

COVER PAGE

Document downloaded by @DAEL

Thu Nov 26 19:02:56 2026
For personal use

When automatic English translation is provided, only the original document is authentic.
The EAA cannot be held responsible of any translation error

Bibliographical reference

Deep Neural Network Based Blind Estimation of Reverberation Time Based on Multi-channel Microphones, Myungin Lee and Joon-Hyuk Chang, Acta Acustica **vol. 104** (Number 3), 2018, pp. 486-495

DOI
https://doi.org/10.3813/AAA.919191

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 104 (2018) 486 – 495
DOI 10.3813/AAA.919191

Deep Neural Network Based Blind Estimation of Reverberation Time Based on Multi-channel Microphones

Myungin Lee, Joon-Hyuk Chang
School of Electronic Engineering, Hanyang University, Seoul, 04763, Korea. jchang@hanyang.ac.kr

Summary
Reverberation causes a performance degradation in distinct speech processing. For this reason, quantitatively estimating the amount of reverberation from the signal received by the microphone has been an important task for characterizing room acoustics and compensating for degradation due to an algorithm. In this paper, a novel method that estimates the reverberation time (T_{60}) based on multi-channel microphones using a deep neural network (DNN) is proposed. Each channel's distribution of the decay rates for each frequency and the generalized cross-correlation with phase transform (GCC-PHAT) between the microphones are adopted as the input feature vectors for DNN training. Those refined features enable the DNN composed of multiple nonlinear hidden layers to learn the nonlinear relationship that labels the reverberation time from the input features, which is known to be challenging with low-order features. The proposed algorithm is evaluated with extensive noisy conditions, and the results show the advantage of employing multi-channel signals with spatial features when compared with conventional methods.

PACS no. 43.60.-c, 43.60.Fg, 43.60.Jn, 43.60.Np

1. Introduction

Speech sound signals comprise a variety of information. The human auditory system perceives sound through the ears and obtains selective information, for example, the message, source direction, source distance, and characteristic of the sound. The characteristics of the speech signal enable us to estimate the acoustic information of the room where the speech has been uttered. In order to obtain such estimation, humans consciously and unconsciously learn from their auditory experience by classifying meaningful information from the rich speech signals [1].

For speech signal processing, it would be much helpful if the reverberation time (T_{60}) can be obtained from the signal perceived by the microphone as the information facilitates the algorithms, for example, dereverberation, recognition, acoustic modeling, and beamforming. Recently, a variety of mobile devices and Internet of Things (IoT) devices with multiple microphones, such as smart home hubs, are trending as applications of speech signal processing. Motivated by these circumstances, in this paper, a deep neural network (DNN), which is a machine learning approach, is adopted to let the algorithm learn to estimate reverberation time based on the feature vectors obtained from multiple microphones.

Previous studies have shown that analytical approaches outperform machine learning [2], as blind reverberation time estimation is a challenging task for neural networks with low-order features. However, by adopting higher-order input vectors, which effectively implicate the characteristics of reverberation, the proposed method performs well whether the circumstance is noisy or not. The input vectors, which represent the distribution of decay rates for each frequency bin and the generalized cross-correlation with phase transform (GCC-PHAT) [3], [4] are precisely selected to adequately represent the degree of reverberation of the speech signal. The distribution of decay rates provides the information of decaying envelope, while it does not have any information on the content of the signal. Meanwhile, reverberated speech shows correlation over longer time spans than clean speech. As a result, they represent different aspects of the reverberation. These features operate with a non-linear and complementary relationship that does not rely only on a single feature vector through DNN training procedure. This structure empowers the algorithm, which is demonstrated through the experiment.

An analysis is carried out by comparing different types of noises for various conditions, and a conclusive critique is made describing the advantage of the proposed algorithm.

Received 4 April 2017,
accepted 24 April 2018.

© S. Hirzel Verlag · EAA

2. Review of Blind Reverberation Time Estimation

The quantity of reverberation is determined by the acoustic factors, such as room geometry, the reflectivity of the surfaces, the locations of the source, and the observation point. Ever since Sabine characterized the reverberation time (T_{60}) [5], it has represented the primary characteristic of a reverberant space, where T_{60} is defined as the time taken for a sound level to diminish by 60 dB at cessation. Sabine established a formula to find T_{60} through an experimental approach according to the geometry and the surface absorption of the environment.

The conventional estimation method presented by Schroeder calculates T_{60} from the room impulse response (RIR) of the space [6]. However, this requires to measure RIRs in situ or estimate them with room acoustic modeling techniques, which is not always possible in real-world applications. Therefore, several blind estimation algorithms, which measure T_{60} without room acoustic information, have been devised based on the analytical interpretation of the reverberated speech signal [7]–[12].

2.1. Maximum Likelihood (ML)

T_{60} estimation methods using ML utilize the likelihood function for each decay rate based on the statistical models of the reverberated speech. In order to establish the deterministic model, the model is derived as a function parameterized with a damping constant δ. The model is expressed with the following.

$$y(t) = x(t) * h(t) + w(t), \tag{1}$$

where $x(t)$, $h(t)$, $w(t)$, and $y(t)$ represent the speech signal, the RIR between the source and the microphone, noise, and reverberated speech, respectively. The model developed by Polack [13] assumes non-stationary stochastic RIR in the time-domain as described by

$$h(t) = b(t)\,\mathrm{e}^{-\delta t} \quad \text{for} \quad t \geq 0, \tag{2}$$

where δ is a damping constant and $b(t)$ is a zero-mean Gaussian stationary noise.

Through the definition of T_{60}, δ derives T_{60} as

$$T_{60} = \frac{3}{\delta \log_{10} e}. \tag{3}$$

The dependency between T_{60} and the damping constant δ is only relevant when the sound field in the enclosure is diffusive and the distance between the microphone and the source is larger than the diffuse-field distance [14]. The ML-based method utilizes the log-likelihood function to search the most likely δ value corresponding to the decay rate of the speech sequence. Löllmann *et al.* [10], [11] extended the work of Ratnam *et al.* [7] and improved the computational efficiency and robustness of the estimation. The study by Löllmann uses downsampling and pre- selection of plausible decays before estimation to mitigate the computational efficiency. By using a histogram and smoothing procedure, the variance of the estimation is lowered.

