JASA EXPRESS LETTERS

https://doi.org/10.1121/10.0002144

Blind estimation of the direct-to-reverberant ratio using a beta distribution fit to binaural coherence

Paul Calamia,[a] **Nava Balsam, and Philip Robinson**
Facebook Reality Labs Research, Redmond, Washington 98052, USA
pcalamia@fb.com, nava@fb.com, philrob22@fb.com

Abstract: Knowledge of the direct-to-reverberant ratio (DRR) can be useful in various acoustic and audio applications. While the DRR can be computed easily from a room impulse response (RIR), blind estimation using sources of opportunity is necessary when such RIRs are not available. This paper describes a method for blind estimation of the DRR which involves fitting a beta distribution to the magnitude-squared coherence between two binaural audio signals, aggregated over time and frequency. Validation experiments utilizing speech convolved with binaural RIRs yield DRR estimates that are within the just-noticeable difference for DRRs in the range −15 to +18 dB.
https://doi.org/10.1121/10.0002144

[Editor: S. K. Tang] Pages: EL359–EL364

Received: 7 July 2020 **Accepted:** 22 September 2020 **Published Online:** 14 October 2020

1. Introduction

Knowledge of the direct-to-reverberant ratio (DRR) between an acoustic source and a listening position or receiver can be useful in a number of acoustic and audio applications, such as dereverberation,[1] source distance estimation,[2] and automatic speech recognition.[3] The DRR is also known to be a *perceptual* cue for source distance and the sense of reverberance.[4] Therefore, appropriate DRR values also are important to promote externalization of virtual sources in augmented-reality environments.[5]

Given a room impulse response (RIR), the DRR is defined as, "At a given location, the ratio of the sound pressure level of a direct sound from a directional source to the reverberant sound pressure level simultaneously incident to the same location."[6] Mathematically, it is computed from a discrete-time RIR, $h(n)$, as follows:

$$\mathrm{DRR} = 10 * \log_{10}\left(\frac{\sum_{n=n_d-n_0}^{n_d+n_0} h^2(n)}{\sum_{n=n_d+n_0}^{\infty} h^2(n)}\right), \tag{1}$$

where n_d is the sample index of the peak of the direct sound arrival and n_0 is the number of samples corresponding to a small temporal window, typically covering a range of 1.0 to 2.5 ms.[7] When $h(n)$ is known, computation of the DRR is straightforward from Eq. (1). However, when the RIR is not available, the DRR can be estimated blindly using acoustic sources of opportunity.

Blind estimation of DRR can be separated into a variety of categories, e.g., single-channel[8] and multichannel approaches,[9] traditional signal-processing[10] and machine-learning[11] approaches, and algorithms exploiting various signal features including spectral standard deviation,[12] modulation energy,[13] and coherence.[1] Eaton *et al.* provide a recent summary of DRR estimation methods in the context of the ACE Challenge for blind estimation of room-acoustic parameters.[14]

In this paper, we present a DRR estimation algorithm that exploits the statistics of binaural magnitude-squared coherence (MSC). Binaural MSC values, aggregated over time and frequency, are fit with a beta distribution, and an estimation model is developed from the relationship between the shape parameters of the distribution and the DRR. The performance of the estimator is evaluated both numerically as well as perceptually, the latter with respect to published just-noticeable differences for DRR.

[a]Author to whom correspondence should be addressed, ORCID: 0000-0002-0401-6996.

