}_{x(n)} + \underbrace{\sum_{p=L_e+1}^{L_h} h_n(p)s(n-p)}_{r(n)}, \quad (1)$$

where $h_n(p), p = 1, \ldots, L_h$, are the coefficients of the (possibly time-varying) RIR between the source and the microphone, L_e is the duration of the direct path and early reflections, $s(n)$ is the clean speech signal, $x(n)$ is the direct and early reverbera-

This work was supported by the Swiss National Science Foundation project "MoSpeeDi".

tion component, and $r(n)$ is the late reverberation component[1]. While the duration of the direct path and early reflections is not concisely defined, it is typically considered to be between 10 ms and 80 ms. In the STFT domain, the microphone signal $Y(k,l)$ at frequency bin k and time frame index l is given by

$$Y(k,l) = X(k,l) + R(k,l), \tag{2}$$

with $X(k,l)$ and $R(k,l)$ being the STFTs of $x(n)$ and $r(n)$, respectively. Since early reverberation tends to improve speech intelligibility [1, 2] and late reverberation is the major cause of speech intelligibility degradation, the objective of spectral enhancement techniques is to suppress the late reverberation component $R(k,l)$ and obtain an estimate of $X(k,l)$.

Assuming that the components in (2) are uncorrelated, the PSD of the microphone signal $Y(k,l)$ is given by

$$\Phi_y(k,l) = \mathcal{E}\{|Y(k,l)|^2\} = \Phi_x(k,l) + \Phi_r(k,l), \tag{3}$$

with $\mathcal{E}$ denoting expected value, $\Phi_x(k,l) = \mathcal{E}\{|X(k,l)|^2\}$ denoting the PSD of the direct and early reverberation component, and $\Phi_r(k,l) = \mathcal{E}\{|R(k,l)|^2\}$ denoting the PSD of the late reverberation component. Given the uncorrelatedness assumption in (3), well-known spectral enhancement techniques such as the Wiener filter can be used to estimate the direct and early reverberation component $X(k,l)$. The Wiener filter obtains a minimum mean-square error (MSE) estimate of the target signal $X(k,l)$ given the microphone signal $Y(k,l)$ as

$$\hat{X}(k,l) = \frac{\xi(k,l)}{\xi(k,l)+1} Y(k,l), \tag{4}$$

with $\xi(k,l)$ denoting the a priori target-to-late reverberation ratio (TRR). The TRR can be estimated using the decision-directed approach as [19]

$$\xi(k,l) = \alpha \frac{|\hat{X}(k,l-1)|^2}{\hat{\Phi}_r(k,l-1)} + (1-\alpha)\max\left[\frac{|Y(k,l)|^2}{\hat{\Phi}_r(k,l)} - 1, 0\right], \tag{5}$$

with α a smoothing factor and $\hat{\Phi}_r(k,l)$ an estimate of the late reverberation PSD. Hence, as can be seen in (4) and (5), an estimate of the late reverberation PSD is required in order to achieve speech dereverberation.

3. Late Reverberation PSD Estimation

In this section, the statistical late reverberation PSD estimator from [7] is briefly reviewed and the proposed DA-based PSD estimator is described.

3.1. Statistical PSD estimator

In [7], the RIR is described as a zero-mean Gaussian random sequence multiplied by an exponential decay Δ given by

$$\Delta = \frac{3\ln(10)}{T_{60}}, \tag{6}$$

with T_{60} the reverberation time. An estimate of the late reverberation PSD is then derived as

$$\hat{\Phi}_r^s(k,l) = e^{-2\Delta L_e/f_s} \Phi_y(k, l - L_e/F), \tag{7}$$

where f_s denotes the sampling frequency and F denotes the frame shift. The PSD $\Phi_y(k,l)$ can be directly computed from the microphone signal as

$$\Phi_y(k,l) = \beta\Phi_y(k,l-1) + (1-\beta)|Y(k,l)|^2, \tag{8}$$

with β a recursive smoothing parameter. As can be observed in (6) and (7), the statistical PSD estimator requires knowledge of the reverberation time T_{60}.