2.2. Spectral Decay Distribution (SDD)

The negative side variance (NSV) has been widely used to estimate T_{60} from reverberated speech [9], [12]. Derived from the decay model of Equation (2), the energy decay relief (EDR) [15] for the corresponding model is

$$\mathcal{E}\left\{h^2(t)\right\} = \sigma_b^2 \mathrm{e}^{-2\delta t} = \sigma_b^2 \mathrm{e}^{-2\delta t} = \sigma_b^2 \mathrm{e}^{\lambda_h t}, \tag{4}$$

where ε denotes expectation, σ_b^2 the variance of b(t), and $\lambda_h = -2\delta$ the decay rate. The model $\mathcal{E}\left\{h^2(t)\right\}$ can be developed for frequency-dependent room decay rates using the short-time Fourier transform (STFT) as

$$\bar{H}(t, f) = P(f)\mathrm{e}^{\lambda_h(f)t} \quad \text{for} \quad t \geq 0, \tag{5}$$

where $\bar{H}(t, f)$ is the energy envelope of the model RIR at time t and frequency f, $\lambda_h(f)$ is the decay rate at f, and $P(f)$ is the initial power spectral density (PSD) of the noisy reverberated speech signal $y(t)$ when $t = 0$. By taking the natural logarithm of (5), the model is linearized as

$$\ln \bar{H}(t, f) = \ln P(f) + \lambda_h(f)t \quad \text{for} \quad t \geq 0. \tag{6}$$

Therefore, by applying linear fitting to the natural logarithm of the energy envelope of the RIR in the STFT domain, the room decay rate $\lambda_h(f)$ can be estimated. To obtain the room decay rate, the SDD method utilizes the decay rate of the reverberated speech by restricting least square linear fitting to the envelope of the log-energy in each STFT domain in a continuous manner. By disregarding the positive-sides of the value from the linear fitting, which are mostly related to speech itself, the NSV σ_{x-}^2 is derived as

$$\sigma_{x-}^2 = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L}\lambda'(k, l)^2, \tag{7}$$

where $\lambda'(k, l)$ denotes the negative decay rate for frequency band k and time frame l among the K frequency bands and L frames. The NSV is used to implement the room decay rates $\lambda_h(f)$ using the data-driven method with a second order polynomial mapping function. The function is trained by a clean speech database with known T_{60}. The successive study offers an algorithm to enhance the noise robustness by selecting time-frequency bins with high likelihoods and applying an *a posteriori* signal-to-noise ratio (SNR) adaptive method [12]. The method is known to demonstrate substantial robustness to additive noises at a realistic level. The performances of the above algorithms are precisely evaluated by the acoustic characterization of environments (ACE) challenge, which is proposed to determine the state- of-the-art in the blind estimation of acoustic parameters [2].

2.3. Single Channel Reverberation Time Estimation using DNN

Since various machine learning methods were devised and widely used for state-of-the-art classifications such as speech recognition [16]-[18] and hand-written character recognition [19], T_{60} estimators using machine learning have been developed [20]-[22]. Still, the precise study made by the ACE challenge states that the suggested machine learning approaches are less successful than the existing analytical approaches [2]. Meanwhile, motivated by the DNN that is proposed by Hinton and Salakhutdinov [23], DNN-single channel features (DNN-SCF), which handles single channel microphone and the DNN, was proposed in a previous study [24]. The estimator adopts speech decay rates in the STFT domain and the distribution of the speech decay rates for each frequency bin as the input feature vectors of the DNN in order to obtain the nonlinear hidden relationships between the features and the reverberation time. The distribution of the speech decay rates for each frequency f is expressed by the following equation

$$P_{\mathrm{SDD}}(f) = \frac{1}{L}\sum_{l=1}^{L}\left(\lambda'(f,l)\right)^2, \tag{8}$$

where $\lambda'(f,l)$ denotes a negative decay rate for the frequency band f and time frame l over the L frames. Likewise, DNN- SCF adopts higher-order input vectors which effectively implicate the characteristics of the reverberation, whereas the previous machine learning approaches adopt lower-order input vectors [20]- [22]. Furthermore, whereas conventional SDD-based methods do not precisely consider the decay rates for each frequency bin by superimposing all decay rates for the input of the polynomial regression [9], [12], DNN-SCF utilizes each distribution value for each frequency bin so that the algorithm can exploit the frequency- dependent characteristics of the reverberation more precisely [24].

3. Proposed DNN-based reverberation time estimation based on multi-channel microphones

Recently, the steady growth of mobile devices and IoT devices, such as smart home hubs, have made multiple microphone-based speech signal processing to be commonplace. Motivated by these circumstances, this paper presents a multi-channel DNN estimator (DNN Multiple Channel Features, DNN-MCF) for the blind estimation of T_{60}. In addition to the conventional features, which are the distribution of the decay rates, the multi-channel microphone-based features are employed as the input features of the DNN structure. As described in the previous studies [9] and [12], the distribution of the decay rates of reverberated speech is highly correlated with the reverberation time. First, the proposed method employs the distribution of the decay rates for each frequency f as input features, which is equivalent to $P_{\mathrm{SDD}}(f)$ from DNN-SCF. In addition, we adopt the generalized cross-correlation with phase transform (GCC-PHAT) between microphones so that the suggested algorithm utilizes the correlated state of the direct sound and the reflected sound.