2. Estimation algorithm

2.1 Binaural coherence

Our DRR estimation relies on the magnitude-squared coherence between two signals collected with binaural microphones.[15] We assume a single sound source at azimuth and elevation angles of (0°,0°) relative to the listener. We compute the time- and frequency-dependent MSC, $|\Gamma_{Y_L Y_R}(k,n)|^2$, using a short-time discrete Fourier transform (STFT), $Y_{L/R}(k,n)$, of the left and right input signals. Specifically, following Zohourian and Martin,[10] we compute

$$\Gamma_{Y_L Y_R}(k,n) = \frac{\hat{\Phi}_{Y_L Y_R}(k,n)}{\sqrt{\hat{\Phi}_{Y_L Y_L}(k,n)\hat{\Phi}_{Y_R Y_R}(k,n)}} \tag{2}$$

from the temporally smoothed cross-spectrum and power spectra

$$\hat{\Phi}_{Y_m Y_{m'}}(k,n) = 0.7\hat{\Phi}_{Y_m Y_{m'}}(k,n-1) + 0.3\,Y_m(k,n)\,Y^*_{m'}(k,n) \tag{3}$$

for $m,m' \in \{L,R\}$, where k is the frequency index and n is the STFT time-frame index. $\hat{\Phi}_{Y_m Y_{m'}}(k,0)$ is computed with only the second term in Eq. (3). Example time/frequency matrices of MSC, computed using a 10-s speech sample convolved with three binaural room impulse responses (BRIRs) with increasing DRR, are shown in Figs. 1(a)–1(c). The same data, collapsed over time and frequency, are shown in Figs. 1(d)–1(f) in histogram format. While there is not a 1:1 mapping between the MSC matrices and the histograms, we exploit the fact that these matrices, when computed from binaural speech samples with similar DRR values, collapse to similar histogram shapes.

2.2 Beta distribution

The natural restriction of MSC values to the range [0, 1], as well as the shapes of the MSC histograms in Figs. 1(d)–1(f), suggest that the collapsed MSC values can be described with a beta distribution, the probability density function (PDF) of which is defined on the interval [0, 1] with a pair of shape parameters a and b,

$$f(x;a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \tag{4}$$

where $B(a, b)$ is the beta function which serves as a normalization constant. Different combinations of the shape parameters allow for skewed PDFs with concentration on either end of the [0, 1] range, as well as a uniform distribution, all of which can approximate the distributions of MSC in different environments with varying DRRs. A scaled beta PDF, fit to the MSC data, is shown with a dashed line in each of Figs. 1(d)–1(f).

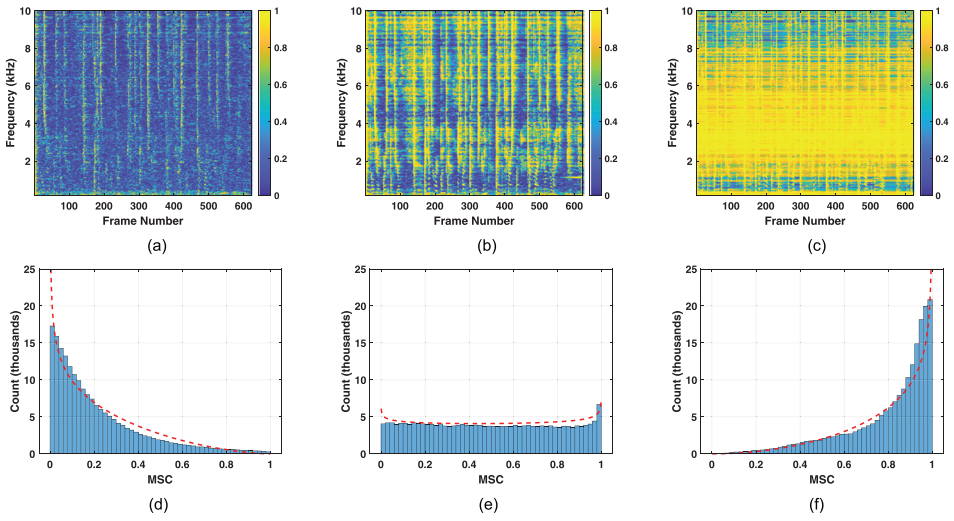

Fig. 1. Example time/frequency matrices and histograms of MSC for speech convolved with BRIRs of different DRR values: (a), (d) −15 dB; (b), (e) 4 dB; (c), (f) 12.5 dB. The dashed line on each histogram represents a scaled beta PDF fit to the MSC data. See Sec. 2.2.

