DIRECT-TO-REVERBERANT RATIO ESTIMATION USING A NULL-STEERED BEAMFORMER

James Eaton[†]*, Alastair H. Moore*[†]*, Patrick A. Naylor*[†]*, and Jan Skoglund*[‡]

[†]Department of Electrical and Electronic Engineering, Imperial College, London, UK
[‡]Google, Inc., Mountain View, CA, USA

{james.eaton, alastair.h.moore, p.naylor}@imperial.ac.uk, jks@google.com

ABSTRACT

Reverberation affects the quality and intelligibility of distant speech recorded in a room. Direct-to-Reverberant Ratio (DRR) is a useful measure for assessing the acoustic configuration and can be used to inform dereverberation algorithms. We describe a novel DRR estimation algorithm applicable where the signal was recorded with two or more microphones, such as mobile communications devices and laptops. The method uses a null-steered beamformer. In simulations the proposed method yields accurate DRR estimates to within ± 4 dB across a wide variety of room sizes, reverberation times and source-receiver distances. It is also shown that the proposed method is more robust to background noise than a baseline approach. The best estimation accuracy is obtained in the region from -5 to 5 dB which is a relevant range for portable devices.

Index Terms— speech enhancement, speech dereverberation, beamforming

1. INTRODUCTION

Determining the acoustic characteristics of an environment is important for speech enhancement and recognition. Speech enhancement algorithm performance can typically be improved if the level of reverberation relative to the speech is known [1]. When the Acoustic Impulse Response (AIR) is available, the Direct-to-Reverberant Ratio (DRR) can be estimated from the impulse response by examining the onset and decay characteristics of the AIR. However, when the AIR is not available the DRR must be estimated from the speech. Portable communications devices such as laptops and smartphones are increasingly incorporating multiple microphones enabling the use of multi-channel algorithms.

Most of the recent approaches to non-intrusive DRR estimation use the spatial coherence between channels to estimate the reverberation, which assumes that all non-coherent energy is reverberation [2, 3, 4, 5]. Falk *et al.* [6] on the other hand uses modulation spectrum features which requires a mapping which is trained on speech. In the related task of reverberation time estimation, Dumortier and Vincent [7] propose using spatial selectivity to enhance the reverberant signal such that its dynamics can be observed more clearly. A similar approach might also be applied to DRR estimation.

The contribution of this paper is to propose a novel DRR estimation approach where we use spatial selectivity to separate direct and reverberant energy and account for noise separately. The formulation considers the response of the beamformer to reverberant sound and the effect of noise. The remainder of this paper is organised as follows: In section 2 we present the method. In section 3 we evaluate the performance, and in section 4 we compare the results with [2]. Finally, we present our conclusions in Section 5.

2. METHOD

2.1. Acoustic model

A continuous speech signal, $s(t)$, radiating from a given position in a room will follow multiple paths to any observation point comprising the direct path as well as reflections from the walls, floor, ceiling and the surfaces of other objects in the room. The reverberant signal $y_m(t)$, captured by the m-th microphone in an array of M microphones in the room is characterised by the AIR, $h_m(t)$, of the acoustic channel between the source and the microphone such that

$$y_m(t) = h_m(t) * s(t) + v_m(t), \tag{1}$$

where $v_m(t)$ is the additive noise at the microphone. The AIR is a function of the geometry of the room, the reflectivity of the surfaces in the room, and the microphone locations. Let

$$h_m(t) = h_{d,m}(t) + h_{r,m}(t), \tag{2}$$

where $h_{d,m}(t)$ and $h_{r,m}(t)$ are the impulse responses of the direct and reverberant paths for the m-th microphone respectively. The DRR at the m-th microphone, $\bar{\eta}_m$, is the ratio of the power arriving directly at the microphone from the source to power arriving after being reflected from one or more surfaces in the room [8]. It can be written as

$$\bar{\eta}_m = \frac{\int |h_{d,m}(t)|^2 dt}{\int |h_{r,m}(t)|^2 dt}. \tag{3}$$

When the impulse response is convolved with a speech signal, the observation at the m-th microphone is the Signal-to-Reverberation Ratio (SRR), γ, given by

$$\gamma_m = \frac{E\{|(h_{d,m}(t))^T * s(t)|^2\}}{E\{|(h_{r,m}(t))^T * s(t)|^2\}}. \quad (4)$$

The SRR is equal to the DRR in the case when $s(t)$ is spectrally white. The aim of non-intrusive or blind DRR estimation is to estimate η_m from the observed signals. In our approach we use spatial selectivity to separate the direct and reverberant components of the sound field.

