Learning to Estimate Reverberation Time in Noisy and Reverberant Rooms

Xiong Xiao[1], *Shengkui Zhao*[2], *Xionghu Zhong*[3],
Douglas L. Jones[2], *Eng Siong Chng*[1,3], *Haizhou Li*[1,3,4,5]

[1]Temasek Laboratories, Nanyang Technological University (NTU), Singapore
[2]Advanced Digital Sciences Center, Singapore
[3]School of Computer Engineering, NTU,Singapore
[4]Human Language Technology Department, Institute for Infocomm Research, Singapore
[5]School of EE & Telecom, University of New South Wales, Australia

xiaoxiong@ntu.edu.sg, shengkui.zhao@adsc.com.sg, xhzhong@ntu.edu.sg

10.21437/Interspeech.2015-680

Abstract

The reverberation time, T_{60}, is an important indicator of the reverberation strength in a room and has many applications in speech processing, such as dereverberation. However, the T_{60} must be blindly estimated if only reverberant speech is available. In this paper, we provide a learning based approach for T_{60} estimation. We treat the T_{60} estimation as a classification problem by dividing the T_{60} range into countable bins (e.g. 19 bins covering 0.1s to 1s with a bin width of 0.05s) and the estimation becomes predicting which bin the true T_{60} falls into for a given speech. We use deep neural networks (DNN) to learn such a mapping from speech to the T_{60}. The DNN is trained on a large amount of reverberant and noisy speech signals generated from various simulated rooms with known reverberations. After training, we observe that the DNN can learn highly sensible features for the T_{60} estimation task. Experimental results on the data from both simulated rooms and real rooms confirmed the effectiveness of the DNN learning based approach. In all the test cases, the DNN method significantly outperforms the state-of-the-art SDD T_{60} estimation method.

Index Terms: T60 estimation, robust reverberation time estimation, deep learning, deep neural networks, dereverberation.

1. Introduction

The estimation of room reverberation time (RT) is an important task for characterising the listening quality of enclosed auditory spaces. This measure is commonly used for hearing aids or other audio processing instruments to assess the amount of reverberation of the listening environments and to apply the most appropriate signal processing strategies [1]. The RT of an environment is defined as the duration for which the sound level attenuates 60 dB below its initial level after the excitation source is switched off. It is often referred to as the T_{60} in the literature.

The T_{60} was explicitly formulated by Sabine [2] where only the geometry of the environment and the absorption coefficients of surfaces are used. When these information is unavailable, a controlled test sound could be generated to estimate the T_{60} based on the interrupted noise method [3] or the Schoreder integration method [4]. However, for many applications such as speech enhancement [5], the controlled test excitation signals are not always available, and it is necessary to determine the T_{60} purely from the recorded speech signals.

Several methods have been proposed for estimating T_{60} from speech signals [6, 7, 8, 9, 10, 11, 12, 13, 14]. In [6], Ratnam presents a blind method based on the maximum likelihood (ML) method by modeling the reverberant tail of a decaying sound using a random Gaussian process modulated by a decaying sequence. In [9], Cox presents a semi-blind neural network approach with a training process on artificial impulse responses with different reverberation times. The study has a limited focus on the pronounced digits and a limited accuracy of 0.1s. In [10], Wen develops a spectral decay distribution (SDD) method. The variance of the negative gradients of speech decay, namely the negative-side variance (NSV), is mapped to decay rate using a polynomial function whose parameters are determined from training data. In [11], Eaton further improves the SDD method by selecting the time-frequency (TF) bins of the speech signals and defining different NSV computation models based on the *a posteriori* signal-to-noise ratio (SNR) estimate. In [15], Gaubitch demonstrates that the T_{60} estimates achieve small errors when the signal-to-noise ratio (SNR) is high, i.e., SNR$\geq$30dB. However, the T_{60} estimates significantly degrade in low SNR levels.