3.2. DA-based PSD estimator

A DA is a neural network trained to reconstruct an N-dimensional target vector $\mathbf{u}$ from an $\tilde{N}$-dimensional corrupted version of it $\tilde{\mathbf{u}}$ [12, 13]. The corrupted vector $\tilde{\mathbf{u}}$ is first mapped to a D-dimensional hidden representation $\mathbf{h}$ as

$$\mathbf{h} = \sigma\{\mathbf{W}_i\tilde{\mathbf{u}} + \mathbf{b}_i\}, \tag{9}$$

with $\sigma\{\cdot\}$ denoting a non-linearity, $\mathbf{W}_i$ denoting a $D \times \tilde{N}$-dimensional matrix of weights, and $\mathbf{b}_i$ denoting the D-dimensional bias vector. For a network with only one hidden layer, the hidden representation $\mathbf{h}$ is then mapped to the N-dimensional reconstructed target vector $\mathbf{z}$ as

$$\mathbf{z} = \mathbf{W}_o\mathbf{h} + \mathbf{b}_o, \tag{10}$$

with $\mathbf{W}_o$ the $N \times D$-dimensional matrix of weights and $\mathbf{b}_o$ the N-dimensional bias vector. The parameters $\mathbf{W}_i$, $\mathbf{b}_i$, $\mathbf{W}_o$, and $\mathbf{b}_o$ are then trained to minimize the MSE between the true target vector $\mathbf{u}$ and the reconstructed target vector $\mathbf{z}$.

For late reverberation PSD estimation, we consider the target vector to be the late reverberation PSD at time frame l across all frequency bins K, i.e.,

$$\mathbf{\Phi}_r(l) = [\Phi_r(1,l)\ \Phi_r(2,l)\ \ldots\ \Phi_r(K,l)]^T. \tag{11}$$

Since the late reverberation PSD in each time frame depends on the microphone signal PSD from the previous time frames, the corrupted input vector to the DA is the TK-dimensional vector $\mathbf{\Phi}_y(l)$ constructed by concatenating the microphone signal PSD of the past T time frames, i.e.,

$$\begin{aligned}\mathbf{\Phi}_y(l) = [&\Phi_y(1,l)\ \ldots\ \Phi_y(K,l)\\ &\Phi_y(1,l-1)\ \ldots\ \Phi_y(K,l-1)\\ &\ldots\\ &\Phi_y(1,l-T+1)\ \ldots\ \Phi_y(K,l-T+1)]^T.\end{aligned} \tag{12}$$

In the experimental results in Section 4, the performance for $T = 5$ and $T = 10$ is investigated. The proposed network architecture is depicted in Fig. 1. The TK-dimensional input $\mathbf{\Phi}_y(l)$ is first mapped to the $(TK+K)$-dimensional hidden representation $\mathbf{h}_1(l)$ using a linear transformation followed by a sigmoid non-linearity as in (9). Experimental analysis suggest that using more than $(TK+K)$ units on the first hidden layer does not yield any performance improvement. Similarly, the hidden representation $\mathbf{h}_1(l)$ is further mapped to the $2K$-dimensional hidden representation $\mathbf{h}_2(l)$. Finally, the hidden representation $\mathbf{h}_2(l)$ is mapped to the K-dimensional target vector $\mathbf{\Phi}_r(l)$ using a linear transformation as in (10). Prior to training, the vectors $\mathbf{\Phi}_r(l)$ and $\mathbf{\Phi}_y(l)$ are transformed to the log-domain and are globally normalized into zero mean and unit variance. The computation of the target late reverberation PSD $\mathbf{\Phi}_r(l)$ for training and evaluation will be discussed in Section 4.

