PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Dereverberation based on deep neural networks with directional feature from spherical microphone array recordings

Jeongmin LIU[1] ; Byeongho JO[2] ; Jung-Woo CHOI[3]

[1,2,3]School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Republic of Korea

ABSTRACT

The dereverberation of a reverberant audio signal can be accomplished through the deconvolution of a room impulse response (RIR) from the reverberant signal. However, the RIR is unknown in most practical situations, which makes the dereverberation as a challenging problem. The previous dereverberation studies utilizing the deep neural network (DNN) have shown that the performance strongly depends on RIRs used for the training data, because the network trained only by RIRs of few number of rooms cannot handle data from various room conditions. To build a more generalized dereverberator, we incorporate directional features extracted by a spherical microphone array recording as the input to the DNN. Since directional cues can provide useful information on the degree of reverberation, it enables the construction of a more generalized dereverberation model. When the DNN model is trained by directional features, all performance measures on the signal-to-noise ratio, speech intelligibility and quality are improved as compared to the same model trained by the single channel feature. Especially, the performance improvement for the test signals recorded in a different reverberation condition and measurement position demonstrates that we can build a more general model with the proposed directional feature.

Keywords: Speech Dereverberation, Deep Neural Networks, Spherical Microphone Array, Acoustic Intensity

1. INTRODUCTION

In many practical situations, the quality and intelligibility of speech signals recorded by the sound system are often deteriorated by the wall reflections, i.e., reverberations. The reverberation is one of the major causes of performance degradation in speech recognition (1) and communication systems (2, 3), and hence, a lot of dereverberation algorithms have been developed.

Recently, deep neural network (DNN) based dereverberation algorithms are being extensively studied (4–6) and show outperforming performances compared to the conventional techniques. These algorithms are able to reduce the reverberation without the prior information of room impulse responses (RIRs), through the repeated training of short time Fourier transform (STFT) of reverberant and dry speech signals.

Previous studies have shown that the dereverberation performance is deteriorated when the reverberation characteristics of the training and test data sets are different, i.e., when the rooms used for the training and test are different. The ways to resolve this problem also have been developed (7, 8). The methods in (7, 8) use the aggregation of the multiple DNN models, each of which is trained by a room of specific reverberation time. By training multiple models using RIRs of different reverberation times and using only the best one in the test stage, they can adapt to various RIR data different from the training dataset. Nevertheless, the aggregation strategy requires huge amount of time and effort for training individual models. Moreover, even when the training and test data sets are obtained from the same room, the performance drop can still occur due to the change of microphone and loudspeaker positions. To build a more generalized DNN model, we propose the use of spatial information as input features of the DNN. The spatial information can provide an important clue of the amount of reverberant field at each time and frequency. For example, the spatial correlation at two distanced positions is reduced in a diffuse sound field (9). Using this characteristic, the sound field measured by multiple microphones are used to predict the diffuseness (9). Knowing the exact amount of reverberant sound energy at each time and frequency is crucial in recovering the direct sound energy of each bin, which is also essential in building a general DNN model that can adapt to various rooms with different reverberation characteristics. To measure the spatial information, microphone arrays should be employed. Recently, various spherical microphone arrays have been developed for virtual reality (VR) applications. Although the spherical microphone arrays can record and analyze spatial information of sound

[1] jeongmin96@kaist.ac.kr
[2] byeongho@kaist.ac.kr
[3] jwoo@kaist.ac.kr

fields using multichannel microphones mounted on a spherical surface, the training of a DNN model using such multichannel data requires high computational effort and long training time. For this reason, we use directional features of which dimension is much reduced compared to that of raw multichannel signals.

In detail, we consider two different types of directional features. The first one is the direction vector (DV) of DirAC (Directional Audio Coding) (10), which has been widely used for both the recording and compression of 3D audio. The other one is the spatially-averaged intensity vector (SIV), which is an extension of the DV using the higher order spherical harmonics. Both features include the spatial information of a sound field in form of the mean active intensity.

To discuss advantages of the DNN model trained with directional features, we first begin with the description of the conventional dereverberation technique based on the spectral mapping in Section 2. Then the directional features are introduced in Section 3, and their use for the DNN training is explained in Section 4. Finally, in Section 5, the performance of DNN models trained with and without the directional features are compared and analyzed.