A recent trend in machine learning is to use flexible deep models, such as deep neural networks (DNN) [16], and large amount of data to solve many learning problems. It is demonstrated that deep models is much more efficient in modeling complex relationship than shallow models, such as neural networks with 1 hidden layer [17]. The combination of DNN and vast amount of training data has led to significant improvement in speech recognition [18], image recognition [19], and direction-of-arrival estimation [20]. In this paper, we treat the T_{60} estimation as a mapping problem, i.e. mapping from speech signal to T_{60} values by using DNN. Unlike the previous methods that use manually designed features, such as NSV [10, 11], we employ DNN to learn suitable features for T_{60} estimation automatically. The T_{60} estimation is casted as a classification problem where the T_{60} range is divided into countable bins. A DNN is used to learn a nonlinear mapping from speech filterbank energy to the T_{60} class posterior probabilities. The DNN is trained from a large amount of noisy and reverberant speech signals generated from simulated room impulse responses. We will show the features learnt by the DNN and evaluate the performance of the DNN on the T_{60} estimation in the experiments.

The rest of the paper is organized as follows. In section 2, the T_{60} estimation problem is formulated and the SDD method is reviewed. In section 3, the proposed DNN based learning approach is described. In section 4, experimental results and discussions are presented. Finally, we conclude in section 5.

Copyright © 2015 ISCA

2. Problem Formulation and the SDD Method

In this section, we provide a brief description of problem of the T_{60} estimation and review the state-of-the-art SDD method. According to Polack's statistical model [21] for the room impulse response (RIR), the short-time Fourier transform (STFT) representation of the room acoustic decay model is

$$H(t,f) = P(f)e^{-\tau(f)t} \tag{1}$$

where $H(t,f)$ is the energy envelope of the RIR at time-frequency bin (t,f), and $P(f)$ is the power spectral density of the modeled stationary process, and $\tau(f)$ is referred to as the decay rate at frequency f. For a frequency f, the decay rate is linked to the reverberation time T_{60} as $T_{60} = \frac{6}{\tau}\ln 10$ when the distance between the speech source and the microphone is larger than the critical distance.

To estimate the decay rate $\tau(f)$, eq. (1) can be reformulated by taking the natural logarithm as

$$\ln H(t,f) = \ln P(f) - \tau(f)t. \tag{2}$$

From eq. (2), the decay rate $\tau(f)$ could thus be estimated by a linear fitting to the natural logarithm of the time-frequency energy envelope $\ln H(t,f)$. Consequently, the reverberation time T_{60} is obtained in a straightforward manner. However, due to the random nature of the fine structure of the RIR, the completely smoothed decay rates can hardly be obtained by the fitting process.

To overcome the above problem, Wen developed a spectral decay distribution (SDD) method based on the spectral decay distributions of the received speech signal in [10]. Let $\hat{\tau}(\omega,k)$ be the negative decay rate estimated at the frequency band ω and time frame k. The NSV σ_-^2 can be modeled as

$$\sigma_-^2 = \frac{1}{LK}\sum_{\omega=1}^{L}\sum_{k=1}^{K}(\hat{\tau}(\omega,k))^2. \tag{3}$$

From eq. (3), a polynomial mapping between the NSV and the decay constant τ can be obtained from a training data set and the T_{60} is calculated based on the relationship $T_{60} = \frac{6}{\tau}\ln 10$. In [11], Eaton reduces the computational complexity of the SDD method by employing the Mel-spaced frequency bands when generating the NSV. To reduce the effect of noise on the T_{60} estimation, Eaton selects the TF bins which are more likely due to speech signals based on the decay rates. In addition, Eaton uses different NSV estimation models based the a posteriori signal-to-noise ratio (SNR) estimates.

From a machine learning point of view, the SDD method uses manually designed features NSV and a polynomial regression model to estimate the T_{60}. The model parameters, i.e. the polynomial coefficients that maps NSV to decay rate, is learnt from a small amount of real data. Although this approach is able to achieve reasonable results at high SNR levels, it is limited in two ways. First, although the NSV feature has good correlation with the decay rate, a single feature is not rich enough to achieve accurate T_{60} estimation in all cases. Second, since the polynomial regression model is simple and contains only several parameters, its parameters can be learnt from a small amount of data. However, the simple polynomial model does not allow us to take advantage of the vast amount of speech data in real life to achieve more accurate and robust T_{60} estimation. In next section, we propose a learning based method that uses deep model (DNN) and large amount of simulated training data to achieve more accurate and robust T_{60} estimation.

