ESTIMATING DIRECT-TO-REVERBERANT ENERGY RATIO BASED ON SPATIAL CORRELATION MODEL SEGREGATING DIRECT SOUND AND REVERBERATION

Yusuke Hioka, Kenta Niwa, Sumitaka Sakauchi, Ken’ichi Furuya, and Yoichi Haneda

NTT Cyber Space Laboratories, NTT Corporation, Tokyo, Japan

ABSTRACT

A new approach for estimating the direct-to-reverberant energy ratio (DRR) using a microphone array is proposed. The method is based on a model of a spatial correlation matrix that segregates direct sound and reverberation. It estimates DRR from the power spectra of both components, which are derived from the correlation matrix of the observed signal. In experiments performed in simulated and actual reverberant environments, the proposed method mostly succeeded in estimating DRR accurately. We also present speech enhancement using binary masking as an example of an application of the estimated DRR. By utilization of the DRR as a factor to discriminate the distances of speakers, separation of speech signals whose sources were located in the same direction but at different distances was achieved.

Index Terms—microphone array, direct-to-reverberant energy ratio, spatial correlation matrix, sound source distance, speech enhancement

1. INTRODUCTION

Estimating the direct-to-reverberant energy ratio (DRR) is quite helpful for determining the features of a reverberant environment because various acoustic parameters, such as reverberation time, diffuseness, etc., can be calculated from DRR [1]. There is also another important aspect in DRR relating to human hearing. Recent research on human hearing has concluded that DRR may provide absolute information, especially in reverberant environments [2].

Several methods are available for estimating DRR. The most primitive way is to calculate DRR directly from the impulse response. However, this is a complicated process because measurement of the impulse response is required. Larsen et al. proposed a method for estimating DRR from simply the short beginning part of the impulse response [3], but it still necessitates prior processing to identify the initial part of the impulse response. Lu et al. also proposed a procedure to estimate DRR. They utilized a binaural input signal and estimated the energy of the reverberant component by eliminating the direct component using an equalization-cancellation (EC) technique [4]. To eliminate direct sound, the EC technique exploits the fact that a large difference between direct sound and reverberation exists in the inter-channel (or spatial) correlation of the binary input signal.

As we mentioned above, knowing DRR is beneficial for discriminating the distances of sound sources because DRR has a one-to-one relation between the sound source distances. Generally, a microphone array is able to handle the spatial characteristics of a sound source. Although several conventional works that use microphone arrays try to emphasize sound sources located at particular distances from the microphones [5][6][7], they basically require a large-aperture microphone array, which is not convenient for implementation. Furthermore, the performances of those works is severely degraded if the environment is highly reverberant.

Our method proposed in this paper estimates DRR by calculating the power spectra of both direct and reverberant components. First, we introduce a model of a spatial correlation matrix that segregates direct and reverberant components. Then, the power spectra of respective components are derived from the spatial correlation matrix of the observed signals. The conventional method [4] focuses on the spatial correlation of only direct sound, whereas the proposed method also focuses on the spatial correlation of reverberation. Furthermore, it is worth applying DRR estimated by the proposed method to speech enhancement located at particular distances due to the proposed method assuming a highly reverberant (or diffuse) condition and the sufficiently small-sized microphone array in the modelling of the spatial correlation matrix.

This paper is organized as follows. In Sec. 2, we first introduce a model of a spatial correlation matrix that segregates the direct and reverberant components and then propose a method for estimating DRR on the basis of this model. From the results of experiments performed in both simulated and actual reverberant environments, we evaluate the performance of the proposed DRR estimation. In Sec. 3, we refer to the results of speech enhancement using binary masking as an example of the application of the estimated DRR. By utilization of the DRR as a factor to discriminate the distances of speakers, separation of speech signals whose sources are located in the same direction but at different distances was achieved. Finally, in Sec. 4, we present our concluding comments.

