RELATIVE TRANSFER FUNCTION ESTIMATION EXPLOITING INSTANTANEOUS SIGNALS AND THE SIGNAL SUBSPACE

Maja Taseska and Emanuël A. P. Habets

International Audio Laboratories, Am Wolfsmantel 33, 91058 Erlangen, Germany*

ABSTRACT

Multichannel noise reduction can be achieved without distorting the desired signals, provided that the relative transfer functions (RTFs) of the sources are known. Many RTF estimators require periods where only one source is active, which limits their applicability in practice. We propose an RTF estimator that does not require such periods. A time-varying RTF is computed per time-frequency (TF) bin that corresponds to the dominant source at that bin. We demonstrate that a minimum variance distortionless response (MVDR) filter based on the proposed RTF estimate can extract multiple sources with low distortion. The MVDR filter has maximum degrees of freedom and hence achieves significantly better noise reduction compared to a linearly constrained minimum variance filter that uses a separate RTF for each source.

***Index Terms*—** Relative transfer function, speech enhancement, noise reduction, MVDR filter

1. INTRODUCTION

Data-dependent multichannel filters provide optimal spatial and spectral response based on the incoming signal statistics. For instance, the multichannel Wiener filter (MWF) computes a minimum-mean square error estimate of the desired signal [1], but it often leads to artefacts and high signal distortion if the signal statistics are inaccurate. Decomposing the MWF filter into a spatial filter and a single-channel filter allows for more flexible control of the overall response, leading to increased robustness against inaccurate signal statistics estimated in practice [2]. The spatial filter is often expressed in terms of the source RTFs, where each RTF describes the coupling between the microphones as a response to a given source. If the RTFs of all sources are known, an MVDR filter (for a single source) and linearly constrained minimum variance (LCMV) filter (for multiple sources) can achieve noise reduction without distorting the desired signals [3].

Different semi-blind methods to estimate the RTFs exist in the literature, which are based on a single-source model. For instance, the RTF estimator in [4] exploits the non-stationarity of speech signals to estimate the RTF of a speech source in noisy conditions. A different, subspace-based approach to RTF estimation was proposed in [5], where the RTF estimate is obtained by solving a generalized eigenvalue (GEV) problem. To estimate the RTFs of multiple sources, isolated periods can be used during which only one source is active [5], so that the single-source model for the mentioned estimators is not violated. Subsequently, the estimated RTFs of the different sources can be employed to compute an LCMV filter with multiple constraints, that extracts the desired sources without distortion, using for instance the framework in [5]. However, having multiple constraints limits the degrees of freedom, and hence the noise and interference reduction capability of the spatial filter.

To mitigate this problem, we propose an RTF estimator suited for noise reduction using an MVDR filter, which can be applied in scenarios with multiple sources, where no information about the number, location, and activity of the sources is available. At each TF bin, a single RTF is estimated which corresponds to the RTF of the dominant source at that TF bin. Such implicit usage of the speech sparsity in the TF domain maximizes the degrees of freedom for noise reduction, while maintaining low distortion of the desired signal, regardless of the number of active sources.

2. PROBLEM FORMULATION

2.1. Signal model

An array consisting of M microphones captures a desired signal and additive noise. The m-th microphone signal is given in the short-time Fourier transform (STFT) domain by

$$Y_m(n,k) = X_m(n,k) + V_m(n,k), \tag{1}$$

where m, n, and k denote the microphone, time, and frequency indices respectively, and X_m and V_m denote the desired signal and the noise. The desired signal contains speech from J sources, i.e. $X_m = \sum_{j=1}^{J} X_{mj}$, where J is arbitrary and unknown. The transfer function between the j-th source and the 1-st microphone is denoted by $H_{1j}(k)$. The RTF vector relative to the first microphone is defined as

$$\boldsymbol{g}_{1j}(k) = [1,\ {}^{H_{2j}(k)}/_{H_{1j}(k)},\ \cdots,\ {}^{H_{Mj}(k)}/_{H_{1j}(k)}]^{\mathrm{T}}. \tag{2}$$