2. SPEECH DEREVERBERATION BASED ON SPECTRAL MAPPING

Most of the conventional dereverberation techniques use the single channel data recorded in a reverberant environment to reconstruct the free-field or anechoic recording without reverberation (4–6). The free-field recording at time τ, denoted by $p_a(\tau)$, can be regarded as the source signal $s(\tau)$ delayed by the propagation delay τ and attenuation by the propagation distance r. That is,

$$p_a(\tau) = \frac{1}{r} s(\tau - \tau_0) \tag{1}$$

On the other hand, the recorded reverberant signal $p_r(\tau)$ is given by the convolution ($*$) of the RIR $h(\tau)$ and source signal $s(\tau)$.

$$p_r(\tau) = h(\tau) * s(\tau) \tag{2}$$

Time-frequency distributions of these signals are given by applying the short time Fourier transform (STFT). First, the spectrogram, i.e., the STFT of the anechoic signal $p_a(\tau)$ and the reverberant signal $p_r(\tau)$ can be decomposed into the magnitude and phase dependent parts as

$$\mathcal{STFT}\left\{p_a(\tau)\right\} = P_a(t,f)\exp\left(i\Phi_a(t,f)\right) \tag{3}$$

$$\mathcal{STFT}\left\{p_r(\tau)\right\} = P_r(t,f)\exp\left(i\Phi_r(t,f)\right) \tag{4}$$

where t and f express indices of the time frame and frequency bin, respectively.

The dereverberation technique based on the spectral mapping tries to estimate the anechoic signal's magnitude spectrogram $P_a(t,f)$ (Figure 1(a)) from the measured spectrogram $P_r(t,f)$ with reverberation (Figure 1(d)) by means of the neural network. The phase spectrogram is usually not reconstructed because the phase spectrum is wrapped within the range [-π, π) and is uniformly distributed without no distinct pattern (unstructured), which makes it difficult to be handled by the neural network (11, 12). For this reason, many spectral mapping techniques utilize the Griffin-Lim algorithm (13) that estimates the anechoic phase $\Phi_a(t,f)$ from the estimated magnitude spectrogram $\hat{P}_a(t,f)$ and the reverberant phase $\Phi_r(t,f)$ (4–8).

3. PROPOSED FEATURE

In this work, we attempt to enhance the dereverberation performance by adding directional feature as the input of the DNN. To this end, two different directional features are considered. The first candidate is the conventional measure: the DV of DirAC (10), and the second is a SIV, which is considered in this study for better estimation of acoustic intensity.

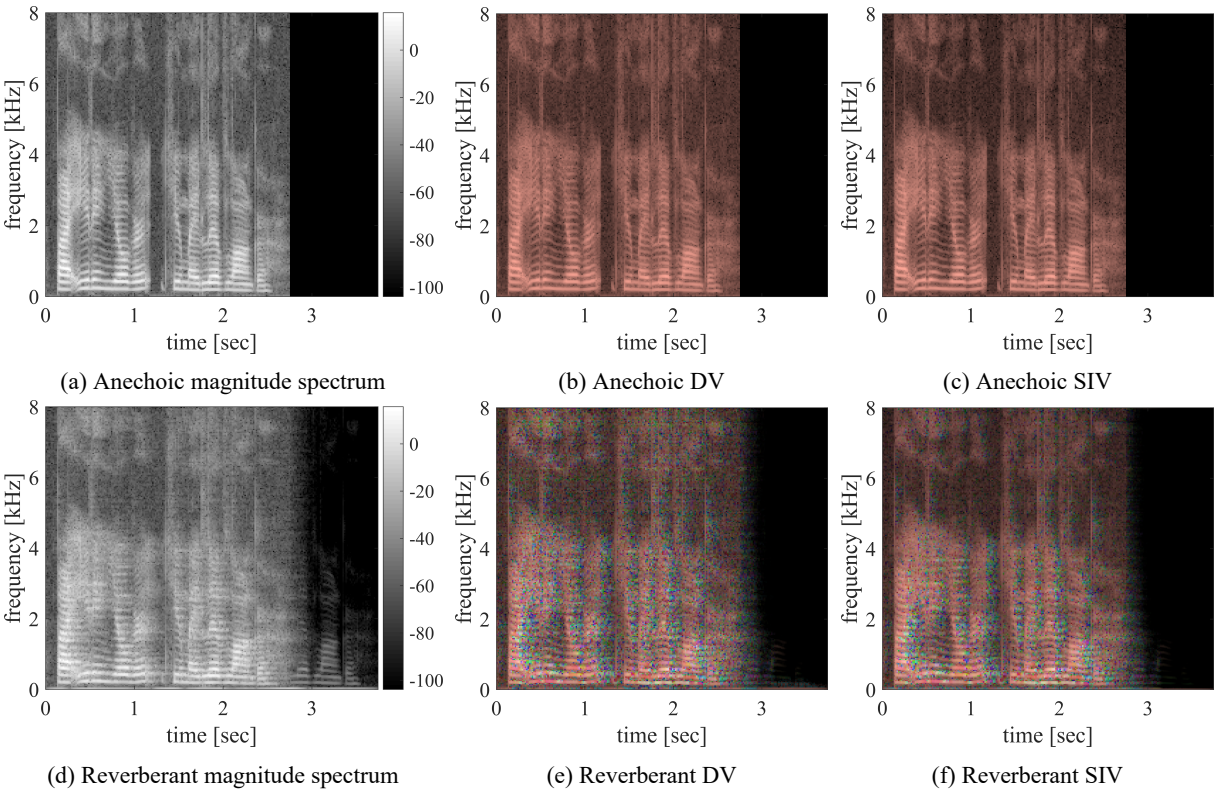
3.1 Direction vector (DV) of DirAC

In DirAC, the acoustic intensity can be estimated from a B-format Ambisonics signal (10). Denoting the STFTs of W, X, Y, Z channels of a B-format signal as A_W, A_X, A_Y, A_Z, the DV $\mathbf{D}$ can be written as

