# MAXIMUM LIKELIHOOD BASED MULTI-CHANNEL ISOTROPIC REVERBERATION REDUCTION FOR HEARING AIDS

*Adam Kuklasiński*[*], *Simon Doclo*[†], *Søren Holdt Jensen*[‡], *Jesper Jensen*[*,‡]

[*]Oticon A/S, 2765 Smørum, Denmark
[†]University of Oldenburg, Department of Medical Physics and Acoustics, Oldenburg, Germany
[‡]Aalborg University, Department of Electronic Systems, 9220 Aalborg, Denmark
adku@oticon.dk, simon.doclo@uni-oldenburg.de,
shj@es.aau.dk, (jsj@oticon.dk, jje@es.aau.dk)

## ABSTRACT

We propose a multi-channel Wiener filter for speech dereverberation in hearing aids. The proposed algorithm uses joint maximum likelihood estimation of the speech and late reverberation spectral variances, under the assumption that the late reverberant sound field is cylindrically isotropic. The dereverberation performance of the algorithm is evaluated using computer simulations with realistic hearing aid microphone signals including head-related effects. The algorithm is shown to work well with signals reverberated both by synthetic and by measured room impulse responses, achieving improvements in the order of 0.5 PESQ points and 5 dB frequency-weighted segmental SNR.

***Index Terms***— multi-channel wiener filter, maximum likelihood, speech dereverberation, isotropic, hearing aids

## 1. INTRODUCTION

Hearing impaired listeners experience increased difficulty in understanding speech in reverberant and noisy conditions [1]. In order to enable them to attain the same speech intelligibility as normal hearing persons, various signal enhancement algorithms are used in Hearing Aids (HAs). Both single- and multi-microphone (spatial) methods are commonly used in HAs, notably spectral modification and beamforming [2].

The Multi-channel Wiener Filter (MWF) [3] is a method which currently receives a lot of attention in the research community, e.g. [4], [5], [6]. Implementation of the MWF requires knowledge of the inter-microphone covariance matrices of the target signal (i.e. speech) and of the interference (e.g. ambient noise or reverberation). Traditionally a Voice Activity Detector (VAD) is used to enable noise covariance matrix estimation during speech pauses, e.g. [6]. This approach is based on the assumption that the interference covariance matrix is constant during speech presence. In reverberant conditions this assumption is not valid, which necessitates on-line estimation of the reverberation covariance matrix.

In the present study, we propose an MWF algorithm for speech dereverberation, which jointly estimates the target and interference spectral variances also during speech presence. The algorithm uses a Maximum Likelihood Estimation (MLE) method presented first in [7] which is novel in the speech dereverberation context. We assume a cylindrically isotropic spatial distribution of the late reverberation and a known speaker direction. Therefore, the structure of the inter-microphone covariance matrices of the speech and reverberation is known and only the time-varying spectral variances (the scaling factors of these matrices) are estimated in the MLE framework.

The proposed algorithm bears some similarities to the one presented in [4]. In both methods an isotropic spatial distribution of the late reverberant field is assumed and the spectral variances of the interference are estimated regardless of speech presence. However, while [4] uses intermediate "reference signals" (based on [5]) to estimate the reverberation variances, we compute these estimates directly form the input covariance matrix (based on [7]). The method presented here is designed for and evaluated in a hearing aid usage scenario and with real room impulse responses, whereas in [4], microphones were assumed to reside in free field and reverberation was simulated using an image model of a rectangular room.

## 2. SIGNAL MODEL AND ASSUMPTIONS

The proposed algorithm operates on $M$ microphone signals represented as complex-valued Short Time Fourier Transform (STFT) coefficients. They are collected in a vector

$$\mathbf{y}(k,n) = [y_1(k,n) \dots y_m(k,n) \dots y_M(k,n)]^T, \quad (1)$$

where $y_m(k,n)$ is the STFT coefficient of the $m$-th microphone signal in the $k$-th frequency sub-band and the $n$-th time frame. Based on the assumption of signal independence between sub-bands, we will operate on them separately. This

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement Nº ITN-GA-2012-316969.

allows us to omit the frequency index $k$ in the following description without loss of generality.

