PERFORMANCE ANALYSIS OF THE COVARIANCE SUBTRACTION METHOD FOR RELATIVE TRANSFER FUNCTION ESTIMATION AND COMPARISON TO THE COVARIANCE WHITENING METHOD

Shmulik Markovich-Golan and Sharon Gannot

Bar-Ilan University, Faculty of Engineering
Ramat-Gan, 5290002, Israel

Shmuel.Markovich@biu.ac.il, Sharon.Gannot@biu.ac.il

ABSTRACT

Microphone array processing utilize spatial separation between the desired speaker and interference signal for speech enhancement. The transfer functions (TFs) relating the speaker component at a reference microphone with all other microphones, denoted as the relative TFs (RTFs), play an important role in beamforming design criteria such as minimum variance distortionless response (MVDR) and speech distortion weighted multichannel Wiener filter (SDW-MWF). Two common methods for estimating the RTF are surveyed here, namely, the covariance subtraction (CS) and the covariance whitening (CW) methods. We analyze the performance of the CS method theoretically and empirically validate the results of the analysis through extensive simulations. Furthermore, empirically comparing the methods performances in various scenarios evidently shows thats the CW method outperforms the CS method.

***Index Terms*—** Relative transfer function, beamforming, MVDR, speech distortion weighted MWF.

1. INTRODUCTION

Beamforming techniques, which process signals from an array of sensors (see [1, 2]), hold great potential for improved performance in speech processing applications (see [3–5]), compared with single channel processing. Several criteria exist for designing the beamformer. The minimum variance distortionless response (MVDR) criterion [6, 7] is designed to minimize the interference power while maintaining the desired speech signal undistorted. Gannot et al. [8] proposed to apply the MVDR-beamformer (BF) in the short-time Fourier transform (STFT) domain and to optimize the MVDR criterion at each frequency bin independently. Doclo et al. [3] proposed the speech distortion weighted multichannel Wiener filter (SDW-MWF) criterion which enables to control the tradeoff between interference reduction and speech distortion. In the limit case of zero-distortion the SDW-MWF and MVDR coincide.

By using acoustic transfer functions (ATFs) in designing the MVDR-BF or SDW-MWF we not only reduce the interference but also cancel or reduce the reverberation effect caused by the room impulse responses (RIRs) (also known as *dereverberation*). Unfortunately, the ATFs are usually unknown and estimating them is a cumbersome task. Instead of using the ATFs, Gannot et al. proposed to use the transfer functions relating the speech component at a reference microphone (one of the microphone signals) with the rest of the microphones, denoted relative transfer functions (RTFs). Correspondingly, the resulting component of the speech obtained at the output of the BF equals the speech component at the reference microphone, with no dereverberation applied, in the MVDR case and with some controlled distortion in the SDW-MWF case. The RTF can also be used for localization [9] and for spatial cue preservation, e.g., binaural cues in hearing aids [10, 11].

Practically, in many scenarios the reverberation level of the enclosure is moderate, and dereverberation is not necessary for speech intelligibility. The existence of a plethora of RTF estimation methods (see [3, 8, 9, 12–15]), makes it a very attractive candidate for BF design. Two of the most common methods for estimating the RTFs, namely the covariance subtraction (CS) [3, 12, 16, 17] and the covariance whitening (CW) [13, 17, 18] methods, utilize estimates of the microphones spatial covariance matrices, obtained during interference-only time-segments and during speech plus interference time-segments. Implementing the CS method is more appealing than the CW method, since it involves simple operations and, opposed to the CW method, does not require any matrix inversion. Unavoidable estimation errors of the RTF are manifested in some excess distortion incurred on the speech component at the output.

In this contribution we analyze the accuracy of the RTF estimate obtained using the CS method. The analysis is verified through extensive simulations. Moreover, we experimentally compare the accuracies of the CS and the CW methods in various scenarios, and empirically show that the CW method outperforms the CS method.

