PSD ESTIMATION IN BEAMSPACE FOR ESTIMATING DIRECT-TO-REVERBERANT RATIO FROM A REVERBERANT SPEECH SIGNAL

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 method for estimation of direct-to-reverberant ratio (DRR) using a microphone array is proposed. The proposed method estimates the power spectral density (PSD) of the direct sound and the reverberation using the algorithm PSD estimation in beamspace with a microphone array and calculates the DRR of the observed signal. The speech corpus of the ACE (Acoustic Characterisation of Environments) Challenge was utilised for evaluating the practical feasibility of the proposed method. The experimental results revealed that the proposed method was able to effectively estimate the DRR from a recording of a reverberant speech signal which included various environmental noise.

Index Terms— direct-to-reverberant ratio, microphone array, speech, power spectral density, beamspace, ACE Challenge

1. INTRODUCTION
Estimation of direct-to-reverberant ratio (DRR) has been attracting interests of researchers in audio and acoustic signal processing since the DRR is one of the important parameters in room acoustics which specifies the acoustic characteristics of a reverberant enclosure [1]. The estimated DRR can also be utilised in various audio signal processing applications such as dereverberation [2], source distance estimation [3, 4, 5, 6], emphasising speech located at a particular distance [7], and spatial audio coding [8].

The room impulse response (RIR) used to be measured to calculate the DRR of a reverberant enclosure. However the measurement of the RIR was actually a hindrance for application users since special equipment and software are needed, which motivated researchers to estimate the DRR in blind manner, i.e. without RIR measurement. Since the propagation properties of the direct sound and the reverberation are quite different, several recent studies utilised more than one microphones (i.e. a microphone array) for the blind DRR estimation problem [9]. Some recent studies achieved blind DRR estimation by introducing a model for the coherence of direct sound and reverberation between two microphones [4, 5, 10, 11] since the magnitude-squared coherence (MSC) is a quantity related to the DRR [4, 5]. The authors also proposed a method based on a similar strategy but using more than two microphones [6]. The method introduced a model for the spatial correlation matrix of a microphone array signal recorded in a reverberant environment which contains the coherence information between each pair of microphones. The method estimated the power spectral density (PSD) [12] of the direct sound and the reverberation simultaneously using the introduced model.

Besides the introduction of the model another novel aspect in this method was in the calculation of the DRR from the estimated PSD of the direct sound and the reverberation. The authors later proposed another DRR estimation method which estimates the PSD using two different beamformers whose beampatterns have an identical shape [13]. It should be noted that these previous methods are semi-blind methods because they assume the direction of the sound source is known a priori. Another similar approach proposed by Thiergart et al. also estimates the PSD for deriving the DRR [14], which uses multiple directional microphones rather than a microphone array.

In this study another PSD based blind DRR estimation method is proposed. The proposed method utilises an algorithm: PSD estimation in beamspace [15, 16], which was originally invented for sound source separation problems by the authors, to estimate the PSD of the direct sound and the reverberation. The method was evaluated on the speech corpus of the ACE (Acoustic Characterisation of Environments) Challenge [17] and its performance was compared to that of the authors’ previous studies [6, 13]. Since the problem specified by the ACE Challenge was a fully blind problem (i.e. no prior information about the source direction is provided), the methods were tested after being combined with conventional direction-of-arrival (DOA) estimation and voice activity detection (VAD) methods. Presenting the performance of the authors’ methods under a fully blind problem is another contribution of this paper.

2. PROBLEM SETUP

2.1. Sound propagation in reverberant environment
An impulse response from a sound source to a microphone (i.e. the sound propagation path) in a reverberant environment is generally described by a model which consists of the direct sound, early reflections, and late reverberation [1]. However as shown in Fig. 1 in this study the impulse response is simply modelled by the two components, the direct sound and the reverberation, where the latter includes both the early reflections and the late reverberation. Given the impulse response in the frequency domain (i.e. the transfer function) is denoted by \( H(\omega) \) where \( \omega \) is the frequency, the model is represented by

\[
H(\omega) = H_D(\omega) + H_R(\omega),
\]

