JOINT ACOUSTIC ECHO CANCELLATION AND BLIND SOURCE EXTRACTION BASED ON INDEPENDENT VECTOR EXTRACTION

Thomas Haubner and Walter Kellermann

Multimedia Communications and Signal Processing,
Friedrich-Alexander-Universität Erlangen-Nürnberg,
Cauerstr. 7, D-91058 Erlangen

Zbyněk Koldovský

Acoustic Signal Analysis and Processing Group,
Technical University of Liberec,
Studentská 2, CZ-46117 Liberec

ABSTRACT

We describe a joint acoustic echo cancellation (AEC) and blind source extraction (BSE) approach for multi-microphone acoustic frontends. The proposed algorithm blindly estimates AEC and beamforming filters by maximizing the statistical independence of a non-Gaussian source of interest and a stationary Gaussian background modeling interfering signals and residual echo. Double talk-robust and fast-converging parameter updates are derived from a global maximum-likelihood objective function resulting in a computationally efficient Newton-type update rule. Evaluation with simulated acoustic data confirms the benefit of the proposed joint AEC and beamforming filter estimation in comparison to updating both filters individually.

***Index Terms*—** Acoustic Echo Cancellation, Blind Source Extraction, Independent Vector Extraction, Step-Size Control

1. INTRODUCTION

Acoustic echo and interference suppression is a crucial part of any modern hands-free speech communication device [1]. It has been tackled by a variety of approaches ranging from traditional adaptive linear filters to sophisticated deep learning-based spectral post filters and beamformers [1–6]. In particular, the combination of acoustic echo cancellation (AEC) and beamforming algorithms have proven to be a powerful approach due to its simultaneous exploitation of spectral and spatial signal characteristics [2, 7]. Yet, sophisticated parameter estimation schemes are required to cope with time-varying echo paths, high-level double-talk and interference that many devices are exposed to [8]. This problem has been addressed by Kalman filter-based inference of the model parameters [4, 9, 10] and machine-learning supported variants [11–13]. Besides Kalman filter-based approaches, also *blind* algorithms originating from independent component analysis (ICA) (see, e.g., [14]) have shown promising results for interference-robust adaptation control [15–18]. They are particularly interesting as they require no prior training and thus are very robust w.r.t. training-testing mismatches which any machine learning-based approach suffers from. So far, semi-supervised blind source separation (BSS) has been mainly used for double-talk robust adaptation control of linear and nonlinear acoustic echo cancellers using various statistical source models and update rules [15–19]. Recently, also independent vector analysis (IVA)-based joint estimation of AEC filters and beamforming filters has been investigated [20]. IVA-based approaches are promising as they have shown competitive multichannel speech enhancement performance when compared to sophisticated deep neural network (DNN) algorithms [21]. Yet, [20] assumes that the number of microphones M is equal to the number of statistically independent point sources. This assumption requires to first estimate M beamforming vectors and subsequently select the desired source of interest (SOI), which becomes computationally complex for large M.

Here, we propose a joint AEC and blind source extraction (BSE) approach which estimates only a single desired SOI out of the observed mixture. For this, we model the frequency-domain SOI as a vector source following a multivariate non-Gaussian source probability density function (PDF) which is independent from the interfering signals [22]. The interference is modeled by a stationary and spatially-correlated circular complex Gaussian [22]. Instead of aiming at a complete separation of the SOI, the loudspeakers signals, and the interferers, we allow the interference estimates to remain mutually mixed [22]. This significantly reduces the number of model parameters to be estimated for large numbers of microphones. The parameters are estimated by a fast-converging Newton-type update which can be interpreted as a joint SOI- and interference-aware batch-normalized least mean squares (BNLMS) echo canceller with a fast independent vector extraction (IVE)-controlled beamformer.

We represent vectors by bold lower-case letters and matrices by bold upper-case letters with $[\cdot]_m$ denoting the mth element of a vector. The M-dimensional identity matrix and the $M \times N$-dimensional zero matrix are denoted by $\boldsymbol{I}_M$ and $\mathbf{0}_{M\times N}$, respectively. Finally, we indicate sampling from a PDF $p(\boldsymbol{z})$ by $\boldsymbol{z} \sim p(\boldsymbol{z})$, equivalency up to a constant by $\stackrel{\mathrm{c}}{=}$, and averaging over time by $\widehat{\mathbb{E}}$.