The paper is structured as follows. We formulate the problem in Sec. 2. In Sec. 3 we present and analyze the CS method, and in Sec. 4 we present the CW method. Results of an extensive simulation study that verifies the validity of the analysis as well as a comparative performance study of the two methods are given in Sec. 5.

2. PROBLEM FORMULATION

In this section we define the considered scenario, i.e. the environment and signals, in Sec. 2.1, and present the estimated second-order statistics (SOS) that will be used in the following sections in Sec. 2.2.

2.1. Environment and signals scenario

A desired speaker and interference signals propagate in a reverberant enclosure and are picked up by an array of M microphones. The received microphone signal at the m-th microphone, denoted x_m

can be formulated in the STFT as:

$$x_m(\ell,k) \triangleq h_m(k)s(\ell,k) + v_m(\ell,k) \tag{1}$$

where ℓ and k are the time-frame and frequency indices, $s(\ell,k)$ denotes the speech source (also known as the *dry* signal), $h_m(k)$ denotes the ATF relating the speaker and the m-th microphone signal and $v_m(\ell,k)$ denotes the interference signal component at the m-th microphone. The ATF is assumed time-invariant, i.e., the speaker is assumed static, hence h_m is not a function of ℓ. Eq. (1) can be extended by using vector notation to formulate the $M \times 1$ vector of microphone signals:

$$\boldsymbol{x}(\ell,k) \triangleq \boldsymbol{h}(k)s(\ell,k) + \boldsymbol{v}(\ell,k) \tag{2}$$

where $\boldsymbol{h}(k)$ denotes the vector of ATFs and $\boldsymbol{v}(\ell,k)$ denotes the vector of interference components. Henceforth, all derivations are considered at a single frequency and therefore k, the frequency index, is omitted for brevity.

Speech signals are highly non-stationary, and amongst the common distributions used to model them in the STFT domain are Gaussian or heavy tailed distributions, such as Laplacian. Here, we adopt the non-stationary Gaussian model, i.e. $s(\ell) \sim \mathcal{CN}\left(0, \beta_s\left(\ell\right)\right)$ where $\beta_s\left(\ell\right) \sim \frac{\alpha_s}{2} \cdot \text{Exp}(\frac{1}{2})$ is a scaled exponential random variable (RV) with an average of α_s. Note that the resulting speech variance is α_s.

The covariance matrix of the received microphones signals is given by:

$$\boldsymbol{R}_x \triangleq \text{E}\left[\boldsymbol{x}(\ell)\boldsymbol{x}^H(\ell)\right] = \boldsymbol{h}\boldsymbol{h}^H\alpha_s + \boldsymbol{R}_v \tag{3}$$

where $\text{E}\left[\bullet\right]$ denotes the expectation operator, $\boldsymbol{R}_v$ denotes the covariance matrix of the interference signal components, i.e., $\boldsymbol{R}_v = \text{E}\left[\boldsymbol{v}(\ell)\boldsymbol{v}^H(\ell)\right]$ and $(\bullet)^H$ denotes the Hermitian operator. Note that the interference signals are assumed stationary.

Assuming that the first microphone is the reference microphone, the vector of RTFs is defined by:

$$\boldsymbol{g} \triangleq \frac{1}{h_1}\boldsymbol{h}. \tag{4}$$

2.2. SOS estimation

Given L_v frames of microphones measurements of an interference-only time-segment, the covariance matrix of the interference signals can be estimated by:

$$\widehat{\boldsymbol{R}}_v \triangleq \frac{1}{L_v}\sum_{\ell=1}^{L_v}\boldsymbol{v}(\ell)\boldsymbol{v}^H(\ell). \tag{5}$$