The input signal $\mathbf{y}(n)$ is assumed to be the sum of the target speech component $\mathbf{s}(n)$ and an interference component $\mathbf{v}(n)$. Both $\mathbf{s}(n)$ and $\mathbf{v}(n)$ are defined similarly to (1). The interference $\mathbf{v}(n)$ is assumed to be late reverberation, ambient noise, or a sum of both. In either case, it is assumed to be uncorrelated to the target speech component $\mathbf{s}(n)$. This allows us to model the covariance matrix of the input as the sum of the covariance matrices of the two signal components:

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(n) &= E\{\mathbf{y}(n)\mathbf{y}^H(n)\} \\ &= E\{\mathbf{s}(n)\mathbf{s}^H(n)\} + E\{\mathbf{v}(n)\mathbf{v}^H(n)\} \\ &= \mathbf{\Phi}_{\mathbf{s}}(n) + \mathbf{\Phi}_{\mathbf{v}}(n). \end{aligned} \tag{2}$$

We model the speaker as a point source and therefore the speech component can be expressed as

$$\mathbf{s}(n) = s(n)\mathbf{d}. \tag{3}$$

The scalar signal $s(n)$ represents the speech signal at a certain reference position, commonly chosen as one of the microphones. Elements of the vector $\mathbf{d}$ represent relative transfer functions of the speech signal between the reference position and all microphones of the array. The vector $\mathbf{d}$ is assumed to be known, and depends primarily on the microphone array geometry and on the direction of the speech source. In the beamforming context, we will refer to $\mathbf{d}$ as a steering vector.

We employ an isotropic model of the interference $\mathbf{v}(n)$. Taking this and (3) into account, (2) can be rewritten as

$$\mathbf{\Phi}_{\mathbf{y}}(n) = \underbrace{\phi_s(n)\mathbf{d}\mathbf{d}^H}_{\mathbf{\Phi}_{\mathbf{s}}(n)} + \underbrace{\phi_v(n)\mathbf{\Gamma}_{\text{iso}}}_{\mathbf{\Phi}_{\mathbf{v}}(n)}, \tag{4}$$

where $\phi_s(n)$ and $\phi_v(n)$ are, respectively, (scalar) spectral variances of the speech and of the interference component of the reference microphone signal. Because, in general, the speech and noise processes are non-stationary, their variances $\phi_s(n)$ and $\phi_v(n)$ are time-varying. The matrix $\mathbf{\Gamma}_{\text{iso}}$ is the normalized covariance matrix of the isotropic sound field, and similarly to $\mathbf{d}$, is assumed to be known and constant.

### 2.1. Discussion of validity of assumptions

The intended application of the proposed algorithm is intelligibility improvement of reverberant and/or noisy speech in HAs. Assumptions with regard to the employed signal model are made to capture aspects of the actual physical signals which are most relevant to this particular task and application.

In reverberant conditions, speech intelligibility is affected primarily by late reverberation, whereas early reflections are believed to be beneficial [8]. For that reason, the model of the interference was chosen to describe properties of specifically the late part of the reverberation.

In [9], the spatial energy distribution of reverberant sound fields was studied. It was shown that all spatial directions are represented in the late reverberant energy, but only few directions are in the energy of early reflections. This supports our assumption that the late reverberant field is isotropic.

An isotropic model of the ambient noise is also ecologically justified, especially in applications where there is no prior knowledge on the spatial distribution of the noise, e.g. in hearing aids. The spatial probability distribution of the noise impinging on the microphone array can reasonably be assumed uniform, i.e. isotropic.

The assumption of $\mathbf{d}$ being known is reasonable in hearing aid design. It is supported by the fact that, in most situations, the hearing aid user is looking at the person he is speaking with (e.g. to facilitate lip reading). Hence, $\mathbf{d}$ corresponds to a target source frontal to the HA user.

In the present work, the interference $\mathbf{v}(n)$ is modeled as independent, and therefore uncorrelated with the speech signal $\mathbf{s}(n)$. This assumption is natural for the ambient noise but is questionable with regard to the reverberation. Our rationale is that the late part of the room impulse responses is considerably disrupted by thermal fluctuations [10] and small movements of the source and microphone array [11]. These instabilities are unavoidable in real use of a HA and effectively decorrelate the late reverberation from the direct sound.