As already mentioned, the proposed DA differs from the DA used in [16–18]. In [16–18], the DA is used to learn a spectral

[1] It should be noted that for the sake of simplicity, a noise-free scenario is assumed in this paper. Nevertheless, the late reverberation PSD estimator proposed in Section 3.2 can also be used in a noisy scenario, as long as an estimate of the noise PSD can be obtained.

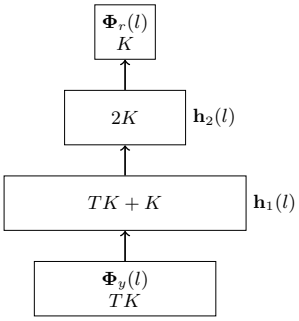

Figure 1: *Proposed DA architecture for late reverberation PSD estimation.*

mapping from the magnitude spectrogram of the microphone signal $|Y(k,l)|$ to the mangitude spectrogram of the direct and early reverberation component $|X(k,l)|$. The estimated magnitude spectrogram of the direct and early reverberation component is then combined with the phase of the received microphone signal in order to achieve speech dereverberation. Differently from [16–18], in the present approach the DA is used as a late reverberation PSD estimator to learn a spectral mapping from the microphone signal PSD $\Phi_y(k,l)$ to the late reverberation PSD $\Phi_r(k,l)$. The estimated late reverberation PSD can then be used in a spectral enhancement technique such as the Wiener filter in order to achieve speech dereverberation.

4. Simulation Results

In this section, the estimation accuracy of the proposed DA-based PSD estimator is experimentally analyzed and compared to the estimation accuracy of the statistical estimator described in Section 3.1. Furthermore, using instrumental performance measures, the dereverberation performance of a Wiener filter when using the DA-based and statistical PSD estimates is extensively compared.

4.1. Datasets and model training

In order to generate the training dataset, 924 clean utterances from the TIMIT training database [20] were used. Reverberant microphone signals were generated by convolving these clean utterances with 10 RIRs, resulting in 9240 training utterances in total. The RIRs were generated using the image-source method [21] and the considered reverberation times ranged from 0.2 s to 2 s with a step size of 0.2 s. The validation dataset was generated using 168 clean utterances from the TIMIT testing database and 9 RIRs, resulting in 1512 validation utterances in total. The RIRs were generated using the image-source method and the considered reverberation times ranged from 0.3 s to 1.9 s with a step size of 0.2 s. Finally, the testing dataset was generated using 167 clean utterances from the TIMIT testing database (different from the clean utterances used for the validation dataset) and 18 RIRs, resulting in 3006 testing utterances in total. The RIRs were generated using the image-source method and the considered reverberation times ranged from 0.35 s to 1.95 s with a step size of 0.1 s.

In order to also evaluate the dereverberation performance in realistic acoustic environments, we additionally consider a realistic testing dataset which is generated by convolving 10 clean utterances from the HINT database [22] with 6 measured RIRs, resulting in 60 realistic testing utterances in total. The reverberation times for the measured RIRs are $T_{60} \in \{0.65\text{ s}, 0.70\text{ s}, 0.75\text{ s}, 0.95\text{ s}, 0.97\text{ s}, 1.25\text{ s}\}$.

The proposed DA was implemented using the PyTorch library [23]. The training was done using the Adam optimizer, with a learning rate of 0.0001 and a batch size of 500. The model was trained for 50 epochs and the model parameters corresponding to the epoch with the lowest validation error were used as the final model parameters.

4.2. Algorithmic settings and performance measures

For all considered datasets, the clean utterances were convolved with the late reflections of the RIRs as in (1) in order to generate the late reverberation components $r(n)$. Since the duration L_e of the early reflections of an RIR is not exactly known, and hence, the start of the late reflections of an RIR is not exactly known, we consider different late reverberation components generated using the reflections of the RIRs starting

$$L_e/f_s \in [0.032\text{ s},\ 0.048\text{ s},\ 0.064\text{ s}] \tag{13}$$

after the direct path arrival. It should be noted that by using different values of L_e to generate the late reverberation components, different target late reverberation PSDs are obtained, and hence, different DA model parameters are obtained. In addition, different values of L_e also yield a different late reverberation PSD estimate when using the statistical estimator, cf. (7).