Similarly to (5), given L_x frames of microphones measurements during a speech and interference time-segment (a different time-segment than the one used for estimating $\widehat{\boldsymbol{R}}_v$), the covariance matrix of the microphones signals is estimated by:

$$\widehat{\boldsymbol{R}}_x \triangleq \frac{1}{L_x}\sum_{\ell=1}^{L_x}\boldsymbol{x}(\ell)\boldsymbol{x}^H(\ell). \tag{6}$$

The latter estimation method is unbiased and its error terms are denoted by:

$$\widetilde{\boldsymbol{R}}_v = \widehat{\boldsymbol{R}}_v - \boldsymbol{R}_v \tag{7a}$$

$$\widetilde{\boldsymbol{R}}_x = \widehat{\boldsymbol{R}}_x - \boldsymbol{R}_x. \tag{7b}$$

For brevity we assume that $L_v = L_x$ and denote the number of frames per segment by L.

3. COVARIANCE SUBTRACTION METHOD

In this section we formulate the CS method (see [3, 12, 16]), and analyze its performance.

3.1. Estimation method

Define the spatial covariance matrix of the speech components:

$$\boldsymbol{R}_\Delta \triangleq \boldsymbol{h}\boldsymbol{h}^H\alpha_s = \boldsymbol{R}_x - \boldsymbol{R}_v. \tag{8}$$

Given the estimated interference covariance matrix $\widehat{\boldsymbol{R}}_v$ and speech plus interference covariance matrix $\widehat{\boldsymbol{R}}_x$, (8) can be estimated by:

$$\widehat{\boldsymbol{R}}_\Delta \triangleq \widehat{\boldsymbol{R}}_x - \widehat{\boldsymbol{R}}_v. \tag{9}$$

By substituting (7a) and (7b), Eq. (9) can be reformulated as:

$$\widehat{\boldsymbol{R}}_\Delta = \boldsymbol{R}_\Delta + \widetilde{\boldsymbol{R}}_\Delta. \tag{10}$$

where

$$\widetilde{\boldsymbol{R}}_\Delta \triangleq \widetilde{\boldsymbol{R}}_x - \widetilde{\boldsymbol{R}}_v. \tag{11}$$

Finally, the estimated RTF using the CS method is given by normalizing the first column of $\widehat{\boldsymbol{R}}_\Delta$ by its first entry (assuming that the first microphone is the reference microphone):

$$\widehat{\boldsymbol{g}}_\Delta \triangleq \frac{\widehat{\boldsymbol{R}}_\Delta\boldsymbol{e}_1}{\boldsymbol{e}_1^H\widehat{\boldsymbol{R}}_\Delta\boldsymbol{e}_1} \tag{12}$$

where $\boldsymbol{e}_1 \triangleq \begin{bmatrix} 1 & \boldsymbol{0}_{1\times(M-1)} \end{bmatrix}^T$ is an $M \times 1$ selection vector and $(\bullet)^T$ denotes the transpose operator. Note, that the CS method assumes: 1) a rank-1 structure for the covariance of speech components; 2) low estimation errors, i.e., $\widetilde{\boldsymbol{R}}_\Delta \approx \boldsymbol{0}$. In practice, the rank-1 approximation of the covariance of the speech components depends on the finite STFT window length and the reverberation time of the enclosure. Cases where the rank-1 approximation does not hold are out of the scope of the current contribution (see Serizel et. al. [17]).

3.2. Performance analysis

Substituting (3), (8), (10) into (12) yields:

$$\widehat{\boldsymbol{g}}_\Delta \triangleq \frac{h_1^*\alpha_s\boldsymbol{h} + \widetilde{\boldsymbol{R}}_\Delta\boldsymbol{e}_1}{|h_1|^2\alpha_s + \boldsymbol{e}_1^H\widetilde{\boldsymbol{R}}_\Delta\boldsymbol{e}_1}. \tag{13}$$

