Instantaneous PSD Estimation for Speech Enhancement based on Generalized Principal Components

Thomas Dietzen, Marc Moonen, Toon van Waterschoot
Dept. of Electrical Engineering (ESAT)
STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics
KU Leuven
Leuven, Belgium
{thomas.dietzen, marc.moonen, toon.vanwaterschoot}@esat.kuleuven.be

***Abstract*—Power spectral density (PSD) estimates of various microphone signal components are essential to many speech enhancement procedures. As speech is highly non-nonstationary, performance improvements may be gained by maintaining time-variations in PSD estimates. In this paper, we propose an instantaneous PSD estimation approach based on generalized principal components. Similarly to other eigenspace-based PSD estimation approaches, we rely on recursive averaging in order to obtain a microphone signal correlation matrix estimate to be decomposed. However, instead of estimating the PSDs directly from the temporally smooth generalized eigenvalues of this matrix, yielding temporally smooth PSD estimates, we propose to estimate the PSDs from newly defined instantaneous generalized eigenvalues, yielding instantaneous PSD estimates. The instantaneous generalized eigenvalues are defined from the generalized principal components, i.e. a generalized eigenvector-based transform of the microphone signals. We further show that the smooth generalized eigenvalues can be understood as a recursive average of the instantaneous generalized eigenvalues. Simulation results comparing the multi-channel Wiener filter (MWF) with smooth and instantaneous PSD estimates indicate better speech enhancement performance for the latter. A MATLAB implementation is available online.**

***Index Terms*—speech enhancement, instantaneous PSD estimation, generalized eigenvalue decomposition, generalized principal components**

I. Introduction

In speech enhancement [1]–[3], recorded microphone signals constitute a mixture of speech, reverberation and noise. In order to enhance the mixture, many approaches rely on power spectral density (PSD) estimates of the various mixture components.

While the problem of PSD estimation has attracted much interest [1], [3]–[12] in speech enhancement, somewhat less attention [4]–[6], [12] is paid to the temporal behavior of PSD estimates. As the PSD is a statistical property defined by means of an expectation operator, its estimation typically involves temporal averaging, which approximates the expectation and requires tuning. Note that while temporal averaging lacks practical alternatives, it causes temporal smoothing and hence may be considered non-ideal in case of speech signals, which are highly non-stationary. Indeed, non-stationarity of speech may even be explicitly exploited in a number of speech enhancement approaches [1], [13]–[16], such that quickly time-varying PSD estimates potentially yield a better performance than slowly time-varying PSD estimates. In literature, quickly time-varying PSD estimates are commonly based on short-term statistics, e.g., the local minima of the smoothed microphone signal spectrum [4] or short-term temporal correlations [5], [6]. In [12], we have proposed to restore non-stationarities by desmoothing the generalized eigenvalues of the temporally smooth microphone signal correlation matrix estimate.

In this paper, we propose a multi-microphone eigenspace-based *instantaneous* PSD[1] estimation approach based on generalized principal components. Similarly to other eigenspace-based PSD estimation approaches [5], [7], [10], [12], we rely on recursive averaging in order to obtain a microphone signal correlation matrix estimate to be decomposed. However, instead of estimating the PSDs directly from the temporally *smooth* generalized eigenvalues of this matrix, yielding temporally smooth PSD estimates, we propose to estimate the PSDs from newly defined *instantaneous* generalized eigenvalues, yielding instantaneous PSD estimates. Here, the instantaneous generalized eigenvalues are defined from the generalized principal components, i.e. a generalized eigenvector-based transform of the microphone signals. As to be shown, the smooth generalized eigenvalues can be understood as a recursive average of the newly defined instantaneous gener-

This work was carried out at the ESAT Laboratory of KU Leuven, in the frame of KU Leuven internal fund C2-16-00449; VLAIO O&O Project no. HBC.2017.0358; EU FP7-PEOPLE Marie Curie Initial Training Network funded by the European Commission under Grant Agreement no. 316969; the European Union's Horizon 2020 research and innovation program/ERC Consolidator Grant no. 773268. This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.

[1]Strictly speaking, the term 'PSD' may be said to be inadequate for the instantaneous quantities estimated in this paper, as our approach partly bypasses the use of an expectation or its approximation by means of temporal averaging. Nonetheless, due to the strong relation to expectation-based PSD estimation, we prefer to maintain the terminology.

alized eigenvalues. Simulation results comparing the speech enhancement performance of the multi-channel Wiener filter (MWF) with smooth and instantaneous PSD estimates indicate better performance for the latter. A MATLAB implementation and audio examples are available online [17].