2.2. Beamforming in the frequency domain

Spatial filtering or beamforming uses a weighted combination of two or more microphone signals to achieve a particular directivity pattern. The output, $Z(j\omega)$, of a beamformer in the complex frequency domain is given by [9]

$$Z(j\omega) = (\mathbf{w}(j\omega))^T \mathbf{y}(j\omega), \quad (5)$$

where $\mathbf{w}(j\omega) = [W_0(j\omega), W_1(j\omega), \ldots, W_{M-1}(j\omega)]^T$ is the vector of complex weights for each microphone, and $\mathbf{y}(j\omega) = [Y_0(j\omega), Y_1(j\omega), \ldots, Y_{M-1}(j\omega)]^T$ is the vector of microphone signals.

Let the signal at the m-th microphone due to a unit plane wave incident on the microphone be $X_m(j\omega, \Omega)$, where $\Omega = (\phi, \theta)$ is the Direction-of-Arrival (DoA), and θ and ϕ are the azimuth and elevation, respectively. The beam-pattern of the beamformer is

$$B(j\omega, \Omega) = (\mathbf{w}(j\omega))^T \mathbf{x}(j\omega, \Omega), \quad (6)$$

where

$$\mathbf{x}(j\omega, \Omega) = [X_0(j\omega, \Omega), X_1(j\omega, \Omega), \ldots, X_{M-1}(j\omega, \Omega)]^T.$$

2.3. Estimation of DRR in the frequency domain

We shall now consider how to use the beamformer to estimate DRR. From (1) and (2), the signal at microphone m in the frequency domain is defined as

$$Y_m(j\omega) = D_m(j\omega) + R_m(j\omega) + V_m(j\omega) \quad (7)$$

where

$$D_m(j\omega) = H_{m,d}(j\omega)S(j\omega),$$

and

$$R_m(j\omega) = H_{m,r}(j\omega)S(j\omega).$$

From (5),

$$Z_y(j\omega) = Z_d(j\omega) + Z_r(j\omega) + Z_v(j\omega), \quad (8)$$

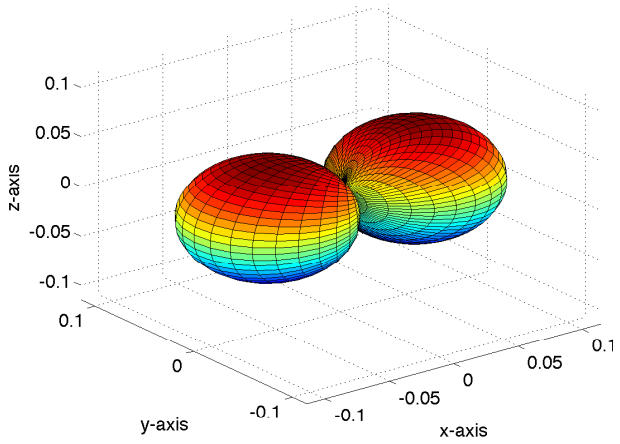

Fig. 1. *2-channel null-steered beamformer gain and directivity pattern at* 200 Hz *with a microphone spacing of* 62 mm. *The maximum gain is* -9.4 dB.

where

$$Z_d(j\omega) = (\mathbf{w}(j\omega))^T \mathbf{d}(j\omega),$$
$$Z_r(j\omega) = (\mathbf{w}(j\omega))^T \mathbf{r}(j\omega),$$
$$Z_v(j\omega) = (\mathbf{w}(j\omega))^T \mathbf{v}(j\omega),$$

and

$$\mathbf{d}(j\omega) = [D_0(j\omega), D_1(j\omega), \ldots, D_{M-1}(j\omega)]^T,$$

and $\mathbf{r}(j\omega)$ and $\mathbf{v}(j\omega)$ are similarly defined. We choose $\mathbf{w}(j\omega)$ such that $Z_d(j\omega) = 0$, thus

$$Z_y(j\omega) = Z_r(j\omega) + Z_v(j\omega). \quad (9)$$

We assume that the reverberant energy is the same at all microphones

$$E\{|R(j\omega)|^2\} = E\{|R_m(j\omega)|^2\} \; \forall m = 1 : M, \quad (10)$$

and that the reverberant sound field is isotropic, i.e. it is composed of plane waves arriving from all directions with equal probability and magnitude. The output of the beamformer due to reverberant energy is thus

