EUSIPCO 2013 1569743791

ON ACOUSTIC CHANNEL IDENTIFICATION IN MULTI-MICROPHONE SYSTEMS VIA ADAPTIVE BLIND SIGNAL ENHANCEMENT TECHNIQUES

Gerald Enzner, Dominic Schmid *

Institute of Communication Acoustics
Ruhr-Universität Bochum, D-44780 Bochum
{gerald.enzner, dominic.schmid}@rub.de

Reinhold Haeb-Umbach

Department of Communications Engineering
University of Paderborn, D-33098 Paderborn
haeb@nt.uni-paderborn.de

ABSTRACT

Among the different configurations of multi-microphone systems, e.g., in applications of speech dereverberation or denoising, we consider the case without *a priori* information of the microphone-array geometry. This naturally invokes explicit or implicit identification of source-receiver transfer functions as an indirect description of the microphone-array configuration. However, this *blind channel identification* (BCI) has been difficult due to the lack of unique identifiability in the presence of observation noise or near-common channel zeros. In this paper, we study the implicit BCI performance of blind signal enhancement techniques such as the adaptive *principal component analysis* (PCA) or the iterative *blind equalization and channel identification* (BENCH). To this end, we make use of a recently proposed metric, the *normalized filter-projection misalignment* (NFPM), which is tailored for BCI evaluation in ill-conditioned (e.g., noisy) scenarios. The resulting understanding of implicit BCI performance can help to judge the behavior of multi-microphone speech enhancement systems and the suitability of implicit BCI to serve channel-based (i.e., channel-informed) enhancement.

1. INTRODUCTION

Figure 1 depicts different ways to obtain an enhanced speech signal $\widehat{s}(k)$ at discrete time k from observed noisy microphone signals $y_i(k)$, $i = 1 \dots P$. The desired source signal, prior to the acoustic transmission to the microphones, is denoted $s(k)$. We first look at the individual building blocks of the figure to briefly describe the context and some previous work in multi-microphone speech enhancement, before we outline the particular focus of this paper.

Let us define the class of *channel-informed signal estimators* which rely on a generative signal model, i.e., a linear expression of the observations $y_i(k)$ in terms of the source $s(k)$, individual source-to-microphone acoustic impulse responses (AIRs) $h_{i,k}$, and observation noises $n_i(k)$. The AIRs are further assumed to be known *a priori*. The multiple-input/output inverse theorem (MINT) [1] then achieves perfect equalization of the acoustic multipath transmission, subject to the absence of common channel zeros. Numerical robustness and the effect of noise on the equalization were, however, not addressed in the MINT context. Recent work, e.g. [2], thus renders a larger picture of channel-informed estimation (i.e., from minimum mean-square error to least-squares criteria) to take noise into account and to suggest a least-squares approximation with guaranteed numerical stability. As an alternative to absolute source-to-microphone transfer functions, it was also proposed to just rely on relative transfer functions (RTFs), e.g., to support adaptive beamforming based on a generalized sidelobe canceller (GSC) structure [3, 4].

*This work was partly supported by grant DFG EN 869/1-3

In commonly time-varying acoustic environments, the required source-to-microphone AIRs or RTFs are not available *a priori*. As a direct consequence, the field of BCI, originally known from communications [5, 6], quickly evolved in the audio and acoustics domain to resolve this lack of information via estimation on the basis of the observed microphone signals [7]. Especially the adaptive approaches based on recursive *cross-relation-error* (CRE) minimization can be applicable for online AIR inference [8], [9], [10]. However, all previous work also recognizes that BCI has to be used with caution since a set of identifiability conditions, e.g., the absence of observation noise and common channel zeros, are typically not met in acoustics. Robust algorithms to overcome the identifiability issues were proposed, e.g., [11], [12]. Recently, it was proven by theory and simulation that RTFs can be well estimated by CRE minimization, despite the violated identifiability conditions [13].

Apart from the two-stage workflow consisting of BCI and channel-informed signal estimation, the literature offers solutions for *direct or blind signal enhancement*. Here, let us refer to the adaptive Frost beamformer [14] which uses a *constrained least mean-square* (LMS) algorithm to find a *minimum-variance distortionless response* (MVDR) solution. Griffiths and Jim then introduced the aforementioned GSC [15] as an unconstrained adaptive MVDR implementation. Since those beamformers at least require direction-of-arrival information, they might still be considered as a variant of the two-stage approach. A truly blind adaptive beamformer can be formulated via PCA as shown by the frequency-domain adaptive algorithms in [16, 4]. The structure of these PCA algorithms is closely related to Oja's online PCA rule in neural networks [17]. Another class of blind signal estimators utilizes the *expectation-maximization* (EM) framework [18] to iteratively estimate the source signal in conjunction with latent variables [19, 20, 21].

