A Simulation Study on Binaural Dereverberation and Noise Reduction based on Diffuse Power Spectral Density Estimators

Ina Kodrasi, Daniel Marquardt, Simon Doclo

University of Oldenburg, Department of Medical Physics and Acoustics
and Cluster of Excellence Hearing4All, Oldenburg, Germany
{ina.kodrasi,daniel.marquardt,simon.doclo}@uni-oldenburg.de

Abstract

Enhancement techniques in binaural hearing aids are crucial to improve speech understanding for hearing impaired persons in reverberation and noise. Since reverberation and noise can be commonly modeled as diffuse sound fields, many state-of-the-art techniques require an estimate of the diffuse power spectral density (PSD). In this paper we evaluate the performance of binaural dereverberation and noise reduction techniques using several diffuse PSD estimators in realistic acoustic scenarios. Two state-of-the-art techniques are considered, i.e., the binaural multi-channel Wiener filter and the binaural minimum variance distortionless response beamformer with partial noise estimation followed by a postfilter. The considered diffuse PSD estimators are blocking matrix-based and eigenvalue decomposition-based estimators. A least-squares generalization of dual-channel blocking matrix-based estimators to the multi-channel case is also presented, yielding the same diffuse PSD estimate as a recently proposed maximum likelihood estimator. Simulation results show the applicability of diffuse PSD estimators for binaural dereverberation and noise reduction, with the eigenvalue decomposition-based estimators always yielding the best performance.

Index Terms: hearing aids, binaural cues, blocking matrix, eigenvalue decomposition

1. Introduction

Dereverberation and noise reduction techniques in binaural hearing aids are crucial to improve speech intelligibility for hearing impaired persons [1]. In addition to reducing the interference, i.e., reverberation and noise, another important objective of such techniques is the preservation of the listener's impression of the acoustical scene by preserving the binaural cues of the speech source and of the interference [2, 3].

In [2] the binaural multi-channel Wiener filter (MWF) has been presented, which can be decomposed into a binaural minimum variance distortionless response (MVDR) beamformer and a single-channel Wiener postfilter. The binaural MWF and MVDR beamformer preserve the binaural cues of the desired speech source, but distort the cues of the interference such that both the speech source and the residual interference are perceived as coming from the same direction [4]. In order to also (partially) preserve the binaural cues of the residual interference, the binaural MWF with partial noise estimation (MWF-N) [4, 5] and the binaural MVDR beamformer with partial noise estimation (MVDR-N) [3] have been proposed, where a trade-off parameter controls the trade-off between interference reduction and cue preservation. The trade-off parameter yielding a desired cue preservation level can be analytically computed only for the binaural MVDR-N beamformer [3], making it a computationally advantageous technique in comparison to the MWF-N. To further increase the interference reduction performance, a single-channel Wiener postfilter can be applied at the output of the MVDR-N beamformer [3]. Since reverberation is commonly modeled as a diffuse sound field [6–11] and since diffuse background noise is commonly encountered in binaural applications, these binaural speech enhancement techniques require (among other parameters) an estimate of the diffuse power spectral density (PSD).

Several multi-channel diffuse PSD estimators have been proposed, such as blocking matrix-based estimators [6, 8, 12–15] and eigenvalue decomposition-based estimators [10, 11]. Blocking matrix-based estimators estimate the diffuse PSD by blocking the target signal using knowledge of the direction of arrival (DOA) of the speech source [6, 8, 15], blind source separation methods [13], or blind system identification methods [14]. The multi-channel blocking matrix-based estimator in [6] uses a maximum likelihood formulation to estimate the diffuse PSD from multiple reference signals, whereas the dual-channel blocking matrix-based estimators in [13–15] estimate the diffuse PSD by solving an equation based on a single reference signal. Eigenvalue decomposition-based estimators on the other hand do not require a blocking matrix and directly estimate the diffuse PSD using the eigenvalues of the prewhitened input PSD matrix.

The objective of this paper is to evaluate the performance of the binaural MVDR beamformer followed by a postfilter (i.e., the binaural MWF) and the binaural MVDR-N beamformer followed by a postfilter using blocking matrix-based and eigenvalue decomposition-based diffuse PSD estimators. In addition, a least-squares generalization of the dual-channel blocking matrix-based estimators from [13–15] to the multi-channel case is presented, which happens to be equivalent to the multi-channel estimator from [6]. The blocking matrix is constructed based on the DOA of the speech source, which is estimated using the binaural DOA estimator proposed in [15]. Simulation results show that the performance of all considered diffuse PSD estimators is high, with the eigenvalue decomposition-based PSD estimators resulting in the best performance. In addition, it is shown that the performance of the blocking matrix-based dual-channel estimator from [15] is very similar to the performance of the blocking matrix-based multi-channel estimator from [6], suggesting that increasing the number of microphones within the blocking matrix-based framework does not increase the diffuse PSD estimation accuracy.