$$\mathbf{V}(t,f) = \begin{bmatrix} A_X(t,f) & A_Y(t,f) & A_Z(t,f) \end{bmatrix}^T \tag{5}$$

$$\mathbf{D}(t,f) = -\operatorname{Re}\left\{A_W(t,f)\mathbf{V}(t,f)^*\right\} \tag{6}$$

The X, Y, Z channels correspond to the particle velocity in the Cartesian coordinates, and the W channel (A_W) is equivalent to the pressure field recorded by an omni-directional microphone, so P_a and P_r defined in Section 2 are the W channel signals recorded in a free-field and reverberant room, respectively. Therefore, the DV can be regarded as the mean active intensity of given time-frequency (TF) bin. In general, averaging through the adjacent time frames is performed in DirAC (10) but omitted in this work to reflect the local intensity fluctuation in time into the dereverberation process. Instead of time averaging, the real part of the complex intensity vector, representing the mean value of the propagating intensity component (mean active intensity), is considered. Using the same



(a) Anechoic magnitude spectrum (b) Anechoic DV (c) Anechoic SIV

(d) Reverberant magnitude spectrum (e) Reverberant DV (f) Reverberant SIV

Figure 1. Visualization of spectrograms and directional features (DV and SIV). (Vector lengths of directional features are scaled in a logarithmic scale)

notation, we can denote the DV recorded in a reverberant room as $\mathbf{D}_r$.

To incorporate DV to a DNN structure, we first concatenate the DV $\mathbf{D}_r$ with the spectrogram P_r in the channel direction. Then the concatenated vector

$$\mathbf{P}_r(t,f) = \begin{bmatrix}\mathbf{D}_r(t,f)^T & P_r(t,f)\end{bmatrix}^T = \begin{bmatrix}D_{r,1}(t,f) & D_{r,2}(t,f) & D_{r,3}(t,f) & P_r(t,f)\end{bmatrix}^T \tag{7}$$

becomes the new input feature: *directional spectrogram*. Using this directional spectrogram, the spectral mapping from $\mathbf{P}_r(t,f)$ to $P_a(t,f)$ is trained by the DNN.

Since we add three more channels, the directional spectrogram can be regarded as a colored image consisting of three color channels and one extra channel. For example, if we use the HSV color space to represent the DV, the length of the DV can be mapped to the Value, and the vector's elevation and azimuth angles in the spherical coordinates are mapped to the Saturation and Hue, respectively (Figure 2). The change in the color pattern of the spectrogram image provide important clues to separate the reverberation from the direct sound. The magnitude spectrogram and the DVs visualized by this color mapping are shown in Figures 1(a) and 1(b), respectively. In this example, the sound from a single direction in an anechoic condition is depicted, so the DV of Figure 1(b) has a single color and only varies in Value.

When the same sound signal is recorded in the reverberant field, reflections from various directions distort the original spectrogram as well as the DV. This can be seen in Figures 1(d) and 1(e). Without the prior knowledge on the reflections, it is hard to predict the original spectrogram (Figure 1(a)) from the reverberant one (Figure 1(d)). In the case of the DV (Figure 1(e)), however, the reflections from different directions are expressed as a colored noise in the spectrogram, which will be beneficial in reconstructing the original spectrogram or DV. Consequently, the dereverberation is equivalent to reconstruction of the original spectrogram image through the denoising of the colored noise. In this regard, we trained DNN with intensity-related input features and compared the dereverberation performances.

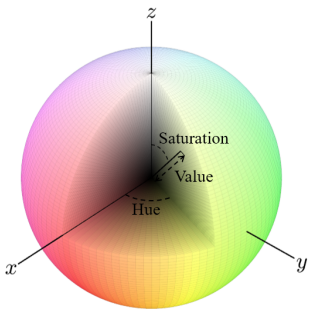


Figure 2. The mapping between the vectors (the DVs and the SIVs) and the colors for each TF bin in Figures 1(b), 1(c), 1(e) and 1(f).