As the processing is done independently at each frequency, we omit the index k wherever possible. Assuming the mul-

[*] A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS.

tiplicative transfer function (MTF) approximation holds [6], the desired signal at any microphone can be related to the desired signal at the first microphone as follows

$$\boldsymbol{x}(n) = \boldsymbol{g}_{11}\, X_{11}(n) + \boldsymbol{g}_{12}\, X_{12}(n) + \ldots + \boldsymbol{g}_{1J}\, X_{1J}(n), \quad (3)$$

where $\boldsymbol{x} = [X_1, \ldots, X_M]^{\mathrm{T}}$. The vectors $\boldsymbol{y}$ and $\boldsymbol{v}$ are defined similarly. The power spectral density (PSD) matrices of the different signals are given by $\boldsymbol{\Phi_y} = \mathrm{E}\left[\boldsymbol{y}\boldsymbol{y}^{\mathrm{H}}\right]$, $\boldsymbol{\Phi_x} = \mathrm{E}\left[\boldsymbol{x}\boldsymbol{x}^{\mathrm{H}}\right]$, and $\boldsymbol{\Phi_v} = \mathrm{E}\left[\boldsymbol{v}\boldsymbol{v}^{\mathrm{H}}\right]$. In the following, it is assumed that X_m and V_m are statistically uncorrelated, such that $\boldsymbol{\Phi_y} = \boldsymbol{\Phi_x} + \boldsymbol{\Phi_v}$.

2.2. Noise reduction

Multichannel noise reduction is achieved by estimating the desired signal received at a given microphone by using all available microphones. The signal estimate is given by $\widehat{X}_1(n) = \boldsymbol{w}^{\mathrm{H}}(n)\, \boldsymbol{y}(n)$, where $\boldsymbol{w}(n)$ represents a spatial filter at a given TF bin. An LCMV filter is obtained by solving

$$\underset{\boldsymbol{w}}{\arg\min}\, \boldsymbol{w}^{\mathrm{H}}\, \boldsymbol{\Phi_v}(n,k)\, \boldsymbol{w}, \quad (4a)$$

$$\text{subject to } \boldsymbol{w}^{\mathrm{H}}\boldsymbol{g}_{1j}(k) = 1, \;\; \text{for } j = 1, \ldots, J. \quad (4b)$$

Clearly, the LCMV filter requires estimation of the possibly time-varying number of sources J and their corresponding RTF vectors $\boldsymbol{g}_{1j}$, which is an extremely challenging task in practice. Moreover, multiple constraints reduce the noise reduction capability of the filter.

The goal in this paper is to estimate the desired signal using only one constraint, regardless of the number of sources. Solving (4) for one constraint yields the MVDR filter [1]

$$\boldsymbol{w}_{\mathrm{MVDR}}(n,k) = \frac{\boldsymbol{\Phi}_{\boldsymbol{v}}^{-1}(n,k)\, \tilde{\boldsymbol{g}}(n,k)}{\tilde{\boldsymbol{g}}^{\mathrm{H}}(n,k)\, \boldsymbol{\Phi}_{\boldsymbol{v}}^{-1}(n,k)\, \tilde{\boldsymbol{g}}(n,k)}. \quad (5)$$

The computation of the time-varying RTF vector $\tilde{\boldsymbol{g}}(n,k)$ represents the main contribution of the paper.

3. STATE-OF-THE-ART RTF ESTIMATION

RTF estimation in single-source scenarios is well studied in the literature. Two established estimators are the subspace-based [5] and the minimum distortion-based estimator [2, 7], which assume the following rank-one model

$$\boldsymbol{x}(n,k) = \boldsymbol{g}_{11}(k)\, X_{11}(n,k) \quad (6a)$$

$$\boldsymbol{\Phi_x}(n,k) = \phi_{x_{11}}(n,k)\, \boldsymbol{g}_{11}(k)\boldsymbol{g}_{11}^{\mathrm{H}}(k), \quad (6b)$$

where $\phi_{x_{11}}$ is the PSD of the source signal at the first microphone. In the following, we briefly review these estimators.