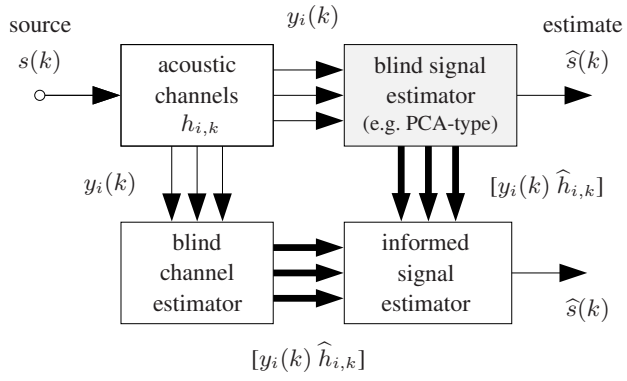

Fig. 1. Three ways from a source signal $s(k)$ to its estimate $\widehat{s}(k)$.

In this paper, we consider the less explored option of obtaining a blind channel estimate $\widehat{h}_{i,k}$ from the algorithms of the direct-signal-enhancement category. Motivated by the fact that good BCI enables good channel-informed signal estimation, conversely, we expect a native blind signal estimator to implicitly perform successful BCI. Indeed, there is already some evidence for this statement, since Oja's early work [22] already highlights a matched-filter characteristic of the PCA with respect to the unknown channel. Furthermore, the EM-type BENCH estimators in [20, 21] make explicit use of the acoustic channel as the latent variable for which the posterior distribution is obtained in conjunction with the signal estimate. According to [17], these EM-type adaptive algorithms, based on a generative stochastic channel model, fall into the class of *factor analysis* techniques and thus form at least some structural relationship but also differences with PCA – see Sec. 2.3. On the other hand, the aforementioned PCA is well known for its ambiguity due to its direct relationship with eigenvector analysis. This makes the PCA a questionable tool to completely solve the BCI task. Our paper therefore applies NFPM [13] as a relaxed impulse-response-distance to evaluate the implicit channel estimates obtained from PCA or BENCH. To achieve diversity of the study, we compare the performance of (a) Oja's classical online rule, (b) the *power iteration* for adaptive eigenvector tracking as used in [4], and (c) the maximum-likelihood BENCH algorithm.

2. SIGNAL MODEL AND ADAPTIVE ALGORITHMS

Figure 2 formally depicts the acoustic transmission from source to microphones as a SIMO system comprising P acoustic impulse responses $h_{i,k}$ driven by the common input signal $s(k)$. Considering the additive observation noises $n_i(k)$, the corresponding linear convolution model for the i-th microphone signal reads

$$y_i(k) = \sum_{\kappa=0}^{L-1} h_{i,\kappa}\, s(k-\kappa) + n_i(k)\ , \quad i = 1, ..., P\ , \qquad (1)$$

where L denotes the number of acoustic channel coefficients. In order to prepare for adaptive signal processing, the most recent samples of the microphone signals are assembled in length R vectors

$$\mathbf{y}_{i,k} = [y_i(k-R+1)\ y_i(k-R+2)\ \cdots\ y_i(k)]^T\ . \qquad (2)$$

Most of the times, we make use of its DFT-domain representation $\underline{\mathbf{y}}_{i,\tau} = \mathbf{F}_M \mathbf{Q} \mathbf{y}_{i,\tau R}$ at frame time τ, where $\mathbf{Q} = [\,\mathbf{0}_{R\times L}\ \mathbf{I}_R\,]^T$ is an $M\times R$ zero-padding and $\mathbf{F}_M$ the size M Fourier matrix, $M = L+R$.

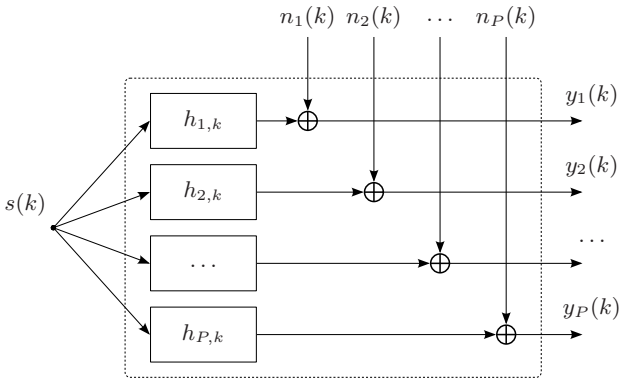

Fig. 2. P-channel single-input multiple-output (SIMO) system.

2.1. Frequency-Domain Online-PCA Algorithm (Oja)

Let $\underline{y}_{i,\tau}[m]$ denote the m-th element (i.e., m-th frequency bin) of the PCA input $\underline{\mathbf{y}}_{i,\tau}$. Since in this section the same operations are carried out for all frequencies, the frequency bin index $[m]$ will be omitted in the following. The mathematical formulation of a frequency-domain PCA then mainly requires the definition of corresponding weights $\underline{w}_i$ to form a complex-valued linear combination of the multiple microphone signals [17], independently for each frequency,

$$\widehat{\underline{s}}(\tau) = \sum_i \underline{w}_i^*\, \underline{y}_{i,\tau} = \underline{\mathbf{w}}^H \underline{\mathbf{y}}_\tau\ , \qquad (3)$$

where $\underline{\mathbf{w}}^H = [\underline{w}_1^*\ \underline{w}_2^*\ \cdots\ \underline{w}_P^*]$ and $\underline{\mathbf{y}}_\tau^H = [\underline{y}_{1,\tau}^*\ \underline{y}_{2,\tau}^*\ \cdots\ \underline{y}_{P,\tau}^*]$ are length P stacked versions of the bin-wise scalars, and asterisk * denotes complex conjugation.