https://doi.org/10.1121/10.0002144

2.3 The DRR estimator

Given the relationship between the DRR and the shape of the best-fit beta distribution to MSC values collapsed over time and frequency, we explored various mappings from combinations of the shape parameters a and b to DRR. One logical candidate is the skewness,

$$\gamma_1 = \frac{2(b-a)\sqrt{a+b+1}}{(a+b+2)\sqrt{ab}}, \tag{5}$$

since it is precisely the asymmetry of the distribution that we want to exploit. Figure 2(a) depicts DRR as a function of beta skewness for 700 binaural signals (see Sec. 3 for details on these signals). Somewhat surprisingly, though, the simpler mapping from b to DRR led to a more accurate estimator.[16] The relationship between b and DRR is shown in Fig. 2(b) where the dashed line represents a two-term power series fit to the data ($p_0 b^{p_1} + p_2$) with $p_0 = -58.17$, $p_1 = 0.22$, and $p_2 = 60.39$. Using training data to learn this fit, we can collect a new sample of binaural data, compute the MSC between the channels over time and frequency, compute the shape parameter b from a beta PDF fit to the MSC distribution, and estimate the DRR for that new sample.

3. Experimental evaluation

BRIRs used to evaluate the DRR prediction model were collected from a variety of publicly available datasets, as well as from internal measurements, as indicated in Table 1. The complete set comprised 70 BRIRs, all measured with the source near 0° azimuth and elevation, spanning a DRR range from −15.1 to 17.8 dB. For each BRIR, DRR values were computed for each channel (ear) according to Eq. (1), and averaged to provide a single value. Monaural, anechoic speech samples from the ACE Challenge dataset[14] were merged into a continuous audio stream which was sliced into 10-s segments, and these segments were convolved with the BRIRs to provide input data for the algorithm. Ten randomly chosen speech samples were used with each BRIR for a total of 700 examples. All BRIRs and speech segments were sampled at 48 kHz.

For each example, the short-time discrete Fourier transform was computed with 32-ms windows, 50% overlap, a Hann window, and FFT size equal to the window size (1536 samples). Frequency bins between 200 Hz and 10 kHz were retained, resulting in 196 560 time/frequency points (315 frequency bins × 624 time frames) for each channel. The MSC between the channels was computed using Eqs. (2) and (3). We originally fit the collapsed MSC data using MATLAB's® *betafit* function to compute the shape parameter b, but found a simple calculation using the mean μ and variance σ^2 of the data to provide equivalent results,

$$b = (1-\mu)\cdot\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right). \tag{6}$$

We then fit a two-term power-series model to a subset of the resulting 700 pairs of DRR and b values, and tested that model with the remaining pairs using a tenfold cross-validation scheme.

4. Results

Example results from our prediction algorithm are shown in Fig. 3. Figure 3(a) contains the DRR estimates for 700 10-s binaural speech signals generated through the process described in Sec. 3: the model was trained with 630 examples (63 of the 70 BRIRs, each convolved with 10 speech samples) and asked to predict the DRR from 70 signals comprising the remaining 7

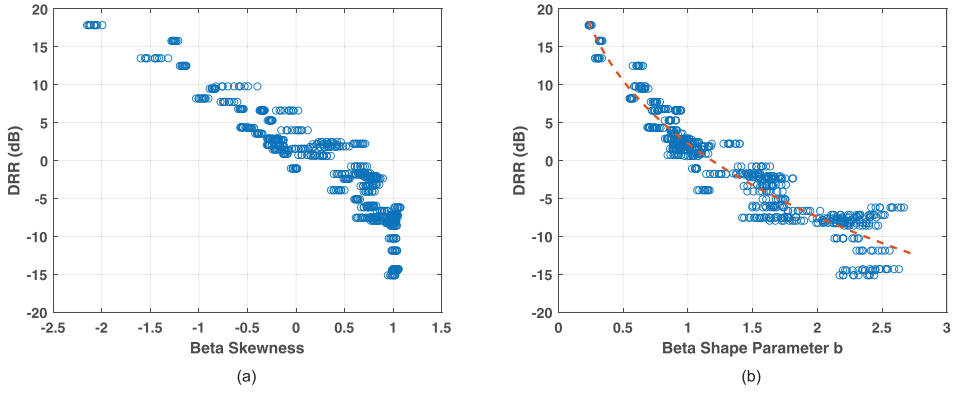

Fig. 2. DRR as a function of the (a) skewness and (b) shape parameter b for a beta distribution with 700 examples. The dashed line in (b) represents a two-term power-series fit to the data.