2. DRR ESTIMATION BASED ON SPATIAL CORRELATION MATRIX SEGREGATING DIRECT SOUND AND REVERBERATION

2.1. Modelling of spatial correlation matrix

First, we decompose the transfer function $H(\omega)$ between a sound source and a microphone into two components, direct component $H_D(\omega)$ and reverberant component $H_R(\omega)$, as described in Fig. 1. Note that the early reflection of the impulse response is also included in $H_R(\omega)$. When we have an $M$-sensors microphone array, the input signal of the $m$-th microphone expressed in the time-frequency

$$H_\omega(\omega)$$

$$H_D(\omega)$$

$$H_R(\omega)$$

Fig. 1. Decomposition of transfer function
domain is given by
\[ X^{(m)}(\omega, t) = \left( \hat{H}^{(m)}_{D}(\omega) + \hat{H}^{(m)}_{R}(\omega) \right) S(\omega, t), \]
where \( t \) denotes the temporal frame index. By this expression, the cross correlation between the \( i \)-th and \( j \)-th microphones is derived as
\[
E[X^{(i)}(\omega, t)X^{(j)*}(\omega, t)] = E\left[|S(\omega, t)|^2 \left( \hat{H}^{(i)}_{D}(\omega)\hat{H}^{(j)*}_{D}(\omega) + \hat{H}^{(i)}_{R}(\omega)\hat{H}^{(j)*}_{R}(\omega) \right) \right],
\]
where \( E[\cdot] \) denotes the expectation. Now under the assumption that the reverberant component is diffuse and the cross correlation between the direct and reverberant components (the third and fourth terms on the right side of Eq. (2)) is sufficiently small, the spatial correlation matrix of the microphone array \( \mathbf{R}(\omega) \) can be approximated by two matrices given by
\[
\mathbf{R}(\omega) = E[\mathbf{X}(\omega, t)\mathbf{X}^H(\omega, t)] \approx \mathbf{P}_D(\omega) + \mathbf{P}_H(\omega)
\]
where
\[
\mathbf{X}(\omega, t) = \begin{bmatrix} X^{(1)}(\omega, t) & X^{(2)}(\omega, t) & \cdots & X^{(M)}(\omega, t) \end{bmatrix}^T,
\]
\[
d_{ij} = \exp\left(j\omega\frac{|r_i - r_j|}{c}\right),
\]
\[
r_{ij} = \frac{\sin\left(\frac{\omega}{c}|r_i - r_j|\right)}{|r_i - r_j|},
\]
and \( r_m \) and \( c \) are the coordinates of the \( m \)-th microphone and sound speed, respectively. Furthermore, \( \mathbf{a}(\theta) = [\sin \theta, \cos \theta]^T \) is the look-direction unit vector of sound that propagates from \( \theta \) when the \( y \)-axis is set to \( 0 \) deg as described in Fig. 2.

On the right side of Eq. (3), the first term expresses the spatial correlation of the direct component. As there exists a time difference of arrival between microphones in the cross correlation of direct sound, the spatial correlation is expressed by simple phase difference. In the modelling of the second term, we utilized the feature that the spatial correlation of diffuse sound can be expressed by a sinc function [8]. In Eq. (3), \( \mathbf{P}_D(\omega) \) and \( \mathbf{P}_H(\omega) \) are defined by
\[
\mathbf{P}_D(\omega) = E[|S(\omega, t)|^2|\hat{H}_D(\omega)|^2],
\]
\[
\mathbf{P}_H(\omega) = E[|S(\omega, t)|^2|\hat{H}_R(\omega)|^2],
\]
where \( E[\cdot] \), \( H^* \), and \( * \) denote the expectation regarding the frame, Hermitian transform, and complex conjugate, respectively. Note that in the derivation of Eq. (3), the aperture size of the microphone array is assumed to be sufficiently small. This means that the array recognizes the received sound as a plain wave and that the magnitude of the transfer function for each microphone can be considered as identical, i.e., \(|\hat{H}^{(i)}_{D}(\omega)||\hat{H}^{(j)*}_{D}(\omega)| = |\hat{H}^{(i)}_{D}(\omega)|^2 \) and \(|\hat{H}^{(i)}_{R}(\omega)||\hat{H}^{(j)*}_{R}(\omega)| = |\hat{H}^{(i)}_{R}(\omega)|^2 \).

