Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function

Shmulik Markovich-Golan[1,2], Sharon Gannot[1] and Walter Kellermann[2]

[1] *Bar-Ilan University, The Faculty of Engineering*, Ramat-Gan, Israel, {shmuel.markovich,sharon.gannot}@biu.ac.il

[2] *Univ. Erlangen-Nuremberg, Multimedia Comm. and Sig. Proc.*, Erlangen, Germany, walter.kellermann@fau.de

***Abstract*—Estimation of the relative transfer functions (RTFs) vector of a desired speech source is a fundamental problem in the design of data-dependent spatial filters. We present two common estimation methods, namely the covariance-whitening (CW) and the covariance-subtraction (CS) methods. The CW method has been shown in prior work to outperform the CS method. However, thus far its performance has not been analyzed. In this paper, we analyze the performance of the CW and CS methods and show that in the cases of spatially white noise and of uniform powers of desired speech source and coherent interference over all microphones, the CW method is superior. The derivations are validated by comparing them to their empirical counterparts in Monte Carlo experiments. In fact, the CW method outperforms the CS method in all tested scenarios, although there may be rare scenarios for which this is not the case.**

***Index Terms*—spatial filter, beamformer, RTF.**

I. Introduction

Spatial filtering is a fundamental operation of modern speech processing applications in which multichannel microphone signals are filtered and summed to extract a desired speech signal. For a survey on spatial filtering techniques for speech signals please refer to [1]–[3]. The design criteria for the minimum variance distortionless response (MVDR) beamformer [4], [5] and for the multichannel Wiener filter (MWF) [1], [6] are widely used for optimizing the signal-to-noise ratio (SNR) at the output. These spatial filters are data-dependent, and their design takes into account the sound fields of the desired signal and noise.

Particularly, these methods rely on the second-order statistics of the noise and desired speech components, given by the noise covariance matrix and the RTFs vector, comprised of the transfer functions (TFs) relating the desired speech component at each microphone to the respective component at the reference microphone. In dynamic scenarios, in which the positions of desired source and the device and the noise field may change over time, correspondingly updating the spatial filter is required to achieve the optimal SNR. To this end, estimation and tracking of the RTFs vector of the desired speech and of the noise covariance matrix is required. In the following we address the problem of estimating the RTF.

A plethora of methods are known for estimating the RTFs vector, e.g. [5]–[11]. Two common methods for estimating it are the CS [6], [8], [12], [13] and CW [9], [12], [14] methods. Batch versions thereof rely on sample covariance matrix (SCM) estimations for two batches of multichannel observations, containing noise only and noisy speech components, respectively. In [15], the performance of the CS method has been analyzed and compared to the empirical performance of the CW method. It has been shown that the CW method outperforms the CS method. However, so far a theoretical analysis of the performance of the CW method has not been derived.

In the following we analyze the performance of the CW and CS methods for RTF estimation and compare them. The structure of the paper is as follows. In Sec. II we formulate the problem and in Secs. III and IV we describe the CW and CS methods for estimating the RTFs vector, respectively, and analyze their performance. In Sec. V, the expressions for the performance of the CW and CS methods are compared and in Sec. VI we verify these expressions through simulation and compare the empirical performance values in multiple Monte Carlo experiments to their respective theoretical values.

II. Problem formulation

Consider the problem of estimating the RTF of a desired speaker. We formulate the problem in the short-time Fourier transform (STFT) domain. Let $s(n,f)$ denote the desired speech signal, where n and f denote the time-frame and frequency-bin indices, respectively. The speech signal is propagating in a reverberant enclosure and is being picked by a microphone array comprising M microphones. The received signals are contaminated by additive multichannel noise, denoted by $\mathbf{w}(n,f)$, and are given by:

$$\mathbf{x}(n,f) = \mathbf{h}(f)s(n,f) + \mathbf{w}(n,f) \tag{1}$$

where $\mathbf{h}(f)$ is a normalized vector of acoustic transfer functions (ATFs) such that the Euclidean norm is given by $\|\mathbf{h}(f)\|^2 = M$. For a simpler notation, hereafter we omit the frequency-bin index f. Let $\phi_s(n) \triangleq \mathrm{E}\left[|s(n)|^2\right]$ denote the variance of $s(n)$, where $\mathrm{E}[\cdot]$ denotes the expectation operator, and let the received noise

$$\mathbf{w}(n) \triangleq \mathbf{h}_i v_i(n) + \mathbf{u}(n) \tag{2}$$

be a superposition of a stationary spatially white noise, denoted $\mathbf{u}(n)$, and a coherent stationary interference component, denoted $v_i(n)$. Their corresponding spectra are denoted ϕ_u and ϕ_i. The vector of ATFs of the coherent interference is