In Sec. II, we present the signal model. In Sec. III, we briefly review the MWF, which serves as an example for the application of PSD estimates and is used to evaluate PSD estimates in this paper. Eigenspace-based PSD estimation is discussed in Sec. IV, where we outline an implementation yielding smooth PSD estimates and propose the alternative approach yielding instantaneous PSD estimates. Both implementations are evaluated in Sec. V.

II. Signal Model

We employ the following notation: vectors are denoted by lower-case boldface letters, matrices by upper-case boldface letters, $\mathbf{I}$ denotes the identity matrix, $\mathbf{A}^T$, $\mathbf{A}^H$, $\mathrm{E}[\mathbf{A}]$, and $\|\mathbf{A}\|_F$ denote the transpose, the complex conjugate transpose, the expected value, and the Frobenius norm of the matrix $\mathbf{A}$. The operation $\mathrm{diag}[\mathbf{A}]$ creates a column vector from the diagonal elements of the matrix $\mathbf{A}$, while $\mathrm{Diag}[\mathbf{a}]$ creates a diagonal matrix from the elements of the vector $\mathbf{a}$. The exponential function with argument a is denoted by $\exp[a]$.

In the short-time Fourier transform (STFT) domain, with m, l, and k indexing the microphone, the frame, and the frequency bin, respectively, and M the number of microphones, let the microphone signals be denoted by $y_m(l,k) \in \mathbb{C}$ with $m = 1,\dots,M$. As we treat all frequency bins independently, the frequency bin index is omitted in the following. We define the stacked microphone signal vector $\mathbf{y}(l) \in \mathbb{C}^M$,

$$\mathbf{y}(l) = \begin{pmatrix} y_1(l) & \cdots & y_M(l) \end{pmatrix}^T, \tag{1}$$

composed of the reverberant speech component $\mathbf{x}(l)$ originating from a single point source and the noise component $\mathbf{v}(l)$,

$$\mathbf{y}(l) = \mathbf{x}(l) + \mathbf{v}(l). \tag{2}$$

The reverberant speech component $\mathbf{x}(l)$ may be decomposed into the early component $\mathbf{x}_{\mathrm{e}}(l)$ containing the direct component and early reflections, and the late reverberant component $\mathbf{x}_\ell(l)$ containing late reflections, i.e.

$$\mathbf{x}(l) = \mathbf{x}_{\mathrm{e}}(l) + \mathbf{x}_\ell(l), \tag{3}$$

which are assumed to have distinct spatial properties as outlined below. Early reflections are assumed to arrive within the same frame, with the early components in $\mathbf{x}_{\mathrm{e}}(l)$ related by the relative early transfer functions (RETFs) in $\mathbf{h} \in \mathbb{C}^M$, i.e.

$$\mathbf{x}_{\mathrm{e}}(l) = \mathbf{h} s(l), \tag{4}$$

Here, $\mathbf{h}$ is assumed to be relative to the first microphone, i.e. $h_1 = 1$, and $s(l) = \mathbf{x}_{\mathrm{e}|1}(l)$ denotes the early component in the first microphone, in the following referred to as early speech source image. We consider $\mathbf{h}$ to be known or previously estimated [3], [12], [18]. We assume that $\mathbf{x}_{\mathrm{e}}(l)$, $\mathbf{x}_\ell(l)$, and $\mathbf{v}(l)$ are mutually uncorrelated [7]–[12]. Let $\mathbf{\Psi}_y(l) = \mathrm{E}[\mathbf{y}(l)\mathbf{y}^H(l)] \in \mathbb{C}^{M\times M}$ denote the microphone signal correlation matrix, and let $\mathbf{\Psi}_{x_{\mathrm{e}}}(l)$, $\mathbf{\Psi}_{x_\ell}(l)$, and $\mathbf{\Psi}_v(l)$ be similarly defined. With (2)–(4), we then find

$$\mathbf{\Psi}_y(l) = \mathbf{\Psi}_{x_{\mathrm{e}}}(l) + \mathbf{\Psi}_{x_\ell}(l) + \mathbf{\Psi}_v(l), \tag{5}$$

wherein $\mathbf{\Psi}_{x_{\mathrm{e}}}(l)$ has rank one and is expressed by

$$\mathbf{\Psi}_{x_{\mathrm{e}}}(l) = \varphi_s(l)\mathbf{h}\mathbf{h}^H, \tag{6}$$