2.2. DRR estimation using power spectra of direct and reverberant components

As the microphone array configuration is generally known \textit{a priori}, and the direction of the sound source can be estimated by various conventional methods [9], \( d_{ij} \) and \( r_{ij} \) in Eq. (3) can be specified. Thus, we estimate the unknown power spectra of both direct and reverberant components, \( \mathbf{P}_D(\omega) \) and \( \mathbf{P}_H(\omega) \), by solving the simultaneous equation given by Eq. (7), which is derived by reformulating Eq. (3).

\[
\begin{bmatrix} 1 & 1 & \cdots & d_{1M} \\ d_{12} & r_{12} & \cdots & r_{1M} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M1} & d_{M2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} P_D(\omega) \\ P_R(\omega) \end{bmatrix} = \begin{bmatrix} R_{11}(\omega) \\ R_{12}(\omega) \\ \vdots \end{bmatrix}
\]

Here \( R_{ij}(\omega) \) in \( \mathbf{R}(\omega) \) denotes the \( i \)-th row and \( j \)-th column components of \( \mathbf{R}(\omega) \), which can be calculated from observed signals. The estimated power spectra of direct and reverberant components are given by solving Eq. (7) using the least-square method given by
\[
\mathbf{P}(\omega) = \mathbf{F}^+(\omega)\mathbf{R}(\omega),
\]
where \( \mathbf{F}^+ \) and \( \mathbf{R} \) are the Moore-Penrose pseudo inverse and estimated value, respectively.

Finally, the estimated DRR is given by using the estimated power spectra \( \mathbf{P}_D(\omega) \) and \( \mathbf{P}_H(\omega) \) in the following Eq. (9).
\[
\text{DRR}_{\text{estimate}} = 10 \log_{10} \left( \frac{\sum_{\omega} \sqrt{P_D(\omega)}}{\sum_{\omega} P_H(\omega)} \right)
\]
2.3. Performance evaluation of DRR estimation

2.3.1. Simulation results

To confirm the effectiveness of the proposed DRR estimation, we performed experiments in simulated reverberant environments whose settings are described in Fig. 3. The sound source was 3-s long Gaussian white noise, and the input signals of an octagonal microphone array were prepared by convolving the simulated impulse response generated by the image method [10]. The estimated DRR when the absorption coefficient, which changes according to the reverberation time of the room, changes is shown in Figs. 4 and 5. For each distance, we applied the DRR estimation for 100 different sound sources. For comparison, DRR directly calculated from the impulse response (DRR_{actual}) defined by

\[
DRR_{actual} = 10 \log_{10} \left( \frac{\sum_{\omega} |H_D(\omega)|^2}{\sum_{\omega} |H_R(\omega)|^2} \right)
\]  

is shown in the figures. When the estimated DRR is identical to DRR_{actual}, the proposed method is considered to be successful in estimating DRR. Furthermore, principle-based DRR in a diffuse sound field (DRR_{diffuse}), given by [8]

\[
DRR_{diffuse} = 10 \log_{10} \left( \frac{S \alpha}{16 \pi D^2} \right),
\]

is also calculated to determine the distance of the actual reverberation from the completely diffuse field. Here, \(S\) and \(\alpha\) are the surface area of the walls and average absorption coefficient, respectively. The lower graphs of Figs. 4 and 5 show the absolute estimation errors of estimated DRR from DRR_{actual}.

The errors, which rapidly increase as the distance from the sound source to the microphone array decreases, are caused by the discrepancy between the modelled and actual spatial correlation of the direct component. This is because the plane wave assumption of the received sound is only valid to the sound sources located in the far-field defined by \(d > \frac{\lambda}{4}\) [11], where \(D\) is the array aperture size. In the case of the microphone array used in this simulation (\(D = 12\) cm), the boundary distance between far-field and near-field was approximately 33 cm at 8 kHz. This is supported by the results obtained where the estimation error started to rapidly increase, when \(d\) became smaller than 30 cm.