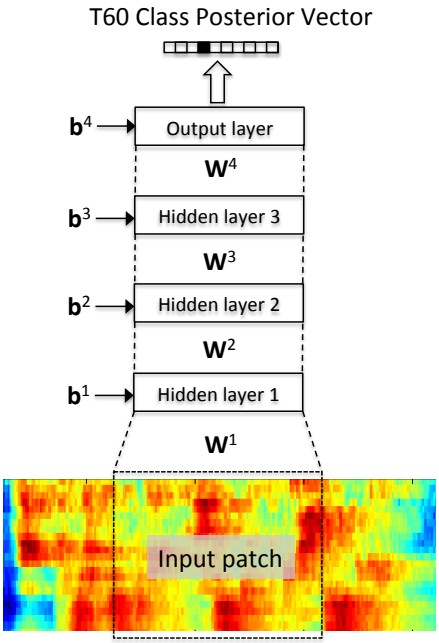

Figure 1: Illustration of T_{60} estimation using DNN based classification.

3. Learning Based T_{60} Estimation Method

In this section, we describe our construction and training of the DNN model to map a given speech input signal to the output of T_{60}. The proposed DNN approach for the T_{60} estimation is illustrated in Fig. 1. At the bottom of the figure is the input features and at the top is the T_{60} classes. Between the input and output, there are several hidden layers that transform the input nonlinearly layer by layer.

The first design consideration is to choose appropriate input features for the DNN. While DNN is a general learning machine, the input features are task dependent. Although the DNN is capable of deriving appropriate features from the raw input signals, it is helpful to use a more compact speech representation with lower dimensions to reduce the training time. Motivated by [10, 11], we choose to use log Mel filterbank energies (LMFBE) as the input features. In this study, we assume the speech signals are sampled at the sampling rate of 16 kHz. The input speech signals are first segmented into frames of 25ms long and each frame is transformed into the frequency domain using a fast Fourier transform (FFT) with length of 512. We then apply 23 Mel-scaled filterbanks to the spectrum to obtain 23 LMFBEs. A comparison of clean and reverberant filterbanks of the same utterance is illustrated in Fig 2. It is seen that the reverberation results in temporal and spectral smearing of the speech pattern. To include enough reverberation information for constructing a successful mapping from an input LMFBE to a T_{60} value, we form a sequence of 51 continuous frames as an input patch to the DNN. It is necessary to use long context in the input as the T_{60} can be as long as a few seconds.

With the input features selected, the second design consideration is to choose the target signal for DNN learning. It is natural to treat the T_{60} estimation task as a regression problem. The DNN can be trained to minimize the mean square error (MSE) between the predicted T_{60} value and the true T_{60} value of the training data. However, our preliminary study shows that

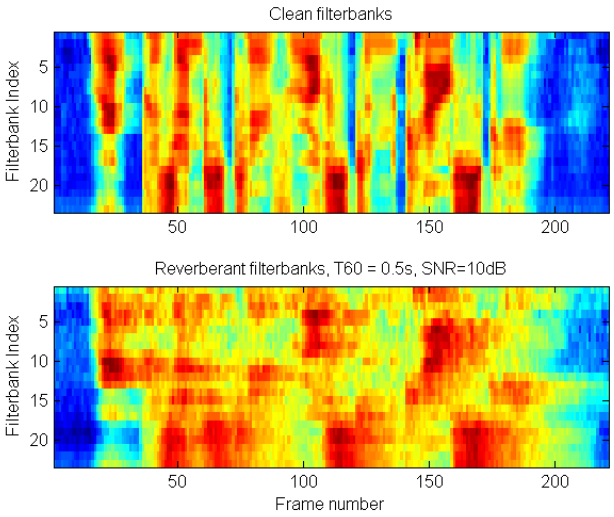

Figure 2: Comparison of clean and reverberant filterbanks of the same utterance.