The GCC-PHAT between the two signals $x_1(n)$ and $x_2(n)$ is defined by

$$G_{\mathrm{PHAT}}(f) = \frac{X_1(f)\left[X_2(f)\right]^*}{\left|X_1(f)\left[X_2(f)\right]^*\right|}, \tag{9}$$

where $X_1(f)$ and $X_2(f)$ refer to the Fourier transform of each signal and * stands for the complex conjugate. By applying the inverse Fourier transform to $G_{\mathrm{PHAT}}(f)$, $R_{\mathrm{PHAT}}(d)$ is obtained, which contains correlational information in the time domain. "d" refers to the time interval between the target signals. Generally, $R_{\mathrm{PHAT}}(d)$ is used to estimate the time delay of arrival (TDOA) between two microphones by using the corresponding relationship between the maximum value of $R_{\mathrm{PHAT}}(d)$ and the TDOA for a particular segment [3]. However, there is no precedent that GCC-PHAT is used for estimating the amount of reverberation. In order to obtain more meaningful characteristics of reverberation from the feature, we observe a range of $R_{\mathrm{PHAT}}(d)$ values considering the fact that the range of d contains valid information of reverberation. Through our experiment, we found that observing range of d from -10 to 10 is enough to present meaningful data. Hence, it is found that even though GCC-PHAT cannot be used for the estimation by itself, it contains the characteristic of reverberation. The property of GCC-PHAT in reverberant conditions can be observed from the spectrum. Figure 1 shows the features derived from $R_{\mathrm{PHAT}}(d)$ for 3 different reverberation times. The horizontal axis refers to the frame number whereas the vertical axis refers to the features derived from $R_{\mathrm{PHAT}}(d)$ among the 4 microphones. The feature is compound from 3 factorial combinations with each $R_{\mathrm{PHAT}}(d)$ values where d is -10 to 10. Therefore, the total number of features derived from GCC-PHAT is $6 \times 21 = 126$. The spectrum shows that in lower reverberation, GCC-PHAT shows distinctive values along the range of the time interval d. Contrarily, in higher reverberation, the spectrum of GCC-PHAT barely changes along the range of d. From this observation, we can expect that the range of $R_{\mathrm{PHAT}}(d)$ will operate positively when the features are trained in DNN.

Meanwhile, noise, which is inevitable in real life, smears the GCC- PHAT values and it generally leads to a severe degradation of the speech signal processing algorithm independent of the reverberation condition. However, for the proposed algorithm, by employing a range of GCC-PHAT, the algorithm utilizes the correlational property of reverberant speech. As we suppose that the noises are uncorrelated, it is unlikely for noise to be highly correlated as the reverberated speech does. Furthermore, as the training-set contains various conditions including low SNRs down to -5 dB and low reverberations down to 0.2 s, severe degradation is prevented.

In the proposed method, as described in Figure 2, the algorithm extracts the distribution of decay rates for each

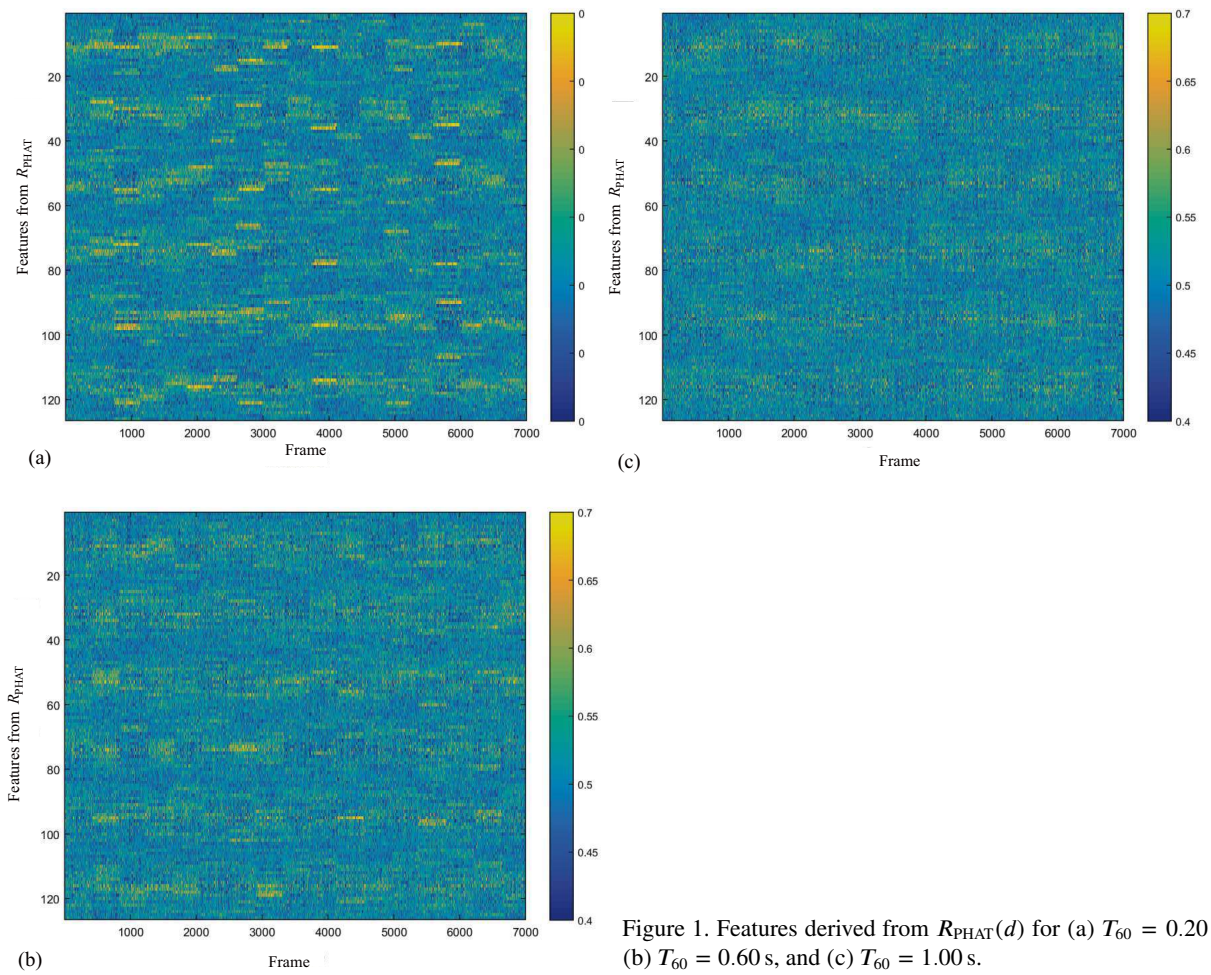


Figure 1. Features derived from $R_{\mathrm{PHAT}}(d)$ for (a) $T_{60} = 0.20\,\mathrm{s}$, (b) $T_{60} = 0.60\,\mathrm{s}$, and (c) $T_{60} = 1.00\,\mathrm{s}$.