with $\varphi_s(l)$ denoting the PSD of the early speech source image $s(l)$. Assuming that $\mathbf{x}_\ell(l)$ and $\mathbf{v}(l)$ may be modeled as diffuse [7]–[11], [19] with coherence matrix $\mathbf{\Gamma} \in \mathbb{C}^{M\times M}$, which may be computed from the microphone array geometry [19] and is therefore considered to be known, we may write $\mathbf{\Psi}_{x_\ell}(l) + \mathbf{\Psi}_v(l)$ as

$$\mathbf{\Psi}_{x_\ell}(l) + \mathbf{\Psi}_v(l) = \varphi_d(l)\mathbf{\Gamma}, \tag{7}$$

$$\text{with} \quad \varphi_d(l) = \varphi_{x_\ell}(l) + \varphi_v(l), \tag{8}$$

and $\varphi_{x_\ell}(l)$ and $\varphi_v(l)$ denoting the PSD of the late reverberant component $\mathbf{x}_\ell(l)$ and the noise component $\mathbf{v}(l)$, respectively. With $s(l)$ representing speech, and in particular if $\mathbf{v}(l)$ represents babble noise, both PSDs $\varphi_s(l)$ and $\varphi_d(l)$ may be considered highly non-stationary, while the associated coherence matrices $\mathbf{h}\mathbf{h}^H$ and $\mathbf{\Gamma}$ are often considered time-invariant [7]–[10].

In the remainder, as we mostly consider the single frame l only, we also drop the frame index for conciseness and refer back to it only where necessary, namely when we differentiate the frames l and $l-1$ in recursive equations.

III. Multi-Channel Wiener Filter

PSD estimates are used in a variety of speech enhancement procedures. In this paper, we evaluate our PSD estimation approach in Sec. V by means of the MWF, which is therefore briefly summarized below.

The MWF $\mathbf{w}_{\mathrm{MWF}}$ is obtained [2], [3] by minimizing the expected error between the filter output and the early speech source image, i.e.

$$\begin{aligned}\mathbf{w}_{\mathrm{MWF}} &= \arg\min_{\mathbf{w}} \mathrm{E}\big[|\mathbf{w}^H\mathbf{y} - s|^2\big] \\ &= \varphi_s \mathbf{\Psi}_y^{-1}\mathbf{h}.\end{aligned} \tag{9}$$

It is well known that the MWF can be decomposed [2], [3] into a minimum variance distortionless response (MVDR) beamformer and a spectral gain as

$$\mathbf{w}_{\mathrm{MWF}} = \underbrace{\frac{\mathbf{\Gamma}^{-1}\mathbf{h}}{\mathbf{h}^H\mathbf{\Gamma}^{-1}\mathbf{h}}}_{\substack{\text{MVDR}\\ \text{beamformer}}} \cdot \underbrace{\frac{\varphi_s}{\varphi_s + \varphi_d\mathbf{h}^H\mathbf{\Gamma}^{-1}\mathbf{h}}}_{\text{spectral gain}}. \tag{10}$$

Hence, if both $\mathbf{\Gamma}$ and $\mathbf{h}$ are assumed to be known or previously estimated, the problem of implementing the MWF reduces to estimating the PSDs φ_s and φ_d. If, on the one hand, the PSD estimates to be obtained are slowly time-varying, the spectral gain will contribute to speech enhancement mostly through variations across frequency. If, on the other hand, instantaneous PSD estimates are obtained, the spectral gain will vary across both frequency and time and thereby act as a spectro-temporal mask [13]–[15].

IV. Eigenspace-based PSD Estimation

Multi-microphone PSD estimation is commonly based on the spatial properties defined in (4)–(8), which may be exploited in an eigenspace decomposition [7], [10], [12]. In Sec. IV-A, we first introduce an eigenspace model of $\mathbf{\Psi}_y$ and $\mathbf{\Gamma}$. In Sec. IV-B, we outline how PSD estimates may be obtained given an eigenvalue and an eigenspace basis estimate. In Sec. IV-C, we consider an implementation based on temporally smooth eigenvalues, and in Sec. IV-D, we propose an implementation based on instantaneous generalized principal components.