The value $\widehat{\underline{s}}(\tau)$ is the principal component of $\underline{\mathbf{y}}_\tau$, if the power of $\widehat{\underline{s}}(\tau)$ is maximally large. This constitutes one possible definition of blind signal estimation in Fig. 1. Because $\widehat{\underline{s}}(\tau)$ could grow without limits if $\underline{\mathbf{w}}$ grows, a unit-norm constraint $||\underline{\mathbf{w}}||^2 = \underline{\mathbf{w}}^H\underline{\mathbf{w}} = 1$ is further imposed. Formally, the power of $\widehat{\underline{s}}(\tau)$ is given by

$$\sigma_s^2 = \mathrm{E}\,\{\widehat{\underline{s}}(\tau)\widehat{\underline{s}}^*(\tau)\} = \mathrm{E}\big\{\underline{\mathbf{w}}^H \underline{\mathbf{y}}_\tau \underline{\mathbf{y}}_\tau^H \underline{\mathbf{w}}\big\} = \underline{\mathbf{w}}^H \mathbf{C}_y \underline{\mathbf{w}}\ , \qquad (4)$$

where $\mathbf{C}_y = \mathrm{E}\{\underline{\mathbf{y}}_\tau \underline{\mathbf{y}}_\tau^H\}$ is known as the data covariance, in general, or as the power spectral density (PSD) matrix of the microphone signals here, and $\mathrm{E}\{\cdot\}$ denotes statistical expectation. The maximization of the output power under the aforementioned constraint is achieved by unconstrained maximization of the objective function

$$\underline{\mathbf{w}}^H \mathbf{C}_y \underline{\mathbf{w}} + \lambda\big(\underline{\mathbf{w}}^H \underline{\mathbf{w}} - 1\big)\ , \qquad (5)$$

where λ is a Lagrange multiplier. Equating the complex derivative w.r.t. $\underline{\mathbf{w}}^*$ to zero then results in a standard eigenvalue problem, i.e.,

$$\mathbf{C}_y \underline{\mathbf{w}} = \widetilde{\lambda} \underline{\mathbf{w}}\ , \qquad (6)$$

$\widetilde{\lambda} = -\lambda \in \mathbb{R}$, explaining the unavoidable $\underline{\mathbf{w}}$-ambiguity in PCA up to an independent complex gain in each frequency bin.

Based on the PCA statement in (3), we can directly apply Oja's celebrated online PCA rule to find the eigenvector for the largest eigenvalue of the matrix $\mathbf{C}_y$ as a solution to the constrained optimization problem. The literature provides both gradient-ascent [22, 17] as well as gradient-descent [23] derivations to arrive at Oja's LMS-type adaptive algorithm for $\underline{\mathbf{w}}$-tracking. Oja's PCA rule applied to each frequency bin reads

$$\widehat{\underline{\mathbf{w}}}_{\tau+1} = \widehat{\underline{\mathbf{w}}}_\tau + \mu \left[\underline{\mathbf{y}}_\tau - \widehat{\underline{\mathbf{w}}}_\tau \widehat{\underline{s}}(\tau)\right] \widehat{\underline{s}}^*(\tau)\ , \qquad (7)$$

where $\widehat{\underline{s}}(\tau) = \widehat{\underline{\mathbf{w}}}_\tau^H \underline{\mathbf{y}}_\tau$ is a principal component estimate using the coefficient estimate $\widehat{\underline{\mathbf{w}}}_\tau$. Unstacking of the vectors finally yields a channel-wise representation of the algorithm, i.e.,

$$\widehat{\underline{w}}_{i,\tau+1} = \widehat{\underline{w}}_{i,\tau} + \mu \left[\underline{y}_{i,\tau} - \widehat{\underline{w}}_{i,\tau} \widehat{\underline{s}}(\tau)\right] \widehat{\underline{s}}^*(\tau)\ , \qquad (8)$$

which still uses $\widehat{\underline{s}}(\tau)$ as a common factor in each frequency bin. More interpretation follows in conjunction with BENCH in Sec. 2.3.

All frequency-domain coefficients $\underline{w}_i$ and their estimates $\widehat{\underline{w}}_{i,\tau}$ naturally can be translated into a time-domain representation. Simply define a DFT-domain column vector $\underline{\mathbf{w}}_i$ with elements $\underline{w}_i[m]$, $m = 1, ..., M$, corresponding to the individual frequency bins. Its counterpart in the time domain is then given by $\mathbf{w}_i = \mathbf{F}_M^{-1} \underline{\mathbf{w}}_i$ and analogously $\widehat{\mathbf{w}}_i = \mathbf{F}_M^{-1} \widehat{\underline{\mathbf{w}}}_i$. This transformation into the time-domain is useful, on the one hand, to arrive at a linear filtering interpretation of the frequency-domain PCA statement and, on the other hand, to enable a comparison – see Sec. 3 – of estimated filter impulse responses $\widehat{\underline{\mathbf{w}}}_\tau$ with the original acoustic channels $h_{i,k}$.