ISBN 978-90-827970-1-5 © EURASIP 2018

normalized, similarly to $\mathbf{h}$. The norms of the actual ATF vectors of the desired and coherent interfering sources which include propagation loss are absorbed into the respective sources, i.e. $s(n)$ and $v_i(n)$, and their variances, i.e. $\phi_s(n)$ and ϕ_i. Define

$$\beta_s(n) \triangleq \phi_s(n)/\phi_u \tag{3a}$$

$$\beta_i \triangleq \phi_i/\phi_u \tag{3b}$$

as the SNR and interference-to-noise ratio (INR), respectively. Selecting the first microphone as reference, we denote the vector of RTFs of the desired speaker as:

$$\mathbf{g} \triangleq \mathbf{h}/\mathbf{e}_1^T\mathbf{h} = \mathbf{h}/h_1 \tag{4}$$

where $\mathbf{e}_1 \triangleq \begin{bmatrix} 1, & 0, & \cdots, & 0 \end{bmatrix}^T$ is an $M \times 1$ vector, which extracts the first element from another $M \times 1$ vector by forming the joint inner product, $\cdot^T$ denotes the transpose operator and h_1 denotes the first element of $\mathbf{h}$.

Since the noise is stationary and assumed to be ergodic, its covariance matrix, defined as

$$\mathbf{\Phi}_w \triangleq \mathrm{E}\left[\mathbf{w}(n)\mathbf{w}^H(n)\right] = \phi_i\mathbf{h}_i\mathbf{h}_i^H + \phi_u\mathbf{I}, \tag{5}$$

where $\cdot^H$ denotes the transpose-conjugate operation, can be estimated with negligible error if averaging over a sufficiently long time interval. Therefore, we assume that $\mathbf{\Phi}_w$ is known and can be used to estimate the RTFs vector. The covariance matrix of the microphones signals at the n-th frame equals:

$$\mathbf{\Phi}_x(n) = \phi_s(n)\mathbf{h}\mathbf{h}^H + \mathbf{\Phi}_w. \tag{6}$$

The problem at hand is to estimate $\mathbf{g}$ given a set of N noisy speech multichannel observations, i.e., $\{\mathbf{x}(n)\}_{n=0}^{N-1}$, also denoted as the *given observation segment*. Define the *average* covariance matrix of the received signals in the given observation segment as:

$$\bar{\mathbf{\Phi}}_x \triangleq \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{\Phi}_x(n) = \bar{\phi}_s\mathbf{h}\mathbf{h}^H + \mathbf{\Phi}_w \tag{7}$$

where $\bar{\phi}_s \triangleq 1/N\sum_{n=0}^{N-1}\phi_s(n)$.

In the following sections we present two common ways for estimating the RTFs vector. The CS and CW methods are presented in Secs. III and IV, respectively.

III. The Covariance-whitening method

A. Description

Define the square-root decomposition of the noise covariance matrix as:

$$\mathbf{\Phi}_w = \mathbf{\Phi}_w^{H/2}\mathbf{\Phi}_w^{1/2} \tag{8}$$

where the matrix $\mathbf{\Phi}_w^{1/2}$ is referred to as the square-root of $\mathbf{\Phi}_w$ and $\mathbf{\Phi}_w^{H/2} \triangleq \left(\mathbf{\Phi}_w^{1/2}\right)^H$. This decomposition is non-unique, and among common square-root decompositions one can find the Cholesky decomposition and the eigenvalue decomposition (EVD) based square-root decomposition. The following derivation is not limited to a certain selection of the square-root operation.

Define the *whitened* microphone signals as:

$$\mathbf{z}(n) \triangleq \mathbf{\Phi}_w^{-H/2}\mathbf{x}(n). \tag{9}$$

It is called *whitened* since the multichannel noise components of the elements of $\mathbf{z}(n)$ are spatially white with unit variance. Given a set of multichannel observations, the covariance matrix of $\mathbf{z}(n)$ is estimated using the SCM method:

$$\hat{\mathbf{\Phi}}_z \triangleq \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{z}(n)\mathbf{z}^H(n). \tag{10}$$