A. Eigenspace Model

We define the generalized eigenvalue decomposition (GEVD) [7], [10], [12], [18] of $\mathbf{\Psi}_y$ and the diffuse coherence matrix $\mathbf{\Gamma}$, cf. (7), i.e.

$$\mathbf{\Psi}_y\mathbf{P} = \mathbf{\Gamma}\mathbf{P}\,\mathrm{Diag}[\boldsymbol{\lambda}_y], \tag{11}$$

where $\boldsymbol{\lambda}_y \in \mathbb{R}^M$ comprises the generalized eigenvalues $\lambda_{y|m}$, and the columns $\mathbf{p}_m$ of $\mathbf{P} \in \mathbb{C}^{M\times M}$ comprise the associated generalized eigenvectors. The generalized eigenvectors in $\mathbf{P}$ are uniquely defined up to a scaling factor and, for any factorization $\mathbf{\Gamma} = \mathbf{\Gamma}^{1/2}\mathbf{\Gamma}^{H/2}$, may be chosen such that $\mathbf{\Gamma}^{H/2}\mathbf{P}$ becomes unitary due to $\mathbf{\Psi}_y$ and $\mathbf{\Gamma}$ being Hermitian. The matrices $\mathbf{\Psi}_y$ and $\mathbf{\Gamma}$ are then diagonalized by

$$\mathbf{P}^H\mathbf{\Psi}_y\mathbf{P} = \mathrm{Diag}[\boldsymbol{\lambda}_y], \tag{12}$$

$$\mathbf{P}^H\mathbf{\Gamma}\mathbf{P} = \mathbf{I}, \tag{13}$$

cf. also (11).

While the eigenspace basis $\mathbf{P}$ varies with the spatial coherence matrices $\mathbf{h}\mathbf{h}^H$ and $\mathbf{\Gamma}$ only and is therefore time-invariant in the assumed spatially stationary scenario, the generalized eigenvalues in $\boldsymbol{\lambda}_y$ vary with the PSDs φ_s and φ_d and hence over time. Using (5) and (7) in (12)–(13) yields

$$\mathrm{Diag}[\boldsymbol{\lambda}_y] = \mathrm{Diag}[\boldsymbol{\lambda}_{x_e}] + \mathrm{Diag}[\boldsymbol{\lambda}_d], \tag{14}$$

$$\text{with}\quad \mathrm{Diag}[\boldsymbol{\lambda}_{x_e}] = \mathbf{P}^H\mathbf{\Psi}_{x_e}\mathbf{P}, \tag{15}$$

$$\mathrm{Diag}[\boldsymbol{\lambda}_d] = \varphi_d\mathbf{I}. \tag{16}$$

In (15), $\mathbf{\Psi}_{x_e}$ and therefore $\mathrm{Diag}[\boldsymbol{\lambda}_{x_e}]$ have rank one. Provided that the generalized eigenvalues and eigenvectors are sorted such that $\lambda_{y|1}$ is the largest generalized eigenvalue, $\boldsymbol{\lambda}_{x_e}$ hence takes the form

$$\boldsymbol{\lambda}_{x_e} = \begin{pmatrix}\lambda_{x_e|1} & 0 & \cdots & 0\end{pmatrix}^T. \tag{17}$$

From (14)–(17) it then follows that $\lambda_{y|1} = \lambda_{x_e|1} + \varphi_d$ and $\lambda_{y|m} = \varphi_d$ for $m > 1$ [10].

B. Eigenspace-based PSD Estimation

Assume that an estimate $\hat{\mathbf{\Psi}}_y$ is available, from which the eigenvalue and eigenspace basis estimates $\hat{\boldsymbol{\lambda}}_y$ and $\hat{\mathbf{P}}$ are obtained. Further, assume that the RETF $\mathbf{h}$ is known or previously estimated. Estimates of $\hat{\varphi}_s$ and $\hat{\varphi}_d$ can then be obtained in the following manner.

Given $\hat{\boldsymbol{\lambda}}_y$, we first obtain $\hat{\varphi}_d$ and $\hat{\lambda}_{x_e|1}$ according to (14)–(17) [10] as

$$\hat{\varphi}_d = \frac{1}{M-1}\sum_{m=2}^{M}\hat{\lambda}_{y|m}, \tag{18}$$

$$\hat{\lambda}_{x_e|1} = \hat{\lambda}_{y|1} - \hat{\varphi}_d. \tag{19}$$

where the averaging in (18) accounts for modeling and estimation errors and (19) is guaranteed non-negative. Noting that $\hat{\mathbf{P}}^{-1} = \hat{\mathbf{P}}^H\mathbf{\Gamma}$ according to (13), we can define a rank-one estimate $\hat{\mathbf{\Psi}}_{x_e}$ [12] as