2. Configuration and Notation

We consider a binaural hearing aid configuration consisting of $M = M_{\mathrm{L}} + M_{\mathrm{R}}$ microphones, with M_{L} denoting the number of microphones of the left hearing aid and M_{R} denoting the number of microphones of the right hearing aid. In the short-time Fourier transform domain, the M-dimensional vector of the re-

This work was supported in part by the Cluster of Excellence 1077 "Hearing4All", funded by the German Research Foundation (DFG), and the joint Lower Saxony-Israeli Project ATHENA, funded by the State of Lower Saxony.

ceived microphone signals at frequency index k and frame index l can be written as

$$\mathbf{y}(k,l) = [Y_{\mathrm{L},1}(k,l) \ \dots \ Y_{\mathrm{L},M_\mathrm{L}}(k,l) \\ Y_{\mathrm{R},1}(k,l) \ \dots \ Y_{\mathrm{R},M_\mathrm{R}}(k,l)]^T, \tag{1}$$

with $Y_{\{\mathrm{L,R}\},m}(k,l)$ the m-th microphone signal of the left and right hearing aid. In a reverberant and noisy acoustic scenario, $\mathbf{y}(k,l)$ is given by

$$\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{d}(k,l) + \mathbf{v}(k,l), \tag{2}$$

with $\mathbf{x}(k,l)$ the direct and early reverberation speech component, $\mathbf{d}(k,l)$ the diffuse sound component, and $\mathbf{v}(k,l)$ the noise component. The diffuse sound component $\mathbf{d}(k,l)$ models the late reverberation [6–11] as well as any noise which can be well approximated by a diffuse sound field, such as background noise in large crowded rooms. The noise component $\mathbf{v}(k,l)$ represents any remaining noise which cannot be modeled by a diffuse sound field, such as uncorrelated sensor noise. For conciseness, the frequency index k will be omitted in the remainder of this paper.

For a single-source scenario, the direct and early reverberation speech component $\mathbf{x}(l)$ can be expressed in terms of the target signals $S_\mathrm{L}(l)$ and $S_\mathrm{R}(l)$ (i.e., direct and early reverberation speech components) in the reference microphones of the left and right hearing aids as

$$\mathbf{x}(l) = S_{\{\mathrm{L,R}\}}(l)\mathbf{a}_{\{\mathrm{L,R}\}}(l), \tag{3}$$

with $\mathbf{a}_\mathrm{L}(l)$ and $\mathbf{a}_\mathrm{R}(l)$ the M-dimensional vectors of relative early transfer functions (RETFs) of the target signals from the reference microphones to all M microphones. The target signal $S_{\{\mathrm{L,R}\}}(l)$ is often defined as the direct speech component only [6–11], such that the vector $\mathbf{a}_{\{\mathrm{L,R}\}}(l)$ can be constructed based on a DOA estimate and head models or measurements of anechoic acoustic transfer functions (ATFs). The PSD matrix of the microphone signals is defined as

$$\mathbf{\Phi}_\mathbf{y}(l) = \mathcal{E}\{\mathbf{y}(l)\mathbf{y}^H(l)\}, \tag{4}$$

with $\mathcal{E}\{\cdot\}$ the expected value operator. As in many speech enhancement techniques, in the following it is assumed that the components in (2) are mutually uncorrelated, such that $\mathbf{\Phi}_\mathbf{y}(l)$ can be written as

$$\mathbf{\Phi}_\mathbf{y}(l) = \mathbf{\Phi}_\mathbf{x}(l) + \mathbf{\Phi}_\mathbf{d}(l) + \mathbf{\Phi}_\mathbf{v}(l), \tag{5}$$

with $\mathbf{\Phi}_\mathbf{x}(l)$, $\mathbf{\Phi}_\mathbf{d}(l)$, and $\mathbf{\Phi}_\mathbf{v}(l)$ denoting the PSD matrices of $\mathbf{x}(l)$, $\mathbf{d}(l)$, and $\mathbf{v}(l)$, respectively. Using (3), $\mathbf{\Phi}_\mathbf{y}(l)$ can be expressed as

$$\mathbf{\Phi}_\mathbf{y}(l) = \underbrace{\Phi_{S_{\{\mathrm{L,R}\}}}(l)\mathbf{a}_{\{\mathrm{L,R}\}}(l)\mathbf{a}^H_{\{\mathrm{L,R}\}}(l)}_{\mathbf{\Phi}_\mathbf{x}(l)} + \underbrace{\Phi_\mathrm{d}(l)\mathbf{\Gamma}}_{\mathbf{\Phi}_\mathbf{d}(l)} + \mathbf{\Phi}_\mathbf{v}(l), \tag{6}$$

with $\Phi_{S_{\{\mathrm{L,R}\}}}(l)$ the time-varying PSD of the target signal, i.e., $\Phi_{S_{\{\mathrm{L,R}\}}}(l) = \mathcal{E}\{|S_{\{\mathrm{L,R}\}}(l)|^2\}$, $\Phi_\mathrm{d}(l)$ the time-varying PSD of the diffuse sound component, and $\mathbf{\Gamma}$ the time-invariant spatial coherence matrix of the diffuse sound field. The spatial coherence matrix $\mathbf{\Gamma}$ is assumed to be known, since it can be constructed based on head models [16] or measurements of anechoic ATFs [9, 17]. In order to simplify the notation, in the following we define the interference component $\mathbf{u}(l) = \mathbf{d}(l) + \mathbf{v}(l)$ and the interference PSD matrix

$$\mathbf{\Phi}_\mathbf{u}(l) = \Phi_\mathrm{d}(l)\mathbf{\Gamma} + \mathbf{\Phi}_\mathbf{v}(l). \tag{7}$$

The objective of binaural speech enhancement techniques is to suppress the interference and obtain estimates of the target signals $\hat{S}_\mathrm{L}(l)$ and $\hat{S}_\mathrm{R}(l)$ by applying M-dimensional filter vectors $\mathbf{w}_\mathrm{L}(l)$ and $\mathbf{w}_\mathrm{R}(l)$ to all microphone signals (cf. Section 3), i.e.,

$$\hat{S}_{\{\mathrm{L,R}\}}(l) = \mathbf{w}^H_{\{\mathrm{L,R}\}}(l)\mathbf{y}(l). \tag{8}$$

