DUAL-CHANNEL MODULATION ENERGY METRIC FOR DIRECT-TO-REVERBERATION RATIO ESTIMATION

Sebastian Braun[1], *João F. Santos*[2], *Emanuël A. P. Habets*[1] , *Tiago H. Falk*[2]

[1] International Audio Laboratories Erlangen
[2] INRS-EMT, University of Quebec, Montreal, QC, Canada

ABSTRACT

Non-intrusive estimators for acoustic parameters like the direct-to-reverberation ratio (DRR) are useful tools but still perform weakly as shown in the acoustic characterization of environments (ACE) challenge. In this paper, we develop a novel dual-channel metric based on the modulation energy domain for DRR estimation. In contrast to established modulation based single-channel metrics like the speech-to-reverberation modulation energy ratio (SRMR), we exploit the spatial information from two microphones as well as the temporal dynamics in the modulation energy domain. The developed metric shows a strong linear correlation to the DRR, which allows a simple mapping. It is shown that the metric is robust against the microphone array configuration, room characteristics and the speech signal. The proposed metric is compared to a reference method based on the spectral variance of the room transfer functions, and both metrics are evaluated using simulated and measured data. In our experiments, the proposed metric achieved a higher correlation and lower RMSE compared the reference method, and outperforms existing SRMR based DRR estimators.

Index Terms— Direct-to-reverberation ratio, speech modulation energy

1. INTRODUCTION

The acoustic properties of rooms and measured positions in a room are often characterized in terms of acoustic measures such as the reverberation time and direct-to-reverberation ratio (DRR). The DRR is a highly relevant and useful measure to describe the relative amount of received reverberation at a microphone, which can be used for example for dereverberation, or as a control parameter for various applications like acoustic scene classification or automatic speech recognition. In general, these acoustic parameters are computed from the room impulse response (RIR) [1]. However, in many applications, the RIR is not known and the parameters have to be estimated non-intrusively from the observed signals.

The acoustic characterization of environments (ACE) challenge [2] provided a framework to evaluate blind estimators for reverberation time (T_{60}) and DRR. Whereas the general performance of submitted T_{60} estimators was satisfying, the performance of the submitted DRR estimators left room for improvement. The main classes of DRR estimators defined in [2] are analytical methods mainly based on spatial information [3, 4], single spectral features [5, 6, 7], or machine learning approaches using multiple features [8].

Non-intrusive single-channel metrics based on the speech modulation energy have been developed to measure speech quality or reverberation time [5, 9]. In [6], the speech-to-reverberation modulation energy ratio (SRMR) was modified to measure the DRR. Although these metrics have been applied to multichannel data, the spatial information was only exploited by averaging the single-channel measure over all channels. In [10], a dual-channel metric for DRR estimation based on the spectral variance of the relative room transfer function was proposed that was not included in the ACE challenge.

In this paper, we aim at developing a modulation energy based metric, which exploits multichannel information in addition to temporal dynamics within the modulation energy domain. We propose to use the normalized modulation energy difference between two microphone signals as a metric. It is shown that the temporal variance of this metric has a high correlation with the DRR and can be therefore be mapped to the DRR using a simple linear function. The proposed metric is compared to the spectral variance based metric [10] and evaluated using the ACE challenge data.

2. SIGNAL MODEL AND PROBLEM FORMULATION

We assume that the m-th omnidirectional microphone signal $x_m(t)$ can be described by a convolution of the anechoic speech signal $s(t)$ with the RIR $h_m(t)$, i. e.

$$x_m(t) = h_m(t) \ * \ s(t), \tag{1}$$

where t denotes the discrete time index and $*$ is the convolution operator. The DRR at the m-th microphone is given by

$$\mathrm{DRR}_m = \frac{\sum_{t=t_\mathrm{d}-t_0}^{t_\mathrm{d}+t_0} h_m^2(t)}{\sum_{t=0}^{t_\mathrm{d}-t_0} h_m^2(t) + \sum_{t=t_\mathrm{d}+t_0}^{\infty} h_m^2(t)}, \tag{2}$$

where t_d denotes the time index of the direct sound peak and t_0 corresponds to 2.5 ms as defined in [2].