$$\begin{aligned}\hat{\mathbf{\Psi}}_{x_e} &= \mathbf{\Gamma}\hat{\mathbf{P}}\,\mathrm{Diag}[\hat{\boldsymbol{\lambda}}_{x_e}]\hat{\mathbf{P}}^H\mathbf{\Gamma}\\ &= \hat{\lambda}_{x_e|1}\mathbf{\Gamma}\hat{\mathbf{p}}_1\hat{\mathbf{p}}_1^H\mathbf{\Gamma}\end{aligned} \tag{20}$$

with $\hat{\boldsymbol{\lambda}}_{x_e}$ similar to (17). An estimate $\hat{\varphi}_s$ may then be obtained by minimizing the difference[2] between $\hat{\mathbf{\Psi}}_{x_e}$ and $\varphi_s\mathbf{h}\mathbf{h}^H$ according to (6) [12], i.e.

$$\begin{aligned}\hat{\varphi}_s &= \arg\min_{\varphi_s}\|\varphi_s\mathbf{h}\mathbf{h}^H - \hat{\mathbf{\Psi}}_{x_e}\|_F^2\\ &= \hat{\lambda}_{x_e|1}\left|\frac{\mathbf{h}^H\mathbf{\Gamma}\hat{\mathbf{p}}_1}{\mathbf{h}^H\mathbf{h}}\right|^2.\end{aligned} \tag{21}$$

Note that the temporal characteristics of the estimates $\hat{\varphi}_s$ and $\hat{\varphi}_d$ directly depend upon the temporal characteristics of $\hat{\boldsymbol{\lambda}}_y$.

C. Smooth Eigenvalue-based Implementation

A temporally smooth estimate of $\mathbf{\Psi}_y = \mathrm{E}[\mathbf{y}\mathbf{y}^H]$, in the following denoted by $\hat{\mathbf{\Psi}}_{y|sm}$, is typically obtained by recursively averaging $\mathbf{y}\mathbf{y}^H$ using some pre-defined forgetting factor $\zeta \in (0,\,1)$, namely by

$$\hat{\mathbf{\Psi}}_{y|sm}(l) = \zeta\hat{\mathbf{\Psi}}_{y|sm}(l-1) + (1-\zeta)\mathbf{y}(l)\mathbf{y}^H(l). \tag{22}$$

The forgetting factor ζ may be expressed in terms of a time constant τ as

$$\zeta = \exp\left[-R/f_s\tau\right], \tag{23}$$

where R is the STFT frame shift in samples, f_s is the sampling rate and τ may be thought of as an equivalent window length.

Given $\hat{\mathbf{\Psi}}_{y|sm}$, we can perform the GEVD $\hat{\mathbf{\Psi}}_{y|sm}\hat{\mathbf{P}} = \mathbf{\Gamma}\hat{\mathbf{P}}\,\mathrm{Diag}[\hat{\boldsymbol{\lambda}}_{y|sm}]$ similar to (11)–(13) in each frame l. Here, $\hat{\mathbf{P}}$ slightly fluctuates over time due to modeling and estimation errors (while $\mathbf{P}$ itself is time-invariant, cf. Sec. IV-A), and $\hat{\boldsymbol{\lambda}}_{y|sm}$ is a smooth estimate of $\boldsymbol{\lambda}_y$. Consequently, if we estimate the PSDs φ_s and φ_d directly from $\hat{\boldsymbol{\lambda}}_{y|sm}$ according to Sec. IV-B, we obtain equally smooth estimates $\hat{\varphi}_{s|sm}$ and $\hat{\varphi}_{d|sm}$. Note that in order to span all M eigenspace dimensions and hence to obtain a meaningful decomposition, $\hat{\mathbf{\Psi}}_{y|sm}$ needs to be well-conditioned, and so τ should scale with M and must be sufficiently large.

[2]Since $h_1 = 1$, cf. Sec. II, one may alternatively obtain an estimate $\hat{\varphi}_s$ directly from the upper left element of $\hat{\mathbf{\Psi}}_{x_e}$ [12]. During speech pauses, however, where $\hat{\mathbf{\Psi}}_{x_e}$ deviates from zero due to modeling and estimation errors only, the estimator in (21) is more robust.

D. Instantaneous Principal Component-based Implementation

In order to obtain instantaneous eigenspace-based PSD estimates while still relying on recursive averaging as in (22) with a sufficiently large time constant τ, we propose to compute instantaneous generalized eigenvalues $\hat{\boldsymbol{\lambda}}_{y|inst}$ based on generalized principal components instead of using the smooth generalized eigenvalues $\hat{\boldsymbol{\lambda}}_{y|sm}$ directly.

