ESTIMATING DIRECT-TO-REVERBERANT RATIO MAPPED FROM POWER SPECTRAL DENSITY USING DEEP NEURAL NETWORK

Yusuke Hioka

University of Auckland
Department of Mechanical Engineering
Private Bag 92019, Auckland 1142, New Zealand
yusuke.hioka@ieee.org

Kenta Niwa

NTT Corporation
NTT Media Intelligence Laboratories
3-9-11 Midori-cho, Musashino,
Tokyo 180-8585, Japan
niwa.kenta@lab.ntt.co.jp

ABSTRACT

A new attempt for estimating the direct-to-reverberant ratio (DRR) by mapping the power spectral density (PSD) of the direct sound and reverberation using the deep neural network is reported. The method finds the correct DRR from the PSD estimated with an algorithm using a microphone array. The experimental results using a recording of a reverberant speech signal, which included various environmental noise, reveal that the proposed method is effective in improving the accuracy of DRR estimation and robust against various noise.

Index Terms— direct-to-reverberant ratio, deep neural network, microphone array, power spectral density, beamspace

1. INTRODUCTION

In recent years, the estimation of the direct-to-reverberant ratio (DRR) has been attracting interests in acoustic signal processing due to its wide variety of applications [1, 2, 3, 4, 5, 6, 7]. By reflecting the growth of interest, in 2015 the Acoustic Characterisation of Environments (ACE) Challenge was organised by the IEEE Audio and Acoustic Signal Processing Technical Committee, which included a task of evaluating methods for estimating the DRR [8].

The room impulse response (RIR) had to be measured for calculating the DRR of a reverberant enclosure. However the measurement of the RIR is actually a burden for application users since special equipment and software are needed. This motivated researchers to address the *blind* estimation of the DRR, which does not require RIR measurement. Due to the difference in the propagation properties of the direct sound and reverberation, many current methods use a microphone array to utilise the spatial properties of the propagation. The coherence of direct sound and reverberation between two microphones has been one of the most commonly utilised properties in previous studies on DRR estimation [3, 4, 6, 9, 10, 11]. Due to various errors in practical recordings, coherence-based methods sometimes derive a complex final DRR estimate, which is not a realistic value for the DRR.

In the meantime the authors also proposed a few DRR estimation methods that derive the DRR by taking the ratio of the power spectral density (PSD) of the direct sound and reverberation being estimated by using multiple beamformings [12, 13]. A similar approach using directional microphones instead of beamforming has also been reported [14]. Although one of the advantages with this approach is that it does not result in a complex DRR estimate, improving the estimation accuracy of the DRR to some extent, some significant offsets and variances were observed from the estimation results when a method was applied to normal speech signals [13]. It is hypothesised that those offsets occurred because the estimated PSD had been affected by several errors. Such errors include the assumptions introduced to the properties of the direct sound and reverberation as well as the statistical uncertainty of the observed signals. Thus, estimation accuracy can be improved if a mapping method that can absorb the errors in the estimated PSD is introduced.

In the last few years, the deep neural network (DNN) [15] has been widely applied to various fields because of its very robust ability to map a *feature* to another piece of information. It is anticipated that the DNN can mitigate the detrimental effects of errors by mapping an effective feature that represents the DRR, e.g. coherence and PSD, to the correct DRR. Being motivated by this thought, this study utilises the DNN for effectively mapping the PSD estimated with the previous method [13] to the correct DRR. The effect of introducing the DNN is investigated by conducting experiments using speech signals recorded in different acoustic environments. In contrast to other previous studies, this study is the first attempt to apply a statistical mapping method to an existing feature to improve DRR estimation accuracy.

2. ESTIMATING PSD OF DIRECT SOUND AND REVERBERATION USING MICROPHONE ARRAY

Like the previous method [13], the proposed method also estimates the PSD of the direct sound and reverberation using a microphone array. The principle of PSD estimation is briefly explained in this section.

2.1. Microphone array observation in reverberant environment

Given that the transfer function from a sound source to the m-th microphone of an M-sensor microphone array in a reverberant room is denoted as $H^{(m)}(\omega)$ with ω being the frequency, it can be separated into two components, as in (1); direct sound and reverberation, where the latter is assumed to include both the early reflections and late reverberation.

$$H^{(m)}(\omega) = H_{\mathrm{D}}^{(m)}(\omega) + H_{\mathrm{R}}^{(m)}(\omega), \tag{1}$$

Here $H_{\mathrm{D}}^{(m)}(\omega)$ and $H_{\mathrm{R}}^{(m)}(\omega)$ are the transfer functions of the direct sound and reverberation, respectively.