In so-called *blind* estimation scenarios, the RIR $h_m(t)$ is unknown such that the DRR has to be estimated from the observed reverberant signals $x_m(t)$. The aim of this paper is to develop a modulation energy-based metric that has a high correlation with the DRR, and is robust against influences of the microphone array configuration, the room characteristics, and the speech signal.

3. REVIEW OF DUAL-CHANNEL SPECTRAL VARIANCE-BASED DRR ESTIMATOR

In [10] a DRR estimator was proposed using two microphones based on the spectral variance of the logarithmic room transfer function difference, i. e. the relative transfer function between the

This work was carried out as part of a research visit at the MuSAE Lab at INRS-EMT.

microphones. The m-th microphone signal in the frequency domain $X_m(\omega)$ is assumed to be

$$X_m(\omega) = H_m(\omega)S(\omega), \tag{3}$$

where $S(\omega)$ is the frequency domain speech signal, $H_m(\omega)$ is the m-th microphone acoustic transfer function, and ω is the angular frequency. The logarithmic subtraction between first and second microphone signal therefore equals the subtraction between the room transfer functions, i. e.

$$\Delta_X(\omega) = 20\log_{10}|X_1(\omega)| - 20\log_{10}|X_2(\omega)| \tag{4}$$
$$= 20\log_{10}|H_1(\omega)| - 20\log_{10}|H_2(\omega)|. \tag{5}$$

Finally the standard deviation σ_X, i. e. the spectral variability, of $\Delta_X(\omega)$ is computed over the frequency range of interest.

In [10], a polynomial mapping from σ_X to the DRR is proposed, which has been trained on reverberant speech. To ensure a fair comparison, we investigate the metric σ_X itself and then find a suitable mapping based on our used training data.

4. PROPOSED MODULATION ENERGY METRIC

4.1. Modulation energy analysis

The microphone signals $x_m(t)$ are analyzed in J subbands by applying bandpass filters $a_j(t)$ to each microphone signal, i. e.

$$x_m(j,t) = a_j(t) * x_m(t) \tag{6}$$

for $j = \{1,\dots,J\}$. Then, the temporal envelope of $x_m(j,t)$ is computed via the Hilbert transform $\mathcal{H}\{\cdot\}$ by

$$e_m(j,t) = \sqrt{x_m^2(j,t) + \mathcal{H}\{x_m(j,t)\}^2}. \tag{7}$$

The modulation spectrum energy per time frame n is computed by applying K modulation bandpass filters $b_k(t)$ to the subband envelopes $e_m(j,t)$, where k is the modulation subband index, applying a window $w(t)$ and summing the energy in each frame by

$$E_m(j,k,n) = \sum_{t=nT}^{(n+1)T-1} \Big(w(t-nT)\ \ [b_k(t) * e_m(j,t)]\Big)^2, \tag{8}$$

where T is the frame length and $k = \{1,\dots,K\}$.

4.2. Proposed metric

To build a metric exploiting the spatial information between the microphones, we propose to utilize the per frame normalized modulation energy difference between two microphones, i. e.

$$\Delta_E(j,k,n) = \frac{E_1(j,k,n) - E_2(j,k,n)}{\sqrt{E_1^2(j,k,n) + E_2^2(j,k,n)}}. \tag{9}$$

In preliminary experiments, we observed that the temporal variability of $\Delta_E(j,k,n)$ over n is related to the DRR. Note that $\Delta_E(j,k,n)$ is dependent on the microphone spacing d: for small d, the observed modulation energy becomes more similar between the microphones, and for $d \to 0$ we have $\Delta_E \to 0$.