$$E\{|Z_r(j\omega)|^2\} = G^2(j\omega) E\{|R(j\omega)|^2\}, \quad (11)$$

where $E\{\cdot\}$ is the expectation operator, and from (6),

$$G^2(j\omega) = \int_\Omega |B(j\omega, \Omega)|^2 \, d\Omega. \quad (12)$$

Substituting (9) into (11) and rearranging, assuming that the reverberation and noise signals are uncorrelated, gives

$$E\{|R(j\omega)|^2\} = \frac{1}{G^2(j\omega)} \left(E\{|Z_y(j\omega)|^2\} - E\{|Z_v(j\omega)|^2\} \right). \quad (13)$$

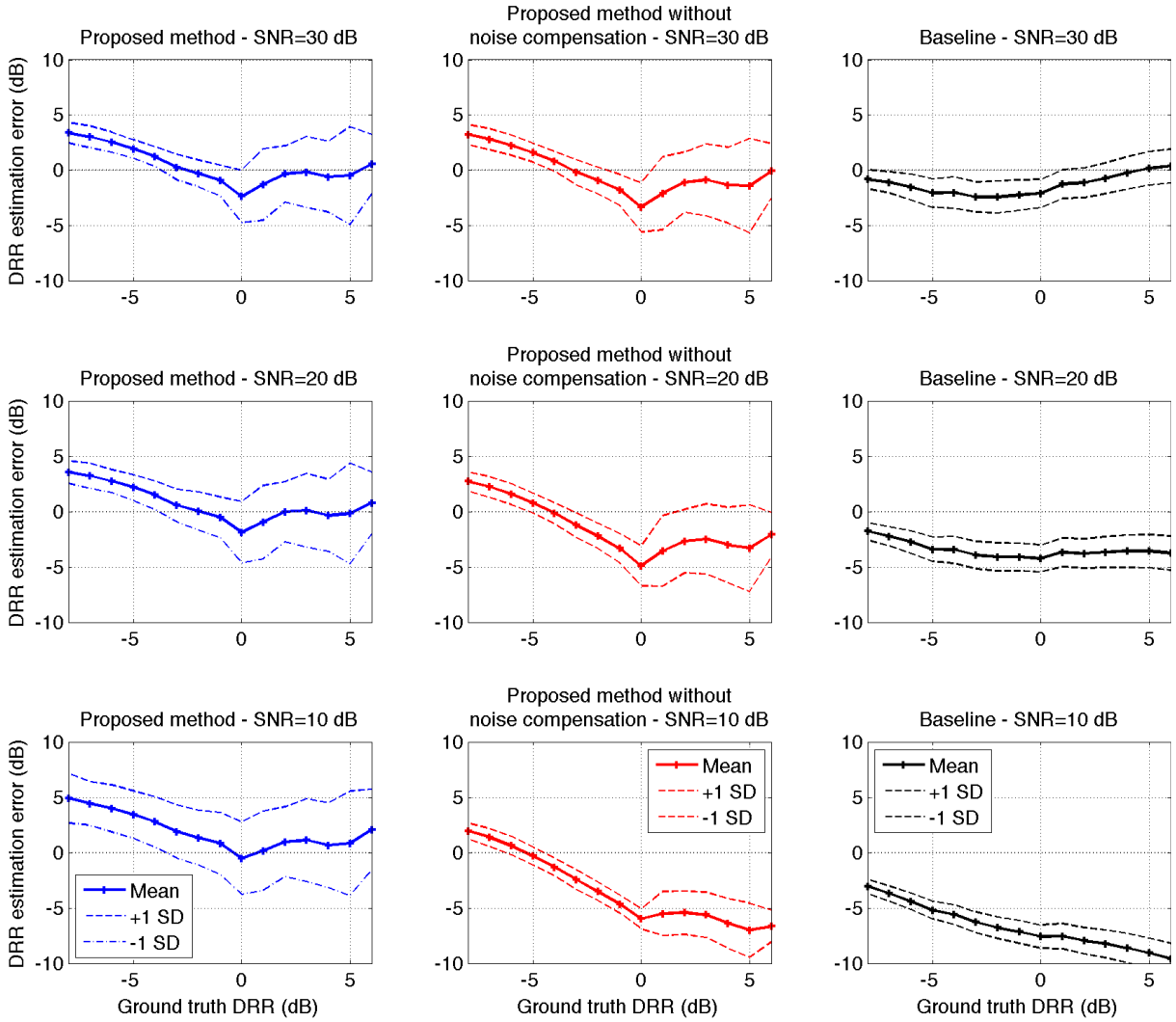

Fig. 2. *Results for the proposed method, proposed method without noise compensation, and the baseline at* $10, 20$ *and* 30 dB *SNR. Results have been aggregated into* 1 dB *bins.*

