Direct-to-Reverberant Energy Ratio Estimation using a First Order Microphone

Hanchi Chen, Thushara D. Abhayapala, Prasanga N. Samarasinghe, Wen Zhang

Research School of Engineering
College of Engineering and Computer Science
The Australian National University
Canberra, Australia

***Abstract*—The Direct-to-Reverberant Ratio (DRR) is an important characterization of a reverberant environment. This work presents a novel blind DRR estimation method based on the coherence function between the sound pressure and particle velocity at a point. First, a general expression of coherence function and DRR is derived in the spherical harmonic domain, without imposing assumptions on the reverberation. In this work, DRR is expressed in terms of the coherence function as well as two parameters which are related to statistical characteristics of the reverberant environment. Then, a method to estimate the values of these two parameters using a microphone system capable of capturing first order spherical harmonics is proposed, under three assumptions which are more realistic than the diffuse field model. Furthermore, a theoretical analysis on the use of plane wave model for direct path signal and its effect on DRR estimation is presented, and a rule of thumb is provided for determining whether the point source model should be used for the direct path signal. Finally, the ACE Challenge Dataset is used to validate the proposed DRR estimation method. The results show that the average full band estimation error is within 2 dB, with no clear trend of bias.**

***Index Terms*—Direct-to-Reverberant Energy Ratio, higher order microphone, spherical harmonics, spherical microphone array.**

I. Introduction

The direct-to-reverberation energy ratio (DRR), defined as the energy ratio between direct signal and its reverberations, is an important parameter to characterize a reverberant environment, along with other parameters such as reverberation time. Since reverberation energy affects the speech signal's clarity [1], the DRR has an influence on the algorithms for various applications such as speech dereverberation [2], teleconferencing [3] and hearing aids [4], both in terms of algorithm performance and strategy. The minimum audible difference in DRR has been investigated in [5]. In [6], DRR is utilized for parametric spatial audio coding. DRR also finds its application in the field of psychoacoustics, where it is believed that DRR helps human to determine the distance of the sound source [7], [8], [1].

DRR estimation methods based on estimating room impulse responses have been presented by Larsen *et al.* [9] and Falk *et al.* [10], However, pre-processing is required for both methods. Mosayyebpour *et al.* [11] presented a method for blind DRR estimation based on higher order statistics, where the inverse filter of the room impulse response is estimated using the skewness of the speech signal. Parada et. al. presented a single channel DRR estimation method base on a neural network learning algorithm [12].

Methods for blind DRR estimation using multiple sensors have also been proposed in literature. With the goal of estimating source distance, Lu [13] presented a DRR estimation algorithm using the equalization-cancellation method, where a binaural microphone system is used to capture sound signal. The coherence function framework was first introduced by Vesa [14] for estimating source distance using binaural signals, where the coherence function of the two input signals was used as a characterization of source distance. Later, the coherence function framework was also used by Jeub [15] to develop a DRR estimation algorithm. In this work, the DRR is estimated by comparing coherence value computed from two microphone inputs with theoretical coherence functions in a diffuse sound field. Thiergart [16] also developed a DRR estimation algorithm based on the complex coherence function of two omnidirectional microphones. In [17] a DRR estimation method based on spectra standard deviation of two microphones was proposed.

Directional or beam forming microphone arrays have also been used to estimate DRR, such as the methods presented in [18] and [19]. In both of these works, the power spectral density (PSD) of the reverberant field were used to estimate DRR. Another method [20] uses a circular microphone array to estimate DRR, the method relies on the spatial correlation matrix of the microphones' received signals. The reverberation is modelled as a diffuse field in that work, while the direct path is assumed to be a plane wave. The DRR is solved using the least mean square method. Kuster [21] presented a method based on coherence function of sound pressure and particle velocity at the receiver position, measured by a differential microphone array.

In recent years, the use of higher order microphones and the technique of spherical harmonic decomposition [22] have become popular in the field of room acoustic analysis. Jarrett *et al.* [23] proposed a method to estimate Signal-to-Diffuse Ratio

Thanks to Australian Research Councils Discovery Projects funding scheme (project no. DP140103412).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2601222, IEEE/ACM Transactions on Audio, Speech, and Language Processing

(SDR, equivalent to DRR when assuming diffuse reverberation field) utilizing spherical harmonic coefficients captured by a higher order microphone. It is shown that this method minimizes the SDR estimation bias. In our previous work [24], we implemented Kuster's method [21] in spherical harmonic domain, utilizing the first order spherical harmonic coefficients to estimate DRR.

In many of the previous works, such as [21], [23], [16] and our previous work [24], the direct path signal is assumed to be a plane wave, and the reverberant sound field is assumed to be diffuse field. In real-life reverberant environments where these assumptions may not hold, the DRR estimation accuracy of these algorithms may degrade. For example, the DRR estimated using Kuster's method tend to be higher than the ground truth in reverberant rooms [21].

In this work, we first develop a general expression for DRR estimation using the coherence function of the sound pressure and particle velocity, using a point source model for the direct path signal, and without applying any assumptions for the reverberation field. Using the relationship between spherical harmonic coefficients and acoustic particle velocity, we develop the framework in the spherical harmonics domain. Then, for the direct path model, we provide a detailed analysis on the error in DRR estimation results when using the plane wave model. We propose a rule of thumb for determining whether the plane wave model can be used without introducing significant error, based on the source-to-microphone distance and target frequency. For the reverberation sound field, we show that the reverberation characteristics related to DRR estimation can be expressed using two parameters, and that under the diffused field assumptions, the values of these parameters can be determined theoretically, which results in the simplified DRR solutions in [21] and [23]. We also provide a theoretical analysis on the two parameters, their physical meanings, and their impact on the DRR estimation, which explains the positive bias phenomenon of Kuster's method [21]. Furthermore, we propose a method to estimate these two parameters for a given reverberant environment, using a first order microphone, under a number of assumptions on the reverberant field which are less strict than the diffuse field model. The DRR can then be calculated using the estimated parameters.

The performance of the proposed DRR estimation algorithm is verified using the ACE Challenge Dataset [25]. It is shown that the results agree with the theoretical analysis, and that the proposed method addresses the positive bias problem of Kuster's method [21], and the mean DRR estimation error is less than 2 dB for all recording scenes in the ACE Challenge Dataset.

II. Background theory

A. Spatial soundfield decomposition

In this work, the spherical harmonic decomposition is used to describe the spatial sound field in the proximity of the microphone system. The sound pressure within a free space region at a point (r, θ, ϕ) with respect to an origin O of the spherical co-ordinate system can be written as [26]

$$P(r, \theta, \phi, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \beta_{nm}(k) j_n(kr) Y_{nm}(\theta, \phi), \quad (1)$$

where $\beta_{nm}(k)$ are soundfield coefficients, $k = 2\pi f/c$ is the wave number, f is the frequency, c is the speed of sound propagation, $j_n(kr)$ is the nth order spherical Bessel function of the first kind, $Y_{nm}(\theta, \phi)$ are the spherical harmonics, defined by

$$Y_{nm}(\theta, \phi) \triangleq \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}} P_{n|m|}(\cos\theta) e^{im\phi}, \quad (2)$$

where $P_{n|m|}(\cos\theta)$ are the associated Legendre functions, and $i = \sqrt{-1}$. In literature, the representation (1) is referred to as *spherical harmonic expansion*, *wave-domain representation*, *modal expansion*, or *multi-pole expansion* of a wavefield.

B. First order microphone and acoustic particle velocity

In the general sense, microphones with certain directional beam patterns, such as cardioid microphones and differential microphones are commonly referred to as first order microphones. In the context of spatial sound field recording based on spherical harmonics decomposition, a first order microphone is a microphone system which is capable of acquiring the 0th and 1st order spherical harmonic coefficients of its surrounding sound field, namely, $\beta_{00}, \beta_{11}, \beta_{10}$ and $\beta_{1,-1}$ in (1). Specific directionalities can be realized through applying beam-forming algorithms on the 0th and 1st order coefficients.

First order microphones are known to have the ability to pick up the velocity component of the impinging sound [27], this velocity component is commonly described using the concept of particle velocity, which refers to the velocity of particle movement in the medium during wave propagation. Here, we derive the expressions that relate the 1st order spherical harmonic coefficients to the acoustic particle velocity in the x, y and z directions. These expressions are used later in the paper.