To measure the temporal variability of $\Delta_E(j,k,n)$, we compute its standard deviation by

$$\sigma_E(j,k) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\Big(\Delta_E(j,k,n) - \mu_E(j,k)\Big)^2} \tag{10}$$

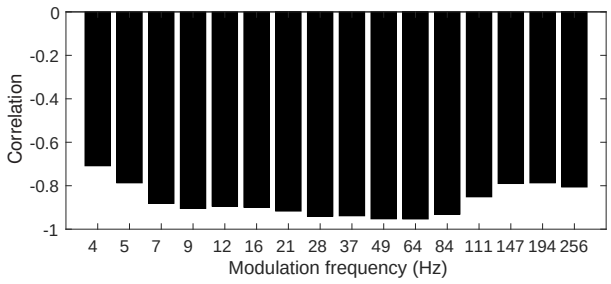

Fig. 1. Pearson correlation coefficients between acoustic frequency averaged $\sigma_E(k)$ per modulation band and the DRR for simulated RIRs with varying distances, reverberation times, source angles and speech signals, and a fixed microphone spacing $d = 10$ cm.

where $\mu_E(j,k)$ is the mean of $\Delta_E(j,k,n)$, and N is the number of frames. In order to craft a one-dimensional metric, we consider only acoustic and modulation subbands with a high correlation with the DRR, before computing the final averaged metric

$$\sigma_E = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\sigma_E(j,k), \tag{11}$$

where J and K are the number of acoustic and modulation subbands, respectively. The choice of the acoustic and modulation filterbanks is discussed in Sec. 4.3.

By inspecting the metrics σ_X and σ_E, both are based on a standard deviation, where σ_X is the spectral inter-channel variability, and $\sigma_E(j,k)$ is the inter-channel temporal variability in the modulation domain for each subband separately.

4.3. Metric analysis and parameter selection

In this section, we present preliminary investigations using simulated RIRs based on the image method [11] to find optimal parameters for the acoustic and modulation filterbanks. While in preliminary experiments, the metric $\sigma_E(j,k)$ showed rather consistent correlation with the DRR over varying source angles (array rotation), T_{60}s and speech signals, we noticed a clear dependence of the microphone spacing.

We convolved speech signals with different RIRs by varying the simulation parameters in the following ranges: $T_{60} = [0.2, 1.4]$ s, a source distance range of $[1, 4]$ m, 3 different source angles and two speech signals. For this investigation, we used acoustic and modulation filterbanks with a wide frequency range and high resolution, i. e., a Gammatone filterbank [12] with $J = 23$ equivalent rectangular bandwidth (ERB) bands between 125 Hz and 16 kHz, and $K = 16$ modulation bands logarithmically spaced between 4 Hz and 256 Hz [5, 13].

Modulation frequency range: As preliminary experiments showed an influence of the microphone spacing on the acoustic subbands, but not on the modulation subbands, we first select the optimal modulation frequency range for a given microphone spacing. The correlation between the DRR and the acoustic frequency averaged metric $\sigma_E(k)$ for a wide range of modulation frequencies is shown in Fig. 1 for a fixed microphone spacing $d = 10$ cm. We can see that a strong correlation exists for modulation center frequencies between 10 and 64 Hz.

Acoustic frequency range: To investigate the dependence of $\sigma_E(j,k)$ on the microphone spacing in the acoustic subbands, the

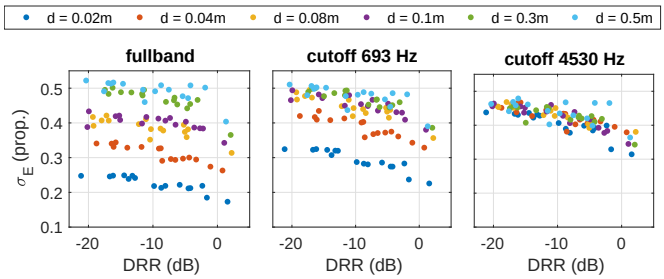

Fig. 2. Dependency of the proposed metric for different microphone spacings and varying lower cutoff frequency using simulated RIRs.

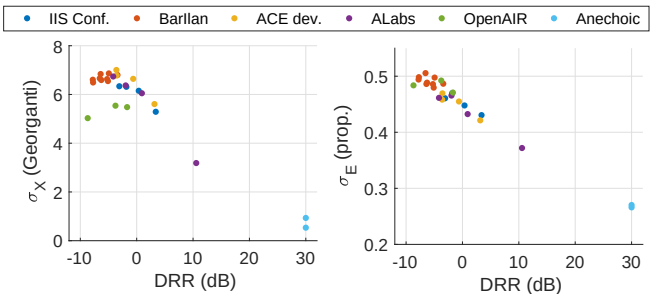

Fig. 3. Relation between metrics and the DRR for different RIR datasets and a single speaker.

simulated RIRs dataset with varying parameters of source distance, source angle, T_{60} and speech signals was extended to varying microphone spacings between $d = [2, 50]$ cm. The standard deviation $\sigma_E(j,k)$ was then averaged for modulation bands between 10 and 64 Hz and over various acoustic frequency ranges. The results are shown in Fig. 2 for a fullband frequency average (left subplot), excluding all frequencies below 693 Hz (middle subplot), and excluding all frequencies below 4530 Hz (right subplot).