2.2. Power-Iteration for Adaptive Eigenvector Tracking

Another approach to solve the eigenvalue problem in (6) is known as the *power iteration* [24]. In the context here, it relies on a recursive estimate $\widehat{\mathbf{C}}_{y,\tau}$ of the PSD matrix $\mathbf{C}_y$, i.e.,

$$\widehat{\mathbf{C}}_{y,\tau} = \alpha\widehat{\mathbf{C}}_{y,\tau-1} + (1-\alpha)\underline{\mathbf{y}}_\tau\underline{\mathbf{y}}_\tau^H \tag{9}$$

in each frequency bin with a smoothing factor $0 < \alpha < 1$. The actual iteration for eigenvector estimation at time τ is started with an estimate $\underline{\widehat{\mathbf{w}}}_{\tau,old}$ and then uses the structure of (6) to update it as

$$\underline{\widehat{\mathbf{w}}}_{\tau,new} = \widehat{\mathbf{C}}_{y,\tau}\underline{\widehat{\mathbf{w}}}_{\tau,old} \;, \tag{10}$$

where the norm of $\underline{\widehat{\mathbf{w}}}_{\tau,new}$ is adjusted to unity after each iteration.

Fast convergence of the power iteration was reported for large eigenvalue spread of the data covariance, e.g., with a near rank-one matrix $\mathbf{C}_y$ in case of a strong direct acoustic path to the microphones [4]. In particular, it converges faster than the covariance recursion in (9). As a result, the iteration can be started with $\underline{\widehat{\mathbf{w}}}_{\tau,old} = \underline{\widehat{\mathbf{w}}}_{\tau-1,new}$ and one iteration per time instant τ was found to be sufficient [4]. While our statements here implied uncorrelated observation noise at the microphones, [4] also presented generalized eigenvector tracking with the help of an additional noise PSD matrix.

2.3. Model-Based ML-BENCH Adaptive Algorithm

The ML-BENCH algorithm [20] is supported by a generative model

$$\mathbf{y}_{i,\tau R} = \mathbf{Q}^T\mathbf{F}_M^{-1}\underline{\mathbf{H}}_i\underline{\mathbf{s}}_\tau + \mathbf{n}_{i,\tau R} \tag{11}$$

for our length R vectors $\mathbf{y}_{i,\tau R}$ of observed microphone signals. This frame-based model represents the strictly linear acoustic channels from Fig. 2 via overlap-save convolution [25]. To this end, a DFT-domain version $\underline{\mathbf{s}}_\tau = \mathbf{F}_M\mathbf{s}_{\tau R}$ of a length M source vector

$$\mathbf{s}_k = [\, s(k-M+1)\; s(k-M+2)\; \ldots\; s(k)\,]^T \tag{12}$$

is employed and linked here with diagonal matrices $\underline{\mathbf{H}}_i = \mathrm{diag}\{\underline{\mathbf{h}}_i\}$ that comprise the zero-padded acoustic channels in the DFT-domain, i.e., $\underline{\mathbf{h}}_i = \mathbf{F}_M\left[\mathbf{h}_i^T\; \mathbf{0}_{R\times 1}^T\right]^T$ and $\mathbf{h}_i = [h_{i,0}\; h_{i,1}\; \cdots\; h_{i,L-1}]^T$. The term $\mathbf{Q}^T\mathbf{F}_M^{-1}$ then achieves linearization of the cyclic convolution by projection into the time-domain and appropriate selection of samples. Neither $\underline{\mathbf{H}}_i$ nor $\underline{\mathbf{s}}_\tau$ are known *a priori*.

According to the EM algorithm [18, 26], the unknown acoustic channel can be modeled as a hidden state variable of the system and a *maximum-likelihood* (ML) estimation of the source signal can be achieved by utilizing the expectation of the log-likelihood of the complete data (observations + hidden variables), i.e.,

$$Q\left(\underline{\mathbf{s}}_\tau, \underline{\mathbf{s}}_{\tau,old}\right) = \mathrm{E}\{\ln\, p(\underline{\mathbf{y}}_\tau, \underline{\mathbf{h}}|\, \underline{\mathbf{s}}_\tau)\}_{q^{old}(\underline{\mathbf{h}})} \;, \tag{13}$$

where $\underline{\mathbf{h}}^T = [\underline{\mathbf{h}}_1^T\; \underline{\mathbf{h}}_2^T\; \cdots\; \underline{\mathbf{h}}_P^T]$ and $\underline{\mathbf{y}}_\tau^T = [\underline{\mathbf{y}}_{1,\tau}^T\; \underline{\mathbf{y}}_{2,\tau}^T\; \cdots\; \underline{\mathbf{y}}_{P,\tau}^T]$, for maximization in the M-step instead of the likelihood $p(\underline{\mathbf{y}}_\tau|\, \underline{\mathbf{s}}_\tau)$. In doing so, the expectation is w.r.t. the joint state posterior (all i)