2. SIGNAL MODEL

We consider a multi-microphone acoustic echo and interference control scenario comprising a single SOI, e.g., a desired near-end

This work was partly supported by The Czech Science Foundation through Project No. 20-17720S.

speaker, a single-channel far-end loudspeaker signal, and various point interferers. The M-channel short-time Fourier transform (STFT)-domain microphone signal $\boldsymbol{x}_{f,\tau}$ at frequency index f and frame index τ is modeled as a linear superposition of the SOI image $\boldsymbol{s}_{f,\tau}$, the echo image $\boldsymbol{d}_{f,\tau}$, and the interference image $\boldsymbol{n}_{f,\tau}$ as follows:

$$\boldsymbol{x}_{f,\tau} = \boldsymbol{s}_{f,\tau} + \boldsymbol{d}_{f,\tau} + \boldsymbol{n}_{f,\tau} \in \mathbb{C}^M. \tag{1}$$

We assume a linear time-invariant multiplicative narrowband transfer function model to describe the acoustic propagation from the SOI and the loudspeaker to the microphones

$$\boldsymbol{s}_{f,\tau} = \boldsymbol{a}_{\text{soi},f}\, s_{f,\tau} \quad \text{and} \quad \boldsymbol{d}_{f,\tau} = \boldsymbol{h}_f\, u_{f,\tau}, \tag{2}$$

with the acoustic transfer functions (ATFs) $\boldsymbol{a}_{\text{soi},f}$ and $\boldsymbol{h}_f$ and the scalar desired source signal $s_{f,\tau}$ and the scalar far-end loudspeaker signal $u_{f,\tau}$. Furthermore, we assume the interfering signal image $\boldsymbol{n}_{f,\tau}$ to be composed of $M-1$ point sources $[\boldsymbol{q}_{f,\tau}]_1$, $\dots$, $[\boldsymbol{q}_{f,\tau}]_{M-1}$, whose acoustic propagation to the microphones is modeled by

$$\boldsymbol{n}_{f,\tau} = \boldsymbol{A}_{bg,f}\, \boldsymbol{q}_{f,\tau} \tag{3}$$

with the mixing matrix $\boldsymbol{A}_{bg,f} \in \mathbb{C}^{M \times M-1}$. Note that this assumption expresses a determined mixing model, i.e., an equal number of observations and latent point sources, which allows for a detailed model analysis in Sec. 3.1.

3. JOINT ACOUSTIC ECHO CANCELLATION AND BLIND SOURCE EXTRACTION

We will now describe the proposed joint AEC and BSE algorithm by first introducing the proposed demixing model and then the blind estimation of its coefficients by maximizing the mutual statistical independence of the SOI estimate, the interference estimate and the loudspeaker signal.

3.1. Demixing Model

To estimate the SOI $s_{f,\tau}$, we use the semi-supervised demixing model [15]

$$\begin{pmatrix} \hat{s}_{f,\tau} \\ \hat{\boldsymbol{z}}_{f,\tau} \\ u_{f,\tau} \end{pmatrix} = \boldsymbol{W}_f \begin{pmatrix} \boldsymbol{x}_{f,\tau} \\ u_{f,\tau} \end{pmatrix} \tag{4}$$

with the SOI estimate $\hat{s}_{f,\tau}$, the background estimate $\hat{\boldsymbol{z}}_{f,\tau}$, modelling interference and residual echo, and the demixing matrix

$$\boldsymbol{W}_f = \begin{pmatrix} \boldsymbol{w}_{\text{bse},f}^{\text{H}} & w_{\text{aec},f}^* \\ \boldsymbol{B}_f & \boldsymbol{h}_{\text{bg},f} \\ \boldsymbol{0}_{1\times M} & 1 \end{pmatrix} \in \mathbb{C}^{M+1 \times M+1} \tag{5}$$

which is composed of the parameters $\boldsymbol{w}_{\text{bse},f} \in \mathbb{C}^M$, $w_{\text{aec},f} \in \mathbb{C}$, $\boldsymbol{h}_{\text{bg},f} \in \mathbb{C}^{M-1}$ and $\boldsymbol{B}_f \in \mathbb{C}^{M-1\times M}$. We analyze the signal estimation problem by computing the overall transmission model

$$\begin{pmatrix} \hat{s}_{f,\tau} \\ \hat{\boldsymbol{z}}_{f,\tau} \\ u_{f,\tau} \end{pmatrix} = \boldsymbol{V}_f \begin{pmatrix} s_{f,\tau} \\ \boldsymbol{q}_{f,\tau} \\ u_{f,\tau} \end{pmatrix} \tag{6}$$

between the source signals $s_{f,\tau}$, $\boldsymbol{q}_{f,\tau}$ and $u_{f,\tau}$ and the signal estimates $\hat{s}_{f,\tau}$, $\hat{\boldsymbol{z}}_{f,\tau}$ and $u_{f,\tau}$ with the transmission matrix

$$\boldsymbol{V}_f = \begin{pmatrix} \boldsymbol{w}_{\text{bse},f}^{\text{H}}\, \boldsymbol{a}_{\text{soi},f} & \boldsymbol{w}_{\text{bse},f}^{\text{H}}\, \boldsymbol{A}_{bg,f} & \boldsymbol{w}_{\text{bse},f}^{\text{H}}\, \boldsymbol{h}_f + w_{\text{aec},f}^* \\ \boldsymbol{B}_f\, \boldsymbol{a}_{\text{soi},f} & \boldsymbol{B}_f\, \boldsymbol{A}_{bg,f} & \boldsymbol{B}_f\, \boldsymbol{h}_f + \boldsymbol{h}_{\text{bg},f} \\ 0 & \boldsymbol{0}_{1\times M-1} & 1 \end{pmatrix}. \tag{7}$$