Let $X^{(m)}(\omega, t)$ be the observed signal of the m-th microphone in the time-frequency domain where t is a frame index. Using (1),

$X^{(m)}(\omega,t)$ can be modelled by

$$X^{(m)}(\omega,t) := \left(H_{\mathrm{D}}^{(m)}(\omega) + H_{\mathrm{R}}^{(m)}(\omega)\right) S(\omega,t), \tag{2}$$

where $S(\omega,t)$ is the spectrum of a sound source.

By further decomposing the transfer functions in (2) into two components, i.e. the transfer function from the sound source to a reference point located close to the microphone array (e.g. the centre of the microphone array), and that from the reference point to each microphone, the transfer function becomes

$$H_{\mathrm{D}}^{(m)}(\omega) = H_{\mathrm{Dref}}(\omega) e^{-j\omega\tau_{\Omega_{\mathrm{D}}}^{(m)}}, \tag{3}$$

$$H_{\mathrm{R}}^{(m)}(\omega) = \int_{\Omega} H_{\mathrm{Rref},\Omega}(\omega) e^{-j\omega\tau_{\Omega}^{(m)}} d\Omega, \tag{4}$$

where $H_{\mathrm{Dref}}(\omega)$ and $H_{\mathrm{Rref},\Omega}(\omega)$ are the transfer functions from the sound source to the reference point with regard to the direct sound and reverberation, respectively. The term $\tau_{\Omega}^{(m)}$ is the time delay of arrival compared to the reference point.

Using the steering vector [16] of an array for the solid angle $\Omega = \{\theta,\phi\}$, where θ is the azimuth and ϕ is the zenith angles ($\theta \in [0,2\pi)$, $\phi \in [0,\pi]$), $\mathbf{a}_{\Omega}(\omega) = [e^{-j\omega\tau_{\Omega}^{(1)}},\cdots,e^{-j\omega\tau_{\Omega}^{(M)}}]^T$, the observation vector of the microphone array is defined as

$$\begin{aligned}\mathbf{x}(\omega,t) &= [X^{(1)}(\omega,t),\cdots,X^{(M)}(\omega,t)]^T \\ &= \mathbf{a}_{\Omega_{\mathrm{D}}}(\omega) S_{\mathrm{D}}(\omega,t) + \int_{\Omega} \mathbf{a}_{\Omega}(\omega) S_{\mathrm{R},\Omega}(\omega,t) d\Omega,\end{aligned} \tag{5}$$

where T denotes the transpose of a vector or a matrix.

$$S_{\mathrm{D}}(\omega,t) = H_{\mathrm{Dref}}(\omega) S(\omega,t), \tag{6}$$

$$S_{\mathrm{R},\Omega}(\omega,t) = H_{\mathrm{Rref},\Omega}(\omega) S(\omega,t), \tag{7}$$

are the direct sound and reverberation arriving from the angle Ω observed at the reference point, respectively.

2.2. Beamforming output

Assuming that an arbitrary beamformer l is applied to the microphone array observation vector $\mathbf{x}(\omega,t)$, the output signal of the beamformer is represented by

$$Y_{\mathrm{BF},l}(\omega) = \mathbf{w}_l^H(\omega)\mathbf{x}(\omega,t), \tag{8}$$

where $\mathbf{w}_l(\omega)$ is the weight vector of beamformer l defined by $\mathbf{w}_l(\omega) := [W_l^{(1)}(\omega),\cdots,W_l^{(M)}(\omega)]^T$.

The PSD of the beamformer's output is then approximated by the summation of the PSD of the direct sound and reverberation multiplied by the gain of the beamformer:

$$\begin{aligned}P_{\mathrm{BF},l}(\omega) &= E[|Y_{\mathrm{BF},l}(\omega)|^2]_t \\ &\approx G_{l,\Omega_{\mathrm{D}}}(\omega) P_{\mathrm{D}}(\omega) + \int_{\Omega} G_{l,\Omega}(\omega) P_{\mathrm{R},\Omega}(\omega) d\Omega,\end{aligned} \tag{9}$$

where $P_{\mathrm{D}}(\omega)$ and $P_{\mathrm{R},\Omega}(\omega)$ are the PSDs of $S_{\mathrm{D}}(\omega,t)$ and $S_{\mathrm{R},\Omega}(\omega,t)$, respectively; $E[\cdot]_t$ is the expectation over frames that can be approximated by the average of several frames and $G_{l,\Omega}(\omega)$ is the gain of the beamformer for the angle Ω defined by $G_{l,\Omega}(\omega) = |\mathbf{w}_l^H(\omega)\mathbf{a}_{\Omega}(\omega)|^2$. In the derivation of (9), the direct sound and reverberation are assumed to be mutually uncorrelated.