$$q^{old}(\underline{\mathbf{h}}) = p(\underline{\mathbf{h}}|\, \underline{\mathbf{y}}_\tau, \underline{\mathbf{s}}_{\tau,old}) = \mathcal{N}(\underline{\widehat{\mathbf{w}}}, \underline{\mathbf{P}}) \tag{14}$$

that relies on a previous signal estimate $\underline{\mathbf{s}}_{\tau,old}$ and all current observations $\underline{\mathbf{y}}_\tau$. Assuming Gaussian observation noise in writing out the Q-function and by equating its complex derivative w.r.t. $\underline{\mathbf{s}}_\tau^*$ to zero, the multi-microphone ML estimate of the source signal is found in this *maximization-step* (M-step) as [20]

$$\underline{\widehat{\mathbf{s}}}_{\tau,new} = \left[\sum_i \left(\underline{\widehat{\mathbf{W}}}_i^H\underline{\widehat{\mathbf{W}}}_i + \underline{\mathbf{P}}_i\right)\right]^{-1}\sum_i \underline{\widehat{\mathbf{W}}}_i^H\,\underline{\mathbf{y}}_i \;, \tag{15}$$

where $\underline{\widehat{\mathbf{W}}}_i = \mathrm{diag}\,\{\underline{\widehat{\mathbf{w}}}_i\}$ and $\underline{\mathbf{P}}_i$ signify means and covariances of individual channel posteriors $p(\underline{\mathbf{h}}_i|\, \underline{\mathbf{y}}_{i,\tau}, \underline{\mathbf{s}}_{\tau,old})$ at time τ.

At this point, those acoustic-channel posteriors still have to be determined in the *expectation-step* (E-step). To this end, consider an equivalent representation of the observation model in (11), i.e.,

$$\mathbf{y}_{i,\tau R} = \mathbf{Q}^T\mathbf{F}_M^{-1}\underline{\mathbf{S}}_\tau\underline{\mathbf{h}}_i + \mathbf{n}_{i,\tau R} \;, \tag{16}$$

where $\underline{\mathbf{S}}_\tau = \mathrm{diag}\,\{\underline{\mathbf{s}}_\tau\}$. Then augment a first-order Markov model to describe a slow variability of the acoustic channels $\underline{\mathbf{h}}_i$, i.e.,

$$\underline{\mathbf{h}}_{i,\tau} = A\cdot\underline{\mathbf{h}}_{i,\tau-1} + \Delta\underline{\mathbf{h}}_{i,\tau} \;, \tag{17}$$

where $0 < A < 1$ denotes the state-transition coefficient and $\Delta\underline{\mathbf{h}}_{i,\tau}$ is a zero-mean and frame-wise uncorrelated Gaussian process noise vector with diagonal covariance $\underline{\mathbf{\Psi}}_i^\Delta = \mathrm{E}\left\{\Delta\underline{\mathbf{h}}_{i,\tau}\Delta\underline{\mathbf{h}}_{i,\tau}^H\right\}$. Based on the dynamical model in (16) and (17), the posterior $q^{old}(\underline{\mathbf{h}}_{i,\tau}) = p(\underline{\mathbf{h}}_{i,\tau}|\, \underline{\mathbf{y}}_{i,\tau}, \underline{\mathbf{s}}_{\tau,old})$ can be learned efficiently via the Kalman filter. In particular, we make use of a stable implementation in form of the state-space frequency-domain adaptive filter (SSFDAF) [27, 28] to denote our E-step for the i-th channel:

$$\underline{\widehat{\mathbf{w}}}_{i,\tau-1}^+ = A\cdot\underline{\widehat{\mathbf{w}}}_{i,\tau-1} \tag{18}$$

$$\underline{\mathbf{P}}_{i,\tau-1}^+ = A^2\cdot\underline{\mathbf{P}}_{i,\tau-1} + \underline{\mathbf{\Psi}}_i^\Delta \tag{19}$$

$$\underline{\boldsymbol{\mu}}_{i,\tau} = \underline{\mathbf{P}}_{i,\tau-1}^+\left[\underline{\widehat{\mathbf{S}}}_{\tau,old}\underline{\mathbf{P}}_{i,\tau-1}^+\underline{\widehat{\mathbf{S}}}_{\tau,old}^H + \frac{M}{R}\underline{\mathbf{\Psi}}_i^{\mathbf{n}}\right]^{-1} \tag{20}$$

$$\underline{\mathbf{e}}_{i,\tau} = \underline{\mathbf{y}}_{i,\tau} - \mathbf{F}_M\mathbf{Q}\mathbf{Q}^T\mathbf{F}_M^{-1}\,\underline{\widehat{\mathbf{S}}}_{\tau,old}\,\underline{\widehat{\mathbf{w}}}_{i,\tau-1}^+ \tag{21}$$

$$\underline{\widehat{\mathbf{w}}}_{i,\tau} = \underline{\widehat{\mathbf{w}}}_{i,\tau-1}^+ + \underline{\boldsymbol{\mu}}_{i,\tau}\underline{\widehat{\mathbf{S}}}_{\tau,old}^H\,\underline{\mathbf{e}}_{i,\tau} \tag{22}$$