The time-varying input interaural coherence (IC) of the interference is defined as

$$\mathrm{IC}_\mathrm{in}(l) = \frac{\mathbf{e}_\mathrm{L}^T\mathbf{\Phi}_\mathbf{u}(l)\mathbf{e}_\mathrm{R}}{\sqrt{\mathbf{e}_\mathrm{L}^T\mathbf{\Phi}_\mathbf{u}(l)\mathbf{e}_\mathrm{L}\mathbf{e}_\mathrm{R}^T\mathbf{\Phi}_\mathbf{u}(l)\mathbf{e}_\mathrm{R}}}, \tag{9}$$

with $\mathbf{e}_{\{\mathrm{L,R}\}}$ an M-dimensional selector vector with one element equal to 1 and all other elements equal to 0 such that $\mathbf{e}^T_{\{\mathrm{L,R}\}}\mathbf{a}_{\{\mathrm{L,R}\}}(l) = 1$. The time-varying output IC of the interference is defined as

$$\mathrm{IC}_\mathrm{out}(l) = \frac{\mathbf{w}_\mathrm{L}^H(l)\mathbf{\Phi}_\mathbf{u}(l)\mathbf{w}_\mathrm{R}(l)}{\sqrt{\mathbf{w}_\mathrm{L}^H(l)\mathbf{\Phi}_\mathbf{u}(l)\mathbf{w}_\mathrm{L}(l)\mathbf{w}_\mathrm{R}^H(l)\mathbf{\Phi}_\mathbf{u}(l)\mathbf{w}_\mathrm{R}(l)}}. \tag{10}$$

Since the IC is complex-valued, binaural speech enhancement techniques typically aim at preserving the real-valued magnitude-squared coherence (MSC) of the interference, defined as

$$\mathrm{MSC}(l) = |\mathrm{IC}(l)|^2. \tag{11}$$

3. Binaural Speech Enhancement

In this section the derivation of the filter $\mathbf{w}_{\{\mathrm{L,R}\}}(l)$ based on the binaural MVDR and MVDR-N beamformers followed by a Wiener postfilter is briefly discussed.

3.1. Binaural MVDR and MVDR-N beamformers

The binaural MVDR beamformer [2] aims at minimizing the output PSD of the interference while preserving the target signal in the left and right reference microphones. The binaural MVDR beamformer can be computed as

$$\mathbf{w}^\mathrm{MVDR}_{\{\mathrm{L,R}\}}(l) = \frac{\mathbf{\Phi}_\mathbf{u}^{-1}(l)\mathbf{a}_{\{\mathrm{L,R}\}}(l)}{\mathbf{a}^H_{\{\mathrm{L,R}\}}(l)\mathbf{\Phi}_\mathbf{u}^{-1}(l)\mathbf{a}_{\{\mathrm{L,R}\}}(l)}. \tag{12}$$

As shown in [4], the beamformer in (12) preserves the binaural cues of the speech source but distorts the output MSC of the interference such that both the speech source and the residual interference are perceived as coming from the same direction. In order to better preserve the interference output MSC, and hence, the impression of the acoustical scene, the binaural MVDR-N beamformer has been proposed [3]. Aiming at preserving both the target signal as well as a scaled version of the interference in the left and right reference microphones, the binaural MVDR-N beamformer can be computed as

$$\mathbf{w}^\mathrm{MVDR\text{-}N}_{\{\mathrm{L,R}\}}(l) = [1-\eta(l)]\mathbf{w}^\mathrm{MVDR}_{\{\mathrm{L,R}\}}(l) + \eta(l)\mathbf{e}_{\{\mathrm{L,R}\}}, \tag{13}$$

where $\eta(l)$ denotes a (real-valued) scaling parameter between 0 and 1 which provides a trade-off between interference reduction and MSC preservation. The value of the parameter $\eta(l)$ yielding a desired user-defined interference output MSC can be computed analytically [3].

3.2. Wiener postfilter

In order to further increase the interference reduction performance, a single-channel Wiener postfilter can be applied at the output of the MVDR and MVDR-N beamformers [3, 18], i.e.,

$$G_{\{\mathrm{L,R}\}}(l) = \frac{\xi_{\{\mathrm{L,R}\}}(l)}{1+\xi_{\{\mathrm{L,R}\}}(l)}, \tag{14}$$

with $\xi_{\{\mathrm{L,R}\}}(l)$ the a-priori signal-to-interference ratio (SIR) at the beamformer output in the left and right hearing aid. The a-priori SIR can be estimated using the decision-directed approach based on an estimate of the interference PSD at the beamformer output [19]. The interference PSD at the beamformer output can be computed as

$$\Phi^{\mathrm{out}}_{\{\mathrm{L,R}\},\mathrm{u}}(l) = \mathbf{w}^H_{\{\mathrm{L,R}\}}(l)\boldsymbol{\Phi}_{\mathbf{u}}(l)\mathbf{w}_{\{\mathrm{L,R}\}}(l), \tag{15}$$

with $\mathbf{w}_{\{\mathrm{L,R}\}}(l)$ the MVDR beamformer in (12) or the MVDR-N beamformer in (13). In order to preserve the binaural cues of the speech source and interference, a common postfilter $G(l)$ is applied to both hearing aids, with

$$G(l) = \sqrt{G_{\mathrm{L}}(l)G_{\mathrm{R}}(l)}. \tag{16}$$

In summary, in Section 5 we consider two different methods for computing the filter $\mathbf{w}_{\{\mathrm{L,R}\}}(l)$, i.e.,

1. using an MVDR beamformer and a Wiener postfilter:

$$\mathbf{w}_{\{\mathrm{L,R}\}}(l) = \mathbf{w}^{\mathrm{MVDR}}_{\{\mathrm{L,R}\}}(l)G(l) \tag{17}$$

2. using an MVDR-N beamformer and a Wiener postfilter:

$$\mathbf{w}_{\{\mathrm{L,R}\}}(l) = \mathbf{w}^{\mathrm{MVDR\text{-}N}}_{\{\mathrm{L,R}\}}(l)G(l) \tag{18}$$

Computing the filters in (17) and (18) requires estimates of the diffuse PSD $\Phi_{\mathrm{d}}(l)$, noise PSD matrix $\boldsymbol{\Phi}_{\mathbf{v}}(l)$, and RETF vector $\mathbf{a}_{\{\mathrm{L,R}\}}(l)$.