## 3. MULTI-CHANNEL WIENER FILTER

It is well known that the MWF is the Linear Minimum Mean Square Error (LMMSE) solution to the problem of signal estimation in a setup presented in Section 2, [3]. It is also well known that the MWF can be factorized into a Minimum Variance Distortionless Response (MVDR) beamformer and a Single Channel (SC) Wiener filter [3].

The structure of the proposed MWF-type algorithm is depicted in Fig. 1. The signal resulting from the MWF is the LMMSE estimate of the target speech signal at the reference position and may be written as

$$\hat{s}(n) = \mathbf{w}_{\text{mwf}}^H(n)\mathbf{y}(n), \text{ where} \tag{5a}$$

$$\mathbf{w}_{\text{mwf}}(n) = \underbrace{\left[\frac{\phi_{s_o}(n)}{\phi_{s_o}(n)+\phi_{v_o}(n)}\right]}_{g_{\text{sc}}(n)} \underbrace{\frac{\mathbf{\Gamma}_{\text{iso}}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{\Gamma}_{\text{iso}}^{-1}\mathbf{d}}}_{\mathbf{w}_{\text{mvdr}}}. \tag{5b}$$

In (5a–5b) the vector of MVDR beamformer coefficients and the SC Wiener filter gain have been denoted as $\mathbf{w}_{\text{mvdr}}$ and $g_{\text{sc}}(n)$, respectively. $\phi_{s_o}(n)$ and $\phi_{v_o}(n)$ denote the spectral variances of the speech and the interference at the output of the MVDR beamformer. They can be expressed as

$$\phi_{s_o}(n) = \phi_s(n), \tag{6a}$$

$$\phi_{v_o}(n) = \phi_v(n)(\mathbf{d}^H\mathbf{\Gamma}_{\text{iso}}^{-1}\mathbf{d})^{-1}. \tag{6b}$$

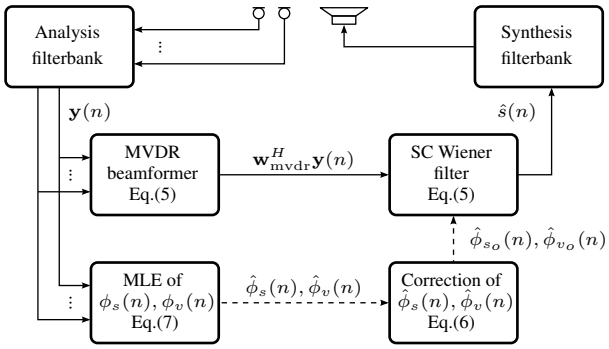


**Fig. 1**. Block diagram of the proposed algorithm.

The MVDR beamformer does not distort the variance of the speech (6a), but the variance of the interference has to be corrected by the beamformer suppression factor (6b) [3]. It is important to note that $\mathbf{w}_{\mathrm{mvdr}}$ depends only on $\mathbf{\Gamma}_{\mathrm{iso}}$ and $\mathbf{d}$. Because we assume that these are known and constant, the beamformer coefficients $\mathbf{w}_{\mathrm{mvdr}}$ can be calculated beforehand.

The SC Wiener filter gain $g_{\mathrm{sc}}(n)$ is time-varying and depends on the spectral variances $\phi_s(n)$ and $\phi_v(n)$. They are unknown and have to be estimated from the noisy/reverberant observations $\mathbf{y}(n)$ for each time frame and frequency bin.

Several methods exist for estimating $\phi_s(n)$ and $\phi_v(n)$, e.g. [4], [5], [7]. The proposed algorithm uses MLEs which were derived by Ye and DeGroat [7] for a similar signal model to the one employed in the present study, although in a non-acoustic context. These MLEs may be expressed as

$$\hat{\phi}_v(n) = \frac{1}{M-1} \operatorname{tr}\left\{ \left( \mathbf{I} - \mathbf{d}\mathbf{w}_{\mathrm{mvdr}}^H \right) \hat{\mathbf{\Phi}}_{\mathbf{y}}(n)\, \mathbf{\Gamma}_{\mathrm{iso}}^{-1} \right\}, \quad (7a)$$

$$\hat{\phi}_s(n) = \mathbf{w}_{\mathrm{mvdr}}^H \left( \hat{\mathbf{\Phi}}_{\mathbf{y}}(n) - \hat{\phi}_v(n)\, \mathbf{\Gamma}_{\mathrm{iso}} \right) \mathbf{w}_{\mathrm{mvdr}}, \quad (7b)$$

where $\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)$ is the estimate of the covariance matrix of the input signal, and $\mathrm{tr}\{\cdot\}$ denotes the matrix trace operator.