$$\underline{\mathbf{P}}_{i,\tau} = \left[\mathbf{I}_M - \frac{R}{M}\underline{\boldsymbol{\mu}}_{i,\tau}\underline{\widehat{\mathbf{S}}}_{\tau,old}^H\,\underline{\widehat{\mathbf{S}}}_{\tau,old}\right]\underline{\mathbf{P}}_{i,\tau-1}^+ \;, \tag{23}$$

where $\underline{\boldsymbol{\mu}}_{i,\tau}$, $\underline{\mathbf{e}}_{i,\tau}$, and $\underline{\mathbf{P}}_{i,\tau}$ are the time-varying Kalman stepsize, the error signal, and the state-error covariance, respectively. The superscript $^+$ denotes prediction terms and $\underline{\mathbf{\Psi}}_i^{\mathbf{n}} = \mathrm{E}\left\{\underline{\mathbf{n}}_{i,\tau}\underline{\mathbf{n}}_{i,\tau}^H\right\}$ is the diagonal covariance of zero-mean and normally distributed observation noise in the DFT domain.

In order to start the EM iteration of the channel-wise Kalman filters (18)-(23) and the multi-channel ML estimator (15), simply let $\underline{\widehat{\mathbf{S}}}_{\tau,old} = \mathrm{diag}\{\underline{\widehat{\mathbf{s}}}_{\tau,old}\} = \mathrm{diag}\{\underline{\widehat{\mathbf{s}}}_{\tau-1,new}\}$, i.e., utilize the estimated source signal obtained from (15) at the previous time instant $\tau-1$. According to our observations, it is then sufficient to apply just one iteration of E- and M-step per time τ.

Looking back at Fig. 1, ML-BENCH can be seen as a native blind signal enhancer, centered in (15) and receiving acoustic channel estimation support via (18)-(23), but also as a blind recursive Bayesian channel estimator, centered in the Kalman filter recursions (18)-(23) and relying on source signal estimation via (15). The usually difficult blind equalization and identification tasks are thus solved iteratively (i.e., jointly) to the advantage of each other.

Regarding the relationship with PCA, it can be seen that the equalizer in (15) implies the PCA statement as shown by (3) and used in (7). The additional inverse part in (15) represents a frequency-dependent normalization (or single-channel correction) of the multi-channel PCA part. Furthermore, the channel update in (22) resembles the basic online PCA recursion in (7) with additional optimum stepsize $\underline{\boldsymbol{\mu}}_{i,\tau}$ and with overlap-save constraining in the calculation of the error signal $\underline{\mathbf{e}}_{i,\tau}$. Still, the model-based inference via ML-BENCH differs significantly from PCA, i.e., ML-BENCH rather falls into the class of factor analysis techniques, cf., e.g., [17].

3. EXPERIMENTAL RESULTS

3.1. Normalized Filter-Projection Misalignment (NFPM)

The NFPM is a multichannel Euclidean distance between estimated impulse responses $\widehat{\mathbf{w}}^T = [\,\widehat{\mathbf{w}}_1^T\ \widehat{\mathbf{w}}_2^T\ \ldots\ \widehat{\mathbf{w}}_P^T\,]$ and true impulse response coefficients in $\mathbf{H}^T = [\,\mathbf{H}_1^T\ \mathbf{H}_2^T\ \ldots\ \mathbf{H}_P^T\,]$, where

$$\mathbf{H}_i = \begin{bmatrix} h_{i,0} & 0 & \cdots & \cdots & \cdots & 0 \\ h_{i,1} & h_{i,0} & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & h_{i,L-1} & h_{i,L-2} \\ 0 & \cdots & \cdots & \cdots & 0 & h_{i,L-1} \end{bmatrix}. \tag{24}$$

is a linear convolution matrix. The NFPM [13] in particular applies the same (i.e., common) correction filter $\mathbf{f} = [\,f_0\ f_1\ \ldots\ f_{2D_f}\,]^T$ to all channels to effectively minimize the distance between $\tilde{\mathbf{h}} = \mathbf{H}\mathbf{f}$ and a zero-padded estimate $\widehat{\mathbf{w}}_z^T = [\,\widehat{\mathbf{w}}_{1,z}^T\ \widehat{\mathbf{w}}_{2,z}^T\ \ldots\ \widehat{\mathbf{w}}_{P,z}^T\,]$, e.g.,

$$\widehat{\mathbf{w}}_{i,z} = \big[\underbrace{0\ \ldots\ 0}_{D_f}\ \widehat{\mathbf{w}}_i^T\ \underbrace{0\ \ldots\ 0}_{D_f}\big]^T. \tag{25}$$

The least-squares NFPM solution over all possible correction filters $\mathbf{f}$ is then given by [13]

$$\mathrm{NFPM}(\mathbf{h}, \widehat{\mathbf{w}}) = \min_{\mathbf{f}} \frac{\left\|\widehat{\mathbf{w}}_z - \tilde{\mathbf{h}}\right\|^2}{\left\|\widehat{\mathbf{w}}_z\right\|^2} = \frac{\left\|\widehat{\mathbf{w}}_z - \mathbf{H}\,(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\widehat{\mathbf{w}}_z\right\|^2}{\left\|\widehat{\mathbf{w}}_z\right\|^2}.$$

It evaluates how accurate P channel impulse responses are identified by an algorithm – up to a common filter error $\mathbf{f}$, which is absorbed by the measure – or, in other words, how accurate relative impulse responses are found. NFPM was developed and well-justified in the context of explicit CRE-based BCI. It is used here intuitively to evaluate implicit BCI of the described signal enhancement techniques.