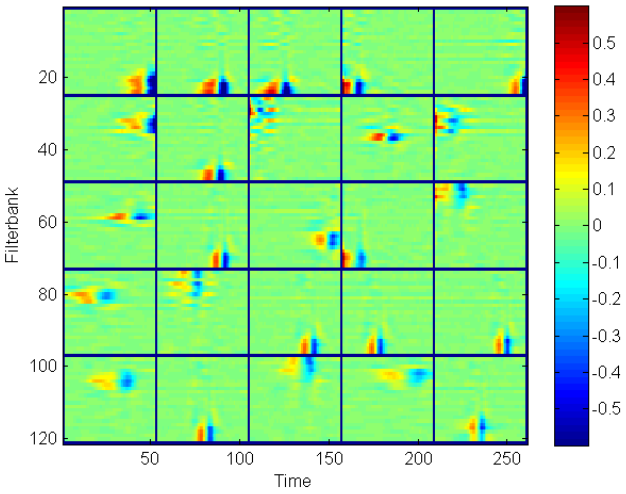

Figure 3: Feature detectors of 25 selected hidden nodes in the first hidden layer. In each cell, the weights of a hidden node is reshaped to a 51×23 matrix, the same size as the input filterbank patch.

casting the T_{60} estimation as a regression problem leads to big estimation errors at the boundary of the T_{60} ranges. For example, when the DNN is trained to handle reverberant speech with T_{60} values ranging from 0.1s to 1.0s, after the training the DNN tends to overestimate T_{60} when the true T_{60} is low and underestimates the T_{60} when the true value is high. That is, the predicted T_{60} value is biased towards the center of the T_{60} range of the training data. This problem is a direct result of treating T_{60} estimation as a regression problem and using MSE as the cost function to train the DNN. To avoid the bias in estimation, we treat the T_{60} estimation as a classification problem. That is, we divide the T_{60} values ranging from 0.1s to 1.0s into 19 bins, one class for one bin. Thus the DNN is trained to predict which class the true T_{60} value belongs to.

Once the input and output of the DNN are defined, the DNN is trained to learn the mapping from the input to the output. We describe our training process using a DNN construction with 3 hidden layers illustrated in Fig. 1. The input vector is affine transformed by $\mathbf{W}^1$ and $\mathbf{b}^1$ first, and then pass through the sigmoid activation function to obtain the output of the hidden layer 1. The output of the hidden layer 1 is again affine transformed by the weighting matrix $\mathbf{W}^2$ and the bias vector $\mathbf{b}^2$ and is passed through the activation function to get the output of the hidden layer 2. The same process is done for the hidden layer 3. Finally, at the top of the model the output layer generates the output vector of the DNN using a softmax function. The output of the DNN is a posterior vector $\mathbf{p}_t = [p(1|\mathbf{o}_t), ..., p(N|\mathbf{o}_t)]^T$ where N is the number of T_{60} classes/bins considered in the system. $p(i|\mathbf{o}_t)$ is the posterior probabilities of the i^{th} bin given the input features $\mathbf{o}_t$ and t is the frame index. The value of T_{60} is predicted as follows:

$$\hat{T}_{60} = c_j \quad (4)$$

$$j = \arg\max_i p(i|\mathbf{o}_t), \quad i \in [1, N] \quad (5)$$

where c_i is the center of the i^{th} bin. In practice, the DNN generates one posterior vector for each speech frame. As the T_{60} is usually estimated for an utterance or a segment of speech, the posteriors used in (5) is usually the average posterior over an utterance or a segment.

There is a tradeoff between the bin width and the number of classes. The larger the bin width, the more accurate the classification, but the poorer the resolution of T_{60} estimation. Our preliminary study shows that it is a reasonable choice to use 19 bins to cover 0.1s to 1.0s with a bin width of 0.05s. To improve the T_{60} estimation resolution, the final T_{60} estimate is obtained as:

$$\hat{T}_{60} = \frac{\sum_{i=j-1}^{j+1} p(i|\mathbf{o}_t) c_i}{\sum_{i=j-1}^{j+1} p(i|\mathbf{o}_t)} \quad (6)$$

Eq. (6) performs a weighted sum of the T_{60} bin centers around the predicted T_{60} class. We observed that the weighted sum improves the T_{60} estimation accuracy consistently.

4. Experiments

4.1. Experimental Settings

A simulated training data set is generated and used to train the DNN T_{60} estimator. We first convolved $7,861$ clean sentences of 92 speakers from the WSJCAM0 [22] training set with simulated RIRs obtained using the image method [23]. Then the additive noises taken from the REVERB Challenge 2014 corpus [24] were added to the training data with the SNR levels randomly chosen from 0dB to 30dB. For each sentence, we randomly pick a room size, reflection rate, source to microphone distance, and SNR to obtain diverse training data. We pass the training set through our simulation algorithm 5 times to generate totally $39,305$ reverberant and noisy sentences for DNN training. As the RIRs are simulated, the true T_{60} of every sentence is known and this information is used for DNN training and also for evaluating performance of T_{60} estimation.