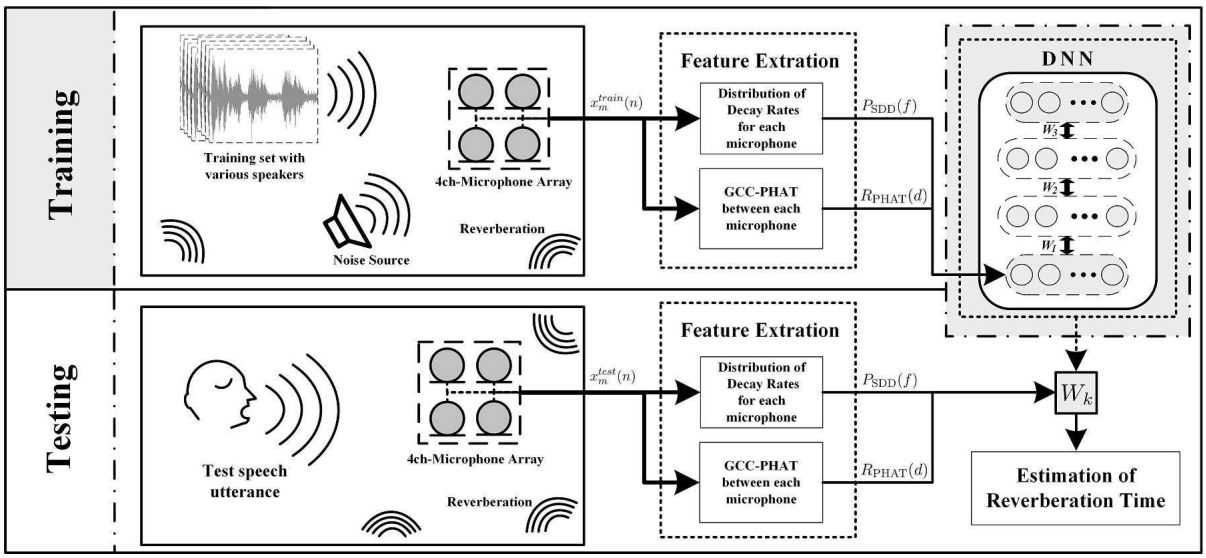


Figure 2. Concept of the proposed DNN-based T_{60} estimation based on multi-channel microphone (DNN-MCF, DNN-Multiple Channel Features).

single channel signal $x_m(n)$ as well as the GCC-PHAT between signals from the microphones, where m represents the channel number from 1 to 4. The features are stacked to form the input vectors of the DNN structure. The DNN with hidden layers, including adequate inside units, models the non-linear relationship between a large number of input features and the target. The specific setups about the input features and DNN are discussed in section 4, 5.1 and 5.2.

4. Synthesis of Reverberated Speech and Noise

A well-organized training step for machine learning requires an appropriate database that characterizes the problem. In order to obtain a representative database for the reverberation time, the synthesis of reverberated speech and noise is created with various combinations. Distance, size, position, and direction are manipulated to represent various circumstance. Especially, the geometry of the microphone, room geometry, range of the T_{60}, and noise condition are selected considering the normal circumstance for which the smart home hub may encounter through the operation in real life. The geometry of the microphone array used in the proposed method is described in Figure 3.

Firstly, the sources are located at every 18° azimuth with different distances and room dimensions. The positions of the speech source and the noise source are determined independently. Since our room configuration is a rectangular parallelepiped, when the spatial gap between sources is same, there is a possibility that the relative positioning of microphone array and sources in the room is duplicated due to the symmetric property of the room. This situation decreases the diversity of the dataset and may cause overfitting to a certain configuration during the training procedure. To avoid the unwanted case, the positioning of 20 sources, which was basically distributed at every 18° azimuth, is manipulated within a range preventing the overlap. As a result, the angle between the kth source and the microphone array, θ_{k}, is

$$\theta_{\mathrm{k}} = 18k + \theta_{\mathrm{thres}}, \tag{10}$$

where θ_{thres} is a random constant between 1 to 17 with uniform distribution.

In our experiment, the horizontal distances between the microphone array and the sources were between 200 cm and 250 cm. The heights of sources were between 150 cm to 190 cm. The width, length, and height of the room teach varied up to 0.2 m from the initial size 9 m ×8 m ×2.5 m. As a result, 20 set of positions are established. Additionally, we made 3 supersets of the 20 set to enlarge the geometry's diversity which leads to 60 different room sizes and placements.

The geometry of the microphone array is determined in a practical manner. Mobile and IoT devices make multichannel microphone arrays accessible; in this study, a square-shaped 4 channel microphone array with 8 cm distance in between the sensors is used.

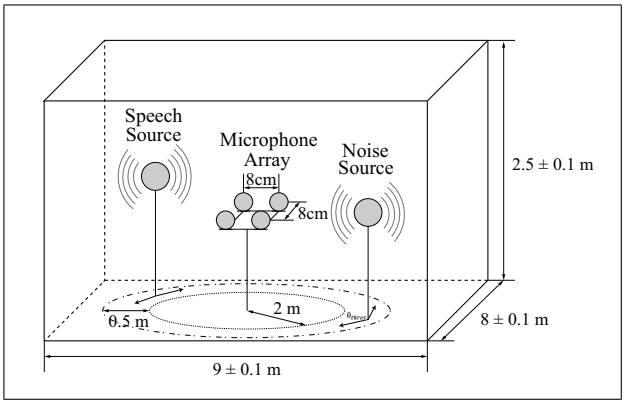


Figure 3. Geometry of the microphone array for the synthesis of the reverberated speeches and noises. The database is synthesized with various angles and positions with random manipulation constants.