Compute the EVD of $\hat{\mathbf{\Phi}}_z$, defined as $\hat{\mathbf{\Phi}}_z = \hat{\mathbf{\Psi}}\hat{\mathbf{\Omega}}\hat{\mathbf{\Psi}}^H$, where $\hat{\mathbf{\Psi}}$ is a matrix, the columns of which are the eigenvectors of $\hat{\mathbf{\Phi}}_z$, and where $\hat{\mathbf{\Omega}}$ is a matrix whose diagonal elements are the eigenvalues of $\hat{\mathbf{\Phi}}_z$, and denote the principal eigenvector of (10) as $\hat{\boldsymbol{\psi}}$. The CW based estimator for the RTFs vector is then given by:

$$\hat{\mathbf{g}}_{\mathrm{CW}} \triangleq \frac{\mathbf{\Phi}_w^{H/2}\hat{\boldsymbol{\psi}}}{\mathbf{e}_1^T\mathbf{\Phi}_w^{H/2}\hat{\boldsymbol{\psi}}}. \tag{11}$$

In the following section we analyze the performance of this estimator.

B. Analysis

Define:

$$\alpha \triangleq \sqrt{\mathbf{h}^H\mathbf{\Phi}_w^{-1}\mathbf{h}} \tag{12a}$$

$$\boldsymbol{\psi} \triangleq \alpha^{-1}\mathbf{\Phi}_w^{-H/2}\mathbf{h} \tag{12b}$$

$$\omega(n) \triangleq |\alpha|^2\phi_s(n) + 1 \tag{12c}$$

$$\bar{\mathbf{\Phi}}_z \triangleq \mathrm{E}\left[\hat{\mathbf{\Phi}}_z\right]. \tag{12d}$$

By substituting (7), (8) and (9) in (10), $\bar{\mathbf{\Phi}}_z$ can be shown to equal:

$$\bar{\mathbf{\Phi}}_z = (\bar{\omega} - 1)\boldsymbol{\psi}\boldsymbol{\psi}^H + \mathbf{I} \tag{13}$$

where $\boldsymbol{\psi}$ and $\bar{\omega}$ are the principal eigenvector and eigenvalue of the EVD of $\bar{\mathbf{\Phi}}_z$, i.e. $\bar{\mathbf{\Phi}}_z = \mathbf{\Psi}\bar{\mathbf{\Omega}}\mathbf{\Psi}^H$, respectively, where $\mathbf{\Psi}$ and $\bar{\mathbf{\Omega}}$ are the eigenvector and eigenvalue matrices of $\bar{\mathbf{\Phi}}_z$, respectively, and

$$\bar{\omega} = |\alpha|^2\bar{\phi}_s + 1. \tag{14}$$

Let us consider (11). In [16], [17], the EVD of a SCM is analyzed and expressions for the first and second-order moments of its eigenvectors and eigenvalues are derived. It is shown that the estimated eigenvalues are unbiased and that for eigenvalues with a multiplicity of 1, the estimates of their respective eigenvectors are also unbiased. Following these derivations, we find that the mean of the estimated principal eigenvector is $\mathrm{E}\left[\hat{\boldsymbol{\psi}}\right] = \boldsymbol{\psi}$ and its corresponding covariance matrix is:

$$\begin{aligned}\mathbf{\Theta}_\psi \triangleq& \frac{\bar{\Omega}_{1,1}}{N}\sum_{m\neq 1}\frac{\bar{\Omega}_{m,m}}{\left(\bar{\Omega}_{1,1} - \bar{\Omega}_{m,m}\right)^2}\left(\mathbf{\Psi}\mathbf{e}_m\right)\left(\mathbf{\Psi}\mathbf{e}_m\right)^H \\ =& \frac{\bar{\omega}}{N\left(\bar{\omega} - 1\right)^2}\left(\mathbf{I} - \boldsymbol{\psi}\boldsymbol{\psi}^H\right)\end{aligned} \tag{15}$$

ISBN 978-90-827970-1-5 © EURASIP 2018

where in the last step we substituted the eigenvalues of the matrix $\bar{\boldsymbol{\Phi}}_z$, i.e., $\bar{\Omega}_{1,1} = \bar{\omega}$ and $\bar{\Omega}_{m,m} = 1$ for $m = 2, \ldots, M$, and $\mathbf{e}_m$ is a selection vector with 1 as its m-th element and 0 for all other elements. Note that in [16], [17], the mean and covariance are derived for stationary signals whereas, here, the desired speech is non-stationary. We conjecture that similar expressions can be derived using the *average* whitened covariance matrix, defined in (13).