4. Diffuse PSD Estimators

In this section it is assumed that estimates of the noise PSD matrix $\boldsymbol{\Phi}_{\mathbf{v}}(l)$ and RETF vector $\mathbf{a}_{\{\mathrm{L,R}\}}(l)$ are available, such that only the diffuse PSD $\Phi_{\mathrm{d}}(l)$ needs to be estimated. The noise PSD matrix $\boldsymbol{\Phi}_{\mathbf{v}}(l)$ can in practice be estimated from the microphone signals using e.g. a multi-channel speech presence probability estimator [20]. The RETF vector $\mathbf{a}_{\{\mathrm{L,R}\}}(l)$ can in practice be estimated as in Section 5, i.e., using a DOA estimator and measurements of anechoic ATFs [15]. To estimate the diffuse PSD $\Phi_{\mathrm{d}}(l)$, we consider blocking matrix-based and eigenvalue decomposition-based estimators.

4.1. Blocking matrix-based estimators

In [13–15] dual-channel (i.e., $M = 2$) diffuse PSD estimators using a single reference signal at the output of a blocking matrix have been proposed. In the following, a least-squares generalization of these estimators for $M > 2$ is presented.

In order to estimate the diffuse PSD, an $M \times (M-1)$-dimensional blocking matrix $\mathbf{B}(l)$ can be used to generate a set of $M-1$ reference signals containing only the interference component, i.e.,

$$\tilde{\mathbf{u}}(l) = \mathbf{B}^H(l)\mathbf{y}(l), \tag{19}$$

with $\mathbf{B}(l)$ such that $\mathbf{B}^H(l)\mathbf{a}_{\mathrm{L}}(l) = \mathbf{0}$ or $\mathbf{B}^H(l)\mathbf{a}_{\mathrm{R}}(l) = \mathbf{0}$. Using $\mathbf{a}_{\mathrm{L}}(l)$, a blocking matrix can be computed from the first $M-1$ columns of the matrix $\mathbf{T}(l)$ defined as

$$\mathbf{T}(l) = \mathbf{I} - \frac{\mathbf{a}_{\mathrm{L}}(l)\mathbf{a}^H_{\mathrm{L}}(l)}{\|\mathbf{a}_{\mathrm{L}}(l)\|^2}, \tag{20}$$

where $\mathbf{I}$ denotes the $M \times M$-dimensional identity matrix. It should be noted that many blocking matrices exist and one can also be computed using $\mathbf{a}_{\mathrm{R}}(l)$ instead of $\mathbf{a}_{\mathrm{L}}(l)$ in (20). Based on (6), the PSD matrix of the $M-1$ reference signals at the blocking matrix output is equal to

$$\boldsymbol{\Phi}_{\tilde{\mathbf{u}}}(l) = \Phi_{\mathrm{d}}(l)\underbrace{\mathbf{B}^H(l)\boldsymbol{\Gamma}\mathbf{B}(l)}_{\tilde{\boldsymbol{\Gamma}}(l)} + \underbrace{\mathbf{B}^H(l)\boldsymbol{\Phi}_{\mathbf{v}}(l)\mathbf{B}(l)}_{\boldsymbol{\Phi}_{\tilde{\mathbf{v}}}(l)}. \tag{21}$$

The PSD matrix $\boldsymbol{\Phi}_{\tilde{\mathbf{u}}}(l)$ can be directly estimated from $\tilde{\mathbf{u}}(l)$, whereas the matrices $\tilde{\boldsymbol{\Gamma}}(l)$ and $\boldsymbol{\Phi}_{\tilde{\mathbf{v}}}(l)$ can be computed using the available diffuse coherence matrix $\boldsymbol{\Gamma}$ and the available noise PSD matrix $\boldsymbol{\Phi}_{\mathbf{v}}(l)$. Since the only unknown quantity is the diffuse PSD $\Phi_{\mathrm{d}}(l)$, the system of equations in (21) represents an overdetermined system of equations. A least-squares estimate of the diffuse PSD can be obtained by minimizing the cost function

$$J(l) = \|\boldsymbol{\Phi}_{\tilde{\mathbf{u}}}(l) - \boldsymbol{\Phi}_{\tilde{\mathbf{v}}}(l) - \Phi_{\mathrm{d}}(l)\tilde{\boldsymbol{\Gamma}}(l)\|^2_F, \tag{22}$$

where $\|\cdot\|_F$ denotes the matrix Frobenious norm. Setting the derivative of (22) with respect to $\Phi_{\mathrm{d}}(l)$ equal to 0, the least-squares estimate of the diffuse PSD can be computed as

$$\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d}}(l) = \frac{\mathrm{trace}\{[\boldsymbol{\Phi}_{\tilde{\mathbf{u}}}(l) - \boldsymbol{\Phi}_{\tilde{\mathbf{v}}}(l)]^H\tilde{\boldsymbol{\Gamma}}(l)\}}{\mathrm{trace}\{\tilde{\boldsymbol{\Gamma}}^H(l)\tilde{\boldsymbol{\Gamma}}(l)\}}, \tag{23}$$

where $\mathrm{trace}\{\cdot\}$ denotes the trace operator. For $M = 2$, $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d}}(l)$ is equal to the PSD estimate derived in [13–15]. Interestingly, for $M > 2$, $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d}}(l)$ is equal to the maximum likelihood PSD estimate derived in [6].