Table 1. BRIR datasets used for algorithm evaluation.

BRIR Source	Head type	No. Rooms	No. BRIRs	DRR Range (dB)	RT (s)
Internal	KEMAR	1	6	−7.9 to −1.9	0.83
BRAS (Ref. 17)	FABIAN	4	17	−8.2 to 2.4	1.17–1.93
AIR (Ref. 18)	Head Acoustics HMS2	5	14	−7.5 to 9.7	0.31–0.94
Oldenburg (Ref. 19)	Brüel & Kjær Type 4128 C	3	6	−5.1 to 17.8	0.06–0.39
Pori (Ref. 20)	Brüel & Kjær HATS custom fit with DPA Type 4053 mics	1	12	−15.1 to 2.2	1.66
IoSR (Ref. 21)	Cortex MKII	1	10	0.9 to 4.4	0.23
Salford/BBC (Ref. 22)	Brüel & Kjær HATS	1	5	2.9 to 12.5	0.24

BRIRs each convolved with 10 samples of unseen speech. 1000 iterations of this process (train on a randomly selected set of 63 BRIRs, test with the remaining 7) yielded an average root-mean-square error (RMSE) of 2.30 dB (min. RMSE = 2.23 dB, max. RMSE = 2.48 dB, median RMSE = 2.30 dB). The shaded gray area represents the just-noticeable difference (JND) for DRR as reported by Larsen *et al.*[4] (see their Fig. 3). The black line represents perfect performance. Prediction performance can be seen to deteriorate slightly at DRRs below approximately −10 dB [as a result of the variance of DRR values around $b \approx 2.5$ in Fig. 2(b)], however, the results remain well within the JND.

While the performance of the algorithm appears to be, for the most part, independent of the BRIR dataset, four of the BRIRs for which the estimates have high error [cyan data points in Fig. 3(a) near measured DRRs of −15, −6.2, and 2.2 dB] come from the Pori dataset.[20] All of these BRIRs were measured in a concert hall, the largest of the rooms within our collection of BRIRs. Strong early reflections from a nearby sidewall may be responsible for skewing the MSC distributions toward higher values and thus causing an overestimation of DRR for some of these cases, but further analysis is necessary to confirm what geometric and/or acoustic characteristics of this room, and the specific measurement positions within it, may be responsible for the reduced estimation performance.

In Fig. 3(b), we compare our results to those from *our implementation* of the estimator by Zohourian and Martin[10] applied to the same 700 examples. Note the improved performance below 0 dB DRR, where their model tends to over-predict the DRR and the error exceeds the JND for DRRs below approximately −5 dB. This characteristic of over-predicting low DRRs also is evident in their Fig. 4 (albeit when the test samples were speech convolved with *simulated* BRIRs with the source at 45° azimuth).

Figure 3(c) shows the performance of the model, in terms of RMSE, as a function of speech sample length. Each box represents 10 iterations of predictions with 70 BRIRs, each convolved with 10 samples of speech of the designated length. The red, central mark indicates the median value, and the box extents indicate the 25th and 75th percentiles. The whiskers indicate the full data extents excluding outliers, which are marked with red "+" symbols. These results suggest that speech samples as short as 4 s, and possibly 2 s, may provide sufficient data for successful DRR predictions, although evaluation with respect to the JND should be done to confirm