3.1. Subspace-based RTF estimation

The subspace-based RTF estimator proposed in [5], is based on the GEV problem for the matrix pencil $(\boldsymbol{\Phi_y}, \boldsymbol{\Phi_v})$

$$(\phi_{x_{11}}\, \boldsymbol{g}_{11}\, \boldsymbol{g}_{11}^{\mathrm{H}} + \boldsymbol{\Phi_v})\, \boldsymbol{u} = \lambda\, \boldsymbol{\Phi_v}\, \boldsymbol{u}, \quad (7)$$

where λ and $\boldsymbol{u}$ denote an eigenvalue and eigenvector pair, and the PSD matrix $\boldsymbol{\Phi_y}$ of the microphone signals is given by

$$\boldsymbol{\Phi_y}(n,k) = \phi_{x_{11}}(n,k)\, \boldsymbol{g}_{11}(k)\boldsymbol{g}_{11}^{\mathrm{H}}(k) + \boldsymbol{\Phi_v}(n,k). \quad (8)$$

As the matrix $\boldsymbol{\Phi_x}$ is of rank one, there is only one generalized eigenvalue λ that is larger than one and $\boldsymbol{g}_{11}$ is a scaled and rotated version of the corresponding eigenvector $\boldsymbol{u}$. By definition, the first entry of $\boldsymbol{g}_{11}$ is equal to one. Hence the RTF can be obtained by the following normalization

$$\boldsymbol{g}_{11} = \frac{\boldsymbol{\Phi_v}\, \boldsymbol{u}}{\boldsymbol{e}_1\, \boldsymbol{\Phi_v}\, \boldsymbol{u}}, \;\; \text{with } \boldsymbol{e}_1 = [1, 0 \ldots, 0]. \quad (9)$$

In practice, the PSD matrices are estimated by temporal averaging, hence even when a single source is present, the rank one assumption is not valid. Nevertheless, selecting the eigenvector that corresponds to the largest eigenvalue represents a good estimate of the source RTF vector, provided that the speech component is significantly stronger than the noise.

3.2. Minimum distortion-based RTF estimation

Based on the rank one model, the RTF vector $\boldsymbol{g}_{11}$ is equal to the first column of the PSD matrix $\boldsymbol{\Phi_x}$, normalized by the signal power at the first microphone, i.e.

$$\boldsymbol{g}_{11}(n,k) = \frac{\boldsymbol{\Phi_x}(n,k)\; \boldsymbol{e}_1}{\boldsymbol{e}_1^{\mathrm{H}}\boldsymbol{\Phi_x}(n,k)\; \boldsymbol{e}_1}. \quad (10)$$

The same solution arises as the so called spatial prediction vector which is found, in the minimum mean-squared error sense, by solving the following optimization problem [2, 7]

$$\underset{\boldsymbol{g}_{11}}{\arg\min}\, \mathrm{E}\left[(\boldsymbol{x} - \boldsymbol{g}_{11}X_1)^{\mathrm{H}}(\boldsymbol{x} - \boldsymbol{g}_{11}X_1)\right]. \quad (11)$$

To compute (10), the PSD matrix $\boldsymbol{\Phi_x}$ of the desired signal is required, which can for instance be obtained as the difference $\boldsymbol{\Phi_y} - \boldsymbol{\Phi_v}$, if an estimate of the noise PSD matrix is available.