Since the reverberation is commonly assumed as a diffuse sound, an isotropy can be imposed for its propagation. Thus, the PSD of the reverberation can be replaced with a constant value that holds for all Ω, i.e.

$$P_{\mathrm{R},\Omega}(\omega) = \overline{P}_{\mathrm{R}}(\omega) = \text{const.} \quad \forall\Omega. \tag{10}$$

Thus, the PSD of the beamformer output in (9) becomes

$$P_{\mathrm{BF},l}(\omega) = G_{l,\Omega_{\mathrm{D}}}(\omega) P_{\mathrm{D}}(\omega) + \overline{P}_{\mathrm{R}}(\omega) \int_{\Omega} G_{l,\Omega}(\omega) d\Omega. \tag{11}$$

Given that there are two beamformers, which have different directivity patterns being applied to the microphone array observation, the output PSD of the two beamformers can be formulated in a matrix form expressed by

$$\underbrace{\begin{bmatrix} P_{\mathrm{BF},1}(\omega) \\ P_{\mathrm{BF},2}(\omega) \end{bmatrix}}_{\mathbf{P}_{\mathrm{BF}}(\omega)} = \underbrace{\begin{bmatrix} G_{1,\Omega_{\mathrm{D}}}(\omega) & \int_{\Omega} G_{1,\Omega}(\omega) d\Omega \\ G_{2,\Omega_{\mathrm{D}}}(\omega) & \int_{\Omega} G_{2,\Omega}(\omega) d\Omega \end{bmatrix}}_{\mathbf{G}(\omega)} \underbrace{\begin{bmatrix} P_{\mathrm{D}}(\omega) \\ \overline{P}_{\mathrm{R}}(\omega) \end{bmatrix}}_{\mathbf{P}_{\mathrm{cmp}}(\omega)} \tag{12}$$

Because the elements in $\mathbf{P}_{\mathrm{BF}}(\omega)$ are derived from the microphone array's observation and that in $\mathbf{G}(\omega)$ are known *a priori*, the PSD of the direct sound and reverberation can be estimated by solving the simultaneous equation using the least squares method

$$\hat{\mathbf{P}}_{\mathrm{cmp}}(\omega) = \mathbf{G}^{-1}(\omega)\mathbf{P}_{\mathrm{BF}}(\omega), \tag{13}$$

where $\hat{\cdot}$ denotes an estimated value.

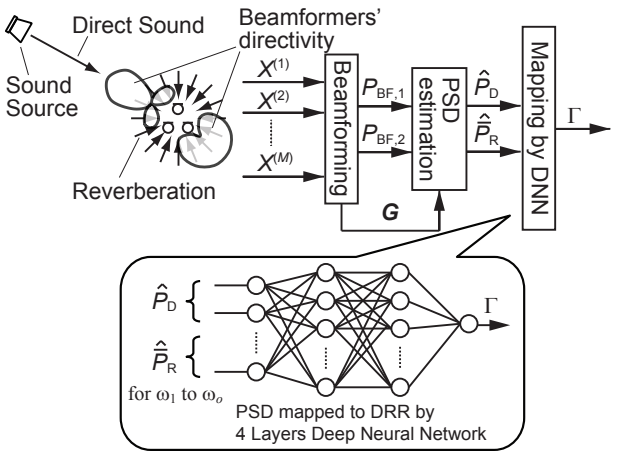

Fig. 1. Block diagram of proposed method

3. DNN-BASED DRR ESTIMATION USING ESTIMATED PSDS IN BEAMSPACE

In the previous study [13], the DRR was derived by simply calculating the ratio of the estimated PSD given by (13), i.e. DRR [dB] $= 10\log_{10}\frac{\sum_{\omega}\hat{P}_{\mathrm{D}}(\omega)}{\sum_{\omega}\hat{P}_{\mathrm{R}}(\omega)}$. Obviously the overall DRR estimation process is deterministic, however since audio signals observed in a practical indoor environment are most often statistical, such deterministic approach sometimes does not provide an accurate estimate of the DRR. For minimising the variance in the estimation errors of the DRR, the proposed method introduces a statistical approach that optimises the mapping between the estimated PSD and targeted DRR.