3.2 Spatially-averaged intensity vector (SIV)

We propose to use an intensity-like feature called SIV. Since the DV of Equation (6) only represents the intensity vector at a single position, we need a new measure that can express the global intensity evaluated over a wide surface. The SIV is defined as the mean active intensity vector averaged over a spherical surface of the microphone array. That is,

$$\bar{\mathbf{I}}(t,f) \triangleq \int_0^{2\pi}\int_0^{\pi} \mathbf{I}(t,f,\mathbf{r}) \sin\theta d\theta d\phi = \int_0^{2\pi}\int_0^{\pi} \frac{1}{2}\,\mathrm{Re}\left\{p(t,f,\mathbf{r})\mathbf{v}(t,f,\mathbf{r})^*\right\} \sin\theta d\theta d\phi, \tag{8}$$

where $\mathbf{I}(t,f,\mathbf{r})$ denotes the mean active intensity at the time frame t, frequency bin f and microphone position $\mathbf{r} = [\, r \quad \theta \quad \phi \,]^T$. The pressure and velocity fields at the same spatio-temporal location are denoted by $p(t,f,\mathbf{r})$ and $\mathbf{v}(t,f,\mathbf{r})$, respectively.

The SIV can be directly calculated from the spherical harmonic domain (SHD) signal. In detail, the pressure signal $p_{nm}(t,f)$ in SHD can be obtained by taking the spherical Fourier transform of pressure signals $p(t,f,\mathbf{r})$ measured from the microphone array (14). Then, the particle velocity signals $\mathbf{v}_{nm}(t,f)$ in SHD can be derived from $p_{nm}(t,f)$ by using the recurrence relations of spherical harmonics (15). According to the Parseval's relation of the spherical harmonics (14), the angular average of multiplied signals in space domain is equal to the multiplication and sum of two SHD signals. That is,

$$\int_0^{2\pi}\int_0^{\pi} p(t,f,\mathbf{r})\mathbf{v}(t,f,\mathbf{r})^* \sin\theta d\theta d\phi = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} p_{nm}(t,f)\mathbf{v}_{nm}(t,f)^*. \tag{9}$$

Therefore, we can calculate the spatially-averaged active intensity from the SHD signals.

$$\bar{\mathbf{I}}(t,f) = \frac{1}{2}\,\mathrm{Re}\left\{\sum_{n=0}^{\infty}\sum_{m=-n}^{n} p_{nm}(t,f)\mathbf{v}_{nm}(t,f)^*\right\}. \tag{10}$$

We concatenate the x,y and z components of the SIV $\bar{\mathbf{I}}_r$ measured in a reverberant room in the channel direction. The concatenated vector

$$\mathbf{P}_r(t,f) = \left[\bar{\mathbf{I}}_r(t,f)^T \quad P_r(t,f)\right]^T = \left[\bar{I}_{r,1}(t,f) \quad \bar{I}_{r,2}(t,f) \quad \bar{I}_{r,3}(t,f) \quad P_r(t,f)\right]^T \tag{11}$$

becomes an input feature of the DNN.

The SIV can also be depicted as a color image. An example measured in an anechoic condition is shown in Figure 1(c), and that measured in a reverberant environment is shown in Figure 1(f). The DVs in Figures 1(b) and 1(e) and the SIV in Figures 1(c) and 1(f) show similar trends.

4. DEEP NEURAL NETWORK

To utilize the proposed input features for DNN, the features should be scaled and normalized first. Then a DNN structure that suits to the input features should be designed. In case of the multi-layer perceptron (MLP), the number of context frames should be determined in advance due to the structural limitation of the fully-connected layer. Accordingly, the amount of information that can be fed into the DNN at once is also limited by the number of context frames, and this limitation affects the estimation performance (4). However, we treat the spectrogram as a single image, so the convolutional layer can be adopted. The deconvolution problem can be recasted as the image-to-image transform problem, so the fully convolutional network (FCN) can be used for this problem (6). The FCN does not utilize any fully-connected layer, and hence, input and output data of various sizes can be handled. Therefore, information of reverberation can be collected over long and wide TF bins. The directional features are arranged as different color channels, so the network structures for the image signal processing can also be trained with the directional spectrogram input.

4.1 Log-magnitude scaling and normalization

The directional feature is a vector quantity consisting of direction and length. For efficient training of DNN, the vector's length is scaled in a logarithmic scale as

$$g(x) = \log_{10}\left(\frac{x+\epsilon}{\epsilon}\right) \tag{12}$$

$$\tilde{D}_{r,i}(t,f) = g\left(\|\mathbf{D}_r(t,f)\|\right)\frac{D_{r,i}(t,f)}{\|\mathbf{D}_r(t,f)\|},\ (i \in \{1,2,3\}) \tag{13}$$

where ϵ is a regularization constant for avoiding the singularity near $x \sim 0$. Through the log-magnitude conversion, the vector's direction remain unchanged. The same log-magnitude conversion using the function $g(x)$ is applied to the spectrogram $P_a(t,f)$ and $P_r(t,f)$. $\tilde{P}_a(t,f)$ and $\tilde{P}_r(t,f)$ denote the log-scaled version of $P_a(t,f)$ and $P_r(t,f)$.