Define the estimation error of $\boldsymbol{\psi}$ as:

$$\dot{\boldsymbol{\psi}} \triangleq \hat{\boldsymbol{\psi}} - \boldsymbol{\psi} \tag{16}$$

Consider the denominator of the CW based RTF estimator in (11). By substituting (16) in $\mathbf{e}_1^T \boldsymbol{\Phi}_w^{H/2} \hat{\boldsymbol{\psi}}$, the latter can be expressed as:

$$\mathbf{e}_1^T \boldsymbol{\Phi}_w^{H/2} \hat{\boldsymbol{\psi}} = \mathbf{e}_1^T \boldsymbol{\Phi}_w^{H/2} \boldsymbol{\psi} \left(1 + \frac{\mathbf{e}_1^T \boldsymbol{\Phi}_w^{H/2} \dot{\boldsymbol{\psi}}}{\mathbf{e}_1^T \boldsymbol{\Phi}_w^{H/2} \boldsymbol{\psi}} \right) \tag{17}$$

and assuming a sufficiently large segment (i.e., $N \gg 1$), we can approximate $\sqrt{\mathrm{E}\left[|\mathbf{e}_1^T \boldsymbol{\Phi}_w^{H/2} \dot{\boldsymbol{\psi}}|^2 \right]} \ll |\mathbf{e}_1^T \boldsymbol{\Phi}_w^{H/2} \boldsymbol{\psi}|$ and the reciprocal of (17) can be approximated using first order Taylor series expansion as:

$$\frac{1}{\mathbf{e}_1^T \boldsymbol{\Phi}_w^{H/2} \hat{\boldsymbol{\psi}}} \approx \frac{1}{\mathbf{e}_1^T \boldsymbol{\Phi}_w^{H/2} \boldsymbol{\psi}} \left(1 - \frac{\mathbf{e}_1^T \boldsymbol{\Phi}_w^{H/2} \dot{\boldsymbol{\psi}}}{\mathbf{e}_1^T \boldsymbol{\Phi}_w^{H/2} \boldsymbol{\psi}} \right). \tag{18}$$

By substituting (18) back to (11) and neglecting second order terms of $\dot{\boldsymbol{\psi}}$, the estimated RTF can be approximated as:

$$\hat{\mathbf{g}}_{\text{CW}} \approx \mathbf{g} + \frac{\alpha}{h_1} \left(\mathbf{I} - \frac{\mathbf{h}\mathbf{e}_1^T}{h_1} \right) \dot{\boldsymbol{\psi}}. \tag{19}$$

Note that since $\mathrm{E}\left[\dot{\boldsymbol{\psi}} \right] = \mathbf{0}$, when the latter approximation of the CW estimator is valid it is unbiased. The covariance matrix of $\hat{\mathbf{g}}_{\text{CW}}$, denoted as $\boldsymbol{\Theta}_{\text{CW}}$, can therefore be approximated as:

$$\begin{aligned}\boldsymbol{\Theta}_{\text{CW}} \approx& \frac{|\alpha|^2}{|h_1|^2} \left(\mathbf{I} - \frac{\mathbf{h}\mathbf{e}_1^T}{h_1} \right) \boldsymbol{\Phi}_w^{H/2} \boldsymbol{\Theta}_{\psi} \boldsymbol{\Phi}_w^{1/2} \left(\mathbf{I} - \frac{\mathbf{h}\mathbf{e}_1^T}{h_1} \right)^H \\ =& \frac{1}{N \bar{\beta}_s |h_1|^2} \left(1 + \frac{1}{M \bar{\beta}_s \left(1 - \frac{M \beta_i |\rho|^2}{M \beta_i + 1} \right)} \right) \boldsymbol{\Gamma} \end{aligned} \tag{20}$$

where

$$\begin{aligned}\boldsymbol{\Gamma} \triangleq& \beta_i \mathbf{h}_i \mathbf{h}_i^H + \mathbf{I} + \frac{\beta_i |h_{i,1}|^2 + 1}{|h_1|^2} \mathbf{h}\mathbf{h}^H \\ & - \left(\frac{\beta_i h_{i,1} \mathbf{h}\mathbf{h}_i^H + \mathbf{h}\mathbf{e}_1^T}{h_1} + \frac{\left(\beta_i h_{i,1} \mathbf{h}\mathbf{h}_i^H + \mathbf{h}\mathbf{e}_1^T \right)^H}{h_1^*} \right) \end{aligned} \tag{21}$$

and

$$\bar{\beta}_s \triangleq \bar{\phi}_s / \phi_u \tag{22a}$$

$$\rho \triangleq \mathbf{h}^H \mathbf{h}_i / M \tag{22b}$$

are the *average* SNR and the angle between $\mathbf{h}$ and $\mathbf{h}_i$, respectively.