3.2. Experimental Configuration

In our simulations, we use room impulse responses that were generated with the image method [29] for a single source and a linear array with $P = 10$ microphones inside a room with dimensions 7 m $\times$ 5 m $\times$ 4 m ($x \times y \times z$) and a reverberation time $T_{60} = 0.2$ s. The source was positioned at (5 m, 1.5 m, 1.5 m), the first array microphone was located at (2 m, 4 m, 1.5 m), whereas all other microphones were placed at distances of 0.1 m in positive x-direction from the first microphone. In order to simulate realistic conditions, we selected a long filter length of 3200 coefficients at a sampling rate of $f_s = 16$ kHz. The microphone signals were then obtained by convolving sentences from 10 different speakers with each impulse response before adding zero-mean white observation noise at different signal-to-noise ratios (SNR).

3.3. Results

Figures 3 to 5 depict the NFPM corresponding to the described algorithms as a function of time. The NFPM calculation uses time-domain coefficients $\widehat{\mathbf{w}}_{i,\tau} = \mathbf{F}_M^{-1}\underline{\widehat{\mathbf{w}}}_{i,\tau}$ for comparison with the acoustic impulse responses $\mathbf{h}_i = [h_{i,0}\ h_{i,1}\ \cdots\ h_{i,L-1}]^T$. Oja's gradient descent rule obviously converges slowly to the common-filter compensated ground-truth $\tilde{\mathbf{h}} = \mathbf{H}\mathbf{f}$, which is plausible for this LMS-type algorithm. ML-BENCH (A=0.9997) and power iteration ($\alpha = 0.95$) converge much faster, with advantages for the power iteration at low SNR=10 dB, for ML-BENCH at high SNR=50 dB, and with a tie at moderate SNR= 30 dB. The common saturation of NFPM is due to finite length $D_f = L$ of the FIR correction $\mathbf{f}$ [13]. In an absolute sense, the fast algorithms converge within a few seconds despite the long AIRs used here. This result is very uncommon, since BCI has been applied for the identification of much shorter impulse responses in most of the previous work, e.g., [7, 9].

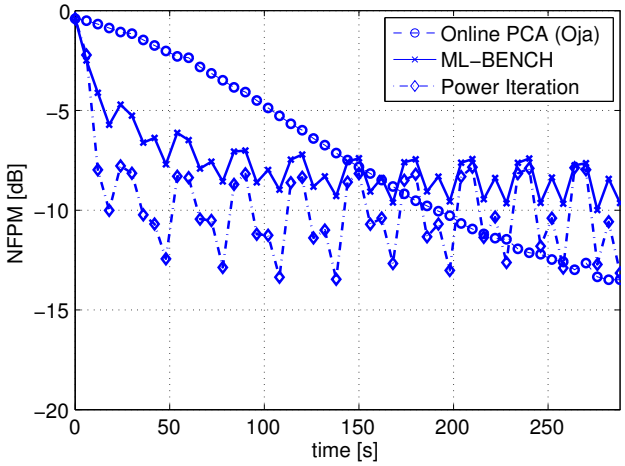

Fig. 3. NFPM comparison of BCI performance at SNR=10 dB.

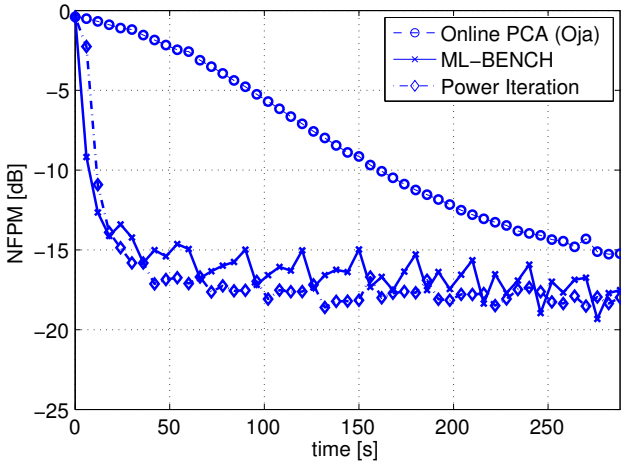

Fig. 4. NFPM comparison of BCI performance at SNR=30 dB.