Defining the spherical coordinate system (r, θ, ϕ) in relation to the Cartesian coordinate system, the particle velocity at the origin is related to the spherical harmonic coefficients by the following theorem:

Theorem 1: The acoustic particle velocity at the point $\mathbf{0} \equiv (0, 0, 0)$ along the x, y and z axes at a particular frequency k can be expressed using the first order spherical harmonic coefficients,

$$V_x(\mathbf{0}, k) = \frac{i\rho_0 c}{\sqrt{24\pi}} (\beta_{11}(k) + \beta_{1,-1}(k)) \quad (3)$$

$$V_y(\mathbf{0}, k) = \frac{-\rho_0 c}{\sqrt{24\pi}} (\beta_{11}(k) - \beta_{1,-1}(k)) \quad (4)$$

$$V_z(\mathbf{0}, k) = \frac{i\rho_0 c}{\sqrt{12\pi}} \beta_{10}(k), \quad (5)$$

where ρ_0 is the density of the medium. Proof of Theorem 1 is given in Appendix A.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2601222, IEEE/ACM Transactions on Audio, Speech, and Language Processing

III. DRR ESTIMATION BASED ON COHERENCE MEASUREMENTS

A. Representation of reverberant sound field

For convenience, the spherical coordinate system is defined such that its origin is at the position of the microphone, and its positive z axis points towards the impinging direction of the direct path signal. In many scenarios, the natural coordinate system may have a different orientation than our definition. In such cases, the spherical harmonics defined under a different coordinate system (with the same origin) can be transformed into our desired coordinate system using the spherical harmonic rotation, which is described in Appendix B.

The sound pressure at a point (r, θ, ϕ) close to the origin can be decomposed using (1). For the direct path, we have

$$P_D(r,\theta,\phi,k) = \sum_{n=0}^{1}\sum_{m=-n}^{n} B_{nm}(k) j_n(kr) Y_{nm}(\theta,\phi), \quad (6)$$

where only the first order sound field is considered.

For the sound field due to reverberation, we have

$$P_R(r,\theta,\phi,k) = \sum_{n=0}^{1}\sum_{m=-n}^{n} \alpha_{nm}(k) j_n(kr) Y_{nm}(\theta,\phi), \quad (7)$$

where $B_{nm}(k)$ and $\alpha_{nm}(k)$ represent the coefficients of the direct path and the reverberant sound field, respectively.

The following assumptions are made regarding the direct path sound:

1: The direct path is due to a point source located at $(r_0, \vartheta, \varphi)$.

2: The direct path signal $P_D(r,\theta,\phi,k)$ is uncorrelated with the reverberant sound field $P_R(r,\theta,\phi,k)$. Using (6) and (7), this assumption can be expressed as

$$E\{B_{nm}\alpha^*_{n'm'}\} = 0, \text{ for all } n \text{ and } m. \quad (8)$$

where $E\{\cdot\}$ denotes the expectation operator.

Since the direct path signal is modelled as sound waves emitted by a point source, $B_{nm}(k)$ can be written using the following expression [28]

$$B_{nm}(k) = A_D\, ik\, h_n^{(1)}(kr_0) Y^*_{nm}(\vartheta,\varphi), \quad (9)$$

where A_D indicates the magnitude of the impinging sound, $h_n^{(1)}(kr_0)$ is the nth order spherical Hankel function of the first kind, r_0 is the distance between the point source and the microphone with $r_0 > r$, $(r_0, \vartheta, \varphi)$ denotes the position of the point source, and $(\cdot)^*$ represents complex conjugate. Since the coordinate system is defined such that $\vartheta = 0$, and due to the fact that $Y_{11}(0,\varphi) = Y_{1,-1}(0,\varphi) = 0$, we have $B_{11}(k) = B_{1,-1}(k) = 0$. Thus the combined sound field coefficients $\beta_{nm}(k)$ can be expressed as follows

$$\beta_{00}(k) = A_D ik\, h_0^{(1)}(kr_0) Y^*_{00}(0,0) + \alpha_{00}(k), \quad (10)$$

$$\beta_{10}(k) = A_D ik\, h_1^{(1)}(kr_0) Y^*_{10}(0,0) + \alpha_{10}(k), \quad (11)$$

$$\beta_{11}(k) = \alpha_{11}(k), \quad (12)$$

$$\beta_{1,-1}(k) = \alpha_{1,-1}(k). \quad (13)$$

Equations (10)-(13) shows that in the coordinate system defined in this section, the direct path signal is only present in $\beta_{00}(k)$ and $\beta_{10}(k)$, but not in $\beta_{11}(k)$ and $\beta_{1,-1}(k)$.

B. Representation of DRR using coherence function

The coherence function between the sound pressure $P(\mathbf{0},k)$ and particle velocity $V_z(\mathbf{0},k)$ along the z direction can be defined as [21],

$$\gamma^2 \triangleq \frac{|E\{P(\mathbf{0},k)V_z(\mathbf{0},k)^*\}|^2}{E\{|P(\mathbf{0},k)|^2\}E\{|V_z(\mathbf{0},k)|^2\}}. \quad (14)$$

Note that $P(\mathbf{0},k) = \beta_{00}Y_{00}(0,0) = 1/\sqrt{2\pi}\beta_{00}$ and $V_z(\mathbf{0},k)$ is proportional to $\beta_{10}\cdot i$[1] in (5). Substituting $P(\mathbf{0},k)$ and (5) into (14), and applying (10) (11), we have

$$\gamma^2 = \frac{|E\{\beta_{00}(\beta_{10}\cdot i)^*\}|^2}{E\{|\beta_{00}|^2\}E\{|\beta_{10}|^2\}} \quad (15)$$

$$= \frac{|E\{H_0(H_1 i)^*\} + E\{\alpha_{00}(\alpha_{10} i)^*\}|^2}{(E\{|H_0|^2\} + E\{|\alpha_{00}|^2\})(E\{|H_1|^2\} + E\{|\alpha_{10}|^2\})}, \quad (16)$$

where the assumption that the direct path is uncorrelated with the reverberations (8) is used, and we denote

$$H_0 \triangleq A_D ik\, h_0^{(1)}(kr_0) Y^*_{00}, \quad (17)$$

$$H_1 \triangleq A_D ik\, h_1^{(1)}(kr_0) Y^*_{10}. \quad (18)$$

Note that the angle arguments $(\vartheta = 0, \varphi = 0)$ of $Y_{nm}(\vartheta,\varphi)$ and the frequency arguments (k) of β_{nm} and α_{nm} have been omitted for simplicity.

The linear scale direct-to-reverberant energy ratio is defined here to be the ratio of measured acoustic energy at the position of measurement due to the direct path and reverberation, since $P(\mathbf{0}) = \beta_{00}Y_{00}$, we have

$$\text{DRR} = \frac{E\{|P_D(\mathbf{0})|^2\}}{E\{|P_R(\mathbf{0})|^2\}} = \frac{E\{|B_{00}|^2\}}{E\{|\alpha_{00}|^2\}}. \quad (19)$$

Using (9) to express B_{00} in (19), we have

$$\text{DRR} = \frac{E\{|A_D ik\, h_0^{(1)}(kr_0) Y^*_{00}|^2\}}{E\{|\alpha_{00}|^2\}} = \frac{E\{|H_0|^2\}}{E\{|\alpha_{00}|^2\}}. \quad (20)$$

Substituting (20) into (15) yields

$$\gamma^2 = \frac{\Big| -\text{DRR}\cdot i\big(\frac{h_1^{(1)}(kr_0)Y^*_{10}}{h_0^{(1)}(kr_0)Y^*_{00}}\big)^* + \frac{E\{\alpha_{00}(\alpha_{10}i)^*\}}{E\{|\alpha_{00}|^2\}}\Big|^2}{(\text{DRR}+1)(\text{DRR}\big|\frac{h_1^{(1)}(kr_0)Y_{10}}{h_0^{(1)}(kr_0)^2 Y_{00}}\big|^2 + \frac{E\{|\alpha_{10}|^2\}}{E\{|\alpha_{00}|^2\}})} \quad (21)$$

which relates the coherence value γ^2 to the DRR of the room.

For convenience, we define

$$R_1 \triangleq \frac{E\{\alpha_{00}(\alpha_{10}\cdot i)^*\}}{E\{|\alpha_{00}|^2\}}\frac{Y_{00}}{Y_{10}} \quad (22)$$

$$R_2 \triangleq \frac{E\{|\alpha_{10}|^2\}}{E\{|\alpha_{00}|^2\}}\frac{Y_{00}^2}{Y_{10}^2} \quad (23)$$

as the reverberation parameters, and

$$H \triangleq \Big(\frac{h_1^{(1)}(kr_0)}{h_0^{(1)}(kr_0)}\Big)^*. \quad (24)$$

[1]Although removing the imaginary argument i here does not affect γ^2, we keep i for the derivation of further expressions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2601222, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Then (21) can be simplified as

$$\gamma^2 = \frac{|-\text{DRR}\cdot i\cdot H + R_1|^2}{(\text{DRR}+1)(\text{DRR}|H|^2+R_2)} \tag{25}$$

$$= \frac{|\text{DRR}|^2|H|^2 + 2\text{DRR}\cdot Im\{HR_1^*\} + |R_1|^2}{(\text{DRR}+1)(\text{DRR}|H|^2+R_2)}, \tag{26}$$

where $Im\{\cdot\}$ denotes imaginary part of the argument. From (25) it can be seen the characteristics of reverberation which affects DRR estimation using coherence method can be expressed using two parameters R_1 and R_2.