4.2. Eigenvalue decomposition-based estimators

While the estimator in Section 4.1 requires knowledge of the RETF vector, an RETF-independent eigenvalue decomposition-based PSD estimator is proposed in [10, 11]. This estimator requires knowledge of the PSD matrix $\boldsymbol{\Phi}_{\mathbf{c}}(l) = \boldsymbol{\Phi}_{\mathbf{x}}(l) + \boldsymbol{\Phi}_{\mathbf{d}}(l)$, which can be computed as

$$\boldsymbol{\Phi}_{\mathbf{c}}(l) = \boldsymbol{\Phi}_{\mathbf{y}}(l) - \boldsymbol{\Phi}_{\mathbf{v}}(l), \tag{24}$$

with $\boldsymbol{\Phi}_{\mathbf{y}}(l)$ directly estimated from the microphone signals. Based on (6), the prewhitened PSD matrix $\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\mathbf{c}}(l)$ is equal to the sum of a rank-1 matrix and a scaled identity matrix, i.e.,

$$\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\mathbf{c}}(l) = \Phi_{S_{\{\mathrm{L,R}\}}}(l)\boldsymbol{\Gamma}^{-1}\mathbf{a}_{\{\mathrm{L,R}\}}(l)\mathbf{a}^H_{\{\mathrm{L,R}\}}(l) + \Phi_{\mathrm{d}}(l)\mathbf{I}. \tag{25}$$

As a result, the eigenvalues of $\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\mathbf{c}}(l)$ are equal to

$$\lambda_1\{\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\mathbf{c}}(l)\} = \sigma(l) + \Phi_{\mathrm{d}}(l), \tag{26}$$

$$\lambda_j\{\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\mathbf{c}}(l)\} = \Phi_{\mathrm{d}}(l), \;\; j = 2, \ldots, M, \tag{27}$$

with $\sigma(l)$ the only non-zero eigenvalue of the rank-1 term in (25). In [11] it is proposed to estimate the diffuse PSD using any of the last $M-1$ eigenvalues $\lambda_j\{\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\mathbf{c}}(l)\}$, $j = 2, \ldots, M$. Due to signal model violations and estimation errors in $\boldsymbol{\Phi}_{\mathbf{c}}(l)$, the last $M-1$ eigenvalues of $\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\mathbf{c}}(l)$ are not equal in practice. In this paper we consider two alternative eigenvalue decomposition-based PSD estimates $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}(l)$ and $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_2}(l)$, with $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}(l)$ computed as the mean of the last $M-1$ eigenvalues and $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_2}(l)$ computed as the second eigenvalue, i.e.,

$$\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}(l) = \frac{\mathrm{trace}\{\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\mathbf{c}}(l)\} - \lambda_1\{\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\mathbf{c}}(l)\}}{M-1}, \tag{28}$$

$$\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_2}(l) = \lambda_2\{\boldsymbol{\Gamma}^{-1}\boldsymbol{\Phi}_{\mathbf{c}}(l)\}. \tag{29}$$

Using any diffuse PSD estimate in (23), (28), or (29), the available coherence matrix $\boldsymbol{\Gamma}$, and the available noise PSD matrix $\boldsymbol{\Phi}_{\mathbf{v}}(l)$, an estimate of the interference PSD matrix $\boldsymbol{\Phi}_{\mathbf{u}}(l)$ in (7) can now be computed.

5. Experimental Results

In this section the dereverberation and noise reduction performance using the filters in (17) and (18) is investigated for different reverberation times and signal-to-noise ratios (SNRs). In addition, the performance is investigated for a stationary speaker as well as for a moving speaker. In order to focus on the diffuse sound suppression, in the following it is assumed that the microphone signals consist only of a direct and early reverberation speech component and a diffuse sound component (i.e., late reverberation and diffuse background noise), i.e., $\mathbf{v}(l) = \mathbf{0}$ and $\mathbf{\Phi_u}(l) = \Phi_{\mathrm{d}}(l)\mathbf{\Gamma}$. For $\mathbf{\Phi_u}(l) = \Phi_{\mathrm{d}}(l)\mathbf{\Gamma}$, the MVDR and MVDR-N beamformers in (12) and (13) can be constructed using only the diffuse spatial coherence matrix $\mathbf{\Gamma}$ (i.e., the scalar $\Phi_{\mathrm{d}}(l)$ cancels out).

5.1. Setup

Signals were recorded in a laboratory with variable acoustics at the University of Oldenburg using two 2-channel behind-the-ear hearing aid dummies placed on the ears of a head-and-torso simulator (HATS), i.e., $M_{\mathrm{L}} = 2$, $M_{\mathrm{R}} = 2$, and $M = 4$. The stationary speaker was simulated by playing back clean speech from a loudspeaker placed at a distance of 2 m from the center of the head. Two stationary speaker scenarios were generated by placing the loudspeaker at two different angles θ_1 and θ_2, with $\theta_1 = 35^\circ$ and $\theta_2 = -35^\circ$. The considered reverberation times for the stationary speaker scenarios were $T_{60} \in \{0.5\ \mathrm{s}, 0.75\ \mathrm{s}, 1\ \mathrm{s}\}$. The moving speaker was a human speaker naturally walking in the frontal hemisphere of the HATS. The considered reverberation time for the moving speaker scenario was $T_{60} \approx 1$ s. To simulate a diffuse noise field, the background noise was generated by placing four loudspeakers facing the corners of the laboratory playing back uncorrelated multi-talker noise. It should be noted that although this background noise was not perfectly diffuse, its MSC was rather similar to the MSC of a diffuse noise field. The speech and the noise signals were recorded separately such that we were able to mix them at different input SNRs (iSNRs) afterward. The considered iSNRs are $\mathrm{iSNR} \in \{0\ \mathrm{dB}, 5\ \mathrm{dB}, \ldots, 20\ \mathrm{dB}\}$.