## 4. EXPERIMENTAL SETUP

In order to evaluate the performance of the proposed algorithm, a series of computer simulations was conducted, Technical details on these simulations are described in Sections 4.1–4.2 and the evaluation results are discussed in Section 5

### 4.1. Speech signals and room impulse responses

Recorded speech utterances of male and female native English speakers were obtained from the TIMIT database [12]. Individual utterances were concatenated into longer sequences and artificially reverberated by convolving them with either synthetic or measured multi-channel Room Impulse Responses (RIRs). Each RIR consisted of 4 channels corresponding to the microphones of a pair of 2-microphone Oticon Epoq Behind-The-Ear (BTE) HAs placed on the ears of a Brüel&Kjær Head And Torso Simulator (HATS).

**Table 1**. Acoustic parameters of the rooms simulated in the evaluation experiment.

| Room | $\mathrm{T}_{60}$ [s] | $\mathrm{C}_{50}$ [dB] | DRR [dB] |
|---|---|---|---|
| Bathroom | 0.8 | 5.2 | −10.1 |
| Cellar | 1.2 | 5.7 | 2.2 |
| Staircase | 2.3 | 11.0 | 4.1 |
| Office | 1.4 | 8.8 | 2.3 |
| Auditorium | 1.3 | 13.4 | 5.2 |
| Isotropic | 1.0 | 4.7 | −0.4 |

Five RIRs were recorded in real rooms with the source of the probe sound placed in front of the HATS at a distance between 0.9 m and 2 m. The reverberation time $\mathrm{T}_{60}$, the clarity index $\mathrm{C}_{50}$ and the Direct-to-Reverberation Ratio (DRR) calculated from these RIRs are given in the upper part of Table 1. In none of the used rooms the reverberation was truly isotropic and in some of them it was strongly dominated by certain directions (especially in the bathroom and auditorium). In that sense, the used RIRs constitute a fair sample of reverberant conditions a hearing aid user might encounter.

A sixth, synthetic RIR was designed to measure the performance of the proposed algorithm in conditions completely matching the underlying assumptions on reverberation isotropy. The reverberation tail of the synthetic RIR was modeled by a sum of 72 exponentially decaying independent white noise sequences, filtered through anechoic Head Related Transfer Functions (HRTFs) measured for 72 evenly spaced positions on the horizontal circle of the HATS. The direct path component of this RIR was computed from the HRTF of a frontally placed sound source. The HRTFs were recorded with an equivalent HA/HATS combination as the real RIR measurements. Parameters of the synthesized RIR are given in the last row of Table 1 (denoted as "Isotropic").

### 4.2. Implementation of the proposed algorithm

The simulated reverberant microphone signals were transformed into time-frequency samples $y_m(k,n)$ using an STFT filterbank. An inverse STFT combined with an overlap-add procedure was used to resynthesize the output signal (see Fig. 1). The frame length of the analysis was 8 ms with 50% overlap between consecutive frames. Traditionally, longer frame lengths are used in speech processing, however, in hearing aids short processing delay is a strong design constraint. A square root Hann window function was used in both the analysis and the synthesis filterbank. A sampling frequency of 16 kHz was used based on the assumption that frequencies above 8 kHz are negligible in speech perception.

In order to implement the algorithm with (5), (6), and (7), $\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)$, $\mathbf{d}$, and $\mathbf{\Gamma}_{\text{iso}}$ are needed. The input covariance matrix $\hat{\mathbf{\Phi}}_{\mathbf{y}}(n)$ was estimated from $\mathbf{y}(n)$ using recursive averaging with a time constant of 40 ms.

For each reverberant condition a different steering vector $\mathbf{d}$ was calculated from the respective RIR truncated to the part containing only the direct path response. Vectors $\mathbf{d}$ were computed by discrete Fourier transformation of the truncated RIRs after appropriate zero-padding. In the synthetic reverberation condition, $\mathbf{d}$ was computed from the anechoic impulse response of the target direction.