Figure 2 reveals that for frequencies above the transition frequency between plane wave and spherical wave propagation, i. e.

$$f_d = \frac{c}{2\pi d}, \tag{12}$$

where c is the speed of sound, the metric $\sigma_E(j,k)$ is independent of the microphone spacing and all points lie approximately on the same monotonic line in relation to the DRR for all $d \gg \frac{c}{2\pi f_d}$.

By averaging $\sigma_E(j,k)$ in (11) only over the acoustic subbands above a certain cutoff frequency f_d, the microphone spacing influence can be completely removed. We can observe that in Fig. 2 in the middle and right plot, the lowest frequency for the frequency averaging was chosen such that the condition (12) is fulfilled for $d > 0.08$ m and for $d > 0.02$ m, respectively. Interestingly, in [14] it was also found that some spectral metrics for distance detection work better for high-pass filtered signals.

Proposed parameter choice: Based on the two previous investigations, we propose to use $K = 8$ modulation bands, logarithmically spaced between 8 and 64 Hz. The filters $a_j(t)$ are implemented as a Gammatone filterbank with $J = 23$ bands on a ERB scale between 125 Hz and 16 kHz. Based on the microphone distance d, in (11) we consider only $J^* \leq J$ bands above the transition frequency f_d given by (12).

5. EXPERIMENTAL SETUP AND RESULTS

5.1. Database and setup

For evaluation, we used multichannel RIRs from different datasets, therefore including various rooms with the reverberation times T_{60}, microphone spacings and source positions summarized in Table 1, resulting in a total of 26 multichannel RIRs, where only two channels from each condition were used. We used 3 male and 3 female speech signals from [15]. In addition, we used the ACE challenge dataset and framework [2] in the last experiment in Sec. 5.4.

The signals were processed using a sampling frequency of 16 kHz. The proposed modulation based dual-channel metric used the ERB and modulation-filterbanks as described at the end of Sec. 4.3 on frames of 256 ms length with a Hamming window and a frame shift of 32 ms. The dual-channel spectral variance metric was implemented using 1 s long frames with a Hann window and a frame shift of 64 ms.

5.2. Noise-free single speaker performance assessment

Figure 3 shows the relations between the proposed and benchmark metrics σ_E and σ_X, respectively, and the true DRR for different datasets and a single speaker. We can observe that the proposed metric σ_E has a clear linear relation to the DRR, while σ_X follows a non-linear but still monotonic function. It is interesting to note that towards low DRR the distribution curve of σ_X saturates and the variance increases. This behavior can be expected as the spectral variance of the room transfer functions depends on the density of spectral peaks and notches caused by reflections. However, with decreasing DRR, the density of the spectral peaks and notches saturates such that the spectral variance does not increase further. In contrast, the modulation based metric shows a clear linear relation with DRR for a wide range of tested DRR values. The DRR of the anechoic signals is limited to 30 dB for practical reasons.

5.3. Noisy multiple speaker performance assessment

The influence of noise and the variability introduced by multiple speakers is shown in Fig. 4 for the metrics σ_E and σ_X. Different signal-to-noise ratios (SNRs) are encoded as colors, whereas points with the same color (denoting SNR) and the same x-value (same DRR) are different speakers.

We can see that for σ_X, the variance of the relation increases at low DRR, whereas for σ_E, the variance of the linear relation increases at high DRR. The relations in Fig. 4 again suggest a linear mapping for σ_E and a non-linear mapping for σ_X to the DRR. We chose to use a 3rd order polynomial mapping for σ_X, since the originally proposed 5th order polynomial mapping [10] yielded worse results using our data. We computed the mapping functions as a least-squares fit from the mean of each metric per DRR condition to the true DRR. Exemplary mapping functions obtained from the data in Fig. 4 are shown as black lines.