The signals are processed using a weighted overlap-add framework with a hamming window and an overlap of 50 % at a sampling frequency $f_s = 16$ kHz. The frame size is 512 samples, resulting in $K = 257$. The microphone signal PSD $\Phi_y(k,l)$ is computed as in (8) using $\beta = 0.67$, which corresponds to a time constant of 40 ms. The late reverberation PSD $\Phi_r(k,l)$ is computed from the late reverberation component $R(k,l)$ similarly as in (8) with $\beta = 0.67$. For the statistical estimator, an estimate of the reverberation time T_{60} is required, cf. (6). In the following simulations, it is assumed that the reverberation time is perfectly known. In practice however, also the reverberation time needs to be estimated, using e.g. [24]. For the Wiener filter implementation in (4), a minimum gain of -10 dB is used.

The estimation accuracy of the considered PSD estimators is evaluated using the PSD estimation error ϵ defined as [25]

$$\epsilon = \frac{1}{LK}\sum_{l=1}^{L}\sum_{k=1}^{K}\left|10\log_{10}\frac{\Phi_{\text{r}}(k,l)}{\hat{\Phi}_{\text{r}}(k,l)}\right|, \tag{14}$$

with L being the total number of time frames in the utterance. It should be noted that for different values of L_e, different target late reverberation PSDs $\Phi_r(k,l)$ in (14) are obtained.

In order to evaluate the dereverberation performance, we use the improvement in frequency-weighted segmental signal-to-noise ratio (ΔfwSSNR) [26], in speech-to-reverberation modulation energy ratio (ΔSRMR) [27], and in cepstral distance (ΔCD) [26] between the processed and unprocessed microphone signals. While the SRMR measure is a non-intrusive measure which does not require a reference signal, the fwSSNR and CD measures are intrusive measures generating a similarity score between a test signal and a reference signal. The reference signal used in this paper is the clean speech signal $s(n)$. It should be noted that positive values of ΔfwSSNR and ΔSRMR and negative values of ΔCD indicate a performance improvement.

4.3. Estimation accuracy of the DA-based and statistical PSD estimators

In the following, the estimation accuracy of the proposed DA-based estimator is compared to the estimation accuracy of the

Table 1: *Average estimation error ϵ [dB] for the proposed and statistical PSD estimators on the training, validation, and testing datasets for different values of L_e.*

	Training dataset			Validation dataset			Testing dataset		
L_e/f_s	0.032 s	0.048 s	0.064 s	0.032 s	0.048 s	0.064 s	0.032 s	0.048 s	0.064 s
$\hat{\Phi}_r^{\text{d5}}(k,l)$	1.42	1.88	2.15	2.05	2.86	3.58	2.08	2.75	3.45
$\hat{\Phi}_r^{\text{d10}}(k,l)$	**1.37**	**1.73**	**1.83**	**2.01**	**2.73**	**3.39**	**2.05**	**2.66**	**3.30**
$\hat{\Phi}_r^{\text{s}}(k,l)$	3.77	5.32	6.53	3.40	4.68	6.01	3.44	4.65	5.93