4. PROPOSED RTF ESTIMATION

When multiple sources are present, the rank of the PSD matrix $\boldsymbol{\Phi_x}$ increases, and the signal has the general form given by (3). To illustrate that the standard subspace-based method described in Section 3.1 is inapplicable in this case, we look at the GEV problem for a two-source scenario

$$(\phi_{x_{11}}\, \boldsymbol{g}_{11}\, \boldsymbol{g}_{11}^{\mathrm{H}} + \phi_{x_{12}}\, \boldsymbol{g}_{12}\, \boldsymbol{g}_{12}^{\mathrm{H}} + \boldsymbol{\Phi_v})\, \boldsymbol{u} = \lambda\, \boldsymbol{\Phi_v}\, \boldsymbol{u}. \quad (12)$$

Equation (12) can be rearranged as follows

$$c_1\, \boldsymbol{g}_{11} + c_2\, \boldsymbol{g}_{12} = (\lambda - 1)\, \boldsymbol{\Phi_v}\, \boldsymbol{u}, \quad (13)$$

$$\text{where } \; c_1 = (\phi_{x_{11}}\boldsymbol{g}_{11}^{\mathrm{H}}\boldsymbol{u})^{-1}, \; c_2 = (\phi_{x_{12}}\boldsymbol{g}_{12}^{\mathrm{H}}\boldsymbol{u})^{-1}.$$

It is clear from (13) that the eigenvectors that correspond to the eigenvalues larger than one, provide two distinct linear combinations of the RTF vectors. Hence, they represent a basis for the signal subspace, but do not provide estimates of the separate RTFs required in the LCMV filter computation.

To extract the desired signal in multi-source scenarios, we propose an estimator that computes the RTF of the dominant source at each TF bin, by projecting the instantaneous vector $\boldsymbol{y}$ onto the multi-dimensional signal subspace. The projection is performed at each TF bin where speech is present, requiring a narrowband voice activity detector (VAD) mechanism. In this manner, given a single RTF per bin, an MVDR filter can be used to extract the desired sources with low distortion. In Section 4.1 we briefly review the VAD method employed in this work, and in Section 4.2 we describe the proposed RTF estimator.

4.1. Narrowband voice activity detector

Assuming that the commonly used Gaussian signal model for the STFT coefficients [8, 9] is valid, the PSD matrices $\boldsymbol{\Phi_v}$, $\boldsymbol{\Phi_x}$, and the signal vector $\boldsymbol{y}$ suffice to compute a narrowband VAD. We define the hypotheses $\mathcal{H}_x$ and $\mathcal{H}_v$ to indicate speech presence and speech absence, respectively. The likelihoods under the different hypotheses are then given by

$$\begin{aligned} p(\boldsymbol{y}\,|\,\mathcal{H}_v) &= \frac{1}{\pi^M \det[\boldsymbol{\Phi_v}]}\, \mathrm{e}^{-\boldsymbol{y}^{\mathrm{H}}\boldsymbol{\Phi_v}^{-1}\boldsymbol{y}}, \\ p(\boldsymbol{y}\,|\,\mathcal{H}_x) &= \frac{1}{\pi^M \det[\boldsymbol{\Phi_v}+\boldsymbol{\Phi_x}]}\, \mathrm{e}^{-\boldsymbol{y}^{\mathrm{H}}[\boldsymbol{\Phi_v}+\boldsymbol{\Phi_x}]^{-1}\boldsymbol{y}}. \end{aligned} \tag{14}$$

The speech presence probability (SPP) $p(\mathcal{H}_x\,|\,\boldsymbol{y})$, can be obtained by applying the Bayes theorem as follows

$$p(\mathcal{H}_x\,|\,\boldsymbol{y}) = \frac{p(\boldsymbol{y}\,|\,\mathcal{H}_x)\cdot p(\mathcal{H}_x)}{p(\boldsymbol{y}\,|\,\mathcal{H}_x)\cdot p(\mathcal{H}_x) + p(\boldsymbol{y}\,|\,\mathcal{H}_v)\cdot p(\mathcal{H}_v)}, \tag{15}$$

where $p(\mathcal{H}_x) = 1 - p(\mathcal{H}_v)$ denotes the a-priori SPP. In this work we use a direct-to-diffuse ratio (DDR)-based a-priori SPP [10], but other fixed or signal-dependent a-priori SPP can be employed as well. Subsequently, a VAD $\mathcal{I}_x$ can be computed for each TF bin by setting a threshold p_{thr}

$$\mathcal{I}_x = \begin{cases} 1 & \text{if } p(\mathcal{H}_x\,|\,\boldsymbol{y}) > p_{\mathrm{thr}} \\ 0 & \text{otherwise .} \end{cases} \tag{16}$$