In order to introduce the generalized principal components and establish its relation to the generalized eigenvalues, let us reconsider the GEVD in (11)–(13). From the generalized eigenvectors in $\mathbf{P}$, we can define the generalized principal components of $\mathbf{y}$ as

$$\boldsymbol{\vartheta} = \mathbf{P}^H \mathbf{y}. \tag{24}$$

Note that with $\boldsymbol{\Psi}_y = \mathrm{E}[\mathbf{y}\mathbf{y}^H]$, the generalized principal components in (24) are related to the generalized eigenvalues in (12) by

$$\boldsymbol{\lambda}_y = \mathrm{diag}\big[\mathrm{E}[\boldsymbol{\vartheta}\boldsymbol{\vartheta}^H]\big]. \tag{25}$$

Now, assume that we have obtained $\hat{\boldsymbol{\Psi}}_{y|sm}$ and its generalized eigenvectors in $\hat{\mathbf{P}}$ as described in Sec. IV-C. Then, with $\hat{\boldsymbol{\vartheta}} = \hat{\mathbf{P}}\mathbf{y}$, we define the instantaneous generalized eigenvalues

$$\hat{\boldsymbol{\lambda}}_{y|inst} = \mathrm{diag}[\hat{\boldsymbol{\vartheta}}\hat{\boldsymbol{\vartheta}}^H], \tag{26}$$

which maintain non-stationarities as they directly depend on the microphone signal $\mathbf{y}$, cf. (24). Based on $\hat{\boldsymbol{\lambda}}_{y|inst}$, we can then obtain instantaneous PSD estimates $\hat{\varphi}_{s|inst}$ and $\hat{\varphi}_{d|inst}$ according to Sec. IV-B.

Note that we may also establish a relation between the instantaneous generalized eigenvalues in (26) and the smooth generalized eigenvalues obtained in Sec. IV-C. With $\hat{\boldsymbol{\lambda}}_{y|sm} = \mathrm{diag}[\hat{\mathbf{P}}^H \hat{\boldsymbol{\Psi}}_{y|sm} \hat{\mathbf{P}}]$ according to (12), inserting $\hat{\boldsymbol{\Psi}}_{y|sm}$ from (22) and using (24), (26), we find

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{y|sm}(l) = \zeta\, \mathrm{diag}[\hat{\mathbf{P}}(l)^H \hat{\boldsymbol{\Psi}}_{y|sm}(l-1)\hat{\mathbf{P}}(l)] \\ + (1-\zeta)\hat{\boldsymbol{\lambda}}_{y|inst}(l),\end{aligned} \tag{27}$$

where any time variations in $\hat{\mathbf{P}}(l)$ are due to modeling and estimation errors only, cf. Sec. IV-C, such that $\mathrm{diag}[\hat{\mathbf{P}}(l)^H \hat{\boldsymbol{\Psi}}_{y|sm}(l-1)\hat{\mathbf{P}}(l)] \approx \hat{\boldsymbol{\lambda}}_{y|sm}(l-1)$. The smooth generalized eigenvalues $\hat{\boldsymbol{\lambda}}_{y|sm}$ therefore nearly correspond to a recursive average of the instantaneous generalized eigenvalues $\hat{\boldsymbol{\lambda}}_{y|inst}$.

V. Simulations

In this section, we compare the speech enhancement performance of the MWF with smooth PSD estimates $\hat{\varphi}_{s|sm}$, $\hat{\varphi}_{d|sm}$ according to Sec. IV-C and the MWF with instantaneous PSD estimates $\hat{\varphi}_{s|inst}$, $\hat{\varphi}_{d|inst}$ according to Sec. IV-D as a function of the time constant τ.