Since we assume that the reverberation power is the same at all microphones, from (7) and (10), we can write

$$E\{|D_m(j\omega)|^2\} = E\{|Y_m(j\omega)|^2\} - E\{|V_m(j\omega)|^2\} - E\{|R(j\omega)|^2\}. \quad (14)$$

The frequency dependent DRR follows from (3) as

$$\eta_m(j\omega) = \frac{E\{|D_m(j\omega)|^2\}}{E\{|R(j\omega)|^2\}}. \quad (15)$$

Substituting (13) and (14) into (15) gives

$$\eta_m(j\omega) = \frac{E\{|Y_m(j\omega)|^2\} - E\{|V_m(j\omega)|^2\}}{\frac{1}{G^2(j\omega)}(E\{|Z_y(j\omega)|^2\} - E\{|Z_v(j\omega)|^2\})} - 1. \quad (16)$$

The overall DRR is then given by

$$\bar{\eta}_m = \frac{1}{\omega_2 - \omega_1} \int_{\omega_1}^{\omega_2} \eta_m(j\omega)\, d\omega, \quad (17)$$

where $\omega_1 \leqslant \omega \leqslant \omega_2$ is the frequency range of interest.

3. PERFORMANCE EVALUATION

Speech signals were randomly selected from the test partitions of TIMIT [10]. These were convolved with AIRs generated using the source-image method [11, 12] for rooms with dimensions $\{3, 4,$ and $5\} \times 6 \times 3$ m, each with Reverberation Time (T_{60}) values from 0.2 to 1 s in 0.1 s intervals. In each room, four locations and rotations of the microphone array were chosen at random from a uniform distribution, and the source positioned perpendicular to the array at distances of

$0.05, 0.10, 0.50, 1, 2$, and 3 m. No microphone or source was allowed to be less than 0.5 m from any wall.

A two-element microphone array was used with a spacing of 62 mm to simulate the microphones on a typical laptop. Beamformer weights were chosen using a delay and subtract scheme to steer a null towards the DoA of the direct path. Since all source positions were equidistant from the two microphones this reduces to a simple subtraction giving the familiar dipole beam pattern shown in Fig. 1. In practical applications time difference of arrival estimation using, for example, GCC-PHAT [13], would be required to set the delay.

Ground truth DRR was estimated for each room, T_{60}, microphone and source position directly from the simulated AIRs. White Gaussian noise was added independently for each microphone at SNRs of $10, 20$, and 30 dB where the clean speech power was determined using an implementation of ITU-T P.56 [14] [15]. In the first experiment the proposed method in the case where oracle values for $E\{|V_m(j\omega)|^2\}$ and $E\{|Z_v(j\omega)|^2\}$ are used is compared with our formulation where noise is ignored (SNR assumed to be ∞ dB), and with the baseline method. In a practical application it is assumed that a noise estimator robust to reverberation will be used. In order to evaluate the effects of noise estimation errors on the accuracy of the DRR estimator, a second experiment was conducted with ± 1.5 dB added to each of $E\{|V_m(j\omega)|^2\}$ and $E\{|Z_v(j\omega)|^2\}$ in (16).

The baseline method for comparison was that of Jeub *et al.* [2]. It returns a vector of estimated DRR by frequency, and the mean of the values $> -\infty$ was used in our comparison.

4. RESULTS AND DISCUSSION

The DRR estimation accuracy of the proposed and baseline algorithms at SNRs of 10 dB, 20 dB, and 30 dB is shown in Fig. 2. To calculate the first and second order statistics, DRR estimates were binned according to their ground truth DRR in 1 dB intervals.