5.4. Experimental results

We conducted two experiments to evaluate both metrics under test for DRR estimation. In the first experiment, we used simulated RIRs to train the third order polynomial mapping for σ_X and a linear mapping for σ_E. For training, we simulated rooms with T_{60} between 0.2 s and 1.1 s, three source angles, four source distances, two female and two male speech signals, and added pink noise with

Database	T_{60} (ms)	source distance (m)	DRR (dB)	mic spacing (m)
Large conference room, Fraunhofer IIS	700	[2.50, 5.50]	[-3.0, 3.4]	0.09
Variable acoustic lab, Bar-Ilan University	{480, 630, 940}	[1.30, 3.90]	[-7.7, -3.4]	0.17
ACE devel. set, cruciform array (Office1, Lobby) [2]	330, 640	[1.10, 2.90]	[-3.6, 3.1]	0.25
Small acoustic lab, AudioLabs Erlangen	280	[0.45, 2.80]	[-4.1, 10.5]	0.09
OpenAIR (Warehouse, Reaktor, Church) [16]	{2300, 3500, 15000}	{28, 15, 26}	[-8.6, 1.7]	0.17
Anechoic (simulation)	0	2	30	0.20

Table 1. Description of RIR databases

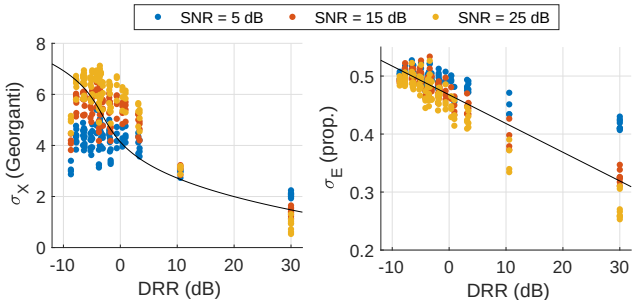

Fig. 4. Relation of the metrics using the RIRs from Tab. 1 with pink noise and various speech signals, and fitted mapping functions.

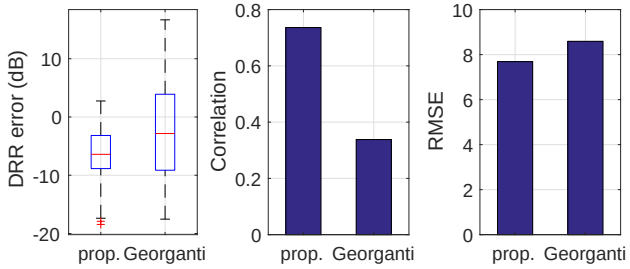

Fig. 5. Results for training on simulated RIRs and evaluation on measured RIRs.

three SNR levels, namely $\{5, 15, 25\}$ dB. For evaluation, we used the measured RIR datasets described in Table 1, one female and one male speech signal and added recorded cafeteria noise [17] with the SNRs $\{2, 12, 22\}$ dB. Figure 5 shows the results of testing between estimated and true DRR, namely the Pearson correlation coefficient and the root-mean-squared error (RMSE). We can observe that the proposed estimator yields a high correlation and lower RMSE than Georganti's metric. However, the boxplot shows that the DRR error median of both estimators is biased, i. e. both show a tendency to underestimate.

As second experiment, we used the data provided by the ACE challenge [2] that is divided into a development and an evaluation dataset. For training, we used the ACE development set together with the measured RIR dataset shown in Table 1 as the number of DRR conditions was very limited in the ACE development set with only four different DRRs. We used two microphones of the cruciform array with a microphone spacing of 0.25 m. The results for testing on the ACE evaluation set are shown in Fig. 6. We can observe that both estimators yield a much lower correlation to the DRR provided by ACE than in the first experiment. The proposed modulation based DRR estimator yields a slightly higher correlation and a

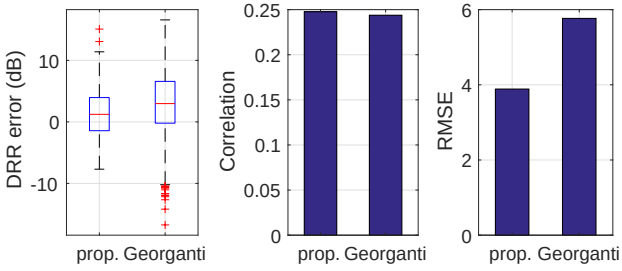

Fig. 6. Results for ACE challenge

37% lower RMSE than Georganti's spectral variance based estimator. However, the correlation of both estimators is lower compared to our experiments shown in Fig. 5.