The signals are processed using a weighted overlap-add framework with a frame size of 512 samples and an overlap of 50% at a sampling frequency $f_s = 16$ kHz. The first microphone of each hearing aid is arbitrarily selected as the reference microphone. The DOA of the speech source is estimated using the binaural DOA estimator in [15]. It should be noted that the DOA estimate obtained in all considered reverberant and noisy scenarios is highly accurate. Using the estimated DOA, the RETF vector $\mathbf{a}_{\mathrm{L,R}}(l)$ is computed from anechoic ATFs measured on the same dummy head [21]. The diffuse coherence matrix $\mathbf{\Gamma}$ is calculated using spatially averaged auto- and cross-correlations of the anechoic ATFs measured for angles ranging between -180° to 175°. To compute the parameter $\eta(l)$ for the MVDR-N beamformer, the desired interference output MSC is defined based on the frequency-dependent values proposed in [17], which are psychoacoustically motivated [22] and do not alter the listener's impression of a diffuse sound field. The PSD matrices $\mathbf{\Phi_y}(l)$ and $\mathbf{\Phi_{\tilde{u}}}(l)$ are estimated using recursive averaging with a time constant of 40 ms. The minimum gain of the Wiener postfilter $G(l)$ is -20 dB.

The dereverberation and noise reduction performance is evaluated in terms of the improvement in PESQ (ΔPESQ) [23] and frequency-weighted segmental SNR (ΔfSSNR) [24] between the output signal and the reference microphone signal for each hearing aid. The PESQ and fSSNR measures are intrusive measures comparing the signal being evaluated to a desired signal. The desired signal for each hearing aid is generated by convolving the clean speech signal with the measured anechoic ATFs corresponding to the true DOAs. The clean speech signal for the moving speaker scenario is assumed to be the signal recorded with a close-talk microphone. The ΔPESQ and ΔfSSNR presented in the following are average improvements between the left and right hearing aids.

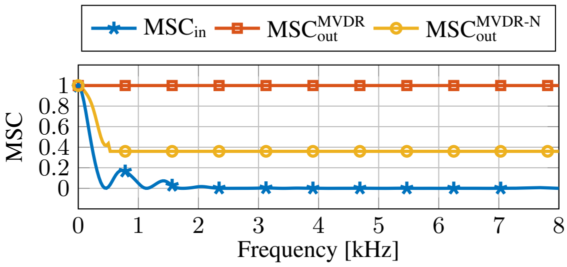

Figure 1: *MSC at the input and output of the MVDR and MVDR-N beamformers.*

The postfilters in (17) and (18) are computed using the blocking matrix-based and eigenvalue decomposition-based diffuse PSD estimators. Two alternative estimates will be investigated for the blocking matrix-based estimator, i.e., $\hat{\Phi}_{\mathrm{d},2}^{\mathrm{BM}}(l)$ denoting the PSD estimate obtained using only the reference microphones on the left and right hearing aids (corresponding to the dual-channel PSD estimator in [15]) and $\hat{\Phi}_{\mathrm{d},4}^{\mathrm{BM}}(l)$ denoting the PSD estimate obtained using all 4 microphones (corresponding to the maximum likelihood PSD estimator in [6]).

5.2. MSC preservation

Since the common Wiener postfilter does not change the binaural cues, to evaluate the interference MSC preservation performance of the considered techniques the MSC is computed at the input and output of the MVDR and MVDR-N beamformers using (9), (10), and (11). Fig. 1 presents the obtained MSC values. Since the interference PSD matrix is modeled by a scaled diffuse coherence matrix, the input MSC is time-invariant and equal to the MSC of a diffuse sound field. Furthermore, the MSC at the output of the MVDR and MVDR-N beamformers is also time-invariant, with the MVDR beamformer always distorting the output MSC and the MVDR-N beamformer always yielding the desired user-defined output MSC. Note that since the late reverberation and the noise are not perfectly diffuse, the interference PSD matrix is not equal to a scaled diffuse coherence matrix in practice. Computing the MSC directly from the signals would yield different results from the ones presented in Fig. 1. However, the presented MSC values do illustrate that in all simulations, the MVDR beamformer distorts the cues of the residual interference whereas the MVDR-N beamformer better preserves them.