Another increase of estimation error found at the distant positions can be considered as the influence of the discrepancy between the modelled and actual spatial correlation of the reverberant component. Although the reverberant component was assumed to be diffuse in the modelling, some early reflections included in the transfer function do not meet this assumption. As the difference between the early reflection and reverberation becomes obvious at distant positions, such modelling error could be conceived as a cause of the estimation errors. Furthermore, the power of the direct component being too small to be accurately detected from the noisy observed signal is another considerable cause of error.

Due to DRR_{actual} shows a completely different value from DRR_{diffuse}, the condition applied to the simulation in Fig. 5 is no longer diffuse. Although the model of a spatial correlation matrix (especially the matrix for the reverberant component) could be invalid in such a semi-anechoic environment, the proposed method still succeeded in estimating correct DRR when the sound source was reasonably distant from the microphone array.

2.3.2. Experimental results

To confirm the effectiveness in an actual environment, we also performed an experiment in a real reverberant room. The room size and position of the microphone array used in this experiment were the same as those given in Fig. 3, except that the loudspeaker was located in the direction of \(\theta = 0^\circ\). The reverberation time of the room was approximately 400 ms. The estimated DRR and DRR_{actual} calculated from the measured impulse response are shown in Fig. 6. The results prove that the proposed method is still effective in actual
reverberant environments.

3. APPLICATION BASED ON ESTIMATED DRR

3.1. Speech emphasis depending on sources’ distances

As an example of an application based on the proposed DRR estimation, we present a speech enhancement based on binary masking [12], which discriminates the distances of speakers using the estimated DRR. Although the binary masking exploits the feature of the power spectrum of a speech signal showing sparseness in the time-frequency domain, the average calculation of the estimated power spectra for a whole frequency band in Eq. (9) does not consider this sparseness feature. To avoid this problem, we modify the calculation of DRR to derive the DRR of a particular frequency band by the weighted average of an estimated power spectra, given by

$$
    \text{DRR}_{\text{estimate}}(\omega) = 10 \log_{10} \left( \frac{\sum_{k=-\infty}^{\infty} \omega \cdot k \cdot P_k(\omega)}{\sum_{k=-\infty}^{\infty} \omega \cdot k \cdot P_k(k)} \right). \tag{12}
$$

Then, the binary mask $G(\omega, t)$ is determined by comparing the estimated DRR with a threshold $T$ given a priori. For example, the binary mask for emphasizing the sound source within a particular distance is given by

$$
    G(\omega, t) = \begin{cases} 
    1 & \text{if } \text{DRR}_{\text{estimate}}(\omega) > T \\
    0 & \text{else} \end{cases}. \tag{13}
$$

Finally, the desired speech $\overline{S}(\omega, t)$ is emphasized by multiplying $G(\omega, t)$ by the microphone input signal $X(\omega, t)$ given by

$$
    \overline{S}(\omega, t) = G(\omega, t) \cdot X(\omega, t). \tag{14}
$$

3.2. Simulation results

The signal waveforms when the speech enhancement was performed in the simulated environment described in Fig. 3 is shown in Fig. 7. In this experiment, a female speaker and a male speaker were located at $d = 0.3$ m and $d = 4.0$ m, respectively, and Eq. (13) was used to determine the binary mask. For the parameters, $T$ and $\Omega$ were set as 0 dB and 3 bins, respectively, and 7-taps hanning window was used for the weight $w_k$. The interference speech was sufficiently suppressed where the improvement of the signal-to-interference ratio was $8.1$ dB while the signal-to-distortion ratio [13] of the output signal was $7.9$ dB. Thus, we have confirmed that the estimated DRR can be utilized as an index to discriminate the sound source distances.

4. CONCLUDING REMARKS

We have proposed a method for estimating DRR using a model of a spatial correlation matrix that segregates direct and reverberant components. From the experimental results performed both in simulated and actual reverberant environments, the proposed method showed good performance in estimating DRR.

We have also shown the results of speech enhancement based on binary masking as an example of an application that utilizes the estimated DRR. Generally, conventional microphone array techniques require sufficiently large aperture size to discriminate the distances of sound sources. However, the proposed method succeeded in separating speech located in the same direction but at different distances without requiring large aperture sizes. Such advantage is due to the proposed method’s ability to estimate DRR accurately using a small-sized microphone array.

5. REFERENCES