The variance of the estimation error of $\hat{\mathbf{g}}_{\text{CW}}$ is defined as

$$\begin{aligned}\theta_{\text{CW}} \triangleq& \mathrm{E}\left[\|\hat{\mathbf{g}}_{\text{CW}} - \mathbf{g}\|^2 \right] = \text{trace}\left(\boldsymbol{\Theta}_{\text{CW}} \right) \\ =& \frac{1}{N \bar{\beta}_s |h_1|^2} \left(1 + \frac{1}{M \bar{\beta}_s \left(1 - \frac{M \beta_i |\rho|^2}{M \beta_i + 1} \right)} \right) \gamma \end{aligned} \tag{23}$$

where $\gamma \triangleq \text{trace}\left(\boldsymbol{\Gamma} \right)$ is given by

$$\gamma = M \left(1 + \frac{|h_{i,1}|^2}{|h_1|^2} - 2 \frac{|h_{i,1}|}{|h_1|} \text{real}\left(\tilde{\rho} \right) \right) + 1 - \frac{2}{M} + \frac{1}{|h_1|^2} \tag{24}$$

and $\tilde{\rho} \triangleq \frac{h_1 h_{i,1}^* \rho}{|h_1| \cdot |h_{i,1}|}$ is the normalized angle between $\mathbf{g}$ and $\mathbf{h}_i / h_{i,1}$. Examining (23) we note that, the estimation error is inversely proportional to number of observations N, so that $\hat{\mathbf{g}}_{\text{CW}}$ is a consistent estimate.

IV. The Covariance-subtraction method

A. Description

Given N multichannel observations, we estimate the *average* covariance matrix of the received signals using SCM:

$$\hat{\boldsymbol{\Phi}}_x \triangleq \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{x}(n) \mathbf{x}^H(n). \tag{25}$$

Define

$$\hat{\boldsymbol{\Phi}}_\triangle \triangleq \hat{\boldsymbol{\Phi}}_x - \boldsymbol{\Phi}_w. \tag{26}$$

The CS based estimator for the RTF vector is then given by:

$$\hat{\mathbf{g}}_{\text{CS}} \triangleq \frac{\hat{\boldsymbol{\Phi}}_\triangle \mathbf{e}_1}{\mathbf{e}_1^T \hat{\boldsymbol{\Phi}}_\triangle \mathbf{e}_1}. \tag{27}$$

B. Analysis

Similarly to the approximated reciprocal of the denominator in (18), the reciprocal of the denominator in (27) can be approximated by

$$\frac{1}{\mathbf{e}_1^T \hat{\boldsymbol{\Phi}}_\triangle \mathbf{e}_1} \approx \frac{1}{\mathbf{e}_1^T \bar{\boldsymbol{\Phi}}_\triangle \mathbf{e}_1} \left(1 - \frac{\mathbf{e}_1^T \dot{\boldsymbol{\Phi}}_x \mathbf{e}_1}{\mathbf{e}_1^T \bar{\boldsymbol{\Phi}}_\triangle \mathbf{e}_1} \right) \tag{28}$$

where $\bar{\boldsymbol{\Phi}}_\triangle \triangleq \bar{\boldsymbol{\Phi}}_x - \boldsymbol{\Phi}_w$. We note that the estimation error of $\bar{\boldsymbol{\Phi}}_\triangle$ equals $\hat{\boldsymbol{\Phi}}_\triangle - \bar{\boldsymbol{\Phi}}_\triangle = \dot{\boldsymbol{\Phi}}_x$ with $\sqrt{\mathrm{E}\left[|\mathbf{e}_1^T \dot{\boldsymbol{\Phi}}_x \mathbf{e}_1|^2 \right]} \ll \mathbf{e}_1^T \bar{\boldsymbol{\Phi}}_\triangle \mathbf{e}_1$.

By substituting (28) into (27) and neglecting second-order terms of the estimation errors $\dot{\boldsymbol{\Phi}}_x \mathbf{e}_1$, the CS based RTF estimator can be approximated, similarly to (19), as

$$\hat{\mathbf{g}}_{\text{CS}} \approx \mathbf{g} + \frac{1}{\bar{\phi}_s |h_1|^2} \left(\mathbf{I} - \frac{\mathbf{h}\mathbf{e}_1^T}{h_1} \right) \dot{\boldsymbol{\Phi}}_x \mathbf{e}_1. \tag{29}$$

Note that since $\mathrm{E}\left[\dot{\boldsymbol{\Phi}}_x \right] = \mathbf{0}$, when the latter approximation of the CS estimator is valid it is unbiased.