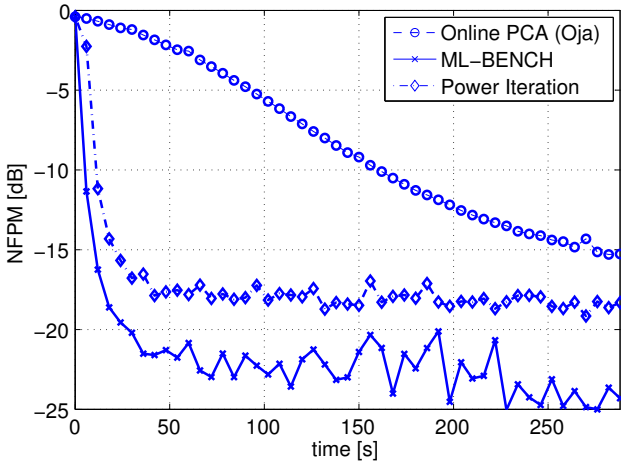

Fig. 5. NFPM comparison of BCI performance at SNR=50 dB.

4. CONCLUSIONS

It was demonstrated that blind signal enhancement techniques can successfully perform BCI in realistic acoustic conditions, up to the truly ill-conditioned part of the problem. Interestingly, the performances of the fast PCA version, i.e., using the power iteration, and the BENCH algorithm are very similar despite their structural differences. The power iteration processes the pair-wise dependencies of all microphones via the PSD matrix. Its formal simplicity seems to be attractive, but the operations related to all frequency bins will accumulate computational load. Implementation of the BENCH algorithm will exhibit higher code complexity due to its richer mathematical structure. However, its inherent source signal estimate practically enables a channel-wise and quasi-supervised system identification with possible advantages in terms of computational load.

5. REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, February 1988.

[2] D. Schmid and G. Enzner, "A parametric least-squares approximation for multichannel equalization of room acoustics," in *Proc. of Intl. Workshop on Acoustic Echo and Noise Control*, Tel-Aviv, September 2010.

[3] 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, August 2001.

[4] A. Krueger, E. Warsitz, and R. Haeb-Umbach, "Speech enhancement with a GSC-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 1, pp. 206 –219, Jan. 2011.

[5] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Trans. Signal Process.*, vol. 43, no. 12, pp. 2982–2993, Dec. 1995.

[6] E. Moulines, P. Duhamel, J.-F. Cardoso, and S. Mayrargue, "Subspace methods for the blind identification of multichannel FIR filters," *IEEE Trans. Signal Process.*, vol. 43, no. 2, pp. 516–525, Feb. 1995.

[7] S. Gannot and M. Moonen, "Subspace methods for multimicrophone speech dereverberation," *EURASIP J. Appl. Signal Process.*, vol. 2003, no. 11, pp. 1074–1090, 2003.

[8] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.

[9] Y. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Proc.*, vol. 82, no. 8, pp. 1127–1138, Aug. 2002.

[10] R. Ahmad, A. W. H. Khong, M. K. Hasan, and P. A. Naylor, "An extended normalized multichannel FLMS algorithm for blind channel identification," in *Proc. European Signal Process. Conf.*, Florence, Italy, Sept. 2006.

[11] N. D. Gaubitch, M. K. Hasan, and P. A. Naylor, "Noise robust adaptive blind channel identification using spectral constraints," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, Toulouse, France, May 2006, pp. 93–96.

[12] N. D. Gaubitch, J. Benesty, and P. A. Naylor, "Adaptive common root estimation and the common zeros problem in blind channel identification," in *Proc. European Signal Process. Conf.*, Antalya, Turkey, Sept. 2005.

[13] D. Schmid and G. Enzner, "Cross-relation-based blind SIMO identifiability in the presence of near-common zeros and noise," *IEEE Trans. Signal Process.*, vol. 60, no. 1, pp. 60–72, Jan. 2012.

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

[15] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[16] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 5, pp. 1529 –1539, July 2007.

[17] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, New York, USA, 2001.

[18] T. Moon, "The EM algorithm in signal processing," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, 1996.

[19] T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying gaussian source model," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 16, no. 8, pp. 1512–1527, Nov. 2008.

[20] D. Schmid, S. Malik, and G. Enzner, "An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain," in *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, March 2012.

[21] D. Schmid, S. Malik, and G. Enzner, "A maximum a posteriori approach to multichannel speech dereverberation and denoising," in *Proc. of Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, September 2012.

[22] E. Oja, "Simplified neuron model as a principal component analyzer," *Journal of Mathematical Biology*, vol. 15, no. 3, pp. 267–273, Nov. 1982.

[23] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Process.*, vol. 41, no. 1, pp. 95–107, January 1995.

[24] J. Karhunen, "Adaptive algorithms for estimating eigenvectors of correlation type matrices," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, Prague, Czech Republic, Mar. 1984, vol. 9, pp. 592–595.

[25] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, Prentice-Hall, Upper Saddle River, New Jersey, 1996.

[26] D. Tzikas, A. Likas, and N. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 25, no. 6, pp. 131–146, Nov. 2008.

[27] G. Enzner and P. Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing, Elsevier*, vol. 86, no. 6, pp. 1140–1156, June 2006.

[28] S. Malik and G. Enzner, "Online maximum-likelihood learning of time-varying dynamical models in block-frequency-domain," in *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, March 2010.

[29] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.