statistical estimator for different definitions of the target late reverberation PSD. The DA-based late reverberation PSD estimate will be referred to as $\hat{\Phi}_r^{\text{d5}}(k,l)$ when using $T = 5$ and as $\hat{\Phi}_r^{\text{d10}}(k,l)$ when using $T = 10$. We analyze the estimation accuracy of $\hat{\Phi}_r^{\text{d5}}(k,l)$, $\hat{\Phi}_r^{\text{d10}}(k,l)$, and $\hat{\Phi}_r^{\text{s}}(k,l)$ on the training, validation, and testing datasets, with the presented estimation error values averaged over all utterances in the datasets. The obtained estimation errors for different values of L_e are presented in Table 1. It can be observed that for all considered datasets and for all values of L_e, the proposed DA-based estimate $\hat{\Phi}_r^{\text{d10}}(k,l)$ yields the lowest estimation error, significantly outperforming the statistical PSD-estimate $\hat{\Phi}_r^{\text{s}}(k,l)$. The average difference between the estimation errors for $\hat{\Phi}_r^{\text{d10}}(k,l)$ and $\hat{\Phi}_r^{\text{s}}(k,l)$ across all datasets and values of L_e is 2.52 dB. Furthermore, it can be observed that the proposed DA-based estimate $\hat{\Phi}_r^{\text{d5}}(k,l)$ also yields a comparable estimation error to $\hat{\Phi}_r^{\text{d10}}(k,l)$, with the average difference between the estimation errors across all datasets and values of L_e being only 0.13 dB. Finally, Table 1 shows that the proposed DA models are capable of generalizing to unseen data for any value of L_e, with the respective PSD estimation errors for $\hat{\Phi}_r^{\text{d5}}(k,l)$ and $\hat{\Phi}_r^{\text{d10}}(k,l)$ being very similar across the validation and testing datasets. In summary, these simulation results show that the proposed DA-based late reverberation PSD estimator is more advantageous than the state-of-the-art statistical PSD estimator, yielding a higher PSD estimation accuracy without additionally requiring knowledge of the reverberation time.

Table 2: *Average dereverberation performance of a Wiener filter on the testing dataset using the proposed and statistical estimators with $L_e/f_s = 0.064$ s.*

Measure	ΔfwSSNR [dB]	ΔSRMR [dB]	ΔCD [dB]
$\hat{\Phi}_r^{\text{d5}}(k,l)$	1.44	**2.01**	**−0.19**
$\hat{\Phi}_r^{\text{d10}}(k,l)$	**1.46**	1.96	**−0.19**
$\hat{\Phi}_r^{\text{s}}(k,l)$	1.16	1.79	−0.16

Table 3: *Average dereverberation performance of a Wiener filter on the realistic testing dataset using the proposed and statistical estimators with $L_e/f_s = 0.064$ s.*

Measure	ΔfwSSNR [dB]	ΔSRMR [dB]	ΔCD [dB]
$\hat{\Phi}_r^{\text{d5}}(k,l)$	**1.46**	1.43	**−0.18**
$\hat{\Phi}_r^{\text{d10}}(k,l)$	1.35	1.37	−0.15
$\hat{\Phi}_r^{\text{s}}(k,l)$	1.38	**1.68**	**−0.18**

4.4. Dereverberation performance of a Wiener filter using the DA-based and statistical PSD estimators

In the following, the dereverberation performance of a Wiener filter using the DA-based and statistical estimators is compared for the testing and realistic testing datasets. Instrumental performance measures are computed for each utterance in the considered dataset, and the presented performance measures are averaged over all utterances in the dataset. Since similar conclusions can be drawn for any value of L_e, we only present the results obtained for $L_e/f_s = 0.064$ s.

Table 2 presents the ΔfwSSNR, ΔSRMR, and ΔCD obtained using a Wiener filter with $\hat{\Phi}_r^{\text{d5}}(k,l)$, $\hat{\Phi}_r^{10}(k,l)$, and $\hat{\Phi}_r^{\text{s}}(k,l)$ on the testing dataset. It can be observed that using the DA-based PSD estimates yields the highest improvement in all instrumental measures. However, the performance differences between using the proposed DA-based PSD estimates and the statistical estimate are rather small.

Table 3 presents the ΔfwSSNR, ΔSRMR, and ΔCD obtained using a Wiener filter with $\hat{\Phi}_r^{\text{d5}}(k,l)$, $\hat{\Phi}_r^{\text{d10}}(k,l)$, and $\hat{\Phi}_r^{\text{s}}(k,l)$ on the realistic testing dataset. It can be observed that using $\hat{\Phi}_r^{\text{d5}}(k,l)$ yields the best performance in terms of ΔfwSSNR, using $\hat{\Phi}_r^{\text{s}}(k,l)$ yields the best performance in terms of ΔSRMR, and using $\hat{\Phi}_r^{\text{d5}}(k,l)$ or $\hat{\Phi}_r^{\text{s}}(k,l)$ yields the best performance in terms of ΔCD. However, similarly as for the testing dataset, the performance differences between the different PSD estimators are rather small.