C. Assumptions for the reverberant sound field

1) Plane wave assumption for the direct path: In previous works, the direct path signal is often assumed to be a plane wave [21], [23]. Under this assumption, the following approximation can be applied (see Appendix C for the proof)

$$\lim_{r_0\to\infty} \frac{h_1^{(1)}(kr_0)}{h_0^{(1)}(kr_0)} \approx -i, \tag{27}$$

and (25) can be simplified into

$$\gamma^2 = \frac{|\text{DRR}+R_1|^2}{(\text{DRR}+1)(\text{DRR}+R_2)} \tag{28}$$

$$= \frac{\text{DRR}^2 + 2\text{DRR}\cdot Re\{R_1\} + |R_1|^2}{(\text{DRR}+1)(\text{DRR}+R_2)}, \tag{29}$$

where $Re\{\cdot\}$ denotes real part of the argument. The plane wave assumption leads to bias in the DRR estimation, primarily for lower frequencies and smaller values of r_0, which is shown in Section IV-B.

2) Diffuse reverberation assumptions in previous works: In many previous works, the sound field due to reverberation is often modelled as diffused field [21], [23], although the exact definition of diffused field may vary. In [23], the diffuse field is defined as an infinite number of uncorrelated plane waves impinging uniformly from the sphere. Under this assumption, it is shown that $E\{\alpha_{00}\alpha_{10}^*\} = 0$, and $E\{|\alpha_{nm}|^2\} = E\{|\alpha_{n'm'}|^2\}$ for all values of n and m [23]. In this case, $R_1 = 0, R_2 = |Y_{00}|^2/|Y_{10}|^2 = 1/3$, and (29) becomes equivalent to the magnitude-squared version of Eq.(18) in [23].[2]

In the case of Kuster's work [21], the reverberant field is assumed to be plane waves whose impinging directions distribute uniformly over $\theta_{\text{in}} \in [0, 2\pi)$, where θ_{in} is the angle between direct path and the plane wave impinging direction. This assumption differs from the reverberant field model used in [23], where plane waves are distributed uniformly over the sphere; this assumption can be fulfilled if the plane waves impinge uniformly over a circle. Under this assumption, Kuster has derived an expression for γ^2 which takes the same form as (29), but with $R_1 = 0$, and $R_2 = 0.5$ [21].

[2]For c_{00} and c_{10}, with $\Omega_{\text{dir}} = (0, 0)$.

3) Assumptions on reverberation used in this work: In many real acoustic environments, the diffused field assumptions for reverberant field made in [23] and [21] often cannot be met, which may lead to inaccuracies in the DRR estimation result. In this work, in order to improve the accuracy of DRR estimation, we relax some assumptions made on the reverberant sound field. In particular, we assume that the reverberant field satisfies the following conditions:

1: The average sound intensity (product of sound pressure and particle velocity) [29] of the reverberant field has the same magnitude in x, y and z directions.

$$|E\{P^{\text{r}}V_z^{\text{r}*}\}| = |E\{P^{\text{r}}V_x^{\text{r}*}\}| = |E\{P^{\text{r}}V_y^{\text{r}*}\}|. \tag{30}$$

where P^{r} and V^{r} denote sound pressure and particle velocity due to reverberation, respectively.

2: The expected energy of the reverberant field particle velocity is constant in x, y and z directions.

$$E\{|V_x^{\text{r}}|^2\} = E\{|V_y^{\text{r}}|^2\} = E\{|V_z^{\text{r}}|^2\}, \tag{31}$$

3: The reverberant field sound intensity is zero mean when averaged over a frequency band.

$$\int_{k_1}^{k_2} E\{P^{\text{r}}(k)V_z^{\text{r}}(k)^*\}dk = 0. \tag{32}$$

where k_1 and k_2 represent the boundary of a frequency band.

In (32), the real part of $P^{\text{r}}V_z^{\text{r}*}$ is often referred to as the active sound intensity, which represents the coherent flow of sound energy in the z direction [29]. The imaginary part of sound intensity, on the other hand, is referred to as the reactive sound intensity, which represents the coherent, but non-propagating, "standing wave" sound energy. A detailed justification of the assumption (32) is given in IV-A.

In a diffuse sound field, both active and reactive components of the sound intensity are equal to zero since the phase of particle velocity varies randomly. The energy of particle velocity can be analytically computed [23], [21]. Applying these results to (28) leads to simplified expressions of γ^2 as shown in [23] and [21].

However, this paper does not assume a diffuse field. Hence the expected energy of the particle velocity and sound intensity cannot be directly computed without the knowledge of the reverberant field. Therefore, a method to estimate these characteristics is needed to compute the DRR. The following subsection describes one such method, using measurements from a first (or higher) order microphone system.

D. Reverberant field estimation

From (12) and (13), it can be observed that the spherical harmonic coefficients β_{11} and β_{1-1} do not contain the direct path signal. In fact, β_{11} and β_{1-1} collectively represent the particle velocity of the reverberations in the directions orthogonal to the direct path. The assumptions on the reverberation (30), (31) and (32) can be expressed using spherical harmonic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2601222, IEEE/ACM Transactions on Audio, Speech, and Language Processing

coefficients as

$$\Big|\frac{E\{\alpha_{00}(\alpha_{10}i)^*\}}{-i\sqrt{12\pi}}\Big| = \Big|\frac{E\{\alpha_{00}(\alpha_{11}i+\alpha_{1,-1}i)^*\}}{-i\sqrt{24\pi}}\Big| = \Big|\frac{E\{\alpha_{00}(\alpha_{11}i-\alpha_{1,-1}i)^*\}}{-\sqrt{24\pi}}\Big|, \quad (33)$$

$$2\cdot E\{|\alpha_{10}|^2\} = E\{|\alpha_{11}+\alpha_{1,-1}|^2\} = E\{|\alpha_{11}-\alpha_{1,-1}|^2\}, \quad (34)$$

and

$$\int_{k_1}^{k_2} E\{\alpha_{00}(k)(\alpha_{10}(k)\cdot i)^*\}dk = 0. \quad (35)$$

Since it is assumed that the direct path signal is uncorrelated with the reverberation signal, substituting (8), (10), (12) and (13) into (33), we can write

$$\sqrt{2}|E\{\alpha_{00}(\alpha_{10}i)^*\}| = |E\{\beta_{00}(\beta_{11}i+\beta_{1,-1}i)^*\}| = |E\{\beta_{00}(\beta_{11}i-\beta_{1,-1}i)^*\}|, \quad (36)$$

which illustrates a way to indirectly estimate the value of $|R_1|$ in (29). Using (10) and (11), the energy of the reverberation can be approximated by

$$E\{|\alpha_{00}|^2\} = E\{|\beta_{00}|^2\} - \frac{Y_{00}^2}{Y_{10}^2|H|^2}(E\{|\beta_{10}|^2\} - E\{|\alpha_{10}|^2\}). \quad (37)$$

If the plane wave model is used for the direct path, (37) can be simplified using (27), as

$$E\{|\alpha_{00}|^2\} \approx \Big(\frac{E\{|\beta_{00}|^2\}}{Y_{00}^2} - \frac{E\{|\beta_{10}|^2\}}{Y_{10}^2} + \frac{E\{|\alpha_{10}|^2\}}{Y_{10}^2}\Big)Y_{00}^2, \quad (38)$$

where $E\{|\alpha_{10}|^2\}$ can be estimated using (34). Substituting (34), (36) and (37) into (22), the estimation expression for $|R_1|$ can be written as

$$|R_1| \approx \frac{1}{2\sqrt{2}}\cdot \frac{|E\{\beta_{00}(\beta_{11}i+\beta_{1,-1})^*\}| + |E\{\beta_{00}(\beta_{11}i-\beta_{1,-1}i)^*\}|}{E\{|\beta_{00}|^2\} - E\{|\beta_{10}|^2\}\frac{Y_{00}^2}{Y_{10}^2H^2} + M_{\text{pwr}}\frac{Y_{00}^2}{Y_{10}^2H^2}}, \quad (39)$$

where we define

$$M_{\text{pwr}} \triangleq \frac{1}{2}(E\{|\beta_{11}+\beta_{1,-1}|^2\} + E\{|\beta_{11}-\beta_{1,-1}|^2\}) \quad (40)$$

similarly, by substituting (12), (13) and (37) into (23), R_2 can be written as

$$R_2 \approx \frac{M_{\text{pwr}}}{E\{|\beta_{00}|^2\} - E\{|\beta_{10}|^2\}\frac{Y_{00}^2}{Y_{10}^2H^2} + M_{\text{pwr}}\frac{Y_{00}^2}{Y_{10}^2H^2}}. \quad (41)$$

Note that all the coefficients required for the calculation can be acquired by a first order microphone array directly. The estimated values of $|R_1|$ and R_2 can be directly substituted into (26) or (29) to estimate DRR using γ^2.