Even though the microphone array results in spatial aliasing in high frequencies above around 2 kHz, the influence is expected to be insignificant from the aspect of T_{60} estimation. First of all, considering the late reverberation, which is mainly composed of low frequency, mostly determines the T_{60}, the influence that spatial aliasing of high frequency affects our experiment is minor. Moreover, as the proposed algorithm is devised for the voice signal, we can expect the potential input signal of GCC-PHAT is a broadband signal. Broadband signals do not experience severe side effects of spatial aliasing unless the signal is composed of strong harmonic components at integer multiples of the fundamental frequency [25]. Still, the exact relation between GCC-PHAT with spatial aliasing and T_{60} estimation has not been studied analytically and the further studies on the microphone array for T_{60} estimation are required.

To simulate the microphone array as a stand-alone device, a unique reverberated speech database for training and testing is required for a range of T_{60} values. The RIRs for each T_{60}, from 0.20 s to 1.20 s at every 0.05 s, are obtained with the image source method from 60 different room sizes and placements [26, 27].

The training and test speech dataset consist of anechoic speech signals from the Aurora 4 database [28] convolved with the numerically simulated RIRs. Noise sources are also synthesized using the same method with randomly distributed positions. The noise database from the Noisex database is employed to represent various noise circumstances [29]. We randomly selected 10% of speech utterances from 7,138 utterances of Aurora 4 for training. Accordingly, the total number of utterances used for training is $60 \times 21 \times 714 = 899,640$. For every utterances, we excluded the non-speech condition so that the validation is made only for the reverberated speech.

In contrast, to achieve unbiased and unsupervised evaluation, a totally different evaluation set is adopted. We used both a native Korean [30] and the TIMIT [31] database to eliminate the speaker, sentence, language, and placement dependency. Furthermore, to evaluate the noise robustness,

various types of noises are synthesized with diverse conditions from SNR of 15 dB to 5 dB.

Still, there are some limitations of the proposed reverberated speech database. We assume that the acoustic parameters are controlled in our expectation to achieve a disciplined experiment for the T_{60} estimation. First, the database is generated with frequency-independent absorption coefficients. Moreover, the absorption coefficients for 6 walls of the room are equivalently determined according to the given T_{60} values. These assumptions are not realistic considering the fact every wall have different frequency-independent absorption coefficients in real life. Besides, even though we adopted 60 different room sizes, the variation of the room size is relatively small as the variation is bounded within 0.2 m. The absorption coefficient and the room size affect the characteristics of reverberation including the ratio of early reflection versus reverberation even for same T_{60} values.

Nevertheless, we consider that the limitation of the absorption coefficient and room size has a minor influence on the proposed method considering the late reverberation is not severely affected by the parameters.

In addition, in the real speech signal processing application, it is common that the speech signals are overlapped because of simultaneous utterances. The influence of speech overlap is not presented in this study. Therefore, the validation for a realistic database with various condition including diverse absorption coefficients, room size, and multiple speakers should be made for a future work.

5. Experiments

5.1. Feature Extraction

In order to extract the feature, the reverberated speeches were sampled at 8 kHz, and features were derived from the STFT domain with a Hamming window of 128 samples, where the FFT size is 128. The size of the observation was sufficiently selected considering the geometry of the microphone array and the acoustic nature of the reflection. The reverberated speech decay rates were obtained for each frequency bin by continuously applying a least-square linear fitting over 30 sequential frames with 25 % overlap in the STFT domain. Likewise, with the corresponding synchronization, GCC-PHAT features were acquired by observing the cross-correlation between the channels from 10 frames ahead and 10 frames behind. The observation range of the correlation value was determined from the experiment. Since the temporal correlation of the signal exiting the diffuse field determines the spatiotemporal correlation [32], the correlation values of the further frames rarely had vivid information. Empirically, whether observing 10 or 20 frames nearby, the option did not affect the performance of the proposed algorithm. However, the complexity of learning and the application was mitigated by cutting off less meaningful information.

5.2. DNN Training Setup

The training setup for the proposed algorithm was built with the distribution of the decay rates $P_{\mathrm{SDD}}(f)$ and GCC-PHAT values $R_{\mathrm{PHAT}}(d)$ abstracted from the synthesized database. The $R_{\mathrm{PHAT}}(d)$ accounted for $6 \times 21 = 126$ dimensions and the $P_{\mathrm{SDD}}(f)$ comprised $4 \times 127 = 508$ dimensions among the 634 total input dimensions. The features work as the representative input data and then the layers of hidden units in the DNN have a role of capturing higher-order correlation from the input data [16]. The DNN had 3 hidden layers and the number of nodes for each hidden layer was set to 1500, which was empirically chosen to adequately describe the intricate relationships whilst avoiding over-fitting to the training setup.

When learning the DNN, each hidden layer was fine-tuned with 1000 epochs. A rectified linear unit (ReLU) was adopted as the activation function of the hidden layer [33]. While the weight matrix and the biases vector on each hidden layer were randomly initialized, the training procedure of the neural network consecutively made the model get stuck in the pool optimal-optima and caused over-fitting when the number of hidden layers and units increased. To address such problems, the dropout technique was applied to the training procedure [34]. Dropout is widely used to reduce over-fitting and improve the regularization process instead of the pre-training procedure due to its training efficiency.

5.3. Simulated Performance Evaluation

The validation was made with unseen speech utterances from TIMIT [31] and a native Korean database [30] with various additive noises from the Aurora 2 noise database [35]. Considering the nature of the multi-channel microphone array, the geometry and the placement of the sources were also unsupervised in order to avoid unwanted supervision by chance.

In order to present a precise observation from various aspects, the performances of the methods were analyzed for each noise type, SNR, and bounded T_{60}. The evaluation on each circumstance was made with three different indicators; the bias, the mean square error (MSE), and the Pearson correlation coefficient (ρ) between the ground truth and the estimation results. These indicators are known to accurately determine the performance of the acoustic parameters. For instance, an estimation algorithm with a low bias and MSE is likely to estimate values close to the median with a low variation. Also, as ρ gets closer to 1, one may assume that the estimation becomes more accurate compared with the case for ρ close to 0 [2].