Among various mapping methods available, the proposed method employs the DNN, a state-of-the-art method that has been

attracting interests in many engineering fields recently. With the proposed method, the DNN is expected to provide an accurate DRR by taking the estimated PSD of the direct sound and reverberation as the *features* and mapping them to the DRR.

Let the following feature vector consisting of the estimated PSD at different frequencies be set to the input layer of a DNN with N-layers

$$\mathbf{q}^{(1)} = [\hat{P}_{\mathrm{D}}(\omega_1), \ldots, \hat{P}_{\mathrm{D}}(\omega_O), \hat{\overline{P}}_{\mathrm{R}}(\omega_1), \ldots, \hat{\overline{P}}_{\mathrm{R}}(\omega_O)]^{\mathrm{T}}, \quad (14)$$

where ω_O is the number of frequency bins where the PSD is available. Given that the network parameter $\mathbf{z}$ includes the weights $\mathbf{Z}^{(2)}, \ldots, \mathbf{Z}^{(N)}$ and the biases $\mathbf{b}^{(2)}, \ldots, \mathbf{b}^{(N)}$, $\mathbf{u}^{(n)}$ and $\mathbf{q}^{(n)}$ are calculated by a recursive update for $N-1$ times expressed by

$$\mathbf{u}^{(n)} = \mathbf{Z}^{(n)}\mathbf{q}^{(n-1)} + \mathbf{b}^{(n)}, \quad (15)$$

$$\mathbf{q}^{(n)} = \mathbf{f}^{(n)}\left(\mathbf{u}^{(n)}\right). \quad (16)$$

Provided that the number of nodes in the n-th layers is denoted as J_n, these parameters are represented by vector forms

$$\mathbf{u}^{(n)} = \left[u_1^{(n)}, \ldots, u_{J_n}^{(n)}\right]^{\mathrm{T}}, \quad (17)$$

$$\mathbf{q}^{(n)} = \left[q_1^{(n)}, \ldots, q_{J_n}^{(n)}\right]^{\mathrm{T}}, \quad (18)$$

$$\mathbf{Z}^{(n)} = \begin{bmatrix} Z_{1,1}^{(n)} & \cdots & Z_{1,J_{n-1}}^{(n)} \\ \vdots & \ddots & \vdots \\ Z_{J_n,1}^{(n)} & \cdots & Z_{J_n,J_{n-1}}^{(n)} \end{bmatrix}, \quad (19)$$

$$\mathbf{b}^{(n)} = \left[b_1^{(n)}, \ldots, b_{J_n}^{(n)}\right]^{\mathrm{T}}, \quad (20)$$

$$\mathbf{f}^{(n)}\left(\mathbf{u}^{(n)}\right) = \left[f^{(n)}\left(u_1^{(n)}\right), \ldots, f^{(n)}\left(u_{J_n}^{(n)}\right)\right]^{\mathrm{T}}. \quad (21)$$

For the activation function $f^{(n)}(\cdot)$, either a sigmoid function or an identity mapping function is chosen depending on the layer.

$$f(u) = \begin{cases} 1/(1+\exp(-u)) & (n = 2, \ldots, N-1) \\ u & (n = N) \end{cases} \quad (22)$$

Given that the number of nodes in the N-th layer is 1, the estimated DRR is expressed by

$$\Gamma = q_1^{(N)}. \quad (23)$$

In the rest of this paper the estimated DRR using $\mathbf{z}$ is denoted as $\Gamma(\mathbf{q}^{(1)}; \mathbf{z})$.

Like other statistical mapping algorithms, the performance of the DNN is greatly affected by the initial values of the network parameters. Thanks to the recent progress in research, a set of appropriate initial values can be found by pre-training based on the deep belief network (DBN) [17]. In the DBN, the initial values of the network parameters are estimated layer-by-layer using stacked restricted Boltzmann machines (RBMs). In this study, the contrastive divergence (CD) [15, 18] is utilised to specify an appropriate update amount for the network parameters of each RBM.