E. DRR estimation procedure

Assuming that the value of DRR is positive, the solution for DRR can be found by solving (25) or (29). For the plane wave model, the solution can be derived as

$$\text{DRR} = \frac{\gamma^2 + R_2\gamma^2 + \sqrt{4|R_1|^2(\gamma^2-1) + \gamma^4(R_2-1)^2 + 4R_2\gamma^2}}{2-2\gamma^2}, \quad (42)$$

where the assumption (32) is used, which leads to $Re\{R_1\} = 0$. The calculated DRR is in linear scale, and the more commonly used log-scale DRR_{log} is defined as

$$\text{DRR}_{\text{log}} = 10\log_{10}\text{DRR}. \quad (43)$$

From our experience in testing the algorithm using the ACE Challenge Development Dataset [25], the estimation of $|R_1|$ and R_2 at a single frequency is often unstable. However, for typical room environments, one can assume that the characteristics of reverberation do not vary rapidly over frequencies since sound waves of similar wavelength are likely to have similar propagation modes. Therefore $|R_1|$ and R_2 can be seen as constant if the frequency band of interest is sufficiently narrow, then one can use the average values of $|R_1|$ and R_2 over a particular frequency band for the calculation of DRR for this frequency band.

For subband and full band DRR estimation, the results are obtained by taking the average of the single frequency DRR estimations within the band, then the values are converted to log scale for convenience.

We recommend the following procedures to estimate the DRR of a particular frequency band from a recording:

Step 1	Determine the direct path impinging direction using a suitable Direction-of-Arrival (DOA) algorithm, which can be done using the signal received by the first order (or higher order) microphone.
Step 2	Use an appropriate algorithm to detect the frames of the recording that contain speech signal and calculate the 0th and 1st order spherical harmonic coefficients for each frequency bin within the frequency band.
Step 3	Rotate the spherical harmonics using the method in Appendix B, such that the z-axis is aligned with the direct path.
Step 4	Calculate $\|R_1\|$ and R_2 for each frequency bin, using (39) and (41), then average over all the frequency bins to obtain an estimation for the whole frequency band.
Step 5	Calculate γ for each frequency and using (25) or (42) with the averaged $\|R_1\|$ and R_2 to estimate the DRR for each frequency.
Step 6	Average the DRR estimations calculated from each frequency bin to obtain the subband or full band DRR estimation. Convert the result to log scale.

A disadvantage of the original coherence method for DRR estimation is that the angle between the direct path and the particle velocity measurement direction is generally unknown, and in a real measurement, the microphone have to be pointed towards the direct path [21]. In our improved method, since we use a first order microphone for measurement, which records the complete sound field, it is possible to derive the velocity measurement in any direction, through rotation of the spherical harmonic coefficients. In addition, the data acquired

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2601222, IEEE/ACM Transactions on Audio, Speech, and Language Processing

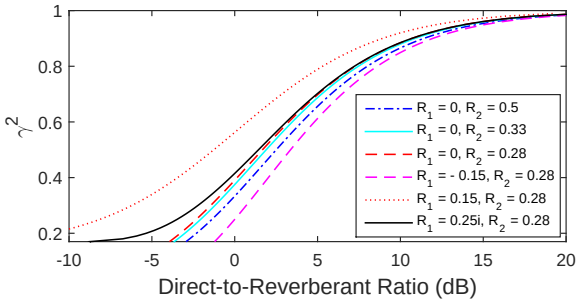


Fig. 1. Theoretical γ^2 versus estimated direct-to-reverberant ratio (DRR) calculated using (42), under various reverberation parameter settings.

by the microphone can be used to perform Direction-of-Arrival (DOA) estimation for the direct path, therefore there is no special requirement for positioning the microphone during measurements.

IV. Impact of parameters on DRR estimation

A. Reverberation parameter

In order to illustrate the impact of R_1 and R_2 on the estimated DRR, we plot the theoretical DRR against γ^2 using (29) with the diffuse field parameter setting proposed by Kuster [21] $(R_1, R_2) = (0, 0.5)$ and Jarrett [23] $(R_1, R_2) = (0, 1/3)$ as well as a number of other values that were commonly found in our experiment $(R_1, R_2) = \{(0, 0.28); (0.15, 0.28); (-0.15, 0.28); (0.25i, 0.28)\}$, as shown in Fig. 1. Note that the assumption (32) is not applied here, in order to illustrate the impact of R_1 on the DRR estimation. It can be seen from Fig. 1 that depending on the values of R_1 and R_2, a deviation of $\pm$3dB in estimated DRR can be observed for low values of γ^2.

From (22), it can be seen that R_1 is equivalent to the sound intensity in the z direction with a certain normalization. Since all normalization factors are real, the real and imaginary part of R_1 correspond to the active and reactive sound intensity, respectively. When $Re\{R_1\} > 0$, it indicates that the net energy flow of reverberation coincides with the direct path signal, and as a result the reverberation will be "added" to the direct path, and as a result contributes to coherence function γ^2 positively. On the other hand, if $Re\{R_1\} < 0$, the net reverberation energy flow in the z direction opposites the direct path, essentially cancelling part of the direct path sound intensity, therefore it contributes to γ^2 negatively. As a result of this, as can be seen in Fig. 1, for the same value of γ^2, a positive $Re\{R_1\}$ corresponds to low value of DRR, and vice versa.

The absolute value of R_1 represents the overall coherence of the reverberant field in the z direction. This includes the reactive part of R_1, which corresponds to the resonating reverberation energy. It can be seen from (29) that $|R_1|$ always contributes to γ^2 positively. Therefore, as seen in Fig. 1, a non-zero value of $|R_1|$ results in lower value of DRR, for the same γ^2, this is especially significant at lower values o γ^2.

Using a first order microphone, it is possible to estimate $|R_1|$ for each frequency bin, if it is assumed that the reverberant sound intensity is uniform in each direction. Unfortunatelly, the sign of $Re\{R_1\}$, which indicates the direction of energy flow, cannot be determined through observation of the sound field in its orthogonal directions. However, by observing the reverberation sound field from the ACE Challenge Development Sataset [25], it was found that both active and reactive sound intensity of the reverberation in the x and y directions have zero mean when averaged over each $1/3$ octave subband, indicating that the energy flow of reverberation changes randomly and rapidly with frequency. Therefore it is reasonable to assume that PV_z^* is also zero mean when observed at multiple frequencies. As as result, when averaging the estimated DRR over each subband, the impact of $Re\{R_1\}$ (and $Im\{HR_1^*\}$ in (26)) on each frequency bin will be cancelled out, and the term can be removed in the derivation of (42), provided that appropriate frequency averaging is performed after calculating DRR for each frequency bin.

As can be seen from Fig. 1, R_2 does not affect the estimated DRR as strongly as R_1, and a lower value of R_2 results in a slightly lower estimation of DRR. From (23) it can be seen that R_2 reflects the expected energy ratio between sound pressure and particle velocity. In Jarrett's diffuse field model [23], the value of R_2 is lower ($R_2 = 1/3$), therefore, we expect Jarrett's method to yield a slightly lower estimation of DRR compared to Kuster's. From our analysis to the ACE Challenge Development Dataset, the value of R_2 typically varies between $0.25 - 0.33$, which is close to Jarrett's model (see Table II).

B. Nearfield sound source

In order to analyze the DRR estimation error due to using a plane wave to approximate the direct path sound field, we compute the difference in the estimated DRR using (25) and (29) ($\Delta\text{DRR} \triangleq 10\log_{10}(\text{DRR}_{\text{plane}}/\text{DRR}_{\text{point}})$). It can be seen by observing (25) that the calculated $\text{DRR}_{\text{point}}$ depends on the product kr_0. Fig. 2 plots ΔDRR as a function of kr_0, for various values of γ^2. In this figure, for simplicity, we assume that $R_1 = 0, R_2 = 0.5$. The selected values of γ^2 $(0.86, 0.65, 0.33$ and $0.19)$ correspond to $\text{DRR}_{\text{plane}} = 10\text{dB}, 5\text{dB}, 0\text{dB}$ and -2.5dB, respectively, using the parameter settings described above.