The proposed method is reasonably accurate with less than ± 3 dB error across DRRs ranging from -5 to 5 dB. As DRR decreases the proposed method tends to overestimate DRR. This is a result of the assumption that reflections arrive from all angles with equal probability. For a particular room and T_{60}, lower DRRs are obtained with larger source-microphone distances. This in turn results in the strong early reflections arriving from directions which are closer to the direct path DoA and are therefore more attenuated by the beamformer null. By under-accounting for these early reflections in (11) the DRR is overestimated.

For positive DRRs the mean estimate is unbiased but has a relatively high variance. Again this is most likely due to the assumption that the reverberation is isotropic. However, the distribution of strong early reflections is more random than for negative DRRs, and so they will be attenuated to a greater or lesser extent by the beam pattern depending on their direction of arrival.

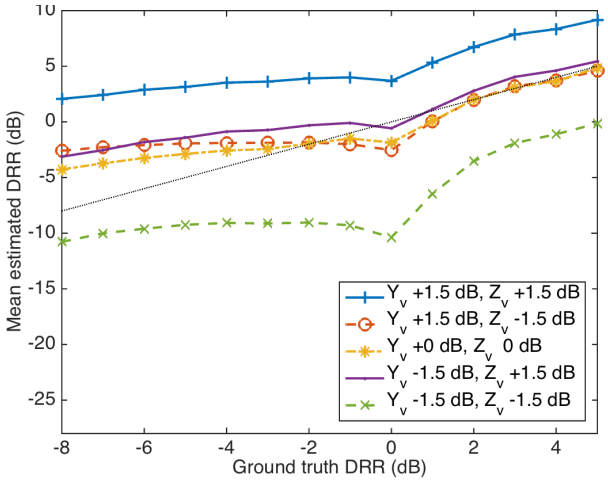

Fig. 3. *Effect of noise estimation errors on mean DRR estimates at* 20 dB *SNR.*

The importance of including noise in the formulation is shown by comparing the proposed method with and without noise compensation to the baseline. Without noise compensation our method follows the tendency of the baseline to underestimate DRR as noise increases. Conversely with the proposed method (with noise included in the formulation) accuracy is consistent across the range of SNRs shown with only a slight increase in the standard deviation of the estimates.

The sensitivity to errors in noise estimation at the reference microphone and at the output of the beamformer is shown in Fig. 3. Where there are errors of opposite polarity affecting the direct and beamformed power, the DRR estimates remain close to the case where there is no error, effectively cancelling each other out. Where the errors are of the same polarity, there is an additive effect with a ± 1.5 dB error on each term leading to a ± 3 dB error overall. This suggests that the method is more sensitive to the bias in a noise estimator than its variance.

5. CONCLUSIONS

We have presented a novel method for estimation of DRR from multi-channel speech taking noise into account, and demonstrated that it is more robust to noise at realistic SNRs than a baseline based on spatial coherence. This is achieved by compensating for the bias caused by the presence of uncorrelated noise at the microphones. Whilst the tests performed were limited to two channel material, the method can be applied to a multi-channel system with an arbitrary number of microphones with the selection of an appropriate beamformer. The formulation returns an estimate of DRR according to frequency, and therefore a frequency dependent DRR could be provided if desired. Also, since the method does not rely on the statistics of speech it could also be applied to music.

6. REFERENCES

[1] E. A. P. Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*, Ph.D. thesis, Technische Universiteit Eindhoven, 2007.

[2] M. Jeub, C.M. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Barcelona, Spain, 2011.

[3] 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. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2374–2384, Nov 2011.

[4] M. Kuster, "Estimating the direct-to-reverberant energy ratio from the coherence between coincident pressure and particle velocity," *J. Acoust. Soc. Am.*, vol. 130, no. 6, pp. 3781–3787, 2011.

[5] O. Thiergart, G. Del Galdo, and E. A. P. Habets, "On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation," *J. Acoust. Soc. Am.*, vol. 132, no. 4, pp. 2337–2346, 2012.

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

[7] B. Dumortier and E. Vincent, "Blind RT60 estimation robust across room sizes and source distances," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 5187–5191.

[8] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[9] J. Benesty, M. Mohan Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*, Springer, 2008.

[10] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, Dec. 1988.

[11] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[12] E. A. P. Habets, "Room impulse response generator for MATLAB," `http://home.tiscali.nl/ehabets/rir_generator.html`, 2003.

[13] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[14] ITU-T, "Objective measurement of active speech level," Mar. 1993.

[15] D. M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 1997–2013.