When the number of available time-frames is large ($L \gg 1$), we can assume small estimation errors (although non-zero) of the interference covariance and speech plus interference covariance matrices. Specifically, the estimation errors of $\boldsymbol{e}_1^H\widehat{\boldsymbol{R}}_v\boldsymbol{e}_1$, $\boldsymbol{e}_1^H\widehat{\boldsymbol{R}}_x\boldsymbol{e}_1$ and correspondingly $\boldsymbol{e}_1^H\widehat{\boldsymbol{R}}_\Delta\boldsymbol{e}_1$ are assumed low, i.e. $\boldsymbol{e}_1^H\widehat{\boldsymbol{R}}_\Delta\boldsymbol{e}_1 \approx \boldsymbol{e}_1^H\boldsymbol{R}_\Delta\boldsymbol{e}_1$:

Assumption 1 $\boldsymbol{e}_1^H\boldsymbol{R}_\Delta\boldsymbol{e}_1 = |h_1|^2\alpha_s \gg \boldsymbol{e}_1^H\widetilde{\boldsymbol{R}}_\Delta\boldsymbol{e}_1$.

Therefore, (13) can be reformulated as:

$$\widehat{\boldsymbol{g}}_\Delta = \left(\boldsymbol{g} + \frac{\widetilde{\boldsymbol{R}}_\Delta\boldsymbol{e}_1}{|h_1|^2\alpha_s}\right)\left(1 - \frac{\boldsymbol{e}_1^H\widetilde{\boldsymbol{R}}_\Delta\boldsymbol{e}_1}{|h_1|^2\alpha_s}\right) \tag{14}$$

where in the last step we used the first-order Taylor series approximation of $\frac{1}{1+\delta} \approx 1 - \delta$ for $|\delta| \ll 1$. Further assuming that the second-order error terms of $\widetilde{\boldsymbol{R}}_\Delta$ are negligible, i.e.:

Assumption 2 $\frac{\widetilde{\boldsymbol{R}}_{\Delta}\boldsymbol{e}_1}{|h_1|^2\alpha_s} \cdot \frac{\boldsymbol{e}_1^H\widetilde{\boldsymbol{R}}_{\Delta}\boldsymbol{e}_1}{|h_1|^2\alpha_s} \approx \boldsymbol{0}_{M\times 1}$

the estimated RTF (14) can be approximated as:

$$\widehat{\boldsymbol{g}}_{\Delta} = \boldsymbol{g} + \widetilde{\boldsymbol{g}}_{\Delta} \tag{15}$$

with the estimation error given by:

$$\widetilde{\boldsymbol{g}}_{\Delta} \triangleq \frac{1}{|h_1|^2\alpha_s}\left(\boldsymbol{I} - \boldsymbol{g}\boldsymbol{e}_1^H\right)\widetilde{\boldsymbol{R}}_{\Delta}\boldsymbol{e}_1 \tag{16}$$

and $\boldsymbol{I}$ denotes the identity matrix of proper dimensions.

It is well-known that the estimation error, denoted $\widetilde{\boldsymbol{R}}$, of the covariance matrix of a Gaussian RV, denoted $\boldsymbol{R}$, obeys a complex Wishart distribution [19], and that the covariance of the errors of the (i_1, j_1) and (i_2, j_2) covariance matrix elements is:

$$\mathrm{E}\left[\widetilde{R}_{i_1,j_1}\widetilde{R}^*_{i_2,j_2}\right] = \frac{R_{i_1,i_2}R^*_{j_1,j_2}}{L}. \tag{17}$$