Next, each channel data are normalized by the mean and standard deviation of the corresponding channel over

the total time duration. In this work, the mean and standard deviation of the normalized data are set to 0 and 0.5, respectively. Denoting the normalization operator as $\mathrm{Norm}(\cdot)$, we can rewrite the normalized directional spectrogram and anechoic spectrogram as

$$\mathbf{P}'_r(t,f) = \begin{bmatrix} D'_{r,1}(t,f) & D'_{r,2}(t,f) & D'_{r,3}(t,f) & P'_r(t,f) \end{bmatrix}^T$$
$$\text{where} \begin{cases} D'_{r,i}(t,f) = \mathrm{Norm}\left(\tilde{D}_{r,i}(t,f)\right), \ (i \in \{1,2,3\}) \\ P'_r(t,f) = \mathrm{Norm}\left(\tilde{P}_r(t,f)\right) \end{cases} \tag{14}$$

$$P'_a(t,f) = \mathrm{Norm}\left(\tilde{P}_a(t,f)\right). \tag{15}$$

For the directional spectrogram with the SIV, the same log-magnitude conversion and normalization are applied.

4.2 Architecture

As a DNN architecture, we employ one of fully-convolutional networks called the FusionNet structure (16) as shown in Figure 3. Basically, the FCN is comprised of the encoder path that sequentially shrinks down the image size using the max pooling layer and generates a bottleneck feature, and the decoder path that increases the image size using the transposed convolutional layer. The conventional U-Net utilizes skip-connections between the encoder and decoder layers to preserve the image details lost by the encoder structure (17), but the FusionNet utilizes residual connections instead, and includes residual blocks in each layer to efficiently resolve the gradient vanishing problem (16). In the designed structure, input features are first processed by the pre-processing block, which plays a role of rearranging input features. Then the output of the processed by the encoder and decoder blocks that extract essential features and reconstruct the anechoic spectrogram, respectively.

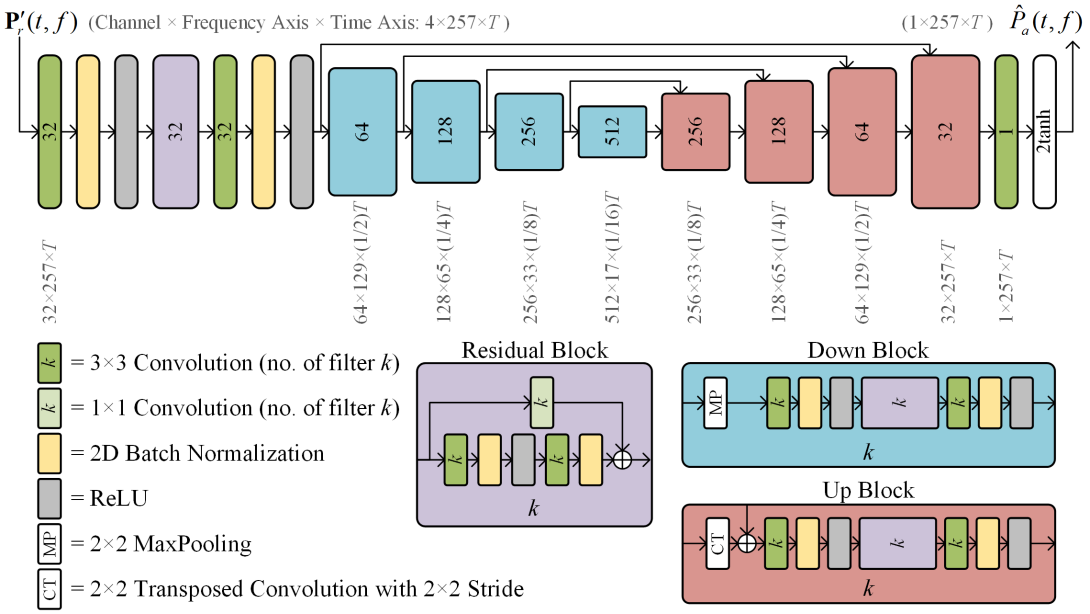


Figure 3. The DNN structure based on FusionNet (16). T_r is the number of time frames in the input $\mathbf{P}'_r$.

The mean squared error (MSE), which has been widely used in the conventional dereverberation studies (4–8), is used as a loss function. The MSE loss function is defined as

$$L_{\mathrm{MSE}}\left(P'_a, \hat{P}'_a\right) = \frac{1}{T_a} \sum_{t=0}^{T_a-1} \sum_{f=0}^{F-1} \left| P'_a(t,f) - \hat{P}'_a(t,f) \right|^2, \tag{16}$$

where $\hat{P}'_a$ is the output of the DNN, T_a is the total number of time frames in the spectrogram P'_a, and F is the number of frequency bins.