5.3. Dereverberation performance for a stationary speaker

In this section the dereverberation performance is investigated for several stationary speaker scenarios with different reverberation times and speaker positions. The presented ΔPESQ and ΔfSSNR are averaged between the considered speaker positions. Fig. 2(a) presents the average ΔPESQ and ΔfSSNR obtained using the MVDR beamformer and a Wiener postfilter with different diffuse PSD estimators. It can be observed that in terms of ΔPESQ, using any diffuse PSD estimator yields a similar improvement, with $\hat{\Phi}_{\mathrm{d},\lambda_1}^{\mathrm{EVD}}$ resulting in a slightly higher ΔPESQ than other PSD estimators. In terms of ΔfSSNR, it can be observed that the eigenvalue decomposition-based estimators yield a larger improvement than the blocking matrix-based es-

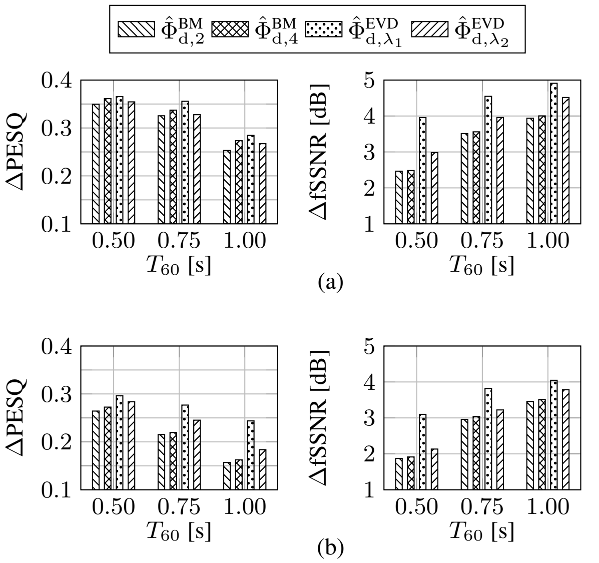

Figure 2: *Dereverberation performance for a stationary speaker using a beamformer and a Wiener postfilter: (a) MVDR and (b) MVDR-N.*

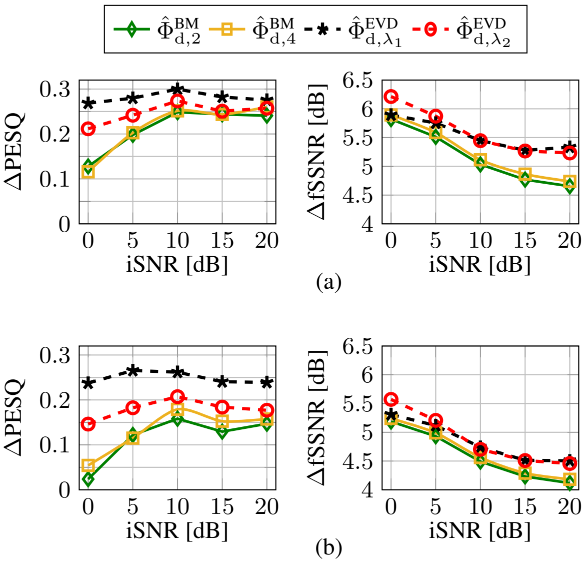

Figure 3: *Dereverberation and noise reduction performance for a stationary speaker using a beamformer and a Wiener postfilter: (a) MVDR and (b) MVDR-N ($T_{60} \approx 1$ s).*

timators, with $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}$ resulting in the best performance. In addition, in terms of both performance measures it appears that the performance obtained using $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},2}$ and $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},4}$ is very similar, suggesting that increasing the number of microphones in the blocking matrix-based framework does not increase the diffuse PSD estimation accuracy. Fig. 2(b) presents the average ΔPESQ and ΔfSSNR obtained using the MVDR-N beamformer and a Wiener postfilter with different diffuse PSD estimators. Overall it can be observed that the performance improvement obtained for all considered reverberation times and diffuse PSD estimators is smaller than in Fig. 2(a). This is to be expected, since the MVDR-N beamformer also (partly) preserves the MSC of the residual interference component (cf. Fig. 1). In terms of both performance measures, it can be observed that the eigenvalue decomposition-based estimators yield a larger improvement than the blocking matrix-based estimators, with $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}$ resulting in the best performance. In addition, similarly to before, the performance obtained using $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},2}$ and $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},4}$ is very similar in terms of both performance measures.

5.4. Dereverberation and noise reduction performance for a stationary speaker

In this section the dereverberation and noise reduction performance is investigated for several stationary speaker scenarios with different iSNRs and speaker positions. The considered reverberation time is $T_{60} \approx 1$ s. The presented ΔPESQ and ΔfSSNR are averaged between the considered speaker positions. Fig. 3(a) presents the average ΔPESQ and ΔfSSNR obtained using the MVDR beamformer and a Wiener postfilter with different diffuse PSD estimators. It can be observed that in terms of both performance measures, the eigenvalue decomposition-based estimators yield a larger improvement than the blocking matrix-based estimators, with $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}$ resulting in the best ΔPESQ and $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_2}$ resulting in the best ΔfSSNR for low iSNRs. In addition, it can be observed that $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},2}$ and $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},4}$ yield a very similar performance in terms of both performance measures. Fig. 3(b) presents the average ΔPESQ and ΔfSSNR obtained using the MVDR-N beamformer and a Wiener postfilter with different diffuse PSD estimators. Overall it can be observed that as expected, the performance improvement obtained for all considered iSNRs and diffuse PSD estimators is lower than in Fig. 3(a). Furthermore, the eigenvalue decomposition-based estimators yield a larger improvement than the blocking matrix-based estimators in terms of both performance measures, with $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}$ resulting in the best ΔPESQ and $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_2}$ resulting in the best ΔfSSNR for low iSNRs. Whereas larger differences can be observed in terms of ΔPESQ between the blocking matrix-based and eigenvalue decomposition-based estimators, the obtained ΔfSSNR for all PSD estimators are rather similar. In addition, similarly to before, the performance obtained using $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},2}$ and $\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},4}$ is very similar.