(1)
where $H_D(\omega)$ and $H_R(\omega)$ are the transfer functions of the direct sound and the reverberation, respectively. The sound propagation paths of these two components are quite different. In general the direct sound is assumed to be spatially coherent, whereas the reverberation is assumed to be diffuse [1]. In other words, as described in Fig. 2, the direct sound arrives into a microphone directly only from the source direction (directional). In contrast, the reverberation arrives from every direction with uniform energy distribution (isotropic). The proposed method focuses on the difference in these spatial properties of the sound propagation paths and segregate the direct sound and the reverberation using a microphone array. The rest of this section discusses the signal modelling for the microphone array observation and the output of the beamforming [18], which is used to estimate the PSD of the direct sound and the reverberation for calculating the DRR.

In the following discussion a few assumptions are made: the source direction from the microphone array is known a priori using any conventional direction of arrival (DOA) estimation techniques [19]; the direct sound and the reverberation are mutually uncorrelated; and both direct sound and the reverberation are modelled by plane waves [18].

### 2.2. Microphone array observation

Let $X^{(m)}(\omega,t)$ be the observed signal of the $m$-th microphone of an $M$-sensors microphone array in the time-frequency domain where $t$ is a frame index. Using the transfer function of the sound propagation path introduced in (1), $X^{(m)}(\omega,t)$ can be modelled by

$$X^{(m)}(\omega,t) := \left(H_D^{(m)}(\omega) + H_R^{(m)}(\omega)\right)S(\omega,t),$$

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

The transfer functions in (2) are further decomposed into two components: the transfer function from the sound source to a reference point located close to the microphone array (e.g. the center of the microphone array), and that from the reference point to each microphone. Since the signals are assumed to be plane waves, the latter transfer function can be approximated by the phase shift caused by the propagation delay. Namely $H_D^{(m)}(\omega)$ and $H_R^{(m)}(\omega)$ are expressed by

$$H_D^{(m)}(\omega) = H_{\text{Dref}}(\omega) e^{-j\omega \tau_D^{(m)}},$$

$$H_R^{(m)}(\omega) = \int_{\Omega} H_{\text{Rref},\Omega}(\omega) e^{-j\omega \tau_R^{(m)}(\omega)} |d\Omega|,$$

where $H_{\text{Dref}}(\omega)$ and $H_{\text{Rref},\Omega}(\omega)$ are the transfer functions from the sound source to the reference point with regard to the direct sound and the reverberation, respectively. $\tau_R^{(m)}(\omega)$ is the propagation delay from the reference point to the microphone $m$ when the sound wave is arriving from a solid angle $\Omega = \{\theta,\phi\}$, where $\theta$ and $\phi$ are the azimuth and the zenith angles, respectively ($\theta \in [0,2\pi)$, $\phi \in [0,\pi]$). Note that $\phi = \pi$ is equal to the plane in parallel to the plane of the microphone array, and $\Omega_D$ denotes the source direction.

The observation vector of the microphone array is defined by

$$x(\omega,t) = [X^{(1)}(\omega,t), \ldots, X^{(M)}(\omega,t)]^T$$

$$= a_{\Omega_D}(\omega) S_D(\omega,t) + \int_{\Omega} a_{\Omega}(\omega) S_{R,\Omega}(\omega,t) |d\Omega|, \quad (5)$$

where $a_{\Omega}(\omega) = [e^{-j\omega \tau_D^{(1)}}, \ldots, e^{-j\omega \tau_D^{(M)}}]^T$ is the steering vector [18] for the angle $\Omega$, and

$$S_D(\omega,t) = H_{\text{Dref}}(\omega) S(\omega,t),$$

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

are the direct sound and the reverberation arriving from the angle $\Omega$ observed at the reference point, respectively. $T$ denotes the transpose of a vector or a matrix.