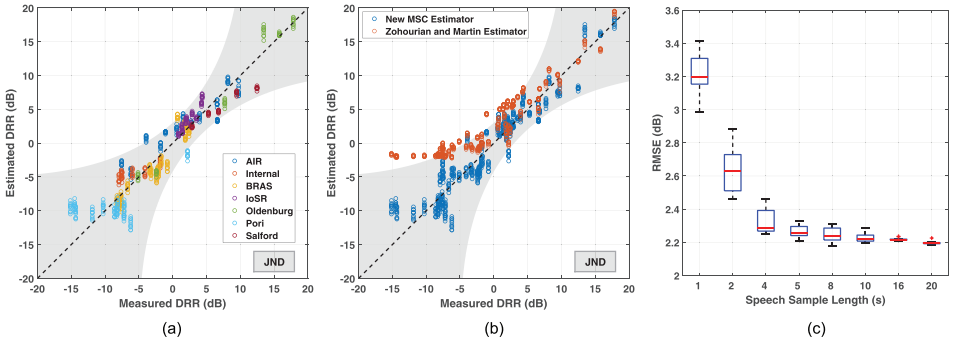

Fig. 3. Example DRR prediction results. (a) tenfold cross validation results using measured BRIRs and 10-s speech samples (train on 630 examples, test on 70 unseen examples). The shaded area indicates the DRR-dependent just-noticeable difference. (b) Same data as in (a) with the addition of estimates from the Zohourian and Martin model (Ref. 10) in red. (c) Prediction performance as a function of speech sample length.

https://doi.org/10.1121/10.0002144

this. One factor that may have influenced the performance of the algorithm, particularly with the shorter speech samples, is the lack of a voice-activity detector (VAD) in the processing pipeline. Without a VAD, the presence of pauses between utterances in the shorter samples may have significantly skewed the MSC distribution and thus degraded the estimates.

5. Summary and conclusions

This paper presents an approach to estimate the direct-to-reverberant ratio blindly from binaural speech signals based on the distribution of magnitude-squared coherence values, computed over time and frequency, between the two channels. Given short segments of speech captured with binaural microphones, we build a prediction model by fitting each set of MSC values with a beta distribution, and map the shape parameter b from the distributions to the DRR with a two-term power series. Initial validation of the model was carried out with 700 examples comprising 70 measured BRIRs, spanning a DRR range from -15 to $+18$ dB, each convolved with ten 10-s speech samples. Tenfold cross validation with these examples yielded results with an overall RMSE of 2.3 dB, nearly all of which fell within the just-noticeable difference for DRR (which is a function of the DRR value). Further evaluation suggests that speech samples as short as 2 s may be sufficient for an acceptable level of estimation performance.

We currently are exploring several directions for future work. First, similar to reverberation time (RT), DRR is a frequency-dependent parameter, so band-limited predictions should be considered. Preliminary work on this aspect (not shown) suggests a small increase in the RMSE of the predictions, although this may be mitigated by the fact that the JND for DRR based on narrow-band signals is higher than that for broadband signals.[4] Second, the performance of this algorithm should be evaluated with varying signal-to-noise ratios, which can be accomplished by adding binaural noise with the appropriate RT and coherence to the training and test examples. Third, this estimator should be tested with live binaural speech signals rather than recorded speech convolved with measured BRIRs. To this end we have implemented the algorithm on the Bela embedded computing platform[23] and currently are planning a data-collection campaign. Fourth, we opted for a model whose input is restricted to acoustic sources at 0° azimuth and elevation to avoid the added complexity of requiring a direction-of-arrival estimate and compensation for interaural differences based on that estimate. However, generalization of our model to arbitrary source locations may be a valuable extension. Fifth, we considered only a small number of mappings from both beta shape parameters a and b to DRR before adopting the model in Sec. 2.3 which utilizes only b. However, there may be a function DRR $= f(a, b)$ that outperforms our current estimator, and we are currently developing a simple machine-learning architecture to learn this function. Finally, there is evidence that DRR and RT are not independent,[11] and there is previous work using coherence to estimate RT,[24] which suggests that joint coherence-based estimation of these two room-acoustics parameters may be beneficial.