The expectation of $\hat{\boldsymbol{\Phi}}_x$ in (25) equals $\bar{\boldsymbol{\Phi}}_x$, see (7). The statistics of a SCM obeys a complex Wishart distribution [18], so that the covariance of the errors of the estimated elements (i_1, j_1) and (i_2, j_2) is:

$$\mathrm{E}\left[\left(\mathbf{e}_{i_1}^T \dot{\mathbf{\Phi}}_x \mathbf{e}_{j_1}\right)\left(\mathbf{e}_{i_2}^T \dot{\mathbf{\Phi}}_x \mathbf{e}_{j_2}\right)^*\right] = \frac{\left(\mathbf{e}_{i_1}^T \bar{\mathbf{\Phi}}_x \mathbf{e}_{i_2}\right)\left(\mathbf{e}_{j_1}^T \bar{\mathbf{\Phi}}_x \mathbf{e}_{j_2}\right)^*}{N}. \tag{30}$$

As a special case of (30), the second-order statistics of the vector $\dot{\mathbf{\Phi}}_x \mathbf{e}_1$ is:

$$\mathrm{E}\left[\left(\dot{\mathbf{\Phi}}_x \mathbf{e}_1\right)\left(\dot{\mathbf{\Phi}}_x \mathbf{e}_1\right)^H\right] = \frac{\mathbf{e}_1^T \bar{\mathbf{\Phi}}_x \mathbf{e}_1}{N} \bar{\mathbf{\Phi}}_x. \tag{31}$$

Next, considering the covariance of the CS estimator in (29) and substituting (31) yields:

$$\mathbf{\Theta}_{\text{CS}} = \frac{\mathbf{e}_1^T \bar{\mathbf{\Phi}}_x \mathbf{e}_1}{\left(\bar{\phi}_s |h_1|^2\right)^2}\left(\mathbf{I} - \frac{\mathbf{h}\mathbf{e}_1^T}{h_1}\right) \bar{\mathbf{\Phi}}_x \left(\mathbf{I} - \frac{\mathbf{h}\mathbf{e}_1^T}{h_1}\right)^H. \tag{32}$$

By reformulating (32) and substituting (3a), (3b) and (21), $\mathbf{\Theta}_{\text{CS}}$ can be expressed as

$$\mathbf{\Theta}_{\text{CS}} = \frac{1}{N|h_1|^2\bar{\beta}_s}\left(1 + \frac{|h_{i,1}|^2\beta_i + 1}{|h_1|^2\bar{\beta}_s}\right)\mathbf{\Gamma}. \tag{33}$$

Similarly to (23), by substituting (24) we derive that

$$\begin{aligned}\theta_{\text{CS}} &\triangleq \mathrm{E}\left[\|\hat{\mathbf{g}}_{\text{CS}} - \mathbf{g}\|^2\right] = \operatorname{trace}\left(\mathbf{\Theta}_{\text{CS}}\right) \\ &= \frac{1}{N|h_1|^2\bar{\beta}_s}\left(1 + \frac{|h_{i,1}|^2\beta_i + 1}{|h_1|^2\bar{\beta}_s}\right)\gamma. \end{aligned} \tag{34}$$

V. Performance comparison

We compare the performance of the CW and CS methods. Note that the covariance matrices of both estimators, (20) and (33), are scaled versions of a common matrix $\frac{1}{N|h_1|^2\bar{\beta}_s}\mathbf{\Gamma}$, i.e. they have the same spatial properties. Therefore, the performance ratio of CW and CS methods equals

$$\begin{aligned}&\xi \triangleq \theta_{\text{CW}}/\theta_{\text{CS}} = \\ &\left(1 + \frac{M\beta_i + 1}{M\bar{\beta}_s\left(M\beta_i\left(1 - |\rho|^2\right) + 1\right)}\right) / \left(1 + \frac{|h_{i,1}|^2\beta_i + 1}{|h_1|^2\bar{\beta}_s}\right). \end{aligned} \tag{35}$$

Let us consider two special scenarios: 1) only spatially white noise, i.e. $\beta_i = 0$; and 2) far-field or uniform powers i.e. $|h_m| = |h_{i,m}| = 1$ for $m = 1, \ldots, M$. For the first case, i.e. only spatially white noise exists, we obtain:

$$\xi_u = \frac{1 + 1/(M\bar{\beta}_s)}{1 + 1/(|h_1|^2\bar{\beta}_s)}. \tag{36}$$

Clearly, in this case the CW estimator outperforms the CS, since $|h_1|^2 \leq \|\mathbf{h}\|^2 = M$. Note that for very high SNR the performance of both estimators is equal, and for very low SNR, i.e. $\bar{\beta}_s \ll 1$, the error of the CW is lower by a factor of $M/|h_1|^2 \geq 1$.