4.2. Proposed RTF estimator

Although the signal subspace obtained using the GEV decomposition of $(\boldsymbol{\Phi_y}, \boldsymbol{\Phi_v})$ does not directly provide the RTF estimates in multi-source scenarios, it is crucial for the proposed RTF estimator. Note that as the PSD matrices are estimated by recursive temporal averaging and the activity of the present sources changes over time, a time-varying signal subspace is obtained, with possibly time-varying dimension, depending on the activity of the sources across time. Experiments indicated that for moderate reverberation levels, and up to four simultaneously active sources, two eigenvectors per frequency bin suffice to represent the signal subspace, regardless of the number of sources. Hence, in this work the subspace dimension is fixed to two. In future work, online dimension selection is to be investigated, for instance by considering all eigenvectors whose corresponding eigenvalues are larger than a pre-defined threshold.

Given the subspace estimate at each TF bin, we compute an orthonormal basis $\boldsymbol{U}_x$ by orthonormalizing the two largest generalized eigenvectors of the matrix pencil $(\widehat{\boldsymbol{\Phi}}_{\boldsymbol{y}}, \widehat{\boldsymbol{\Phi}}_{\boldsymbol{v}})$, where the hat indicates an estimate of the true PSD matrix. Let us denote a projection matrix onto the signal subspace by $\boldsymbol{P}_x = \boldsymbol{U}_x\boldsymbol{U}_x^{\mathrm{H}}$. The key idea of the proposed method is to enforce the instantaneous RTF estimate $\boldsymbol{g}_{\mathrm{inst}}(n,k)$ given by

$$\boldsymbol{g}_{\mathrm{inst}}(n,k) = \frac{\boldsymbol{y}(n,k)\boldsymbol{y}^{\mathrm{H}}(n,k)\boldsymbol{e}_1}{\boldsymbol{e}_1^{\mathrm{H}}\boldsymbol{y}(n,k)\boldsymbol{y}^{\mathrm{H}}(n,k)\boldsymbol{e}_1}, \tag{17}$$

to lie in the estimated signal subspace, by performing the following subspace projection at each TF bin

$$\boldsymbol{g}_{\mathrm{proj}}(n,k) = \frac{\boldsymbol{P}_x(n,k)\,\boldsymbol{g}_{\mathrm{inst}}(n,k)}{\boldsymbol{e}_1^{\mathrm{H}}\,\boldsymbol{P}_x(n,k)\,\boldsymbol{g}_{\mathrm{inst}}(n,k)}, \tag{18}$$

where the denominator normalizes the first element to one. The vector $\boldsymbol{g}_{\mathrm{inst}}$ captures the spatial information of the dominant source, whereas the subspace projection denoises $\boldsymbol{g}_{\mathrm{inst}}$ and confines it onto the current subspace estimate. As a result, when speech is present, the expression (18) provides an estimate of the RTFs vector for the dominant source at TF bin (n,k). The final RTF estimate is obtained as follows

$$\tilde{\boldsymbol{g}}(n,k) = \mathcal{I}_x(n,k)\,\boldsymbol{g}_{\mathrm{proj}}(n,k) + [1 - \mathcal{I}_x(n,k)]\,\tilde{\boldsymbol{g}}(n-1,k), \tag{19}$$

hence, when speech is present (i.e., $\mathcal{I}_x(n,k) = 1$) the RTF estimate is obtained by (18), whereas when speech is absent the RTF estimate from the previous frame is used.

Finally, it should be noted that the proposed RTF estimate can be used in an MVDR filter to extract a sum of multiple speech sources, but cannot be used for source separation where the RTFs of the individual sources are required.