For the rational validation of the proposed algorithm, the indicators derived from 4 different estimation methods, an ML-based method [11], SDD with Mel-spaced frequency bands and selective averaging by Gerkmann (SDDSA-G) [12], DNN-SCF [24], and the multi-channel SDD based DNN estimator (DNN-MSDD), are also presented for comparison. DNN-MSDD is implemented by this study and does not have GCC-PHAT features. DNN-SCF and DNN-MSDD were assessed to observe the influence of the number of channels and the GCC-PHAT features.

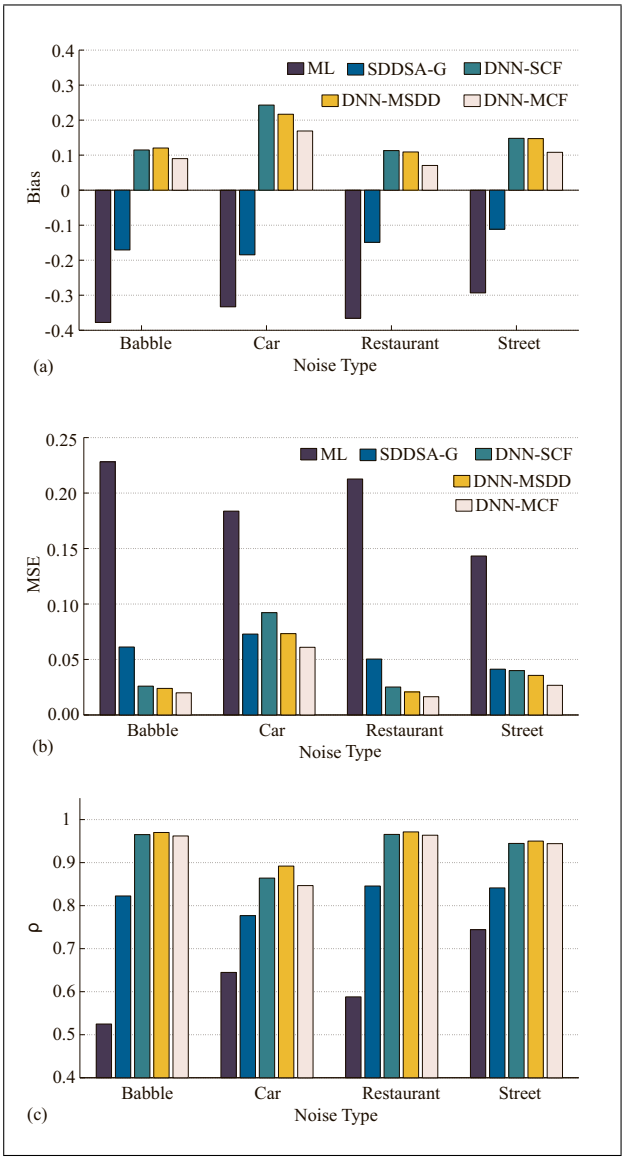


Figure 4. T_{60} estimation algorithm performance in various noise environments for all SNRs: (a) bias (b) MSE, and (c) ρ.

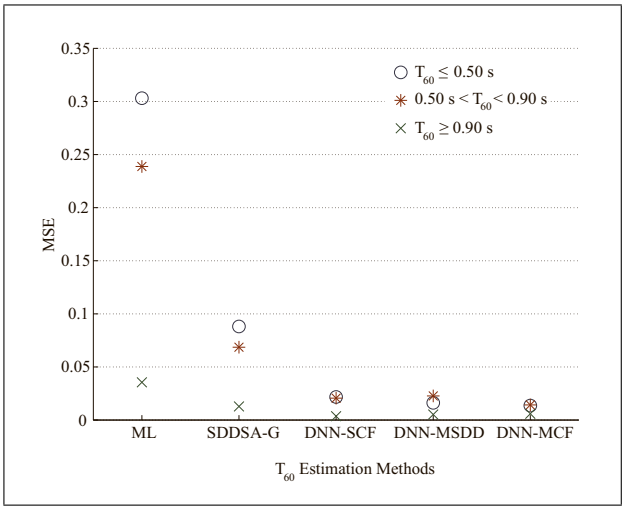


Figure 5. T_{60} estimation algorithm performance (MSE) in all environments and for all SNRs for $T_{60} \leq 0.50s$, $0.50s < T_{60} < 0.90s$, and $T_{60} \geq 0.90s$.

5.4. Results and Discussion

In this section, the performance of the estimators is analyzed for various aspects, including noise type, SNR, and T_{60} range, and the meaning of the results are discussed.

First, Figure 4 presents the performance in various noise types for all SNRs. From the aspects of bias and MSE, the overall comparison shows that DNN-MCF outperformed the other methods regardless of the noise type. The results show that the extension of the microphone channel and the adoption of the correlational parameter improved the performance from the viewpoints of bias and MSE.

Moreover, according to Figure 5, DNN-MCF showed the most balanced MSE value regardless of the T_{60} range. When the irregularity of the performance among the T_{60} range is severe, it is hard for the applications to utilize T_{60} with an optimum result. The stability of DNN-MCF was derived from its complementary characteristic of the features. While T_{60} estimators using decay rates as a feature were known to be vulnerable to small changes of the gradient of PSD for the speech signal, the proposed algorithm overcame the deficiency by utilizing GCC-PHAT, which demonstrates the distinct nature of the reverberation which could not be obtained from the distribution of the decay rates.