5. EXPERIMENTS AND RESULTS

To verify the effectiveness of directional feature, three different DNN models were compared. The RIRs between various loudspeaker positions and spherical microphone array locations were simulated in two virtual rooms of different reverberation conditions. Then DNN models were trained by the speech corpus convolved by these simulated RIRs.

5.1 Experimental settings

The spherical microphone array used in the simulation has a rigid surface of radius 4.2 cm and 32 microphones arranged at the same position as those of the Eigenmike (18), which enables the decomposition of spherical harmonics up to the third order (N=3).

The specifications of two virtual rooms used to synthesize RIRs are as follows; T60 of the Room 1 is 0.31 s, and its dimension is [10.52 m, 7.1 m, 3.02 m]. The Room 2 has longer T60 of 0.66 s and dimension of [6 m, 6 m, 6 m]. The RIRs were simulated by the spherical microphone array impulse response (SMIR) generator (19). To obtain many RIRs required for training DNN, positions of the microphone array and sound source were randomly generated from a uniform random distribution. Exceptional cases are when the source and microphone array were not separated by more than 50 cm from each other or from the walls constituting the room. In such cases, position samples were discarded and repopulated from the uniform distribution. For each room, 20 positions of the microphone array - source pair were used to simulate the seen RIR dataset, and another 20 positions are used to build the unseen RIR dataset.

To calculate the DV from the spherical microphone recording, the real SHD signals of the zeroth and the first order were calculated after compensating the mode strength $b_n(kr)$ of the rigid sphere (20). To avoid excessive amplification during the compensation, the regularized compensation of the following form was used:

$$b_n(kr)^{-1} = \frac{1+\sqrt{\lambda}}{|b_n(kr)|+\sqrt{\lambda}} \exp\left(-i\angle b_n(kr)\right), \tag{17}$$

where $\lambda = (1-\sqrt{1-1/G^2})/(1+\sqrt{1-1/G^2})$ and $G = 10/32$.

The directional spectrogram with the SIV (Equation (11)) was calculated from complex SHD signals. However, the modal strength compensation was omitted in here, because the definition of SIV includes the modal strength. The magnitudes of the input and desired output of the DNN are normalized by the mean and variance at each frequency, so the modal strength compensation itself does not influence the performance of DNN. In the final speech reconstruction stage, the phase spectrogram is reconstructed by applying the Griffin-Lim algorithm (13) to the output magnitude spectrogram of DNN. Before the phase reconstruction, however, the mode strength of the output magnitude spectrogram is compensated such that the output signal of the DNN trained by the SIV can be compared to that trained by the DV.

TIMIT corpus of 16 kHz sampling rate (21) was used as source signals. The dataset obtained by convolving the training data of TIMIT and seen RIR are denoted as total training dataset. Among the total training dataset, randomly selected 7000 samples were used as a training set, and 3000 samples were utilized as a validation set. Likewise, 2500 random samples chosen out of the test data of TIMIT convolved by seen RIRs were used for a seen test set. The other 2500 samples sampled from the convolution of the TIMIT test data and unseen RIRs were used as unseen test set.

For the STFT, the Hann window of 32 ms length and 512-points FFT were used. The hop size was the half of the window size, and 257 frequency bins were calculated (F=257). The parameter used for the log-magnitude conversion function $g(x)$ was $\epsilon = 10^{-10}$. The number of iterations for the Griffin-Lim algorithm (13) was 20.

The training of DNN were carried out for three different input features. Depending on the room and input features used for the training, the models were named as follows:

- Room No.-NoDF: trained without directional features
- Room No.-DV: trained with DV of DirAC
- Room No.-SIV: trained with SIV

The trained models were tested by the seen and unseen test set of the same room. This is for investigating the performance differences against the seen and unseen data of the same reverberant conditions. Next, the model is tested by the unseen test data of the different room as well.

5.2 Results

To evaluate the dereverberation performance, perceptual evaluation of speech quality (PESQ) (22), and short-time objective intelligibility (STOI) (23), frequency-weighted segmental signal-to-noise ratio (fwSegSNR) (24) were calculated. PESQ is a measure of speech quality and varies within the range [-0.5, 4.5]. STOI has a range of [0, 1] and higher number indicates better intelligibility.

The DNN architecture trained by three input features generated in the Room 1 (Room1-NoDF, Room1-DV, Room1-SIV) were evaluated by these performance measures, and the results are depicted in Figures 4(a) to 4(c). The abscissa of each graph indicates the test dataset used for the performance evaluation. The same measures for the case of Room2-NoDF, Room2-DV, and Room2-SIV are presented in Figures 5(a) to 5(c).