The normalized covariance matrix of the isotropic sound field $\mathbf{\Gamma}_{\text{iso}}$ was modeled as

$$\mathbf{\Gamma}_{\text{iso}} = \frac{1}{S}\sum_{s=1}^{S} \mathbf{d}_{\text{hrtf}}(\alpha_s)\mathbf{d}_{\text{hrtf}}^{H}(\alpha_s), \tag{8}$$

where each relative transfer function vector $\mathbf{d}_{\text{hrtf}}(\alpha_s)$ corresponded to the HRTF measured in an anechoic chamber for one of the azimuth angles $\alpha_s \in \{5°, 10°, \ldots, 360°\}$ using the HA/HATS. In this way, $\mathbf{\Gamma}_{\text{iso}}$ represents the frequency-dependent inter-microphone covariance matrix (up to a scalar multiplication) of a cylindrically isotropic sound field.

## 5. PERFORMANCE EVALUATION

The evaluation of the proposed algorithm was based on three objective performance measures: Speech-to-Reverberation Modulation energy Ratio (SRMR) [13], Frequency-Weighted Segmental SNR (FWSegSNR) [14] and Perceptual Evaluation of Speech Quality (PESQ) [14]. Their Matlab implementations were obtained from the 2014 Reverb Challenge [15] website. The evaluation results are presented in Fig. 2.

The three performance measures were calculated for: the unprocessed reverberant signal $y_1(n)$, the signal processed by the beamformer only ($\mathbf{w}_{\text{mvdr}}^{H}\mathbf{y}(n)$), and the reverberant signal enhanced by the full algorithm ($\hat{s}(n)$) (see (5) and Fig. 1). The results calculated from these signals are denoted as "Input", "MVDR", and "MWF", respectively. The proposed algorithm was evaluated for two different microphone array configurations: 4-microphone (using both HAs), and 2-microphone (using only the left HA). In the 4-microphone case we assume that the signals are communicated between the two hearing aids instantly and without error.

The reference signal used to compute FWSegSNR and PESQ was the direct path speech signal $s(n)$. In case of the SRMR, which is a non-intrusive measure, the score of the reference signal was also computed and is presented in Fig. 2(b).

### 5.1. Discussion of results

For the simulations with synthetic isotropic reverberation (denoted as "Isotropic"), the proposed algorithm results in an increase of all considered performance measures. Both the MVDR beamformer and the SC Wiener filter stages of the algorithm contribute positively to that increase. Moreover, the 4-microphone configuration results in a better performance than the 2-microphone configuration. This is an indication, that the proposed method is able to use and benefit from the additional spatial information available in the 4-microphone setup, i.e. when two HAs are used.

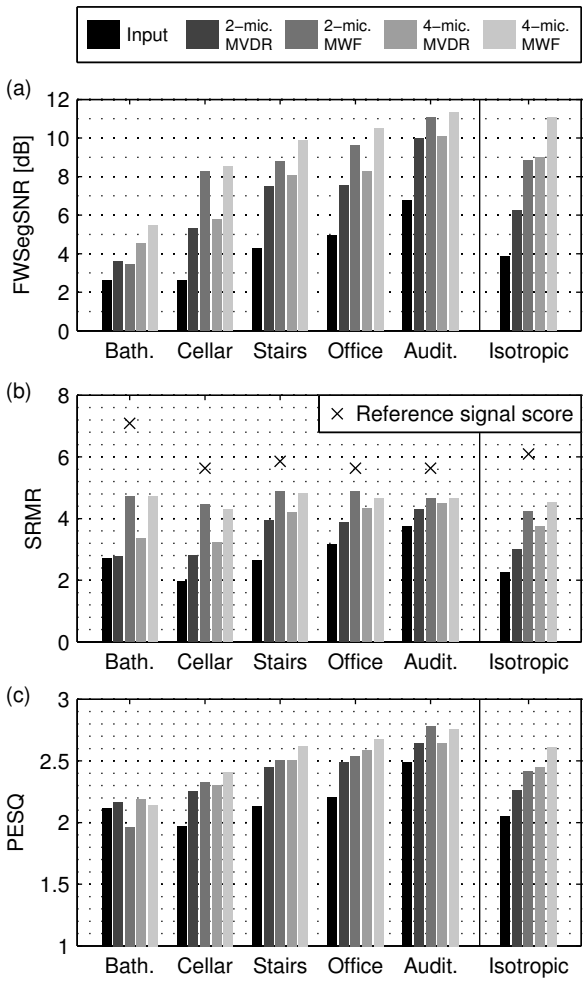


**Fig. 2**. (a) FWSegSNR, (b) SRMR and (c) PESQ scores of the reverberant ("Input"), and processed ("MVDR" and "MWF") signals for different reverberation conditions, and configurations of the microphone array.