5.5. Dereverberation and noise reduction performance for a moving speaker

In this section the dereverberation and noise reduction performance is investigated for a moving speaker scenario with $T_{60} \approx 1$ s and iSNR $= 10$ dB. Since both ΔPESQ and ΔfSSNR show very similar patterns, Table 1 presents only the ΔfSSNR obtained using the MVDR and MVDR-N beamformers and a Wiener postfilter. It can be observed that using the eigenvalue decomposition-based estimate $\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_2}$ results in the best performance. However, the performance obtained using the other considered diffuse PSD estimators is also comparable. In addition, it can be observed that as expected, the improvement obtained for all diffuse PSD estimators when using the MVDR-N beamformer is lower than when using the MVDR beamformer. However, the performance loss is rather insignificant, particularly when using the eigenvalue decomposition-based estimators.

In summary, the simulation results presented in this paper show the applicability of diffuse PSD estimators for binaural dereverberation and noise reduction based on beamforming and spectral filtering. Although all PSD estimators yield a high performance, the eigenvalue decomposition-based estimators result in the best performance for all considered techniques and scenarios. It should be noted that although the considered PSD estimators are based on a diffuse sound field model, the late reverberation and background noise considered in these simu-

Table 1: *Dereverberation and noise reduction performance in terms of ΔfSSNR using an MVDR and MVDR-N beamformer and a Wiener postfilter for a moving speaker ($T_{60} \approx 1$ s, iSNR = 10 dB).*

	$\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},2}$	$\hat{\Phi}^{\mathrm{BM}}_{\mathrm{d},4}$	$\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_1}$	$\hat{\Phi}^{\mathrm{EVD}}_{\mathrm{d},\lambda_2}$
MVDR	7.42	7.55	6.78	**7.86**
MVDR-N	6.83	6.89	6.66	**7.63**

lations were not perfectly diffuse, confirming the applicability of the considered estimators in realistic acoustic environments. Informal listening tests suggest that blocking matrix-based estimators yield a larger interference suppression while causing more signal distortions, whereas eigenvalue decomposition-based estimators yield a smaller interference suppression while introducing less signal distortions. In the future, formal listening tests should be conducted to truly assess the quality of these different late reverberation PSD estimators for binaural dereverberation and noise reduction.

6. Conclusion

In this paper we investigated the dereverberation and noise reduction performance of the binaural MVDR and MVDR-N beamformers followed by a Wiener postfilter when using blocking matrix-based and eigenvalue decomposition-based diffuse PSD estimators. A least-squares generalization of dual-channel blocking matrix-based estimators to the multi-channel case was also presented, yielding the same PSD estimate as a recently proposed multi-channel maximum likelihood estimator. Simulations results show that independently of the technique used, the eigenvalue decomposition-based PSD estimators yield the best performance. Furthermore, it is shown that increasing the number of microphones within the blocking matrix-based framework does not increase the PSD estimation accuracy.

7. References

[1] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 331–342, Jul. 2006.

[2] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. J. R. Liu, Eds. Hoboken, USA: John Wiley & Sons, 2010.

[3] D. Marquardt, "Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques," Ph.D. dissertation, University of Oldenburg, Oldenburg, Germany, Dec. 2015.

[4] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[5] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1579–1585, Apr. 2007.

[6] S. Braun and E. A. P. Habets, "Dereverberation in noisy environments using reference signals and a maximum likelihood estimator," in *Proc. European Signal Processing Conference*, Marrakech, Morocco, Sep. 2013.

[7] S. Braun, M. Torcoli, D. Marquardt, E. A. P. Habets, and S. Doclo, "Multichannel dereverberation for hearing aids with interaural coherence preservation," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Juan les Pins, France, Sep. 2014, pp. 124–128.

[8] O. Schwartz, S. Braun, S. Gannot, and E. A. P. Habets, "Maximum likelihood estimation of the late reverberant power spectral density in noisy environments," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, Oct. 2015.

[9] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1595–1608, Sep. 2016.

[10] I. Kodrasi and S. Doclo, "EVD-based multi-channel dereverberation of a moving speaker using different RETF estimation methods," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, San Francisco, USA, Mar. 2017, pp. 116–120.

[11] ——, "Late reverberant power spectral density estimation based on an eigenvalue decomposition," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, New Orleans, USA, Mar. 2017, pp. 611–615.

[12] L. Wang, T. Gerkmann, and S. Doclo, "Noise power spectral density estimation using MaxNSR blocking matrix," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1493–1508, Sep. 2015.

[13] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A stereophonic acoustic signal extraction scheme for noisy and reverberant environments," *Computer Speech and Language*, vol. 27, no. 3, pp. 726–745, May 2013.

[14] M. Azarpour, G. Enzner, and R. Martin, "Binaural noise PSD estimation for binaural speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 7068–7072.

[15] D. Marquardt and S. Doclo, "Noise power spectral density estimation for binaural noise reduction exploiting direction of arrival estimates," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, Oct. 2017, submitted.

[16] M. Jeub, M. Dorbecker, and P. Vary, "A semi-analytical model for the binaural coherence of noise fields," *IEEE Signal Processing Letters*, vol. 18, no. 3, pp. 197–200, Mar. 2011.

[17] D. Marquardt, V. Hohmann, and S. Doclo, "Interaural coherence preservation in multi-channel Wiener filtering-based noise reduction for binaural hearing aids," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2162–2176, Dec. 2015.

[18] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[19] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[20] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[21] 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 Journal on Advances in Signal Processing*, vol. 2009, no. 1, Jul. 2009.

[22] A. Walther and C. Faller, "Interaural correlation discrimination from diffuse field reference correlations," *The Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1496–1502, Mar. 2013.

[23] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862*, International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

[24] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.