First, we compare the cases with and without the directional features. In the Room 1 (Room1-NoDF VS. Room1-SIV) and Room 2 (Room2-NoDF VS. Room2-SIV), all the measures are enhanced when the model is

trained by the SIV. In addition, all models show worse performance with the unseen test data than with the seen test data, but the degree of performance reduction is much less in the models trained by the SIV. The cause of this improvement by the use of directional feature can be suspected from the fact that both the early reflections and the degree of reverberation can be inferred from the directional feature.

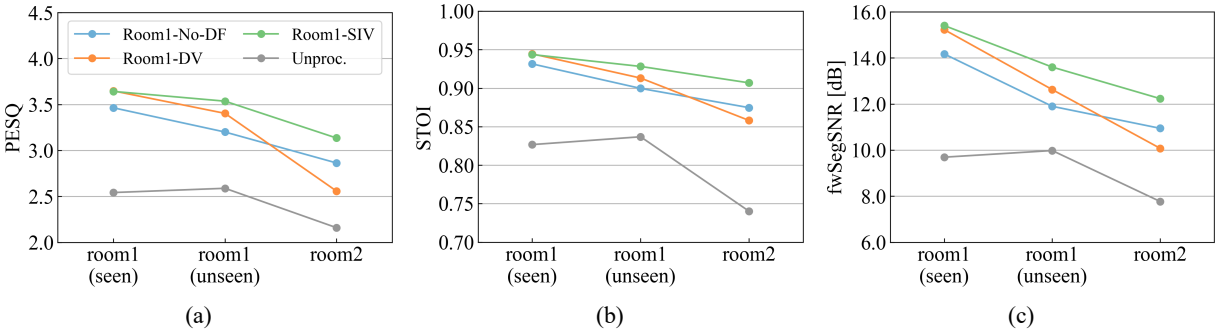


Figure 4. The average PESQ (a), STOI (b), and fwSegSNR (c) of the output of the three models trained by room 1 training set ("Room1-NoDF", "Room1-DV", and "Room1-SIV") and their reverberant input speech ("Unproc.").

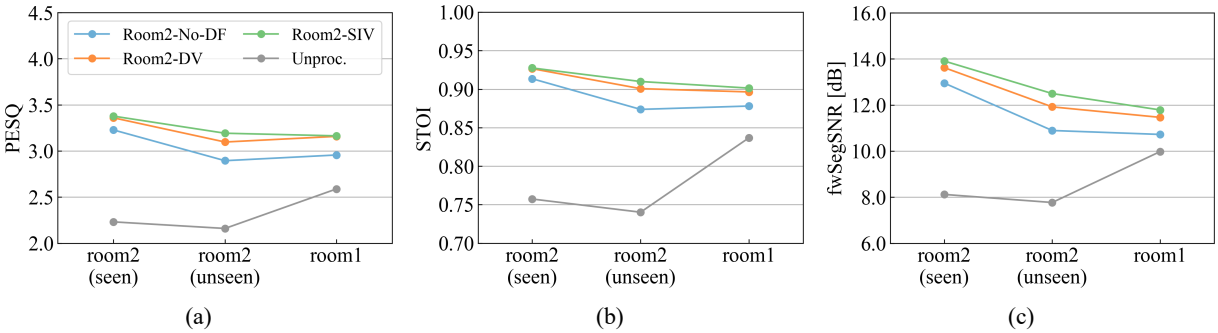


Figure 5. The average PESQ (a), STOI (b), and fwSegSNR (c) of the output of the three models trained by room 2 training set ("Room2-NoDF", "Room2-DV", and "Room2-SIV") and their reverberant input speech ("Unproc.").

When the DV is used instead of the SIV, however, the performance gain is somewhat inconsistent. In detail, the Room1-DV model shows similar performance to the Room1-SIV when tested by the Room 1 seen dataset, but its performance is rapidly decreased when tested by the unseen dataset. Especially, the performance of Room1-DV model tested by the unseen data of Room2 is even lower than that of the model without directional features (Room1-NoDF). The similar trend can be observed with the model trained by the Room 2 dataset (Figure 5). However, unlike the Room 1 cases, the model trained by the DV (Room2-DV) does not show the dramatic performance degradation for the room 1 test dataset. In short, the Room1-DV model has a weakness in processing the data that have longer reverberation than the training dataset. More detailed analysis is necessary, but it is evident that DV models calculating particle velocity using only low order of spherical harmonic coefficients cannot properly estimate the amount of reverberation when trained by the short reverberation dataset.