References and links

[1]A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," IEEE/ACM Trans. Audio Speech Lang. Proc. **23**(6), 1006–1018 (2015).

[2]Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," IEEE Trans. Audio Speech Lang. Proc. **18**(7), 1793–1805 (2010).

[3]A. Brutti and M. Matassoni, "On the use of early-to-late reverberation ratio for ASR in reverberant environments," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2014), pp. 4638–4642.

[4]E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," J. Acoust. Soc. Am. **124**(1), 450–461 (2008).

[5]T. Sporer, S. Werner, and F. Klein, "Adjustment of the direct-to-reverberant-energy-ratio to reach externalization within a binaural synthesis system," in *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality* (2016).

[6]ANSI/ASA S1.1, *Acoustical Terminology* (American National Standards Institute, New York, 2013).

[7]The size of the window around the direct sound is influenced by the initial time-delay gap as well as the type of receiver used in the measurement. For the latter, a longer window may be used with a binaural manikin to allow the direct sound to scatter around the head.

[8]C. S. Doire, M. Brookes, P. A. Naylor, D. Betts, C. M. Hicks, M. A. Dmour, and S. H. Jensen, "Single-channel blind estimation of reverberation parameters," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), pp. 31–35.

[9]Y. Hioka and K. Niwa, "PSD estimation in beamspace for estimating direct-to-reverberant ratio from a reverberant speech signal," arXiv:1510.08963 (2015).

[10]M. Zohourian and R. Martin, "Binaural direct-to-reverberant energy ratio and speaker distance estimation," IEEE/ACM Trans. Audio Speech Lang. Proc. **28**, 92–104 (2019).

[11]F. Xiong, S. Goetze, B. Kollmeier, and B. T. Meyer, "Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals," IEEE/ACM Trans. Audio Speech Lang. Proc. **27**(2), 255–267 (2018).

[12]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)* (2014), pp. 4713–4717.
[13]S. Braun, J. F. Santos, E. A. Habets, and T. H. Falk, "Dual-channel modulation energy metric for direct-to-reverberation ratio estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), pp. 206–210.
[14]J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE Challenge," IEEE/ACM Trans. Audio Speech Lang. Proc. **24**(10), 1681–1693 (2016).
[15]We expect the method will work with other two-microphone configurations but we did not explore them.
[16]The use of skewness in the estimator led to an RMSE in prediction of 2.62 dB, while the use of the shape parameter b led to an RMSE of 2.30 dB. See Sec. 4.
[17]L. Aspöck, F. Brinkmann, D. Ackermann, S. Weinzierl, and M. Vorländer, "BRAS—Benchmark for Room Acoustical Simulation," (2019), http://dx.doi.org/10.14279/depositonce-6726.2 (Last viewed 10/16/2019).
[18]M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *2009 16th International Conference on Digital Signal Processing* (2009), pp. 1–5.
[19]H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," EURASIP J. Adv. Sign. Process. **2009**(1), 298605.
[20]J. Merimaa, T. Peltonen, and T. Lokki, "Concert Hall Impulse Responses—Pori, Finland: Reference" (2005), http://legacy.spa.aalto.fi/projects/poririrs/ (Last viewed 10/16/2019).
[21]J. Francombe, "IoSR listening room multichannel BRIR dataset" (2016), http://epubs.surrey.ac.uk/813511/ (Last viewed 3/13/2020).
[22]D. Satongar, Y. W. Lam, and C. Pike, "Measurement and analysis of a spatially sampled binaural room impulse response dataset," in *21st International Congress on Sound and Vibration* (2014), pp. 1–8.
[23]A. McPherson and V. Zappi, "An environment for submillisecond-latency audio and sensor processing on BeagleBone Black," in *Audio Engineering Society Convention* (2015), p. 138.
[24]R. Scharrer and M. Vorländer, "Blind reverberation time estimation," in *Proceedings of the International Conference on Acoustics*, Sydney, Australia (2010).