Meanwhile, in terms of ρ, DNN-MSDD showed the highest ρ while every DNN based approach was on the edge of the analytical methods. Even though ρ for DNN-MCF was slightly lower than that of DNN-MSDD, the difference between them was minor. For instance, the overall average ρ of DNN-MCF scored 0.89 and MSDD scored 0.90, while DNN-SCF, SDDSA- G, and ML scored 0.86, 0.71, and 0.44, respectively. Tables I, II, and III present detailed experimental results for each noise condition, and the best performances are marked in bold. Bias errors with the lowest absolute value are defined as the best performance. The lowest MSE values and the highest ρ values are identified as the best performance

To summarize, the proposed method showed distinguishable performance in almost every circumstance with three different indexes. This contains suggestions for both prospects of machine learning and room acoustics. Machine learning only works when the data is sufficient to learn the function of the needed level of complexity. Recent T_{60} estimators using machine learning approaches have struggled to determine the estimation function from relatively low-level features [20]–[22]. In order to improve a machine learning based T_{60} estimator with higher accuracy, we believe there are two approaches. The first method is employing a sufficiently powerful machine learning algorithm that can obtain the T_{60} estimation function from the low-level data. The second approach is employing higher-order features that can assist the existing machine learning algorithm to obtain the mapping function easily.

In the proposed method, we refined the speech signal into the distribution of the decay rates and GCC-PHAT. By using the refined features, DNN could obtain the T_{60} esti-

Table I. T_{60} Estimation Algorithm Performance: Bias in Various Noise Environments by SNR.

Noise type	SNR (dB)	ML	SDDSA-G	DNN-SCF	DNN-MSDD	DNN-MCF
Babble	5	-0.4731	-0.2258	0.1948	0.1922	**0.1485**
	10	-0.3793	-0.1747	0.1134	0.1222	**0.0936**
	15	-0.2807	-0.1115	0.0361	0.0466	**0.0282**
Car	5	-0.4296	**-0.2512**	0.3637	0.3259	0.2720
	10	-0.3242	-0.1851	0.2421	0.2154	**0.1658**
	15	-0.2452	-0.1176	0.1234	0.1094	**0.0691**
Restaurant	5	-0.4453	-0.1930	0.1902	0.1780	**0.1269**
	10	-0.3714	-0.1524	0.1123	0.1131	**0.0684**
	15	-0.2804	-0.1012	0.0365	0.0364	**0.0163**
Street	5	-0.3682	**-0.1682**	0.2463	0.2313	0.1730
	10	-0.2851	-0.1108	0.1425	0.1512	**0.1101**
	15	-0.2265	-0.0554	0.0551	0.0595	**0.0413**

Table II. T_{60} Estimation Algorithm Performance: MSE in Various Noise Environments by SNR.

Noise type	SNR (dB)	ML	SDDSA-G	DNN-SCF	DNN-MSDD	DNN-MCF
Babble	5	0.3260	0.0884	0.0502	0.0459	**0.0367**
	10	0.2258	0.0602	0.0205	0.0200	**0.0169**
	15	0.1326	0.0351	0.0072	**0.0058**	0.0062
Car	5	0.2739	**0.1120**	0.1736	0.1380	0.1177
	10	0.1772	0.0690	0.0788	0.0629	**0.0500**
	15	0.1000	0.0377	0.0242	0.0190	**0.0152**
Restaurant	5	0.2912	0.0700	0.0485	0.0394	**0.0300**
	10	0.2164	0.0498	0.0202	0.0179	**0.0128**
	15	0.1304	0.0313	0.0068	0.0049	**0.0064**
Street	5	0.2129	0.0616	0.0794	0.0682	**0.0501**
	10	0.1306	0.0397	0.0309	0.0309	**0.0226**
	15	0.0863	0.0225	0.0096	0.0079	**0.0075**

Table III. T_{60} Estimation Algorithm Performance: ρ in Various Noise Environments by SNR.

Noise type	SNR (dB)	ML	SDDSA-G	DNN-SCF	DNN-MSDD	DNN-MCF
Babble	5	0.2501	0.7764	0.9432	**0.9512**	0.9349
	10	0.5710	0.8233	0.9695	**0.9739**	0.9676
	15	0.7535	0.8679	0.9824	**0.9849**	0.9834
Car	5	0.4531	0.6829	0.7563	**0.8095**	0.7257
	10	0.6589	0.7877	0.8849	**0.9055**	0.8700
	15	0.8224	0.8596	0.9510	**0.9607**	0.9443
Restaurant	5	0.4022	0.8126	0.9448	**0.9581**	0.9409
	10	0.5945	0.8456	0.9694	**0.9720**	0.9693
	15	0.7672	0.8789	0.9825	**0.9841**	0.9810
Street	5	0.6054	0.7984	0.9100	**0.9174**	0.9040
	10	0.7736	0.8372	0.9495	**0.9552**	0.9504
	15	0.8531	0.8880	0.9743	0.9775	**0.9779**

mation function which could not be obtained from the low-level features. Likewise, employing appropriate higher-order features for machine learning can boost the implementation of arcane problems.

For the future work, the validation of real reverberated speech and the adaptation of another machine learning approaches remain. In our experiment, a computer-based simulation was used to evaluate the performance. Even though simulated RIR is useful for restricted and quantitative inspections, still, the validation for real reverberated speech is essential for the real application of the estimator. Furthermore, there is a possibility that the proposed method can be improved with different machine learning approaches.

6. Conclusions

Motivated by recent demands of mobile devices and IoT devices, e.g., smart home hubs, a novel algorithm estimating T_{60} using the DNN based on multi-channel microphone was proposed. The expansion of microphones enabled us to adopt a novel spatial feature, GCC-PHAT, and the adaptation of the feature fortified the previously introduced feature, the distribution of the decay rates. The meaning of each feature from the viewpoint of reverberation was described in detail throughout the paper and the structure of the machine learning was expounded. Considering the nature of the neural network approach, which is highly likely to be decided by the training data, the synthesis of the noisy reverberated speech was also precisely introduced. The previous state-of- the-art approaches were also introduced and compared with the proposed algorithm. Both state-of-the-art conventional algorithms, the ML-based estimation and SDDSA-G, turned out appropriate for comparison for their representative characteristics on the efficiency and noise robustness. In order to present the efficacy of the channel expansion and the correlational feature, DNN-SCF and DNN-MSDD were also assessed. The validation of the algorithms was presented from various aspects and digitized with three different indexes: bias, MSE, and ρ. Through this analysis, the proposed method showed promising performance for almost every circumstances with accuracy. Accordingly, the proposed machine learning approach with refined features had competitive priorities among the existing methods, which are known to be dominated by analytical approaches. Still, achieving actual experimental tests with real reverberated speech for the proposed method remains for our future work and the optimization of specific procedures considering its performance and computational cost should be studied.