### 2.3. Beamforming output

Assume more than one arbitrary but different beamformers are applied to the microphone array observation $x(\omega,t)$, then the output signal of the beamformer $l$ is represented by

$$y_{BF,l}(\omega) = w_l^H(\omega) x(\omega,t)$$

$$= w_l^H(\omega) a_{\Omega_D} S_D(\omega,t) + w_l^H(\omega) \int_{\Omega} a_{\Omega} S_{R,\Omega}(\omega,t) |d\Omega|,$$

where $w_l(\omega)$ is the weight vector of the beamformer $l$ defined by

$$w_l(\omega) = [W_l^{(1)}(\omega), \ldots, W_l^{(M)}(\omega)]^T.$$ \quad (10)

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

$$P_{BF,l}(\omega) = E[|y_{BF,l}(\omega)|^2]$$

$$ \approx w_l^H(\omega) a_{\Omega_D}(\omega) E[|S_D(\omega,t)|^2] a_{\Omega_D}^H(\omega) w_l(\omega)$$

$$+ w_l^H(\omega) \left\{ \int_{\Omega} a_{\Omega}(\omega) E[|S_{R,\Omega}(\omega,t)|^2] |d\Omega| \right\} a_{\Omega}^H(\omega) w_l(\omega)$$

$$= G_{l,\Omega_D}(\omega) P_D(\omega) + \int_{\Omega} G_{l,\Omega}(\omega) P_{R,\Omega}(\omega) |d\Omega|,$$ \quad (13)
Because the elements in solving the simultaneous equation

\[ P_D(\omega) = E[|S_D(\omega, t)|^2]_t, \]

\[ P_{R,\Omega}(\omega) = E[|S_{R,\Omega}(\omega, t)|^2]_t. \]

\(E[\cdot]_t\) is the expectation over frames that can be approximated by the average of several frames, and \(G_{r,\Omega}(\omega)\) is the gain of the beamformer for angle \(\Omega\) defined by

\[ G_{r,\Omega}(\omega) = |w_H^H(\omega)\mathbf{A}_0(\omega)|^2. \]

In the derivation of (13), the uncorrelatedness assumption for the relation between the direct sound and the reverberation, i.e. \(E[S_D^H(\omega, t)S_{R,\Omega}(\omega, t)]_t = 0\), is utilised.

Since the isotropic propagation has been assumed for the reverberation, the PSD of the reverberation can be replaced by a constant value that holds for all \(\Omega\), i.e.

\[ P_{R,\Omega}(\omega) = \overline{P_R}(\omega) = \text{const.} \quad \forall \Omega. \]

Thus, the PSD of the beamformer output in (13) becomes

\[ P_{BF,\Omega}(\omega) = G_{r,\Omega}(\omega)P_D(\omega) + \overline{P_R}(\omega) \int_{\Omega} G_{r,\Omega}(\omega)d\Omega. \]

### 2.4. Calculating DRR from estimated PSD

The DRR is commonly defined by the energy ratio of the impulse response for the direct sound to that for the reverberation [9]. According to Parseval’s theorem and with some approximations, the DRR can also be calculated from the PSD of the direct sound and the reverberation observed at the reference point of the microphone array

\[ \Gamma[\text{dB}] := 10 \log_{10} \left( \frac{\sum_{\omega} |H_D(\omega)|^2}{\sum_{\omega} |H_{R,\Omega}(\omega)|^2} \right) \]

\[ \approx 10 \log_{10} \left( \frac{\sum_{\omega} P_D(\omega)}{\sum_{\omega} \int_{\Omega} P_{R,\Omega}(\omega)d\Omega} \right) \]

\[ = 10 \log_{10} \left( \frac{\sum_{\omega} P_D(\omega)}{4\pi \sum_{\omega} \overline{P_R}(\omega)} \right). \]

In the following section an algorithm to estimate the PSD, i.e. \( P_D(\omega) \) and \( \overline{P_R}(\omega) \), is introduced.