Correspondingly, the covariance matrices of the estimation errors $\widetilde{\boldsymbol{R}}_v\boldsymbol{e}_1$ and $\widetilde{\boldsymbol{R}}_x\boldsymbol{e}_1$ equal:

$$\boldsymbol{C}_v \triangleq \mathrm{E}\left[\widetilde{\boldsymbol{R}}_v\boldsymbol{e}_1\left(\widetilde{\boldsymbol{R}}_v\boldsymbol{e}_1\right)^H\right] = \frac{\boldsymbol{e}_1^H\boldsymbol{R}_v\boldsymbol{e}_1}{L}\boldsymbol{R}_v \tag{18a}$$

$$\boldsymbol{C}_x \triangleq \mathrm{E}\left[\widetilde{\boldsymbol{R}}_x\boldsymbol{e}_1\left(\widetilde{\boldsymbol{R}}_x\boldsymbol{e}_1\right)^H\right] = \frac{\boldsymbol{e}_1^H\boldsymbol{R}_x\boldsymbol{e}_1}{L}\boldsymbol{R}_x. \tag{18b}$$

Now, since different time-segments are used for estimating $\widehat{\boldsymbol{R}}_v$ and $\widehat{\boldsymbol{R}}_x$, their corresponding estimation errors can be assumed statistically independent, and hence:

$$\boldsymbol{C}_{\Delta} \triangleq \mathrm{E}\left[\widetilde{\boldsymbol{R}}_{\Delta}\boldsymbol{e}_1\left(\widetilde{\boldsymbol{R}}_{\Delta}\boldsymbol{e}_1\right)^H\right] = \boldsymbol{C}_v + \boldsymbol{C}_x. \tag{19}$$

The ratio of the squared norm of the RTF estimation error and the squared norm of the RTF, is denoted RTF accuracy and is defined as:

$$\epsilon_{\Delta} \triangleq \frac{\mathrm{E}\left[\|\widetilde{\boldsymbol{g}}_{\Delta}\|^2\right]}{\|\boldsymbol{g}\|^2}. \tag{20}$$

By substitution of (16) and (19) into (20) we have:

$$\epsilon_{\Delta} = \frac{1}{\|\boldsymbol{g}\|^2\left(|h_1|^2\alpha_s\right)^2}\,\mathrm{tr}\left\{\left(\boldsymbol{I} - \boldsymbol{g}\boldsymbol{e}_1^H\right)\boldsymbol{C}_{\Delta}\left(\boldsymbol{I} - \boldsymbol{g}\boldsymbol{e}_1^H\right)^H\right\} \tag{21}$$

where $\mathrm{tr}\{\bullet\}$ denotes the trace operator.

Noting that $\left(\boldsymbol{I} - \boldsymbol{g}\boldsymbol{e}_1^H\right)\boldsymbol{h} = \boldsymbol{0}$ and substituting (4), (18a), (18b) and (19) into (21) yields:

$$\epsilon_{\Delta} = \frac{1}{L}\cdot\frac{1}{\|\boldsymbol{h}\|^2\alpha_s}\cdot\left(1+\frac{2}{\eta}\right)\cdot \cdot\,\mathrm{tr}\left\{\left(\boldsymbol{I} - \boldsymbol{g}\boldsymbol{e}_1^H\right)\boldsymbol{R}_v\left(\boldsymbol{I} - \boldsymbol{g}\boldsymbol{e}_1^H\right)^H\right\} \tag{22}$$

where η is defined as the signal to interference ratio (SIR) at the reference microphone:

$$\eta \triangleq \frac{|h_1|^2\alpha_s}{\boldsymbol{e}_1^H\boldsymbol{R}_v\boldsymbol{e}_1}. \tag{23}$$

Further simplification of (22) yields the final expression for the RTF accuracy using the CS method:

$$\epsilon_{\Delta} = \frac{1}{L}\cdot\frac{1}{\|\boldsymbol{h}\|^2}\cdot\left(1+\frac{2}{\eta}\right)\cdot \cdot\left(\frac{\mathrm{tr}\{\boldsymbol{R}_v\}}{\alpha_s} - \frac{2\,\mathrm{re}\left\{\boldsymbol{e}_1^H\boldsymbol{R}_v\boldsymbol{g}\right\}}{\alpha_s} + \frac{\|\boldsymbol{h}\|^2}{\eta}\right) \tag{24}$$

where $\mathrm{re}\{\bullet\}$ denotes the real operator.