6. CONCLUSIONS

To build a general DNN model that can reduce reverberation of speeches recorded in various room conditions, we proposed to use two directional features, the DV and SIV, as the input feature of DNN. The directional features of three independent components can be treated as independent color channels of a spectrogram image. In this regard, the FusionNet architecture with multiple encoder and decoder blocks were trained by the proposed color spectrogram images. The models trained by datasets from different directional features and rooms were tested using the seen and unseen microphone positions and rooms. Results showed that the model trained by the SIV can handle general datasets including the one recorded in the room of different reverberation time. The model trained by the DV also showed better performance than the model trained by the pressure signal only, but it was inferior when the reverberation time of the test data is longer than that of the training dataset.

ACKNOWLEDGEMENTS

This work was supported by the Korea Research Foundation (KRF) under Contract No. 2019R1A2C1007393 and by the BK21 Plus program through the National Research Foundation (NRF) funded by the Ministry of Edu-

cation of Korea.

REFERENCES

1. Yoshioka T, Sehr A, Delcroix M, Kinoshita K, Maas R, Nakatani T, et al. Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. IEEE Signal Process Mag. 2012; 29(6) 114–126.
2. Kokkinakis K, Hazrati O, Loizou PC. A channel-selection criterion for suppressing reverberation in cochlear implants. J Acoust Soc Am. 2011; 129(5) 3221–3232.
3. Roman N, Woodruff J. Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold. J Acoust Soc Am. 2013; 133(3) 1707–1717.
4. Han K, Wang Y, Wang DL, Woods WS, Merks I, Zhang T. Learning spectral mapping for speech dereverberation and denoising. IEEE/ACM Trans Audio Speech Lang Process. 2015; 23(6) 982–992.
5. Zhao Y, Wang ZQ, Wang D. A two-stage algorithm for noisy and reverberant speech enhancement. Proc IEEE Int Conf Acoust Speech Signal Process; 2017. p. 5580–5584.
6. Ernst O, Chazan SE, Gannot S, Goldberger J. Speech dereverberation using fully convolutional networks. Proc Eur Signal Process Conf EUSIPCO; Rome 2018. p. 390–394.
7. Wu B, Li K, Yang M, Lee CH. A reverberation-time-aware approach to speech dereverberation based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process. 2017; 25(1) 98–107.
8. Lee WJ, Wang SS, Chen F, Lu X, Chien SY, Tsao Y. Speech dereverberation based on integrated deep and ensemble learning algorithm. Proc IEEE Int Conf Acoust Speech Signal Process; 2018. p. 5454–5458.
9. Pierce AD. Acoustics: An introduction to its physical principles and applications (McGraw-Hill series in mechanical engineering). McGraw-Hill Book Company; 1981.
10. Pulkki V, Faller C. Directional audio coding: filterbank and stft-based design. Proc Audio Eng Soc Conv 120; Paris, France 2006. p. 6658.
11. Gerkmann T, Krawczyk-Becker M, Le Roux J. Phase processing for single-channel speech enhancement: history and recent advances. IEEE Signal Process Mag. 2015; 32(2) 55–66.
12. Zheng N, Zhang XL. Phase-aware speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process. 2019; 27(1) 63–76.
13. Griffin D, Lim J. Signal estimation from modified short-time Fourier transform. IEEE Trans Acoust. 1984; 32(2) 236–243.
14. Rafaely B. Fundamentals of Spherical Array Processing. Springer-Verlag Berlin Heidelberg; 2015. Chapter 1: Mathematical Background; 1–30.
15. Jo B, Choi JW. Spherical Harmonic Smoothing for Localizing Coherent Sound Sources. IEEE/ACM Trans Audio Speech Lang Process. 2017; 25(10) 1969–1984.
16. Quan TM, Hilderbrand DGC, Jeong Wk. FusionNet: A deep fully residual convolutional neural network for image segmentation in connectomics. arXiv preprint arXiv:1612.05360. 2016; 1–8.
17. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Proc Med Image Comput Comput Assist Interv; 2015. p. 234–241.
18. EigenStudio User Manual R02C [Internet]. mh acoustics LLC; 2017 [cited 2019 May 31]. Available from: https://mhacoustics.com/sites/default/files/EigenStudio%20User%20Manual%20R02C.pdf
19. Jarrett DP, Habets EAP, Thomas MRP, Naylor PA. Rigid sphere room impulse response simulation: algorithm and applications. J Acoust Soc Am. 2012; 132(3) 1462–1472.
20. Jarrett DP, Habets EAP, Naylor PA. Theory and Applications of Spherical Microphone Array Processing. Springer International Publishing; 2017.
21. Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS. DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Technical Report N. 1993; 93.
22. Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. Proc IEEE Int Conf Acoust Speech Signal Process; Salt Lake City, USA 2001. p. 749–752.
23. Taal CH, Hendriks RC, Heusdens R, Jensen J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. Proc IEEE Int Conf Acoust Speech Signal Process; Dallas, USA 2010. p. 4214–4217.
24. ANSI/ASA S3.5. American National Standard Methods for Calculation of the Speech Intelligibility Index. Acoustical Society of America; 1997.