3. PSD ESTIMATION IN BEAMSPACE

Assume two beamformers which have different directivity patterns are applied to the microphone array observation as shown in Fig. 3. According to (18) the output PSD of the two beamformers can be formulated in a matrix form given by (22).

\[
\begin{bmatrix}
    P_{BF,\Omega}(\omega) \\
    P_{BF,\Omega}(\omega)
\end{bmatrix} =
\begin{bmatrix}
    G_{1,\Omega}(\omega) & \int_{\Omega} G_{1,\Omega}(\omega)d\Omega \\
    G_{2,\Omega}(\omega) & \int_{\Omega} G_{2,\Omega}(\omega)d\Omega
\end{bmatrix}
\begin{bmatrix}
    P_D(\omega) \\
    P_{R}(\omega)
\end{bmatrix} =
\begin{bmatrix}
    \hat{P}_D(\omega) \\
    \overline{P_R}(\omega)
\end{bmatrix}.
\]

Because the elements in \( P_{BF}(\omega) \) and \( G(\omega) \) are known \( a \) \( priori \), the PSD of the direct sound and the reverberation can be estimated by solving the simultaneous equation

\[ \hat{P}_{comp}(\omega) = G^{-1}(\omega) \hat{P}_{BF}(\omega), \]

\[ \hat{P}_{comp}(\omega) = G^{-1}(\omega) \hat{P}_{BF}(\omega), \]

where \( \hat{\cdot} \) denotes an estimated value. The DRR of each sound source is estimated by substituting \( P_D(\omega) \) and \( \overline{P_R}(\omega) \) in (21) by the estimated values found in \( \hat{P}_{comp}(\omega) \).

The previous method invented by the authors [13] also uses two beamformers however their beampatterns have to be identical while their mainlobes are pointing in different directions. It is obvious that the previous method is solving (22) given

\[ \int_{\Omega} G_{1,\Omega}(\omega)d\Omega = \int_{\Omega} G_{2,\Omega}(\omega)d\Omega := \int_{\Omega} G_{\Omega}(\omega)d\Omega \]

\[ G_{1,\Omega}(\omega) \neq G_{2,\Omega}(\omega), \]

which is a specific case of PSD estimation in beamspace.

Then the PSD of the direct sound is estimated from the difference of the output PSD of these beamformers

\[ \hat{P}_D(\omega) = \frac{P_1(\omega) - P_2(\omega)}{G_{1,\Omega}(\omega) - G_{2,\Omega}(\omega)}, \]

whereas the PSD of the reverberation is calculated by the following subtraction of the estimated \( P_D(\omega) \) from the PSD of the microphone observation:

\[ \hat{P}_R(\omega) = P^{(m)}_X(\omega) - \hat{P}_D(\omega), \]

where \( P^{(m)}_X(\omega) \) is derived by

\[ P^{(m)}_X(\omega) = E[|X^{(m)}(\omega, t)|^2]_t. \]

4. PERFORMANCE EVALUATION

The performance of the proposed method was evaluated along with that of the authors’ previous methods [6, 13] using the ACE Challenge corpus [17]. Among various microphone array configurations provided in the corpus, the 3-channel mobile microphone array (right-angled triangular configuration) was selected for the evaluation. The DOA of the direct sound \( \Omega_D = \{\theta_D, \phi_D\} \) was estimated by the steered beamformer based method [19] with a delay-and-sum beamforming [18] used for the beamformer. Since speech signals are generally nonstationary, a voice activity detection (VAD) was applied for extracting frames which include reasonable amount of speech components, then only the extracted frames were used for calculating the expectation in the PSD estimation. The VAD was realised by simply selecting the frames of the short time Fourier transform of a microphone observation where the power of the signal in the frames exceeded a certain threshold, which was determined based on the level of stationary noise estimated by [20].

Figure 3: PSD estimation in beamspace for estimating DRR. Two different beamformers are applied to the microphone array observation to create the beamspace for estimating the DRR.
Two delay-and-sum beamformers whose mainlobes being pointed in $\Omega_1 = \{\theta_D, \phi_D\}$ and $\Omega_2 = \{\theta_D + \frac{\pi}{6}, \phi_D\}$, respectively, were used for PSD estimation in beamspace. The signals (sampled at 16 kHz [17]) were analysed by the short-time Fourier transform whose frame size was 512 samples. The frame was shifted by half the frame size (i.e. 256 samples). The spatial resolution for the DOA estimation was $\frac{\pi}{36}$ rad for the azimuth and $\frac{\pi}{40}$ rad for the zenith angles. Since, in the ACE Challenge, an opportunity was given to calibrate a method using a small dataset (i.e. Dev dataset [17]), a calibration was introduced to the proposed method in order to compensate biases seen in the estimation. The biases would have been caused by the difference of the definitions of DRR between the ACE Challenge corpus and the proposed method. From the model introduced in Fig. 1, all the early reflections were incorporated into the reverberation in the proposed method whereas some very early reflections were accounted to be the direct sound in the corpus [17]. The calibration simply biased all raw estimated DRR by a constant value, which was calculated by taking the average of the estimation errors given in the test with the Dev dataset.