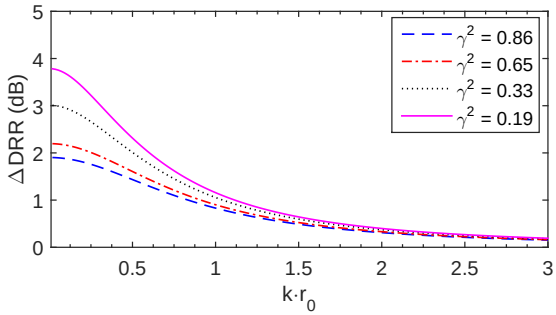


Fig. 2. Plot of theoretical DRR versus kr_0 using plane wave model (29) and point source model (25) with $\gamma^2 = 0.86, 0.65, 0.33$ and 0.19.

From Fig. 2 we can see that the plane wave model results in higher DRR estimations than that of the point source

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2601222, IEEE/ACM Transactions on Audio, Speech, and Language Processing

model for smaller values of kr_0, where $\Delta\text{DRR} \approx 2-4$ dB for $kr_0 = 0$, depending on the value of γ^2. At higher frequencies and larger source-microphone distance (higher kr_0), the difference between the two methods reduce rapidly, at $kr_0 > 3$, the difference in the calculated DRR using the two models becomes negligible.

Comparing the curves corresponding to each value of γ^2, it can be seen that the estimation error of the plane wave model is smaller when γ^2 is larger, corresponding to higher values of DRR. The user may select the appropriate model for their applications, based on the target frequency band and expected source distance. Here, we propose a rule of thumb for determining whether to use the point source model or the plane wave model. When $kr_0 \geqslant 2$, the error caused by plane wave model is less than 0.5 dB for all values of γ^2, as can be observed in Fig. 2. For $kr_0 < 2$, the use of point source model is recommended for improving DRR estimation accuracy.

V. Validation using ACE Challenge Database

A. The ACE Challenge Database

The ACE Challenge Database [25] is used to validate our algorithm. The database consists of two datasets: the Evaluation dataset, and the Development dataset. The Development dataset is provided to the ACE Challenge participants as a training database, using which the participants can train and fine tune their algorithms. The Evaluation dataset is used to evaluate the performance of fine-tuned algorithms.

The Evaluation dataset consists of 4500 synthesized recordings of various configurations. A total of 5 rooms are used to record the room impulse responses, with two recording setups (positions) for each room. The room details are summarized in Table I. We note that although the impulse responses of 7 rooms were recorded according to [25], only 5 of them are used to create the Evaluation dataset; the other two rooms were used to create the Development dataset. The speech and noise setup for the Development dataset differ from that of the Evaluation database, therefore in this work, the Development dataset is only used for developing the DRR algorithm; the results presented in this section are all generated using the Evaluation dataset.

The impulse responses are recorded using an Eigenmike, and the reverberant speech recordings are synthesized by convolving the impulse responses with anechoic speech recordings [25]. The speech recordings consist of voices of 10 talkers, 5 female and 5 male, with 5 separate utterance recordings for each talker. Three different types of noise ("Ambient", "Fan" and "Babble") are recorded separately under the same room setup and mixed into the reverberant speech recordings, each with three SNR settings (-1 dB, 12 dB and 18 dB).

The ground truths for both full band and subband DRR have been provided. For subband DRR, the central frequencies for all bands have been chosen according to the ISO standard [25].

B. Algorithm setup

Since the ground truth for direct path DOA is not given, we have to estimate the DOA for each of the ten scene setups. This is done by segmenting each speech recording into multiple short frames, and selecting the frames that correspond to the beginning of each utterance (where the impinging signal is almost purely due to the direct path). To find the frames containing speech, a simple speech detection algorithm calculates the average signal energy of each frame, and select the frames with higher energy, which are considered to contain the speech signal. If the energy of a frame is significantly higher than the previous one, then this window is considered to contain the beginning of an utterance. We then calculate the spherical harmonic coefficients for each selected frame and for frequencies between 200-2000 Hz, and perform a frequency averaged MUSIC DOA estimation in the spherical harmonic domain [30], [31]. The estimated DOA is used for further calculations.

In order to maintain the highest possible frequency resolution while at the same time to avoid violating the assumption that the direct path signal and reverberations are uncorrelated, we choose the analysis window length to be 10 ms. When fine-tuning our algorithm using the ACE Development Dataset, it was found that a window length shorter than 10 ms does not reduce the average value of γ^2, therefore we assume that the chosen window length is appropriate.

For each speech recording, only the windows that contain the speech signal are used for analysis. For each frequency subband, we calculate the 0th and 1st order spherical harmonic coefficients for each selected window and for all the frequency bins within each subband. We then follow steps 3 through 6 in Section. III-E to estimate DRR for each subband.

Although the ground truth for subband DRR is given for all frequency bands between 20 Hz and 20 kHz [25], the recorded speech signal does not cover the complete spectrum. Therefore, we focus on the subbands with central frequency between 199.52 Hz and 2511.89 Hz, where there is sufficient energy in the speech recordings for DRR estimation. For this reason, we cannot estimate the full band DRR in the complete sense, instead, we calculate the average DRR over the selected subbands, which is used as the full band estimation. The full band ground truth DRR used for comparison is also calculated by averaging the corresponding subband ground truths, instead of using the full band DRR provided by the database.

The exact source-to-microphone distance is not provided by the database. However, according to the organizer of the ACE Challenge, the microphones are placed at no less than 1 m away from the source for all recording scenarios. Since we only focus on frequencies above 199.52, using the rule of thumb proposed in Section IV-B, $kr_0 \approx 3.66 > 2$, therefore the plane wave model is sufficiently accurate, hence used for DRR estimation.

The error of estimated DRR is defined as

$$\text{DRR}_{\text{err}} = 10\log_{10}\big(\frac{\text{DRR}_{\text{est}}}{\text{DRR}_{\text{truth}}}\big). \tag{44}$$

The mean and standard deviation of DRR estimation error is then calculated using DRR_{err} from each recording.

C. Full band results

The full band DRR estimation results for the ACE Database are shown in Fig. 3. In this figure, we plot the mean and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2601222, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE I
ROOM DIMENSIONS (APPROX.) AND MINIMUM/MAXIMUM DRR FOR EACH ROOM RECORDING CONFIGURATION

Room Name	Length (m)	Width (m)	Height(m)	Volume (m^3)	Setup A min DRR	Setup A max DRR	Setup B min DRR	Setup B max DRR
Lecture Room 1	6.9	9.7	3.0	200	-0.82	15	0.87	7.9
Lecture Room 2	13.4	9.2	2.9	360	-0.37	13	-3.7	6.4
Meeting Room 1	6.6	4.7	3.0	92	-2.0	11	-3.1	7.6
Meeting Room 2	10.3	9.2	2.6	250	-2.6	11	1.1	12
Office 2	5.1	3.2	2.9	48	-0.44	13	-2.3	9.5

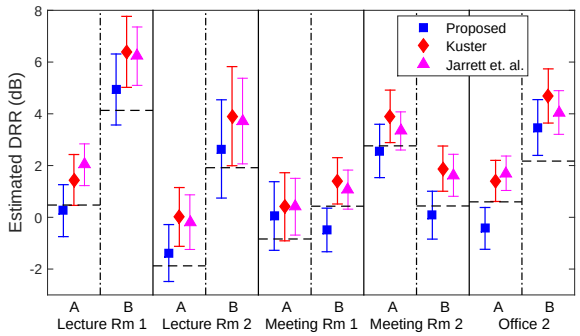


Fig. 3. Mean and standard deviation of estimated DRR using the proposed method (blue), Kuster's method (red) and Jarrett's method (pink) for all 5 rooms and 2 locations (A and B) in each room, with 18 dB SNR, averaged over 3 noise types. Dashed lines indicate ground truth DRR.

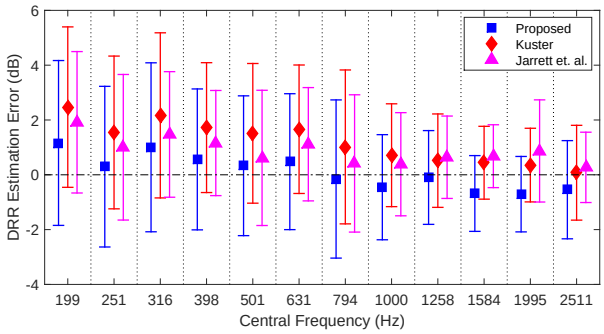


Fig. 4. Mean and standard deviation of subband DRR estimation error for all rooms and configurations with 18 dB SNR, using the proposed method (blue), Kuster's method (red) and Jarrett's method (pink).

standard deviation of the DRR estimations for each of the 10 room configurations. Only the recordings with 18 dB SNR are used for this analysis. In order to better evaluate the performance of the proposed method, both Kuster's [21] and Jarrett's [23] methods were implemented for comparison. In the case of Kuster's method, since the algorithm requires a pair of omnidirectional microphones placed very close to each other for recording, which is not available in the ACE challenge (other microphone array setups used in the ACE Challenge have a minimum spacing of 60 mm [25], which is too large for accurate measurement of particle velocity), we use the 0th and 1st order spherical harmonics in place of the sound pressure and particle velocity in calculation. Since it is shown that the spherical harmonics are equivalent to sound pressure and particle velocity, this implementation is expected to be representative of Kuster's method.