In our simulations, we use a linear array of $M = 5$ microphones spaced by $8\,\mathrm{cm}$. In total, 48 scenarios are generated. The source is positioned $2\,\mathrm{m}$ away at an angle of $\{0,\ 30,\ 60\}^\circ$ relative to the broadside direction of the microphone array, where sound propagation is modeled using measured room impulse responses (RIRs) [20] of $0.61\,\mathrm{s}$ reverberation time. In each source position, both male and female speech are used as source signals, where we select 8 sections of $10\,\mathrm{s}$ from each of the source signal files [21]. Diffuse babble noise [22], [23] is added at a signal-to-noise ratio (SNR) of $5\,\mathrm{dB}$, where the SNR is defined as the power ratio of $\mathbf{x}$ and $\mathbf{v}$ in the time domain. The sampling rate is $f_s = 16\,\mathrm{kHz}$. The STFT processing uses square root Hann windows of 512 samples with $R = 256$ samples overlap. The presumed available estimates of the RETFs in $\mathbf{h}$ are generated based on the directions of arrival, i.e. the estimate corresponds to the free-field steering vector. We measure performance in terms of the perceptual evaluation of speech quality *PESQ* [24] with mean opinion scores $\in [1, 4.5]$, the short-time objective

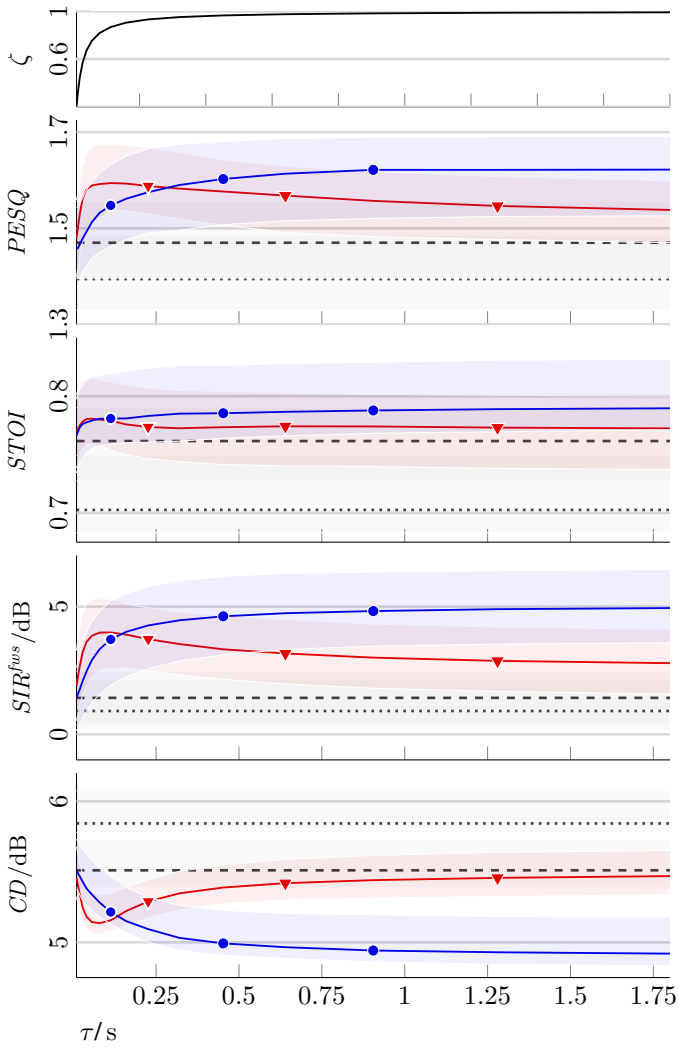

Fig. 1: The forgetting factor ζ [——] and the performance measures *PESQ*, *STOI*, SIR^{fws}, and *CD* versus τ for the first microphone signal [·····], the MVDR [- - -], the MWF with smooth PSD estimates [–▾–] and the MWF with instantaneous PSD estimates [–•–]. The graphs denote the median scores over all scenarios, the shaded areas indicate the range from the first to the third quartile.

intelligibility $STOI$ [25] with scores $\in [0, 1]$, the frequency-weighted segmental signal-to-interference ratio SIR^{fws} [1] in dB and the cepstral distance CD [1] in dB. The clean reference signal is generated by convolving the speech source signal with the early part of the RIR to the first microphone. The computed measures are averaged over all 48 scenarios.