In summary, these simulation results show that the proposed DA-based late reverberation PSD estimator yields a similar or slightly better dereverberation performance as the state-of-the-art statistical PSD estimator, without requiring any additional knowledge such as an estimate of the reverberation time. It should be noted that the PSD estimation accuracy and the dereverberation performance of the statistical estimator might still degrade if the reverberation time is estimated.

5. Conclusion

In this paper we have proposed a novel approach to single-channel late reverberation PSD estimation using a DA. Differently from state-of-the-art speech enhancement techniques which use a DA to learn a spectral mapping from the microphone signal mangitude spectrogram to the desired signal magnitude spectrogram, in this paper the DA is trained to learn a spectral mapping from the microphone signal PSD to the late reverberation PSD. Extensive simulation results have shown that the proposed approach yields a higher PSD estimation accuracy and a similar dereverberation performance as a state-of-the-art statistical estimator, which additionally requires knowledge of the reverberation time. Analyzing the performance of the proposed DA-based estimator in the presence of additive noise as well as extending the proposed approach to jointly estimate the late reverberation and noise PSDs remains a topic for future research.

6. References

[1] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, Jun. 2003.

[2] A. Warzybok, J. Rennies, T. Brand, S. Doclo, and B. Kollmeier, "Effects of spatial and temporal integration of a single early reflection on speech intelligibility," *Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 269–282, Jan. 2013.

[3] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, Jul. 2006.

[4] A. Warzybok, I. Kodrasi, J. O. Jungmann, E. A. P. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, "Subjective speech quality and speech intelligibility evaluation of single-channel dereverberation algorithm," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sep. 2014, pp. 333–337.

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

[6] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, Jun. 2007.

[7] K. Lebart and J. M. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, no. 3, pp. 359–366, May-Jun. 2001.

[8] E. A. P. Habets, S. Gannot, and I. Cohen, "Speech dereverberation using backward estimation of the late reverberant spectral variance," in *IEEE Convention of Electrical and Electronics Engineers in Israel*, Eilat, Israel, Dec. 2008, pp. 384–388.

[9] ——, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–774, Sep. 2009.

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

[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. International Workshop on Acoustic Echo and Noise Control*, Shanghai, China, Sep. 2016.

[12] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. International Conference on Machine Learning*, Helsinki, Finland, Jun. 2008, pp. 1096–1103.

[13] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, Jan. 2009.

[14] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder," in *Proc. 14th Annual Conference of the International Speech Communication Association*, Lyon, France, Aug. 2013, pp. 3512–3516.

[15] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 1759–1763.

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

[17] B. Wu, K. Li, M. Yang, and C. H. Lee, "A study on target feature activation and normalization and their impacts on the performance of DNN based speech dereverberation systems," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Jeju, Korea, Dec. 2016.

[18] ——, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, Jan. 2017.

[19] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1. Web download," 1993.

[21] E. A. P. Habets, "Room impulse response (RIR) generator," available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator/.

[22] M. Nilsson, S. D. Soli, and A. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, Feb. 1994.

[23] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. 31st Conference on Neural Information Processing Systems*, Vancouver, Canada, May 2017, pp. 161–165.

[24] J. Eaton, N. D. Gaubitch, and P. A. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 161–165.

[25] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, Oct 2011, pp. 145–148.

[26] S. Quackenbush, T. Barnwell, and M. Clements, *Objective measures of speech quality*. New Jersey, USA: Prentice-Hall, 1988.

[27] T. H. Falk, C. Zheng, and W. Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.