It can be observed from Fig. 3 that all three methods yield less than 3 dB mean error for all of the 10 room setups. The method proposed by Jarrett *et al.* shows a similar trend as that proposed by Kuster, but with a slightly lower estimation of DRR in most setups, as can be expected from Fig. 1. The proposed method, on the other hand, results in $1-3$ dB lower estimated DRR for most configurations.

A clear trend of DRR overestimation (estimated DRR higher than ground truth) can be observed for both methods that assume diffuse reverberant field. This is consistent with Kuster's observations from his experiments, where his method tend to overestimate DRR in real-life recording setups. The proposed method does not show any clear tendency of overestimation or underestimation, with 5 of the setups having positive mean error and the other 5 setups having negative mean error.

In terms of standard deviation, one would expect that the proposed method would yield higher standard deviation compared to Kuster's method, since in the proposed algorithm, both $|R_1|$ and R_2 need to be estimated for each frequency band, which would add uncertainty to the distribution of estimated DRR. However, from Fig. 3 it can be seen that the proposed algorithm yields almost identical standard deviation as Kuster's method, which indicates that the primary contributor of standard deviation is the coherence function γ^2, which is common for both the proposed method and Kuster's method.

On the other hand, Jarrett's method results in the lowest error standard deviation for all scenarios. The reason for this is that in the other two methods, only the first order spherical harmonics are used to calculate the coherence γ^2, while Jarrett's method utilizes all of the available spherical harmonic coefficients to reach a more consistent estimation of γ^2, which reduces its deviation due to random interference and other sources of error.

D. Subband results

The subband estimation results are shown in Fig. 4. In this figure, we plot the mean and standard deviation of the subband DRR estimation error using the proposed method as well as the two baseline methods. The error mean and standard deviation are averaged over the results from all 10 rooms, and once again only the 18 dB SNR recordings are used for the analysis. Only the DRR for the subbands with central frequency between 199 Hz and 2511 Hz are calculated.

From Fig. 4 it can be seen that in general, the mean error of the proposed method falls within 1 dB of the ground truth for all frequency bands. furthermore, the subband results below 1000 Hz show a different pattern than the subbands above 1000 Hz. Below 1000 Hz, the mean error are all

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2601222, IEEE/ACM Transactions on Audio, Speech, and Language Processing

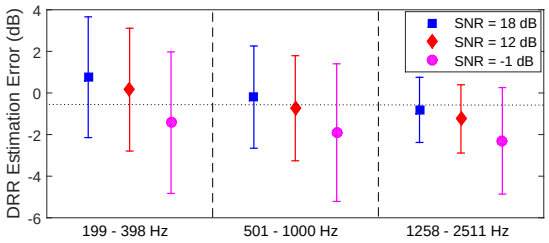


Fig. 5. Mean and standard deviation of DRR estimation error with 18dB, 12dB and −1dB SNR.

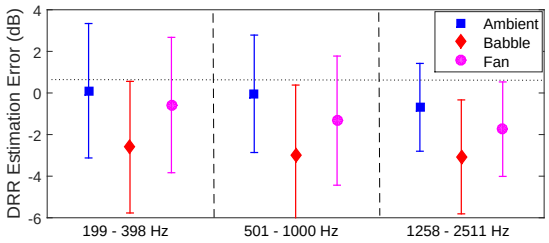


Fig. 6. Mean and standard deviation of DRR estimation error in multiple noisy environments with −1dB SNR.

positive, indicating a slight overestimation of DRR; the error standard deviation is approximately 3 dB for these subbands. On the other hand, for frequency bands above 1000 Hz, the mean error becomes negative; the standard deviation of estimation error reduces to 2 dB at 1000 Hz, and decreases further at higher frequencies. On the other hand, both Kuster's and Jarrett's methods show a clear trend of overestimation, this is especially significant for Kuster's method at lower frequencies. Jarrett's method yields lower DRR estimations compared to Kuster's, and in most frequency bands, have the lowest standard deviation.

Due to the geometry of the Eigenmike, only the 1st order spherical harmonics can be reliably captured for frequencies below 1000 Hz [32]. Below 1000 Hz, the 2nd order spherical harmonics are aliased onto the 1st order coefficients, and the aliasing error increases with frequency; at 1000 Hz and above, our algorithm begins to calculate the 2nd order coefficients, which removes the aliasing and improves the accuracy of the 1st order coefficients. Furthermore, at higher frequencies, the wavelength of the sound becomes closer to the dimension of the Eigenmike (8.4 cm diameter), which further increases the accuracy of 1st order spherical harmonic acquisition. This explains why the error standard deviation decreases gradually at higher frequencies.

Overall it can be seen that compared to the two baseline algorithms, the proposed method produces an unbiased DRR estimation. The standard deviation of the proposed algorithm is on par with Kuster's method, but slightly higher than Jarrett's method.

E. Impact of noise on DRR estimation

In order to examine the impact of noise (interference) on the result of DRR estimation, the algorithm is run for the Evaluation dataset recordings of each SNR setting (18dB, 12dB and −1dB), and we calculate the mean and standard deviation for each SNR setting, the results are shown in Fig. 5. In this figure,the subband results are separated into three frequency ranges: low (199-398 Hz), medium (501-1000 Hz) and high (1258-2511 Hz). Each frequency range covers four subbands, and the subband results are averaged within each frequency range, in order to simplify the data representation.

It can be seen from Fig. 5 that the difference between the estimation results with 18 dB and 12dB SNR is less than 1 dB. At −1 dB SNR, however, the DRR estimation becomes strongly biased towards underestimation. The cause of this

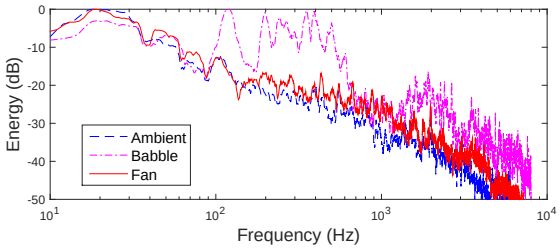


Fig. 7. Normalized power spectrum of the "Ambient", "Babble" and "Fan" noises in the ACE Evaluation Dataset.

phenomenon is that the interference/acoustic noise, which does not have the same impinging direction as the direct path signal, will reduce the coherence between sound pressure and gradient (particle velocity), resulting in a lower value of γ^2, thereby lowering the estimated DRR.

The other impact of high interference level is the increased error standard deviation. When developing and testing our algorithm using the ACE Development Dataset, we noticed that our frequency averaged MUSIC DOA algorithm became much less reliable at −1 dB SNR, compared to 18 dB and 12 dB SNR. A direct result of inaccurate DOA estimation is the decreased consistency of DRR estimations at different utterance/interference configurations in the same room setup, which is reflected by a higher error standard deviation. It is expected that if a more interference-robust DOA algorithm is applied, or if the DOA information can be measured directly, the proposed algorithm would produce more consistent estimations at low SNR.

How different types of interference affect the performance of the DRR estimation is also investigated. The three noise types mixed into the recordings each have different spectral characteristics, and therefore their effects on the subband DRR estimation vary. This is illustrated in Fig. 6, which plots the estimation results for the low, medium and high frequency ranges and for each of the three noise types. The SNR of all recordings used in this analysis are −1 dB.