Fig. 1 reports the simulation results. As to be expected, both versions of the MWF [–▾–, –•–] outperform the MVDR [- - -], which in turn shows some improvement over the unprocessed microphone signal [·····]. The two versions of the MWF however show a different behavior. The MWF with smooth PSD estimates [–▾–] reaches a fairly sharp performance peak at relatively low values of τ, with decreasing performance for larger values, where the spectral gain in (10) becomes less time-variant. This behavior is explained by the fact that when computing smooth PSD estimates according to Sec. IV-C, the time constant τ trades off the accuracy of the eigenspace basis estimate $\hat{\mathbf{P}}$ on the one hand and the degree of non-stationarity maintained in the PSD estimates $\hat{\varphi}_{s|sm}$ and $\hat{\varphi}_{d|sm}$ on the other hand. The MWF with instantaneous PSD estimates [–•–] in contrast shows a monotonous performance increase in τ, which facilitates tuning. This is explained by the fact that when computing instantaneous PSD estimates according to Sec. IV-D, the accuracy of the eigenspace basis estimate $\hat{\mathbf{P}}$ still increases with τ, while the instantaneous PSD estimates $\hat{\varphi}_{s|inst}$ and $\hat{\varphi}_{d|inst}$ maintain non-stationarities independently of τ. At large values of τ, in all measures, the improvement with respect to the MVDR is more than twice as large for MWF with instantaneous PSD estimates as compared to the MWF with smooth PSD estimates. Note that in a spatially dynamic scenario with time-varying RETF $\mathbf{h}$, the eigenspace basis $\mathbf{P}$ becomes time-variant, in which case the performance of the MWF with instantaneous PSD estimates might possibly not increase monotonically in τ anymore, but will presumably show a peak depending on the pace of RETF variations.

VI. Conclusion

In this paper, as an alternative to smooth PSD estimation based on smooth generalized eigenvalues, we have proposed an instantaneous PSD estimation approach based on generalized principal components. The instantaneous PSD estimates maintain non-stationarities and hence potentially outperform smooth PSD estimates for speech enhancement, as exemplarily shown for the MWF.

References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[2] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*. Wiley, 2010, pp. 269–302.

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

[4] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[5] R. C. Hendriks, J. J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.

[6] A. H. Kamkar-Parsi and M. Bouchard, "Instantaneous binaural target PSD estimation for hearing aid noise reduction in complex acoustic environments," *IEEE Trans. Instrument. Meas.*, vol. 60, no. 4, pp. 1141–1154, Apr. 2011.

[7] A. Kuklasiński, S. Doclo, S. H. Jensen, and J. Jensen, "Maximum likelihood PSD estimation for speech enhancement in reverberation and noise," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1599–1612, Sep. 2016.

[8] O. Schwartz, S. Gannot, and E. A. P. Habets, "Joint estimation of late reverberant and speech power spectral densities in noisy environments using Frobenius norm," in *Proc. 24th European Signal Process. Conf. (EUSIPCO 2016)*, Budapest, Hungary, Aug. 2016, pp. 1123–1127.

[9] S. Braun, A. Kuklasiński, O. Schwartz, O. Thiergart, E. A. P. Habets, S. Gannot, S. Doclo, and J. Jensen, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 26, no. 6, pp. 1056–1071, June 2018.

[10] I. Kodrasi and S. Doclo, "Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1102–1114, June 2018.

[11] A. I. Koutrouvelis, R. C. Hendriks, R. Heusdens, and J. Jensen, "Robust joint estimation of multi-microphone signal model parameters," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 7, pp. 1136–1150, July 2019.

[12] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Square root-based multi-source early PSD estimation and recursive RETF update in reverberant environments by means of the orthogonal Procrustes problem," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 755 – 769, Jan. 2020.

[13] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1673–1682, Mar. 2008.

[14] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, no. 4, pp. 2336–2347, Apr. 2009.

[15] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. 2004 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 7092–7096.

[16] T. Dietzen, S. Doclo, M. Moonen, and T. van Waterschoot, "Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone dereverberation, interfering speech cancellation, and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 740 – 754, Jan. 2020.

[17] T. Dietzen, "GitHub repository: instantaneous PSD estimation for speech enhancement based on generalized principal components," https://github.com/tdietzen/INSTANT-PSD, Mar. 2020.

[18] 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 *Proc. 2015 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2015)*, Brisbane, QLD, Australia, Apr. 2015, pp. 544–548.

[19] F. Jacobsen and T. Roisin, "The coherence of reverberant sound fields," *J. Acoust. Soc. Amer.*, vol. 108, no. 1, pp. 204–210, July 2000.

[20] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. 2014 Int. Workshop Acoustic Signal Enhancement (IWAENC 2014)*, Antibes – Juan les Pins, France, Sept. 2014, pp. 313–317.

[21] Bang and Olufsen, "Music for Archimedes," Compact Disc B&O, 1992.

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

[23] Auditec, "Auditory tests (revised)," Compact Disc Auditec, 1997.

[24] ITU-T, "Perceptual evaluation of of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," in *ITU-T Recommendation P.862, Int. Telecommun. Union*, Geneva, Switzerland, Feb. 2001.

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.