5. PERFORMANCE EVALUATION

The proposed RTF estimator was evaluated in a simulated room with dimensions $4.5 \times 4 \times 3$ m^3. The microphone signals were obtained by convolving clean speech with simulated room impulse responses [11]. White sensor noise and diffuse babble noise were added [12]. The experiments were performed using a uniform linear array of 5 omnidirectional microphones, with microphone distance of 3 cm. The method is however applicable to any constellation of co-located or distributed microphones. The sampling rate was 16 kHz and the STFT frame length was 128 ms with 50% overlap.

The evaluation consists of two parts: (i) in Section 5.1, the signals from multiple sources are extracted by an MVDR filter with the proposed RTF estimate and evaluated in terms of noise reduction and speech distortion; (ii) in Section 5.2, the estimated RTF is compared to the ideal RTF using the Hermitian angle between the RTF vectors. The matrices $\boldsymbol{\Phi_y}$ and $\boldsymbol{\Phi_v}$ were obtained by a first-order recursive averaging with a time constant of 30 ms. The SPP threshold p_{thr} was set to 0.7.

First, to compare existing and proposed RTF estimators without the effect of erroneous noise PSD estimates, the system was evaluated using an oracle recursive estimate obtained from the noise signal separately. Clearly, in practice, the noise signal is not available. In addition, the system was evaluated in a fully blind scenario where $\mathbf{\Phi}_{\boldsymbol{v}}$ is estimated from the data using the DDR-based framework proposed in [10].

5.1. Objective extracted signal quality

Scenarios with one to four sources were tested, each located at 1-2 meters from the array. For each number of sources, the results was averaged over 5 source constellations. Signals with 30 seconds of speech were used, with all sources simultaneously active. The signal extraction was evaluated in terms of segmental speech distortion (SD) index ν_{sd}, as defined in [13, Eq. 4.44] and segmental noise reduction (segNR). The sum of all speech signals received at the first microphone served as a desired signal reference. The segNR for frame i of length T samples (corresponding to 30 ms) was computed as follows

$$\text{segNR}(i) = 10 \log_{10} \frac{\left\langle |v_m(t)|^2 \right\rangle}{\left\langle |\hat{v}_m(t)|^2 \right\rangle}, \; (i-1)T \leq t < iT, \quad (20)$$

where $\langle \cdot \rangle$ denotes temporal average, and $\hat{v}_m$ is the filtered noise. The overall segNR is obtained by averaging segNR(i) over i. Reverberation times of 350 ms and 500 ms were tested. The signal to sensor noise ratio was 30 dB. The power of the babble noise was varied to obtain SNRs of 18 dB and 8 dB.

The results are given in Fig. 1. An LCMV filter with an ideal RTF vector for each source was tested, to demonstrate that multiple constraints significantly limit the segNR. The spatial prediction-based, the standard subspace-based, and the proposed RTF estimators, denoted by ”SP”, ”GEVD”, and ”Proposed”, were used in an MVDR filter. Derived from a minimum-distortion principle, the SP-based estimator maintains SD index $\nu_{\text{sd}} < 0.1$ in all scenarios, however, the segNR is often insufficient, and decreases with increasing number of sources. The GEVD-based estimator offers a segNR up to 6 dB larger than the SP-based, at the cost of rapidly increasing SD index as the single source model is violated. The advantage of using multidimensional subspace in the proposed RTF estimator is manifested in the low SD index for any number of sources, while the segNR is only by less than 1 dB worse than the GEVD method. Even in the single source case, the SD index when using the proposed estimator is lower than the GEVD-based, due to the fact that for reverberant environments and short STFT frames, the MTF approximation might not hold [6], hence violating the rank-one model. Although not presented in the figures, we additionally evaluated scenarios with very low SNR (around 0 dB). In severe noise conditions, the advantage of the proposed RTF estimator is lost due to increasing number of misdetections in the VAD, leading to high SD index similar to the GEVD method.