From Fig. 6 it can be seen that the "Ambient" noise type has the least effect on DRR estimation accuracy causing only a small bias towards under estimation, while the "Babble" noise results in more than 3 dB of under estimation for all frequency ranges. The "Fan" noise has slightly more impact than the "Ambient" noise type, but less than that of the "Babble" noise. The cause of this result is due to both the spectral and spatial characteristics of the different noise types.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2601222, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE II
MEAN OF ESTIMATED PARAMETERS IN EACH ROOM CONFIGURATION AND FREQUENCY RANGE

Room	Setup	$\lvert R_1\rvert$			R_2		
		Low	Med	High	Low	Med	High
Lecture Room 1	A	0.280	0.194	0.219	0.288	0.251	0.265
	B	0.277	0.293	0.331	0.290	0.293	0.332
Lecture Room 2	A	0.232	0.201	0.146	0.316	0.268	0.290
	B	0.239	0.189	0.232	0.337	0.277	0.314
Meeting Room 1	A	0.191	0.120	0.157	0.294	0.248	0.253
	B	0.239	0.279	0.118	0.297	0.273	0.291
Meeting Room 2	A	0.226	0.265	0.278	0.201	0.329	0.321
	B	0.211	0.215	0.225	0.241	0.255	0.281
Office 2	A	0.268	0.167	0.174	0.252	0.269	0.286
	B	0.199	0.213	0.193	0.263	0.282	0.278

Fig. 7 plots the normalized power spectrum of the three noise types, the spectra are acquired by manually selecting the sections of recordings that contain purely noise signal. It can be seen that the "Ambient" noise consists of primarily low frequency signals that do not overlap with the speech signal spectrum. Therefore, the subbands of interest are most likely to have higher SNR than the full band SNR of -1 dB. As a result, the ambient noise has the least effect on the accuracy of DRR estimation. On the other hand, the "Babble" noise is essentially a speech recording by itself, therefore it almost completely overlaps with the spectrum of the speech of the talker, resulting in the lowest SNR in the speech spectrum of the three noise types. The "Fan" noise has very similar spectral characteristics as the "Ambient" noise type, although its higher frequency components have more energy than that of the "Ambient" noise, which leads to slightly more impact on DRR estimation.

According to the ACE Challenge description [25], the "Fan" noise is generated using one or two fans inside the recording environment, while the "Babble" noise records the voices of up to 7 people talking around the recording location. The "Ambient" noise is a recording of the ambient noise within the room. Due to the larger number of uncorrelated sources, each with a different DOA, the "Babble" noise is likely to have a lower coherence level than that of the "Fan" noise. Therefore when mixed into the speech recording, the "Babble" noise would lower γ^2 further than the "Fan" noise. Although the nature of the "Ambient" noise is unclear, in typical room environments its source is likely to be AC vents or windows, both of which can be considered as localized sources, thus creating a more coherent sound field than the "Babble" noise. In addition, due to its spectral characteristics, its impact on DRR estimation is the smallest of all three noise types.

F. Estimated parameters from the ACE Evaluation Dataset

The parameters $|R_1|$ and R_2 estimated for each subband of every speech recording in the ACE Evaluation Dataset has been recorded and is presented in Table. II, where we have taken the average values of $|R_1|$ and R_2 for the low, medium and high frequency ranges and for all the recordings from each room configuration, only the data from recordings with 18 dB SNR are used for this calculation.

As can be seen from Table II, although the values of R_1 and R_2 vary for each room configuration and frequency range, in general, $|R_1|$ falls within the range of 0.15-0.25, while R_2 lies in between 0.25-0.33 in the majority of cases. From Fig. 1, it can be seen that the values of $|R_1|$ and R_2 shown in Table II would lead to our proposed algorithm yielding lower DRR estimations than assuming $R_1 = 0, R_2 = 0.5$, which is indeed the case in our estimation results.

From the above results, we believe that setting $|R_1| = 0.2$ and $R_2 = 0.28$ provides a more reasonable and accurate model for a general reverberant sound field within room environments, compared to the diffuse model where it is assumed that $R_1 = 0$, and $R_2 = 1/2$ or $1/3$. It is sometimes easier to acquire or implement differential microphone pairs than complete first order microphone systems (such as the Eigenmike), and when a differential array is to be used to estimate room DRR, we suggest using (25) or (42) to calculate DRR, and assume $|R_1| = 0.2, R_2 = 0.28$, which is likely to yield more accurate estimation results.

VI. CONCLUSION

In this work, we present a novel algorithm for estimating DRR using a first order microphone system. We show that the proposed algorithm is a generalization of previous DRR estimation methods based on sound pressure-particle velocity coherence function. Using the proposed algorithm, it is possible to estimate the characteristics of a reverberant sound field which are relevant to DRR estimation, thereby improving the estimation accuracy of the method. We also show that at low frequency and small source-to-microphone distance, using the plane wave model for the direct path signal can result in a positive bias on the estimated DRR. Through validating the proposed algorithm using the ACE Challenge Dataset, it was found that the proposed algorithm provides ± 2 dB mean estimation error for the frequency range of human speech (200-2500 Hz), and shows no obvious bias.

APPENDIX A
PROOF OF THEOREM 1

The particle velocity $V_{\boldsymbol{x}}(\boldsymbol{x}_0, k)$ at position $\boldsymbol{x}_0$, in the direction $\boldsymbol{x}$, is related to the sound pressure by

$$V_{\boldsymbol{x}}(\boldsymbol{x}_0, k) = \frac{i}{k\rho_0 c} \frac{\partial P(\boldsymbol{x}_0, k)}{\partial \boldsymbol{x}}. \tag{45}$$

For the proof of (3), we consider the sound pressure at a point on the x-axis, whose coordinate in the spherical coordinate system is $(r, \pi/2, 0)$, the sound pressure can be decomposed using (1),

$$P(r, \frac{\pi}{2}, 0, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \beta_{nm}(k) j_n(kr) Y_{nm}(\frac{\pi}{2}, 0). \tag{46}$$

Taking the partial derivative of $P(r, \pi/2, 0, k)$ in the direction of r, which is equivalent to $\frac{\partial P(x,y,z)}{\partial x}$, we have

$$\frac{\partial P(r, \frac{\pi}{2}, 0, k)}{\partial r} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \beta_{nm}(k) \frac{\partial j_n(kr)}{\partial r} Y_{nm}(\frac{\pi}{2}, 0) \tag{47}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2601222, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Since we consider the partial derivative at the origin, we let $r \to 0$. Using the recurrent relationship [33]

$$nj_{n-1}(x) - (n+1)j_{n+1}(x) = (2n+1)\frac{d\ j_n(x)}{dx}, \tag{48}$$

and the fact that

$$j_n(0) = \begin{cases} 1, & \text{if } n = 0 \\ 0, & \text{if } n = 1, 2, 3... \end{cases} \tag{49}$$

It can be shown that

$$\lim_{r\to 0} \frac{\partial j_n(kr)}{\partial r} = \begin{cases} k/3, & \text{if } n = 1 \\ 0, & \text{otherwise.} \end{cases} \tag{50}$$

In addition, $Y_{10}(\pi/2, 0) = 0$. Therefore from (47) we have

$$\lim_{r\to 0} \frac{\partial P(r, \frac{\pi}{2}, 0, k)}{\partial r} = \frac{k}{3}\big(\beta_{11}Y_{11}(\frac{\pi}{2}, 0) + \beta_{1,-1}Y_{1,-1}(\frac{\pi}{2}, 0)\big) \tag{51}$$

Substituting (51) into (45) with the values $Y_{11}(\pi/2, 0) = Y_{1,-1}(\pi/2, 0) = \sqrt{3/8\pi}$ completes the proof.

For the proof of (4), we consider the partial derivative of sound pressure at $(r, \pi/2, \pi/2)$. The derivation is identical to that of $\frac{\partial P}{\partial x}$, except that $Y_{nm}(\pi/2, 0)$ are replaced by $Y_{nm}(\pi/2, \pi/2)$.

In the case of (5), we consider the partial derivative of sound pressure at $(r, 0, \phi)$ along r. Similar to (47), we can write

$$\frac{\partial P(r, 0, \phi, k)}{\partial r} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \beta_{nm}(k)\frac{\partial j_n(kr)}{\partial r}Y_{nm}(0, \phi). \tag{52}$$

Due to the fact that $Y_{11}(0, \phi) = 0$ and $Y_{1,-1}(0, \phi) = 0$, and utilizing (50), we can simplify (52), such that

$$\lim_{r\to 0} \frac{\partial P(r, 0, \phi, k)}{\partial r} = \frac{k}{3}\beta_{10}Y_{10}(0, \phi) \tag{53}$$

Substituting (53) into (45) with $Y_{10}(0, \phi) = \sqrt{3/4\pi}$ into completes the proof.

Appendix B
Spherical harmonics rotation

A method to rotate the calculated spherical harmonics so that the rotated coefficients correspond to the desired coordinate system is given as follows.

Our goal is to derive a transformation matrix $\boldsymbol{M}$, so that the original and transformed coefficients can be expressed using

$$\begin{bmatrix} \beta_{00} \\ \beta_{11} \\ \beta_{10} \\ \beta_{1,-1} \end{bmatrix} = \begin{bmatrix} M_{00}^{00} & M_{11}^{00} & M_{10}^{00} & M_{1,-1}^{00} \\ M_{00}^{11} & M_{11}^{11} & M_{10}^{11} & M_{1,-1}^{11} \\ M_{00}^{10} & M_{11}^{10} & M_{10}^{10} & M_{1,-1}^{10} \\ M_{00}^{1,-1} & M_{11}^{1,-1} & M_{10}^{1,-1} & M_{1,-1}^{1,-1} \end{bmatrix} \begin{bmatrix} C_{00} \\ C_{11} \\ C_{10} \\ C_{1,-1} \end{bmatrix}, \tag{54}$$

where β_{nm} and C_{nm} represent the spherical harmonic coefficients after and before rotation, respectively. The values of $M_{nm}^{n'm'}$ can be calculated using numerical integration [34],

$$M_{nm}^{n'm'} = \int_{\boldsymbol{s}} Y_{n'm'}(\boldsymbol{Rs})Y_{nm}^{*}(\boldsymbol{s})d\boldsymbol{s} \tag{55}$$

where $\boldsymbol{R}$ denotes the rotation matrix for the spherical coordinates. Since we are only concerned with the direction of the z-axis, we can decompose the rotation into two steps: a rotation along the z-axis by $-\varphi$, followed by a rotation along the y-axis by $-\vartheta$, where (ϑ, φ) denote the impinging direction of the direct path in the original coordinate system.