4. COVARIANCE WHITENING METHOD

In this section we formulate the CW method (see [13,18]). Using the Cholesky decomposition (or any other matrix square-root operator), define the square-root of $\widehat{\boldsymbol{R}}_v$ and of its inverse:

$$\widehat{\boldsymbol{R}}_v = \left(\widehat{\boldsymbol{R}}_v^{1/2}\right)^H\widehat{\boldsymbol{R}}_v^{1/2} \tag{25a}$$

$$\widehat{\boldsymbol{R}}_v^{-1} = \left(\widehat{\boldsymbol{R}}_v^{-1/2}\right)^H\widehat{\boldsymbol{R}}_v^{-1/2}. \tag{25b}$$

After obtaining $\widehat{\boldsymbol{R}}_v^{-1/2}$ from an interference-only time-segment, we use it to generate the *whitened* signal, defined as:

$$\boldsymbol{y}(\ell) \triangleq \widehat{\boldsymbol{R}}_v^{-1/2}\boldsymbol{x}(\ell). \tag{26}$$

Substituting the definition of (2) in the latter equation yields:

$$\boldsymbol{y}(\ell) = \boldsymbol{q}d(\ell) + \boldsymbol{u}(\ell) \tag{27}$$

where

$$d(\ell) \triangleq \sqrt{\widehat{\gamma}}\exp(\mathrm{j}\phi)s(\ell) \tag{28a}$$

$$\boldsymbol{u}(\ell) \triangleq \widehat{\boldsymbol{R}}_v^{-1/2}\boldsymbol{v}(\ell) \tag{28b}$$

denote the scaled source signal (with an ambiguity phase shift ϕ and gain $\sqrt{\widehat{\gamma}}$) and the whitened interference signals, respectively, and

$$\boldsymbol{q} \triangleq \frac{\widehat{\boldsymbol{R}}_v^{-1/2}\boldsymbol{h}}{\sqrt{\widehat{\gamma}}}\exp(-\mathrm{j}\phi) \tag{29}$$

denotes the normalized ATF in the whitened domain. The nominal power normalization factor and its estimate are given by:

$$\gamma \triangleq \boldsymbol{h}^H\boldsymbol{R}_v^{-1}\boldsymbol{h} \tag{30a}$$

$$\widehat{\gamma} \triangleq \boldsymbol{h}^H\widehat{\boldsymbol{R}}_v^{-1}\boldsymbol{h}. \tag{30b}$$

Note that it follows from (7a), (25b), (28b) that $\mathrm{E}\left[\boldsymbol{u}(\ell)\boldsymbol{u}^H(\ell)\right] \approx \boldsymbol{I}$. The covariance matrix of $\boldsymbol{y}(\ell)$ is constructed similarly to (5), (6):

$$\widehat{\boldsymbol{R}}_y \triangleq \frac{1}{L}\sum_{\ell=1}^{L}\boldsymbol{y}(\ell)\boldsymbol{y}^H(\ell). \tag{31}$$

Define the eigenvalue decomposition of $\widehat{\boldsymbol{R}}_y$ as $\widehat{\boldsymbol{R}}_y \triangleq \widehat{\boldsymbol{Q}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{Q}}^H$ where $\widehat{\boldsymbol{Q}}$ is an orthogonal matrix comprising of the eigenvectors, and $\widehat{\boldsymbol{\Lambda}}$ is a diagonal matrix with the eigenvalues on its diagonal.