Once the initial values are given to the network parameters, these parameters are then optimised by the back propagation [19] in order to minimise the estimation error of the DRR. Assume K PSD samples with its true DRR used as the supervisory signals are denoted as

$$\{(\mathbf{q}_1^{(1)}, \tilde{\Gamma}_1), (\mathbf{q}_2^{(1)}, \tilde{\Gamma}_2), \ldots, (\mathbf{q}_K^{(1)}, \tilde{\Gamma}_K)\}. \quad (24)$$

where $\tilde{}$ denotes the true value. The procedures in (15) and (16) applied to K samples can be represented by a matrix form given by

$$\mathbf{U}^{(n)} = \mathbf{Z}^{(n)}\mathbf{Q}^{(n-1)} + \mathbf{b}^{(n)}\mathbf{1}_K^{\mathrm{T}}, \quad (25)$$

$$\mathbf{Q}^{(n)} = \mathbf{f}^{(n)}(\mathbf{U}^{(n)}), \quad (26)$$

where

$$\mathbf{U}^{(n)} = [\mathbf{u}_1^{(n)}, \ldots, \mathbf{u}_K^{(n)}], \quad (27)$$

$$\mathbf{Q}^{(n)} = [\mathbf{q}_1^{(n)}, \ldots, \mathbf{q}_K^{(n)}]. \quad (28)$$

A mean square error is used to measure the difference between the true and estimated DRRs.

$$E(\mathbf{z}) = \frac{1}{2}\sum_{k=1}^{K} \parallel \tilde{\Gamma}_k - \Gamma(\mathbf{q}_k^{(1)}; \mathbf{z}) \parallel^2 \quad (29)$$

Using the back propagation, the gradient of the network parameters is recursively calculated from the output layer ($n = N$) towards the input layer ($n = 1$). Given $\tilde{\mathbf{\Gamma}} := [\tilde{\Gamma}_1, \ldots, \tilde{\Gamma}_K]$, the gradient at the n-th layer $\mathbf{\Delta}^{(n)}$ is derived by

$$\mathbf{\Delta}^{(n)} = \begin{cases} \mathbf{f}^{(n)\prime}(\mathbf{U}^{(n)}) \odot (\mathbf{Z}^{(n+1)\mathrm{T}}\mathbf{\Delta}^{(n+1)}) & (n = 2, \cdots, N-1) \\ \tilde{\mathbf{\Gamma}} - \mathbf{Q}^{(n)} & (n = N). \end{cases} \quad (30)$$

Here, $\odot$ denotes an element-wise product of matrices. The gradient of the error functions is derived by

$$\partial\mathbf{Z}^{(n)} = \frac{1}{K}\mathbf{\Delta}^{(n)}\mathbf{Q}^{(n-1)\mathrm{T}}, \quad (31)$$

$$\partial\mathbf{b}^{(n)} = \frac{1}{K}\mathbf{\Delta}^{(n)}\mathbf{1}_K^{\mathrm{T}}, \quad (32)$$

Finally, the network parameters are updated by adding values derived from the gradient

$$\mathbf{Z}^{(n)} \leftarrow \mathbf{Z}^{(n)} + \Delta\mathbf{Z}^{(n)}, \quad (33)$$

$$\mathbf{b}^{(n)} \leftarrow \mathbf{b}^{(n)} + \Delta\mathbf{b}^{(n)}, \quad (34)$$

where the perturbations for each update are calculated by

$$\Delta\mathbf{Z}^{(n)} = \mu\Delta\mathbf{Z}^{(n)*} - \epsilon\left(\partial\mathbf{Z}^{(n)} + \lambda\mathbf{Z}^{(n)}\right), \quad (35)$$

$$\Delta\mathbf{b}^{(n)} = \mu\Delta\mathbf{b}^{(n)*} - \epsilon\partial\mathbf{b}^{(n)}. \quad (36)$$

Here, $\Delta\mathbf{Z}^{(n)*}$ and $\Delta\mathbf{b}^{(n)*}$ are the perturbations of the previous update, ϵ is the learning rate, μ and λ are the momentum coefficient and weight decay, respectively.

4. EXPERIMENTS

The performance of the proposed method was investigated by evaluating the accuracy of the estimated DRR and comparing it to the previous method [13] which does not rely on a statistical mapping method.