The simulated test data is generated in a similar way as the training data but using a different "et_1" speech set of WSJCAM0 [22] which contains 568 sentences of 14 speakers. The SNR levels are predefined from -10dB to 30dB and note that the -10dB SNR level is not seen in the training data. The real test data was recorded by a male speaker in a small (4×3m), a

Table 1: MAE comparison of the DNN and SDD methods for T_{60} estimation on simulated data.

Room	Dist	SNR(dB)					Average
		-10	0	10	20	30	
SDD							
Small	Near	0.358	0.164	0.104	0.095	0.088	0.162
	Far	0.341	0.149	0.100	0.095	0.091	0.155
Medium	Near	0.342	0.171	0.088	0.084	0.110	0.159
	Far	0.341	0.198	0.121	0.101	0.109	0.174
Large	Near	0.259	0.123	0.102	0.119	0.148	0.150
	Far	0.341	0.177	0.124	0.122	0.124	0.178
Average		0.330	0.164	0.107	0.103	0.112	0.163
DNN							
Small	Near	0.140	0.038	0.031	0.037	0.044	0.058
	Far	0.135	0.040	0.028	0.034	0.040	0.056
Medium	Near	0.070	0.033	0.025	0.025	0.027	0.036
	Far	0.093	0.063	0.050	0.039	0.040	0.057
Large	Near	0.183	0.058	0.040	0.041	0.043	0.073
	Far	0.224	0.124	0.096	0.091	0.086	0.124
Average		0.141	0.059	0.045	0.045	0.047	0.067

medium (6×4m), and a large (10×7m) meeting rooms using a generic microphone. The near and far distances are defined as 1.5m and 3m. Two noises (white and babble) are also recorded in each meeting room and added to the speech recordings at different SNR levels to simulate the effect of noise distortion. The true T_{60} time of the real rooms are determined using recorded clapping sounds and the Schroeder's method [4].

4.2. Features learned by DNN

DNN can be seen as a combination of automatic feature detector (the hidden layers) and a logistic regression classifier. By observing the first layer weights of the DNN, we can gain some ideas of what the network has learnt. Each hidden node in the first hidden layer acts like a feature detector, and the weights associated with a hidden node detect certain patterns in the input data. Fig. 3 illustrates the feature detectors (the reshaped weight vectors) of 25 selected hidden nodes in the first hidden layer. It is seen that although the feature detectors have different shapes and locations, most of them are trying to extract the gradient of the speech energy w.r.t. time. The gradient information is known to be useful to detect speech decay rate (see Eq. (2)). In the SDD method [10, 11], the NSV feature is also derived from negative decay rate, which is the gradient of speech energy less than 0. While the previous methods use the gradient information in manually designed feaures, the DNN automatically find similar features via learning from training data. An advantage of the DNN based learning approach is that it learns a large number of such features, which may allow more accurate estimation of T_{60} than the SDD method which only uses one NSV feature.

4.3. T60 Estimation Results

We compare the T_{60} estimation performance of the DNN method with the SDD method [11] on the simulated data in Table 1. The mean absolute error (MAE) between true and predicted T_{60} values is used for all the comparisons. For all test conditions the DNN method outperforms the SDD method significantly, especially at low SNR levels where the performance SDD method degrades dramatically.

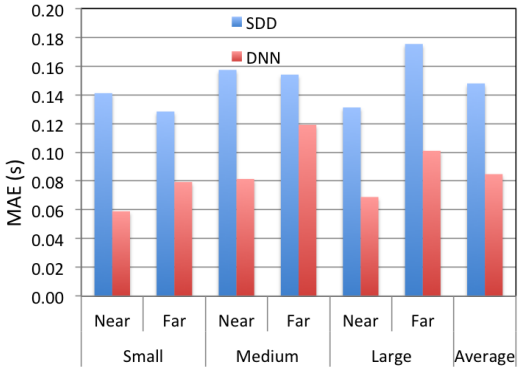

Figure 4: MAE comparison of the DNN and SDD methods on real recordings with near and far distances in 3 types of rooms.