Fig. 4 shows the comparison of the error distribution of the estimated DRR for different types and levels of noise in the tested acoustical environment provided in the full corpus set (i.e. Eval dataset [17]). By looking at the extent of the distribution of the estimation errors (height of the box plots), the proposed method is more accurate compared to the previous methods. It is revealed that the method with coherence based PSD estimation [6] suffers from very large variance of the estimation error compared to the methods based on PSD estimation in beamspace. Presumably the error in [6] would be mainly caused by the fact that the method could not distinguish the reverberation arriving from the same direction as that of the direct sound [13]. The results also show the bias calibration contributed to improve the estimation accuracy.

In order for further investigating the performance of the proposed method in different acoustical environment, a distribution of the estimation error in each acoustical environment included in the Eval dataset (2 different distances in 5 different rooms: Office 2, Meeting Rooms 1&2, Lecture rooms 1&2 [17]) is shown in Fig. 5. From the result it is clearly seen that both the proposed and previous methods have been affected by the change of the acoustical environment. One of possible causes of this adverse effect would be the poor accuracy of the pre-estimated DOA since both the proposed and previous methods strongly rely on the DOA of the speaker. This fact should be a good guidance for future study to develop a more practically sound DRR estimation method which should be less dependent on the DOA information.

5. CONCLUSION

A method for estimating DRR from speech signals has been proposed. The method estimates the PSD of the direct sound and reverberation using the output signal of two beamformers applied to a microphone array observation. The estimated PSD’s are used for calculating the DRR. The proposed method was evaluated by experiments using the speech corpus of the ACE Challenge. Nevertheless the proposed method suffered from its estimates being biased, which may be compensated by some preliminary calibration, it was revealed that the proposed method provides more accurate estimates of the DRR compared to the authors’ previous methods. Further improvement should be sought to make the proposed method more robust to the variation of the acoustical environment and less dependent on the estimated DOA of the speech signal.

---

Figure 4: Comparison of estimation error for different types of noise provided in the ACE speech corpus.

Figure 5: Distribution of the estimation error for different experimental setups provided in the ACE corpus.
6. REFERENCES

[1] H. Kuttruff, *Room Acoustics*, 5th ed. Applied Science Publishers LTD, 1973, ch. 2.

[2] P. Naylor and N. Gaubitch, Eds., *Speech Dereverberation*. Springer, 2010.

[3] 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.

[4] 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.

[5] ———, “Binaural sound source distance learning in rooms,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1498–1507, 2009.

[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] ———, “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)*, March 2010, pp. 149 –152.

[8] M.-V. Laitinen and V. Pulkki, “Utilizing instantaneous direct-to-reverberant ratio in parametric spatial audio coding,” in *Audio Engineering Society Convention 133*, October 2012.

[9] E. Larsen, C. Schmitz, C. Lansing, W. O’Brien, B. Wheeler, and A. Feng, “Acoustic scene analysis using estimated impulse responses,” in *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2004.*, vol. 1, November 2003, 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. Galdo, and E. Habets, “Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 309 –312.

[12] J. Proakis and D. Manolakis, *Digital Signal Processing, Principles, Algorithms, and Applications*, 4th ed. Pearson, 2007, ch. 4.1.2.

[13] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, and Y. Haneda, “Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1240–1250, June 2013.

[14] Y. Hioka and K. Niwa, “PSD estimation in beamspace for source separation in a diffuse noise field,” in *14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, September 2014, pp. 85–88.

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

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

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

[18] K. Niwa, Y. Hioka, and K. Kobayashi, “Post-filter design for speech enhancement in various noisy environments,” in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, September 2014, pp. 35–39.