For the second scenario, the far-field case or the uniform powers case, we obtain:

$$\xi_{\text{ff}} = \left(1 + \frac{\beta_i + 1/M}{\bar{\beta}_s\left(M\beta_i\left(1 - |\rho|^2\right) + 1\right)}\right) / \left(1 + \frac{\beta_i + 1}{\bar{\beta}_s}\right).$$

In this case, the CW estimator also outperforms the CS, since $\beta_i + 1/M < \beta_i + 1$ and $\bar{\beta}_s\left(M\beta_i\left(1 - |\rho|^2\right) + 1\right) \geq \bar{\beta}_s$.

Note, that in the general case of (35) there may be rare cases for which the CS outperforms CW, i.e. $\xi > 1$. This may happen when $|\rho|^2 \to 1$ and $M\beta_i \gg 1$, i.e. when $\mathbf{h}$ and $\mathbf{h}_i$ are almost identical and the INR is very large. In this case, ξ can be approximated as

$$\xi_0 \approx \left(1 + \frac{\beta_i}{\bar{\beta}_s}\right) / \left(1 + \frac{|h_{i,1}|^2\beta_i + 1}{|h_1|^2\bar{\beta}_s}\right). \tag{37}$$

If additionally $|h_{i,1}| < |h_1|$, the CS method will outperform the CW method. Note that since we assumed that $|\rho|^2 \to 1$, we can expect that $|h_{i,1}| \approx |h_1|$ and conclude that even in these rare cases the performance of the CW method will not be much worse than that of the CS method.

VI. Verification

To verify the derived performance expressions for the CW and CS methods we conduct multiple Monte-Carlo experiments for various simulated scenarios and compare the empirical performance values to their theoretical counterparts.

Since all derivations in this paper are frequency independent, we simulate narrowband scenarios and analyze their performance. The following are the baseline values used for various parameters of the simulation. The number of microphones M is set to 10, the SNR and INR are set to 20dB and 10dB, respectively, the number of frames N is set to 1000 and the squared absolute angle between the ATFs of the desired speaker and the coherent interference $|\rho|^2$ is set to 0.5. For each scenario 100 different cases of $\mathbf{h}$ and $\mathbf{h}_i$ are randomly selected (rather than simulating a specific microphone constellation), and for each case 100 time segments of length N are randomly generated and used for constructing the CW and CS estimators. We compute the empirical second-order statistics of aforementioned estimators.

We validate the derivations and examine the effect of changing a single parameter while keeping the others set to their predefined baseline values. Due to space limitations we present only part of the results. The tested parameters are selected from the following ranges: 1) SNR values $\bar{\beta}_s \in \{0\text{dB}, 5\text{dB} \ldots, 40\text{dB}]$; 2) number of microphones $M \in \{2, 4, \ldots, 20\}$; and 3) $|\rho|^2 \in \{0, 0.05, \ldots, 1\}$. The baseline values are selected as intermediate points in the ranges of the various parameters. The results of aforementioned cases are depicted in Figs. 1,2 and 3, respectively. In each figure, we depict the empirical squared absolute error of the CW and CS RTF estimators, denoted as $\hat{\theta}_{\text{CW}}$ and $\hat{\theta}_{\text{CS}}$ in solid lines, respectively, and their theoretical counterpart values denoted as θ_{CW} and θ_{CS} in dashed lines, respectively.

It can be clearly deduced from these figures that the derivations are valid. We note that the performance of both estimators improve as the SNR increases, and for sufficiently high SNR they coincide. For low SNR the CW significantly outperforms the CS method, while for very low SNR values, the approximations used in the derivations are violated (e.g., (18),(28)) and the empirical performance values differ from

ISBN 978-90-827970-1-5 © EURASIP 2018

the theoretical ones. Also, the performance of both estimators improves as $|\rho|^2$ increases (Fig. 3) and as the number of microphones M decreases (Fig. 2). Evidently from these figures, the CW method outperforms the CS method in all cases that were tested.