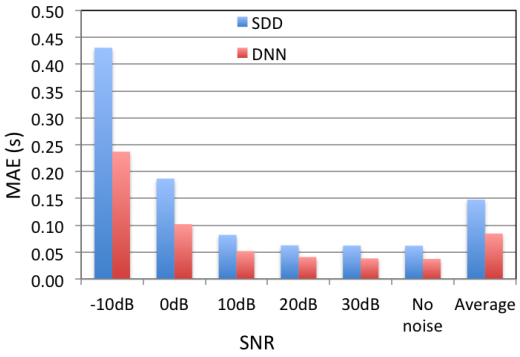

Figure 5: MAE comparison of the DNN and SDD methods for T_{60} estimation on real recordings at different SNR levels.

In Fig. 4 and Fig. 5, we show the T_{60} estimation results of the DNN and SDD method on the real recordings. Although the characteristics of the real rooms are not available when the DNN is trained, the DNN still obtains much smaller MAE than the SDD method in all test cases. The results show that the DNN generalizes well on unseen test data. We believe that the true potential of the proposed DNN learning-based approach could be beyond the results shown in this study. Various ways of using larger amount of training data, or trying to match the training data to the test environment may further improve the performance. We will study them further in our future work.

5. Conclusion

In this paper, we estimated the T_{60} reverberation time by using a DNN trained from a large amount of simulated reverberant and noisy speech data. When training is finished, the DNN automatically learns a group of feature extractors that extract the gradient of speech energy w.r.t. time. Such features are shown useful to T_{60} estimation task. Experimental results on simulated and real speech signals show that the DNN outperforms the SDD method significantly in all test cases and is able to generalize well to unseen test conditions and robust to severe noisy conditions.

6. Acknowledgements

This work is supported by DSO funded project MAISON DSOCL14045 and A*STAR funded HCCS programme.

7. References

[1] H. Kuttruff, *Room acoustics*. 4 edn, Taylor & Francis, 2000.

[2] W. C. Sabine, *Collected Papers on Acoustics.* Harvard U.P., Cambridge, 1922.

[3] *Acoustics-measurement of the reverberation time of rooms with reference to other acoustical parameters*, International Organization for Standardization, Geneva, Switzerland, 1997.

[4] M. R. Schroeder, "New method of measuring reverberation time," *Journal of the Acoustical Society of America*, vol. 37, pp. 409–412, 1965.

[5] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "The NTU-ADSC systems for reverberation challenge 2014," in *Reverberation Challenge Workshop*, Florence, Italy, May 2014.

[6] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. OBrien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[7] R. Ratnam, D. L. Jones, and W. D. OBrien, "Fast algorithms for blind estimation of reverberation time," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 537–540, 2004.

[8] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel, August 2010.

[9] T. J. Cox, F. Li, and P. Darlington, "Extracting room reverberation time from speech using artificial neural networks," *J. Audio Eng. Soc*, vol. 49, no. 4, pp. 219–230, 2001.

[10] J. Wen, E. Habets, and P. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, March 2008, pp. 329–332.

[11] J. Eaton, N. Gaubitch, and P. Naylor, "Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost," in *proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, May 2013, pp. 161–165.

[12] T. Jan and W. Wang, "Blind reverberation time estimation based on laplace distribution," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, Aug 2012, pp. 2050–2054.

[13] T. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, April 2010.

[14] T. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 7, pp. 1766–1774, Sept 2010.

[15] N. D. Gaubitch, H. W. Loellmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC)*, Sept 2012, pp. 1–4.

[16] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[17] Y. Bengio, *Foundations and Trends in Machine Learning*, 2009, ch. Learning deep architectures for AI, pp. 1–127.

[18] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *the 40th IEEE International Conference on Acoustic, Speech, and Signal Processing*, Brisbane, Australia, April 2015.

[21] J. D. Polack, "La transmission de l' énergie sonore dans les salles," PhD thesis, Université du Maine, Le Mans, 1988.

[22] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: a british english speech corpus for large vocabulary continuous speech recognition," 1995, pp. 81–84.

[23] J. B. Allen and D. Berkley, "Image method for efficiently simulating small-room acoust." *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Jul. 1979.

[24] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, 2013.