A corpus of impulse responses measured using a microphone array located in four rooms with different acoustical characteristics was employed in the experiment. The microphone array consisted of three unidirectional microphones; the orientation of each microphone differed from the others by 120 degrees. The sound source location was determined by the combination of the angle and distance

Table 1. Parameters used in experiment

Sampling rate	16 kHz
FFT length	32 ms
# of microphones, M	3
# of rooms	4 (rooms A, B, C, and D)
microphone array locations	centre, close to wall
angle of sources	5 (0, 45, 90, 135, 180 degrees)
distance of sources	6 (0.25, 0.50, 0.75, 1.0, 1.5, 2.0 m)
# of source signals	16 (training), 8 (evaluation)
# of background-noise types	3
# of SNR patterns	7 (20, 15, 10, 5, 0, -5, 10 dB)
# of signals for training, K	80640 (= $4 \cdot 2 \cdot 5 \cdot 6 \cdot 16 \cdot 3 \cdot 7$)
# of signals for evaluation	40320 (= $4 \cdot 2 \cdot 5 \cdot 6 \cdot 8 \cdot 3 \cdot 7$)
# of layers, N	4
# of nodes, J_n	J_1: 514, J_2: 640, J_3: 640, J_4: 1
Learning coefficient, ϵ	0.02, 0.01, 0.005, 0.0025, 0.0001
Iteration number	100 (for each ϵ)
Momentum coefficient, μ	0.5 (first 5), 0.9 (after 6)
Decay weight, λ	0.0002

from the microphone array. Five different angles and six different distances were available in the corpus.

Sentences spoken by male and female speakers for 8 sec were used as the sound source. Sixteen sentences were utilised for the training and eight sentences different from those used for the training were employed for the evaluation (i.e. open test). To simulate a microphone observation, one of the speech signals was convolved with the measured impulse response, then environmental noise recorded separately in the same rooms was added at different signal to noise ratios (SNRs). The environmental noise included three different types typically observed in practical environments: office, exhibition, and shopping centre. The true DRR used as a supervisor in the training process was calculated from an impulse response of the same room measured with an omni-directional microphone using the definition in [20] ranging from 1.5 up to 25.2 dB.

Two beamformers designed using the minimum variance distortionless response (MVDR) [21] were employed for estimating the PSD using (13). The angle of the direct sound Ω_{D} was estimated using the beamforming method [21]. The estimated PSD at each frequency bin was normalised to fit between 0 and 1 then was passed to the network parameter (14) for running the DNN. The network parameters were optimised by the back propagation after applying pre-training. The number of datasets used for the training was balanced across the parameters. The other parameters used in the experiment are summarised in Table 1. The accuracy of estimating the DRR was evaluated by calculating the estimation error of the k-th sample given by $\rho_k = \hat{\Gamma}_k - \tilde{\Gamma}_k$.

Figure 2 shows the distribution of the estimation error in the different rooms for different SNR. The previous method resulted in degraded estimation accuracy when the input SNR was decreased. In contrast, the proposed method exhibited steady performance even when the input signal was severely contaminated by noise. Since the proposed method trained the network parameter to minimise the squared errors across all samples, as in (29), it maintained its performance even if the effect of the noise was significant. The estimation error might have been increased had the number of datasets across SNR been unbalanced which needs a further investigation. Another noticeable trend was that the proposed method was less sensitive to the change in rooms. Overall, these findings imply the great potential of using a statistical mapping method for improving the accuracy of DRR estimation in various acoustic environments.