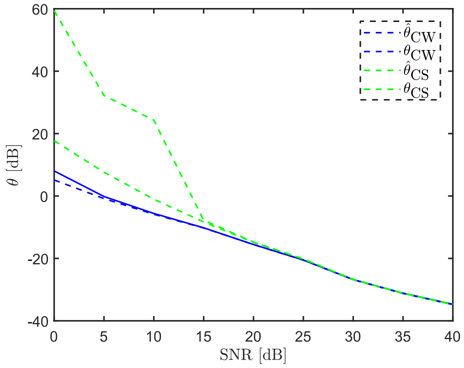

Fig. 1: Performance of CW and CS estimators for different values of SNR.

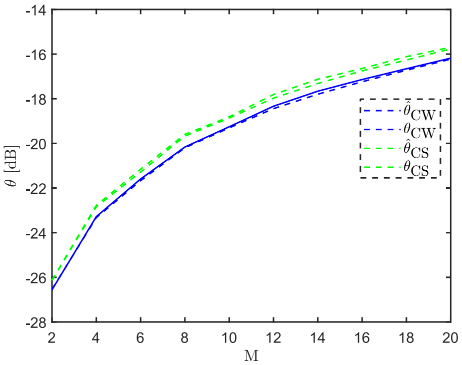

Fig. 2: Performance of CW and CS estimators for different values of M.

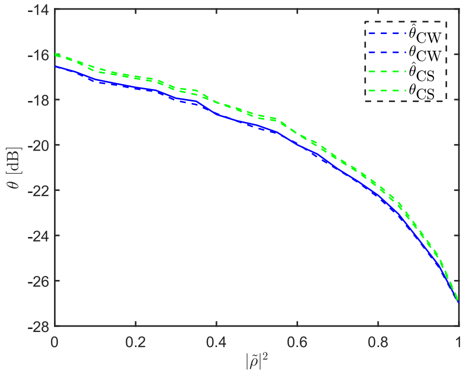

Fig. 3: Performance of CW and CS estimators for different values of $|\rho|^2$.

VII. Conclusion

We have considered the problem of RTF estimation, and have analyzed the performance of two common methods, namely the CW and the CS methods. The derivations have been validated by comparison to empirical values that were obtained in Monte Carlo experiments. Extending the analysis to the case of very low SNR will be treated in future work. The CW method has been shown to outperform the CS method in the cases of a spatially white noise and of uniform powers for desired source and coherent interference over all microphones. In fact, the CW method has outperformed the CS method in all scenarios that were tested, although there may be rare scenarios for which it is not the case.

References

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[2] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing*. Springer, 2003, pp. 155–194.

[3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[4] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[5] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[6] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel Wiener filtering techniques for noise reduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Springer, 2005, pp. 199–228.

[7] T. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[8] I. Cohen, "Relative transfer function identification using speech signals," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 5, pp. 451–459, Sep. 2004.

[9] S. Markovich-Golan, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[10] K. Reindl, S. Markovich-Golan, H. Barfuss, S. Gannot, and W. Kellermann, "Geometrically constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[11] J. Málek and Z. Koldovskỳ, "Sparse target cancellation filters with application to semi-blind noise extraction," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[12] R. Serizel, M. Moonen, B. V. Dijk, and J. Wouters, "Low-rank approximation based multichannel wiener filter algorithms for noise reduction with application in cochlear implants," *IEEE Trans. Audio, Speech and Language Processing*, vol. 22, no. 4, pp. 785–799, Apr. 2014.

[13] C. Zheng, A. Deleforge, X. Li, and W. Kellermann, "Statistical analysis of the multichannel Wiener filter using a bivariate normal distribution for sample covariance matrices," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018, DOI:10.1109/TASLP.2018.2800283.

[14] A. Bertrand and M. Moonen, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 60, pp. 233–246, Jan. 2012.

[15] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 544–548.

[16] G. W. Stewart and J.-g. Sun, *Matrix perturbation theory*, 1st ed. Academic Press, Jul. 1990.

[17] P. Stoica and T. Söderström, "Eigenelement statistics of sample covariance matrix in the correlated data case," *Digital Signal Processing*, vol. 7, no. 2, pp. 136–143, 1997.

[18] N. R. Goodman, "Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction)," *The Annals of mathematical statistics*, vol. 34, no. 1, pp. 152–177, 1963.

ISBN 978-90-827970-1-5 © EURASIP 2018