After an in-depth analysis, it is hypothesized that the reasons for the low performance in the ACE challenge are different for the two estimators: As can be seen in Fig. 4, the variance of the proposed modulation based estimator increases at high DRRs. The ACE data is unfortunately biased towards high DRRs compared to the datasets used in the first experiment. Georganti's estimator performs better in high than low DRR, but its variance increases in noisy conditions. The challenging and sometimes highly non-stationary noise types with low SNR are a further reason for the generally low performance the estimators on the ACE evaluation set.

However, the proposed modulation based DRR estimator outperforms the existing multichannel modulation based DRR estimators that average the normalized speech-to-reverberation modulation ratio (NSRMR) and overall speech-to-reverberation modulation ratio (OSRMR) over all channels by a 312% higher correlation and a 20% lower RMSE [6] on the ACE Cruciform array data.

6. CONCLUSION

We have developed a dual-channel metric in the modulation energy domain for DRR estimation. The proposed metric was analyzed and tuned for a wide variety of acoustic conditions and was shown to be robust against influences of different speakers, rooms, angles of incidence and the microphone spacings. The proposed metric was tested on both simulated and recorded data (i.e., from the ACE Challenge [2]), and shown to outperform a benchmark metric based on the spectral variance of the relative room transfer functions on both datasets. Furthermore, the proposed dual-channel modulation metric outperforms existing modulation based metrics for DRR estimation. Future work could aim at increasing the robustness in noisy conditions and more advanced mapping techniques between the metric and the DRR.

7. REFERENCES

[1] H. Kuttruff, *Room Acoustics*, Taylor & Francis, London, fourth edition, 2000.

[2] 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. Process.*, vol. 24, no. 10, pp. 1681–1693, Oct 2016.

[3] Y. Hioka, K. Furuya, K. Niwa, and Y. Haneda, "Estimation of direct-to-reverberation energy ratio based on isotropic and homogeneous propagation models," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Aachen, Germany, Sept. 2012, pp. 1–4.

[4] J. Eaton, A. H. Moore, P. A. Naylor, and J. Skoglund, "Direct-to-reverberant ratio estimation using a null-steered beamformers," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 46–50.

[5] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.

[6] M. Senoussaoui, J. Santos, and T. H. Falk, "SRMR variants for improved blind room acoustics characterization," in *Proc. ACE Challenge Workshop, Satellite IEEE*, 2015.

[7] T. de M. Prego, A. A. de Lima, R. Zambrano-López, and S. L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decompositions," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

[8] P. P. Parada, T. Sharma, D. van Waterschoot, and P. A. Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ACE challenges," in *Proc. ACE Challenge Workshop, Satellite IEEE*, New Paltz, NY, USA, 2015.

[9] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Antibes, France, Sept. 2014.

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

[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] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Tech. Rep., Apple Computer, Perception Group, Tech. Rep, 1993.

[13] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system. – i. model structure," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3615–3622, 1996.

[14] E. Georganti, T. May, S. van de Par, A. Harma, and J. Mourjopoulos, "Speaker distance detection using a single microphone," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 1949–1961, Sept 2011.

[15] European Broadcasting Union, "Sound quality assessment material recordings for subjective tests," 1988, http://tech.ebu.ch/publications/sqamcd.

[16] D. Murphy and S. Shelley, "Open AIR library," www.openairlib.net, 2017.

[17] J. Thiemann, N. Ito, and E. Vincent, "Diverse Environments Multichannel Acoustic Noise Database (DEMAND)," June 2013, http://parole.loria.fr/DEMAND/.