Appendix C
Proof of Equation (27)

The closed form expression of spherical Hankel functions of the first kind is [33]

$$h_n^{(1)}(z) = i^{-n-1}z^{-1}e^{iz}\sum_{0}^{n}(n + \frac{1}{2}, k)(-2iz)^{-k}. \tag{56}$$

The expression of $h_0^{(1)}(z)$ and $h_1^{(1)}(z)$ can then be written as

$$h_0^{(1)}(z) = -ie^{iz}\frac{1}{z} \tag{57}$$

$$h_1^{(1)}(z) = -e^{iz}\frac{z+i}{z^2}, \tag{58}$$

substituting (57) and (58) into (27) yields

$$\lim_{r_0\to\infty} \frac{h_1^{(1)}(kr_0)}{h_0^{(1)}(kr_0)} = \lim_{r_0\to\infty} \frac{kr_0 + i}{ikr_0} = \lim_{r_0\to\infty} \big(-i + \frac{1}{kr_0}\big) = -i, \tag{59}$$

which completes the proof.

References

[1] D. Griesinger, "The importance of the direct to reverberant ratio in the perception of distance, localization, clarity, and envelopment," in *Audio Engineering Society Convention 126*. Audio Engineering Society, 2009.

[2] K. Lebart, J. M. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[3] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with post-filtering," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, 1998.

[4] D. B. Hawkins and W. S. Yacullo, "Signal-to-noise ratio advantage of binaural hearing aids and directional microphones under different levels of reverberation," *Journal of Speech and Hearing Disorders*, vol. 49, no. 3, pp. 278–286, 1984.

[5] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 450–461, 2008.

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

[7] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: A summary of past and present research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 409–420, 2005.

[8] A. J. Kolarik, S. Cirstea, and S. Pardhan, "Evidence for enhanced discrimination of virtual auditory distance among blind listeners using level and direct-to-reverberant cues," *Experimental brain research*, vol. 224, no. 4, pp. 623–633, 2013.

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

[10] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. on Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, 2010.

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

[12] P. P. Parada, D. Sharma, T. van Waterschoot, and P. A. Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ACE challenge," in *In Proc. ACE Challenge Workshop, a satellite event of WASPAA, New Paltz, NY, USA*, Oct 2015.

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

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

[15] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *19th European Signal Processing Conference*, Aug 2011, pp. 1347–1351.

[16] O. Thiergart, G. Del Galdo, and E. A. Habets, "Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones." in *In Proc. International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 309–312.

[17] E. Georganti, J. Mourjopoulos, and S. van de Par, "Room statistics and direct-to-reverberant ratio estimation from dual-channel signals," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4713–4717.

[18] 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)*, May 2014, pp. 7440–7444.

[19] Y. Hioka and K. Niwa, "PSD estimation in beamspace for estimating direct-to-reverberant ratio from a reverberant speech signal," in *In Proc. ACE Challenge Workshop, a satellite event of WASPAA, New Paltz, NY, USA*, Oct 2015.

[20] 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 Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2374–2384, 2011.

[21] M. Kuster, "Estimating the direct-to-reverberant energy ratio from the coherence between coincident pressure and particle velocity," *The Journal of the Acoustical Society of America*, vol. 130, no. 6, pp. 3781–3787, 2011.

[22] T. D. Abhayapala and D. B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2. IEEE, 2002, pp. II–1949.

[23] D. P. Jarrett, O. Thiergart, E. A. P. Habets, and P. A. Naylor, "Coherence-based diffuseness estimation in the spherical harmonic domain," in *2012 IEEE 27th Convention of Electrical Electronics Engineers in Israel (IEEEI),*, Nov 2012, pp. 1–5.

[24] H. Chen, P. N. Samarasinghe, T. D. Abhayapala, and W. Zhang, "Estimation of the direct-to-reverberant energy ratio using a spherical microphone array." in *In Proc. ACE Challenge Workshop, a satellite event of WASPAA, New Paltz, NY, USA*, Oct 2015.

[25] J. Eaton, A. H. Moore, N. D. Gaubitch, and P. A. Naylor, "The ACE challenge - corpus description and performance evaluation," in *In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA*, Oct 2015.

[26] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 697–707, September 2001.

[27] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*. Springer, Mar. 2007.

[28] E. G. Williams, *Fourier Acoustics: Sound Radiation and Near field Acoustical Holography*. USA: Academic, 1999.

[29] F. Fahy, *Sound Intensity*. London: Elsevier Applied Science, 1989.

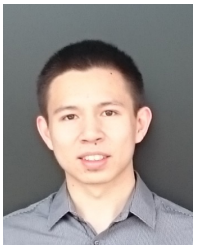
[30] T. D. Abhayapala and H. Bhatta, "Coherent broadband source localization by modal space processing," in *10th International Conference on Telecommunications (ICT 2003)*, vol. 2, 2003, pp. 1617–1623.

[31] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'09)*, Oct 2009, pp. 221–224.

[32] R. A. Kennedy, P. Sadeghi, T. D. Abhayapala, and H. M. Jones, "Intrinsic limits of dimensionality and richness in random multipath fields," *IEEE Trans. on Signal Processing*, vol. 55, no. 6, pp. 2542–2556, 2007.

[33] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation, 1964, no. 55, p. 439.

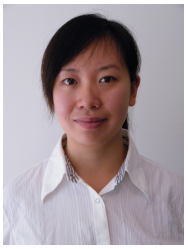
[34] J. Kautz, J. Snyder, and P.-P. J. Sloan, "Fast arbitrary BRDF shading for low-frequency lighting using spherical harmonics." *Rendering Techniques*, vol. 2, pp. 291–296, 2002.



Hanchi Chen received the B.E. degree (with first class honours) in Electronics and Telecommunication Engineering in 2012, from the Australian National University (ANU), Canberra. He is currently a pursuing a Ph.D. degree at ANU in the field of spatial audio signal processing. His research interests include spatial audio and multi-channel processing techniques, especially active noise control.

Prof. Thushara Abhayapala received the B.E. degree (with Honors) in Engineering in 1994 and the Ph.D. degree in Telecommunications Engineering in 1999, both from the Australian National University (ANU), Canberra. Currently he is the Deputy Dean of the College of Engineering & Computer Science, ANU. He was the Director of the Research School of Engineering at ANU from January 2010 to October 2014 and the Leader of the Wireless Signal Processing (WSP) Program at the National ICT Australia (NICTA) from November 2005 to June 2007. His research interests are in the areas of spatial audio and acoustic signal processing, and multi-channel signal processing. He has supervised over 30 PhD students and co-authored over 200 peer reviewed papers. Professor Abhayapala is an Associate Editor of IEEE/ACM Transactions on Audio, Speech, and Language Processing. He is a Member of the Audio and Acoustic Signal Processing Technical Committee (2011-2016) of the IEEE Signal Processing Society. He is a Fellow of the Engineers Australia (IEAust).

Prasanga Samarasinghe received a B.E. degree (with honors) in electronic and electrical engineering from the University of Peradeniya, Sri Lanka, in 2009. She completed a Ph.D. degree at the Australian National University (ANU), Canberra, in 2014. She is currently a Research Fellow at the Research School of Engineering at ANU, and her research interests are spatial sound recording and reproduction, spatial noise cancellation, and array optimization using compressive sensing techniques.



Wen Zhang (S'06–M'09) received the B.E. degree in telecommunication engineering from Xidian University, Xian, China, in 2003, the M.E. degree in electrical engineering (with first class honours) and the Ph.D. degree from the Australian National University, Canberra, in 2005 and 2010, respectively. From 2010 to 2012, she was a Postdoctoral Fellow at CSIRO Process Science and Engineering in Sydney, Australia. She is currently working as a Research Fellow in the College of Engineering and Computer Science at the Australian National University. Her research interests include spatial audio, binaural source localization, and active noise control. She is awarded the Discovery Early Career Researcher Award (DECRA) fellowship by the Australian Research Council in 2015.