5.2. Distance between estimated and ideal RTFs vectors

To measure the deviation of the RTF estimate at each TF bin from the ideal RTF of the dominant source at that bin, we consider for each source j the Hermitian angle between the ideal RTF $\boldsymbol{g}_{j,\text{ideal}}$ and the estimated RTF $\tilde{\boldsymbol{g}}$, computed as

$$\theta_{\text{H}}^{j} = \begin{cases} \arccos \frac{|\boldsymbol{g}_{j,\text{ideal}}^{\text{H}}(k)\, \tilde{\boldsymbol{g}}(n,k)|}{\|\boldsymbol{g}_{j,\text{ideal}}(k)\| \, \|\tilde{\boldsymbol{g}}(n,k)\|} & \text{if source } j \text{ is dominant} \\ 0 & \text{otherwise.} \end{cases} \quad (21)$$

The angle θ_{H}^{j} is relevant only if the source j is dominant, hence it is set to zero otherwise. We illustrate an example with two sources, reverberation time 350 ms and SNR of 18 dB. The fact that the GEVD estimator does not consider instantaneous data is reflected in the distance measure in Fig. 2(a) where the estimated RTF vector is aligned with the source that is on average stronger at a particular frequency. The proposed estimator on the other hand, exploits sparsity and aligns the RTF vector with the dominant source at each bin. The improvement in RTF vector alignment is also visible from the time-averaged results over a 30 seconds signal in Fig. 2(b).

6. CONCLUSIONS

An RTF estimator was proposed that can be used in an MVDR filter to extract multiple desired sources. By using an estimate of the multidimensional signal subspace, multiple sources are considered in the model, whereas by using instantaneous signals, speech sparsity is exploited. The implicit usage of speech sparsity results in only one filter constraint per TF bin, which in turn leads to maximum degrees of freedom for noise reduction. The estimator is suitable for blind scenarios with unknown number of sources where no oracle information about the source activity over time is available. Audio samples are available at http://www.audiolabs-erlangen.de/resources/2015-EUSIPCO-RTF.

REFERENCES

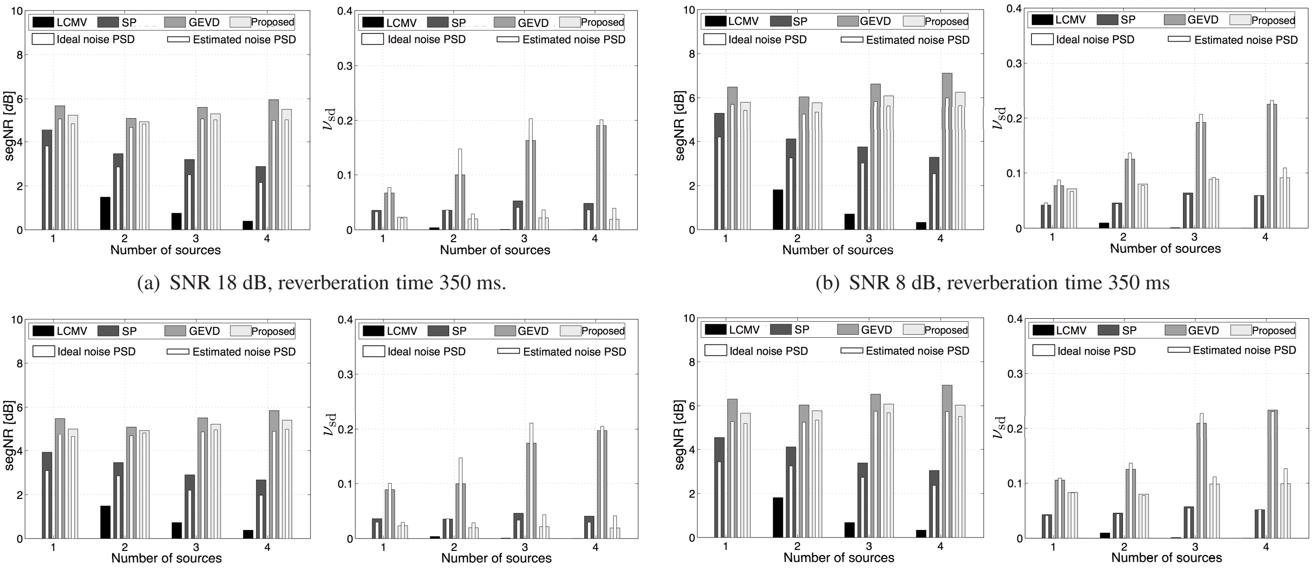
[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, Berlin, Germany, 2008.