Define the *major* eigenvector $\widehat{\boldsymbol{q}}$ as the eigenvector in $\widehat{\boldsymbol{Q}}$ which corresponds to the maximal eigenvalue in $\widehat{\boldsymbol{\Lambda}}$ (associated with the speech component). Finally, the estimated RTF using the CW method is obtained by transforming $\widehat{\boldsymbol{q}}$ back from the whitened domain and scaling:

$$\widehat{\boldsymbol{g}}_{\Theta} \triangleq \frac{\widehat{\boldsymbol{R}}_v^{1/2}\widehat{\boldsymbol{q}}}{\boldsymbol{e}_1^H\widehat{\boldsymbol{R}}_v^{1/2}\widehat{\boldsymbol{q}}}. \tag{32}$$

5. MODEL VERIFICATION AND EXPERIMENTAL STUDY

The theoretical analysis of the accuracy of the CS method is verified by comparing it to the empirical accuracy in various scenarios, averaged over multiple Monte-Carlo experiments. The accuracy of the CS method is also *empirically* compared to the accuracy of the CW method in all scenarios. Note that simulations were performed directly in the STFT domain at a single frequency bin using synthetic non-stationary Gaussian signals and synthetic ATFs (without speech signals or RIRs simulation). In future work, we plan to examine the theoretical performance of RTF estimators with speech signals recorded in real environments.

A circular array comprising M microphones with spacing of half a wavelength is used. The direction of arrival (DOA) of the desired source is denoted θ_d. The interference is comprised of Q stationary coherent sources, arriving from equally spaced DOAs in the range $[0^o, 360^o)$, and of spatially-white sensors noise. The ratios between the variances of the desired speaker and the sensors noise, and between the variances of the desired speaker and the sum of all interferences are denoted signal to noise ratio (SNR) and SIR, respectively. The ATFs of the various sources are modeled as a summation of two components [20]: 1) a direct arrival component, modeling the different distances between the microphones and the speaker; 2) reflection components modeled as a complex normal RV. The ratio between the power of the direct arrival component and the average reflections power is denoted direct to reverberant ratio (DRR). The true RTF is defined as in Eq. (4).

Various scenarios are simulated in order to verify the theoretical formula of the accuracy of the CS based RTF method, given in (24). The DOA of the desired source is selected from the set of directions $\{0^o, 10^o, \ldots, 350^o\}$. In each scenario the reverberant component of the ATFs is randomly selected 10 times. For each instance of the ATFs, the sources are randomly generated in 100 Monte-Carlo experiments. We obtain estimates of the RTFs based on the CS and the CW methods, and average their accuracy measures.

We verify the influence of the number of frames L, SIR and the number of microphones M on the accuracy of the CS estimate. Unless stated otherwise the simulation parameters are set to: $Q = 1$, $M = 6$, SIR $= 0$dB, SNR $= 20$dB, DRR $= 30$dB and $L = 2000$. First, the performance is examined for different values of L, selected from $\{1000, 2000, 4000, 8000\}$. The results are depicted in Fig. 1. Second, the performance is examined for different values of SIR, selected from $\{0\text{dB}, 3\text{dB}, 6\text{dB}, 9\text{dB}\}$, while the number of coherent interfering sources is set to $Q = 3$. The results are depicted in Fig. 2. Finally, the performance is examined for different numbers of microphones M, selected from $\{4, 6, 8\}$, while the number of coherent interfering sources is set to $Q = 6$. The results are depicted in Fig. 3. In all figures, the accuracies are depicted versus the DOA of the speaker. For each scenario, denoted by a different marker (circle, square, diamond and star), we plot the empirical accuracies of the CS and CW methods (denoted in short by *emp.*), denoted by dashed-blue and dotted-red curves respectively, and the theoretical accuracy of CS method (denoted in short by *th.*), denoted by a solid-green curve.

Considering the results, the accuracies of both CS and CW methods improve as the number of frames L and the SIR increases. Also, it seems that the accuracies are not sensitive to the number of microphones M in both methods.

The validity of the theoretical analysis of the accuracy of the CS method is verified from these figures. Moreover, it is also evident that the CW method outperforms the CS method in all tested scenarios.