7. Acknowledgement

This work was supported by the research fund of Signal Intelligence Research Center supervised by the Defense Acquisition Program Administration and the Agency for Defense Development of Korea.

References

[1] R. C. O' Reilly, Y. Munakata: Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain. Cambridge, MA: MIT Press, 2000.

[2] J. Easton, N. D. Gaubitch, A. H. Moore, P. A. Naylor: Estimation of room acoustic parameters: The ACE challenge. IEEE/ACM Transactions on Audio, Speech, & Language Processing **24**(10) (2016) 1681–1693.

[3] C. Knapp, G. Carter: The generalized correlation method for estimation of time delay. IEEE Trans. Acoust., Speech, Signal Process. **24**(4) (1976) 320–327.

[4] M. S. Brandstein, H. F. Silverman:, A robust method for speech signal time-delay estimation in reverberant rooms, Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 1997, Munich, Germany.

[5] W. C. Sabine: Collected Papers on Acoustics. Cambridge, MA: Harvard Univ. Press, 1922.

[6] M. R. Schroeder: New method of measuring reverberation time. J. Acoust. Soc. Am. **37** (1965) 409–412.

[7] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O' Brien, Jr., C. R. Lansing, A. S. Feng: Blind estimation of reverberation time. J. Acoust. Soc. Am. **114**(5) (2003) 2877–2892.

[8] R. Ratnam, D. L. Jones, and W. D. O' Brien, Fast algorithms for blind estimation of reverberation time. IEEE Signal Process. Letters, **11**(6) (2004) 537–540

[9] J. Y. C. Wen, E. A. P. Habets, P. A. Naylor: Blind estimation of reverberation time based on the distribution of signal decay rates. in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2008, Las Vegas, NV, USA.

[10] H. W. Löllmann, P. Vary: Estimation of the reverberation time in noisy environments. Proc. of Int. Workshop on Acoustic Echo and Noise Control (IWAENC), 2008, Seattle, Washington, USA.

[11] H. W. Löllmann, E. Yilmaz, M. Jeub, P. Vary: An improved algorithm for blind reverberation time estimation. Proc. Int. Workshop Acoust. Echo Noise Control, 2010, Tel-Aviv, Israel.

[12] J. Eaton, N. D. Gaubitch, P. A. Naylor: Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2013, Vancouver, BC, Canada.

[13] J. Polack: La transmission de l'énergie sonore dans les salles. Dissertation, Université du Maine, 1988.

[14] H. Kuttruff: Room Acoustics. 5th ed. Spon Press, London, U.K. 2009.

[15] J.-M. Jot: An analysis/synthesis approach to real-time artificial reverberation. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 1992, San Francisco, CA, USA.

[16] A. Mohamed, G. Dahl, G. Hinton: Deep belief networks for phone recognition. Proc. NIPS, 2009, Vancouver, BC, Canada.

[17] A. Mohamed, G. Dahl, G. Hinton: Acoustic modeling using deep belief networks. IEEE Trans. on Audio, Speech and Lang. Proc. **20**(1) (2012) 14–22.

[18] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. **29**(6) (2012) 82–97.

[19] H. Lee, C. Ekanadham, A. Ng: Sparse deep belief net model for visual area V2. Proc. NIPS, 2007, Vancouver, BC, Canada.

[20] F. Xiong, S. Goetze, B. T. Meyer: Blind estimation of reverberation time based on spectro-temporal modulation filtering. Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2013, Vancouver, BC, Canada.

[21] J. F. Santos, T. H. Falk: Blind room acoustics characterization using recurrent neural networks and modulation spectrum dynamics. in Proc. Audio Eng. Soc. Conf., 2016, Leuven, Belgium.

[22] P. P. Parada, D. Sharma, T. van Waterschoot, P. A. Naylor: Evaluating the non-intrusive room acoustics algorithm with the ACE challenge. Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015, New Paltz, NY, USA.

[23] G. Hinton, R. Salakhutdinov: Reducing the dimensionality of data with neural networks. Science **313**(5786) (2006) 504–507.

[24] M. Lee, J.-H. Chang: Blind estimation of reverberation time using deep neural network. IEEE Int. Conf. on Network Infrastructure and Digital Content (IC-NIDC), 2016, Beijing, China.

[25] J. Dmochowski, J. Benesty, S. Affès: On spatial aliasing in microphone arrays. IEEE Trans. Signal Process. **57**(4) (2009) 1383–1395.

[26] J. B. Allen, D. A. Berkley: Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4) (1979) 943–950.

[27] RIR Generator, AudioLabs – Room impulse response generator for MATLAB, https://github.com/ehabets/RIR-Generator, 2016.

[28] N. Parihar, J. Picone, D. Pearce, H. G. Hirsch: Performance analysis of the Aurora large vocabulary baseline system. Proc. of the European Sig. Proc. Conf. (EUSIPCO), 2004, Vienna, Austria.

[29] A. Varga, H. Steeneken: Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication **12**(3) (1993) 241–251.

[30] S.-W. Yoon, H.-G. Kang, Y.-C. Park, D.-H. Yoon: An efficient transcoding algorithm for G.723.1 and G.729A 3461 speech coders: Interoperability between mobile and IP network. Speech Communication **43(1-2)** (2004) 17–31.

[31] J. S. Garofolo: Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, 1988, MD, USA.

[32] B. Rafaely: Spatial-temporal correlation of a diffuse sound field. J. Acoust. Soc. Am. **107** (2000) 3254–3258.

[33] C. Zhang, P. C. Woodland: Parameterised sigmoid and ReLU hidden activation functions for DNN acoustic modelling. Proc. Interspeech, 2015, Dresden, Germany.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov: Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. (JMLR) (2014) 1929–1958.

[35] D. Pearce, H.-G. Hirsch: The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proc. Int. Conf. Spoken Lang. Process. (ICSLP) **4**, 2000, Beijing, China.