In most simulations with RIRs measured in real rooms the increase in the performance measures was lower than in the synthetic isotropic reverberation condition. Nonetheless, in some cases the improvement was of similar magnitude (in the cellar, staircase, and office conditions). This suggests that the isotropic late reverberation model is sufficiently accurate in many real-world reverberant environments and can be used to effectively dereverberate the signal. The increase of the performance scores was smaller in simulations using the RIR of the auditorium, and even negative in the bathroom con-

dition (PESQ and FWSegSNR). Analysis of these two RIRs revealed that the isotropy assumption was not valid in these situations because of isolated specular reflections dominating the reverberation.

The sound quality and speech intelligibility of the processed signals was subjectively assessed through informal listening tests. The perceptual gain from using the algorithm was most pronounced in the simulated isotropic reverberation condition. In the cellar, staircase and the office conditions, the speech was audibly dereverberated and the sound quality was almost unaffected. In the auditorium and particularly in the bathroom conditions, sound artifacts were noticeable.

It is relevant to mention, that the algorithm proposed in this paper is also applicable to target signals other than speech and to interference types other than reverberation. However, it is a prerequisite that the spatial distribution of the interference is isotropic or is otherwise known or estimated. Although the evaluation of the proposed algorithm was conducted in reverberant-only condition, it is reasonable to expect similar performance in an arbitrary isotropic non-stationary noise.

## 6. CONCLUSION

In this paper we have proposed a Multi-channel Wiener Filter (MWF) which uses joint Maximum Likelihood Estimation (MLE) of speech and reverberation spectral variances. The MLE method was adopted from the work of Ye and DeGroat [7]. The proposed MWF algorithm was implemented and its speech dereverberation performance for hearing aids was evaluated. It was shown that the proposed algorithm performs well in both synthetic and realistic reverberation conditions. The performance of the proposed method was best when the assumption on the interference isotropy was close to valid. In non-isotropic reverberation/ambient noise conditions on-line estimation of the interference covariance matrix structure could be used to improve the performance. This is a topic for future research.

## REFERENCES


[1] A.K. Nabelek and J.M. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech, Lang. Hearing Res.*, vol. 17, no. 4, pp. 724–739, 1974.

[2] V. Hamacher et al., "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP J. Appl. Signal Process.*, pp. 2915–2929, 2005.

[3] K.U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays – Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Springer, 2001.

[4] S. Braun and E.A.P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. 21st Eur. Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, 2013.

[5] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Bucharest, Romania, 2012, pp. 295–299.

[6] B. Cornelis, M. Moonen, and J. Wouters, "Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel wiener filtering based noise reduction," *J. Acoust. Soc. Am.*, vol. 131, no. 6, pp. 4743–4755, 2012.

[7] H. Ye and R.D. DeGroat, "Maximum likelihood DOA estimation and asymptotic Cramér-Rao bounds for additive unknown colored noise," *IEEE Trans. Signal Process.*, vol. 43, no. 4, pp. 938–949, 1995.

[8] J.S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244, 2003.

[9] B.N. Gover, J.G. Ryan, and M.R. Stinson, "Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2138–2148, 2004.

[10] G.W. Elko, E. Diethorn, and T. Gänsler, "Room impulse response variation due to temperature fluctuations and its impact on acoustic echo cancellation," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, 2003, pp. 67–70.

[11] J. Mourjopoulos, "On the variation and invertibility of room impulse response functions," *Journal of Sound and Vibration*, vol. 102, no. 2, pp. 217 – 228, 1985.

[12] J.S. Garofolo et al., *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM*, NIST, 1993.

[13] 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, 2010.

[14] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[15] K. Kinoshita et al., "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2013.