[2] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, “Theoretical analysis of binaural multimicrophone noise reduction techniques,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 342–355, Feb. 2010.

[3] S. Gannot and I. Cohen, “Adaptive beamforming and postfiltering,” in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., chapter 47. Springer-Verlag, 2008.

[4] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” *IEEE Trans. Signal Process.*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[5] S. Markovich, S. Gannot, and I. Cohen, “Multichannel eigenspace beamforming in a reverberant noisy environment



(a) SNR 18 dB, reverberation time 350 ms.

(b) SNR 8 dB, reverberation time 350 ms

(c) SNR 18 dB, reverberation time 500 ms

(d) SNR 8 dB, reverberation time 500 ms

Fig. 1: Objective quality of the extracted signal for different reverberation times and high to moderate SNRs.

with multiple interfering speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1071–1086, Aug. 2009.

[6] Y. Avargel and I. Cohen, "On multiplicative transfer function approximation in the short-time Fourier transform domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[7] J. Chen, J Benesty, and Y. Huang, "A minimum distortion noise reduction algorithm with multiple microphones," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 481–493, Mar 2008.

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

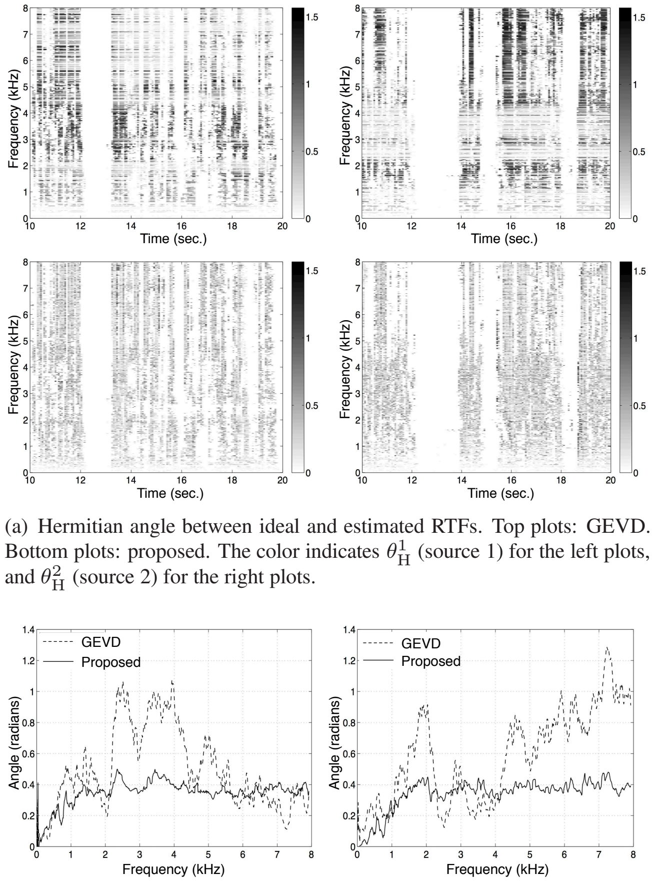
[9] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[10] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sep. 2012.

[11] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2006.

[12] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[13] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, SpringerBriefs in Electrical and Computer Engineering. Springer-Verlag, 2011.



(a) Hermitian angle between ideal and estimated RTFs. Top plots: GEVD. Bottom plots: proposed. The color indicates $\theta_{\rm H}^{1}$ (source 1) for the left plots, and $\theta_{\rm H}^{2}$ (source 2) for the right plots.

(b) Hermitian angle averaged over time. Left: Source 1. Right: Source 2

Fig. 2: Hermitian angle between estimated and ideal RTF vectors. SNR 18 dB, reverberation time 350 ms.