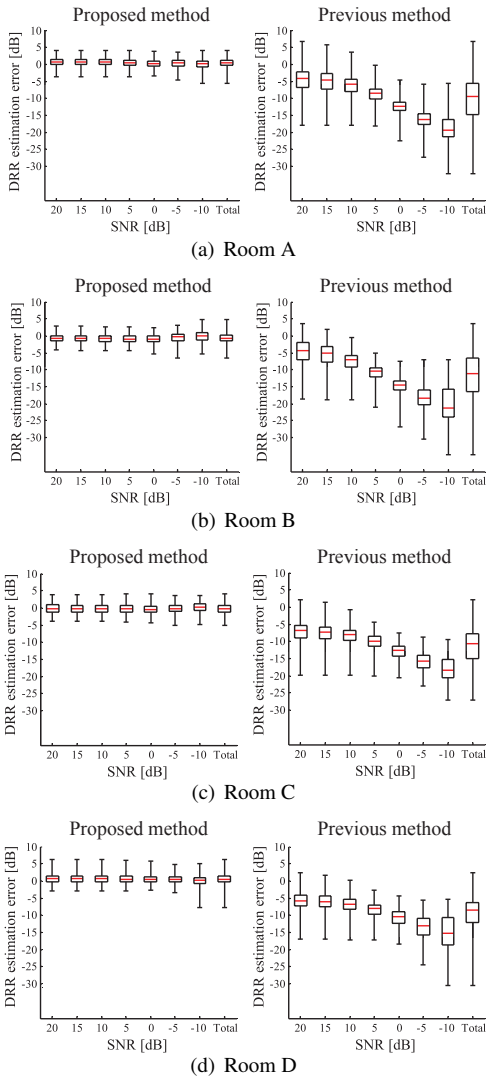

Fig. 2. Experimental results from evaluating accuracy of estimating DRR using proposed and previous methods in different rooms

5. CONCLUSION

This study has reported on a new attempt to use the DNN for improving the accuracy of the DRR estimation problem. The DNN is utilised to map an effective feature, such as the PSD, to the correct DRR to mitigate the errors included in the feature. Although the introduction of the DNN, which requires a training process, makes DRR estimation no longer a blind method, significant improvement in the estimation accuracy is convincing such that the use of the DNN in combination with an effective feature has the potential to provide better DRR estimation accuracy, provided the application allows the training process.

Further experiments using the ACE corpus [8] to investigate the performance of the proposed method and compare it to other methods would be a suggested future work.

6. REFERENCES

[1] Patrick A. Naylor and Nikolay D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[2] Y. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1793–1805, 2010.

[3] S. Vesa, "Sound source distance learning based on binaural signals," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2007, pp. 271–274.

[4] S. Vesa, "Binaural sound source distance learning in rooms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1498–1507, 2009.

[5] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio based on spatial correlation model segregating direct sound and reverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010)*, March 2010, pp. 149 –152.

[6] Y. Hioka, K. Niwa, S. Sakauchi, K. Furuya, and Y. Haneda, "Estimating direct-to-reverberant energy ratio using D/R spatial correlation matrix model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374 – 2384, November 2011.

[7] Mikko-Ville Laitinen and Ville Pulkki, "Utilizing instantaneous direct-to-reverberant ratio in parametric spatial audio coding," in *Audio Engineering Society Convention 133*, October 2012.

[8] J. Eaton, A. H. Moore, N. D. Gaubitch, and P. A. Naylor, "The ACE challenge - corpus description and performance evaluation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

[9] E. Larsen, C.D. Schmitz, C.R. Lansing, W.D. O'Brien, B.C. Wheeler, and A.S. Feng, "Acoustic scene analysis using estimated impulse responses," in *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2004.*, November 2003, vol. 1, pp. 725–729.

[10] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *EUSIPCO 2011*, September 2011.

[11] O. Thiergart, G.D Galdo, and E.A.P Habets, "Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional mirophones," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, March 2012, pp. 309 –312.

[12] Yusuke Hioka, Ken Furuya, Kenta Niwa, Yoichi Haneda, et al., "Estimation of direct-to-reverberation energy ratio based on isotropic and homogeneous propagation model," in *13th International Workshop on Acoustic Signal Enhancement (IWAENC 2012)*, September 2012, pp. 1–4.

[13] Y. Hioka and K. Niwa, "PSD estimation in beamspace for estimating direct-to-reverberant ratio from a reverberant speech signal," in *ACE Challenge Workshop, a satellite event of IEEE-WASPAA 2015*, October 2015.

[14] O. Thiergart, T. Ascherl, and E.A.P. Habets, "Power-based signal-to-diffuse ratio estimation using noisy directional microphones," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, May 2014, pp. 7440–7444.

[15] Yoshua Bengio, "Learning deep architectures for ai," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009.

[16] Don H. Johnson and Dan E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Simon & Schuster, 1992.

[17] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[18] Geoffrey E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, Aug. 2002.

[19] David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group, Eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, MIT Press, 1986.

[20] S. Mosayyebpour, H. Sheikhzadeh, T.A. Gulliver, and M. Esmaeili, "Single-microphone lp residual skewness-based inverse filtering of the room impulse response," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1617–1632, July 2012.

[21] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Digital Signal Processing - Springer-Verlag. Springer, 2001.