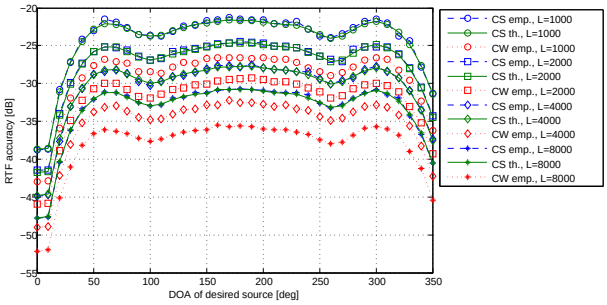

Fig. 1: RTF accuracies for various values of L.

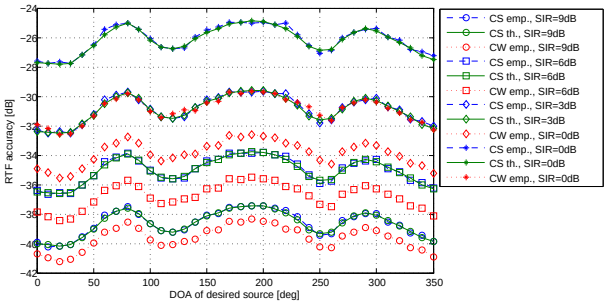

Fig. 2: RTF accuracies for various SIR levels.

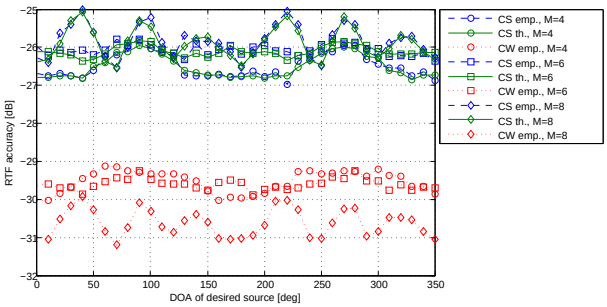

Fig. 3: RTF accuracies for various numbers of microphones, M.

6. CONCLUSIONS

Two common methods for estimating the RTF were surveyed, namely, the CS and the CW methods. A theoretical analysis of the CS was derived. The derivation is based on the complex Wishart distribution of the estimated covariance matrices, assuming that the number of frames used for the estimation is large enough. The derived theoretical model was verified through extensive simulations study. Moreover, from these simulations, it is evident that the CW method outperforms the CS method.

7. 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] H. Cox, R. M. Zeskind, and M. M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 10, Oct. 1987.

[3] 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.

[4] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 945–978.

[5] E. Habets, J. Benesty, S. Gannot, and I. Cohen, "The MVDR beamformer for speech enhancement," in *Speech Processing in Modern Communication*. Springer, 2010, pp. 225–254.

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

[7] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30, pp. 27–34, Jan. 1982.

[8] 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.

[9] 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.

[10] E. Hadad, S. Gannot, and S. Doclo, "Binaural linearly constrained minimum variance beamformer for hearing aid applications," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Sept 2012, pp. 1–4.

[11] D. Marquardt, V. Hohmann, and S. Doclo, "Binaural cue preservation for hearing aids using multi-channel Wiener filter with instantaneous ITF preservation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 21–24.

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

[13] 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.

[14] 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.

[15] 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.

[16] M. Souden, J. Benesty, and S. Affes, "A study of the LCMV and MVDR noise reduction filters," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4925–4935, Sep. 2010.

[17] 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.

[18] A. Bertrand and M. Moonen, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 233–246, Jan. 2012.

[19] N. Goodman, "Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction)," *Annals of mathematical statistics*, pp. 152–177, 1963.

[20] S. Markovich-Golan, S. Gannot, and I. Cohen, "Performance of the SDW-MWF with randomly located microphones in a reverberant enclosure," *IEEE Trans. Audio, Speech and Language Processing*, vol. 21, no. 7, pp. 1513–1523, 2013.