As we aim in BSE at the separation of the SOI estimate $\hat{s}_{f,\tau}$ from the multivariate background estimate $\hat{\boldsymbol{z}}_{f,\tau}$, with potentially still statistically dependent components, and the loudspeaker signal $u_{f,\tau}$, we conclude that the optimum demixing matrix (5) must enforce a block-diagonal transmission matrix $\boldsymbol{V}_f$. Note that a complete diagonalization of the transmission matrix is not necessary as we do not require the background estimates $\hat{\boldsymbol{z}}_{f,\tau}$ to be equal to the interfering source signals $\boldsymbol{q}_{f,\tau}$. The block-diagonality assumption allows to couple the different components of the demixing matrix $\boldsymbol{W}_f$ and the mixing model parameters $\boldsymbol{a}_{\text{soi},f}$, $\boldsymbol{A}_{bg,f}$ and $\boldsymbol{h}_f$ (cf. Sec. 2). In particular, we require that

$$\boldsymbol{w}_{\text{bse},f}^{\text{H}}\, \boldsymbol{a}_{\text{soi},f} = 1 \tag{8}$$

which is well-known as distortionless response constraint by identifying $\boldsymbol{w}_{\text{bse},f}$ as beamforming vector [23]. Furthermore, the SOI should not be contained in the background estimate $\hat{\boldsymbol{z}}_{f,\tau}$ which is equivalent to

$$\boldsymbol{B}_f\, \boldsymbol{a}_{\text{soi},f} = \boldsymbol{0}_{M-1\times 1} \tag{9}$$

with $\boldsymbol{B}_f$ being identified as blocking matrix. A straightforward choice for $\boldsymbol{B}_f$ is given by [22]

$$\boldsymbol{B}_f = \begin{pmatrix} \boldsymbol{g}_f & -\gamma_f\, \boldsymbol{I}_{M-1} \end{pmatrix} \tag{10}$$

which allows to parametrize the blocking matrix $\boldsymbol{B}_f$ entirely in terms of the SOI ATF $\boldsymbol{a}_{\text{soi},f} = \begin{pmatrix} \gamma_f & \boldsymbol{g}_f^{\text{T}} \end{pmatrix}^{\text{T}}$. Finally, to ensure that the echo signal is neither contained in the SOI estimate $\hat{s}_{f,\tau}$ nor in the background estimate $\hat{\boldsymbol{z}}_{f,\tau}$, we must ensure that

$$w_{\text{aec},f}^* = -\, \boldsymbol{w}_{\text{bse},f}^{\text{H}}\, \boldsymbol{h}_f \tag{11}$$

$$\boldsymbol{h}_{\text{bg},f} = -\, \boldsymbol{B}_f\, \boldsymbol{h}_f\,. \tag{12}$$

Note that (11) readily identifies $-w_{\text{aec},f}^*$ as concatenated system of the echo ATF $\boldsymbol{h}_f$ and the beamforming vector $\boldsymbol{w}_{\text{bse},f}$. By replacing $w_{\text{aec},f}$, $\boldsymbol{h}_{\text{bg},f}$ and $\boldsymbol{B}_f$ (cf. Eqs. (10)-(12)) in the upper part of the demixing model (4), we obtain

$$\begin{pmatrix} \hat{s}_{f,\tau} \\ \hat{\boldsymbol{z}}_{f,\tau} \end{pmatrix} = \begin{pmatrix} \boldsymbol{w}_{\text{bse},f}^{\text{H}} \\ \begin{pmatrix} \boldsymbol{g}_f & -\gamma_f\, \boldsymbol{I}_{M-1} \end{pmatrix} \end{pmatrix} \left(\boldsymbol{x}_{f,\tau} - \boldsymbol{h}_f\, u_{f,\tau} \right) \tag{13}$$

with the signal estimator being solely parametrized by $\boldsymbol{w}_{\text{bse},f}$, $\boldsymbol{a}_{\text{soi},f} = \begin{pmatrix} \gamma_f & \boldsymbol{g}_f^{\text{T}} \end{pmatrix}^{\text{T}}$ and $\boldsymbol{h}_f$. We conclude that the considered constraints lead naturally to the traditional approach of first subtracting the echo estimate $\hat{\boldsymbol{d}}_{f\tau} = \boldsymbol{h}_f\, u_{f,\tau}$ from the microphone signal $\boldsymbol{x}_{f,\tau}$ and subsequently applying a beamformer [2]. In the next sections, we will show how to jointly estimate the unknown parameters $\boldsymbol{w}_{\text{bse},f}$, $\boldsymbol{a}_{\text{soi},f}$ and $\boldsymbol{h}_f$.

3.2. Cost Function

Adopting the idea of independent vector extraction (IVE) [22], we suggest to estimate the parameter vector $\boldsymbol{\theta}$, containing $\boldsymbol{w}_{\text{bse},f}$, $\boldsymbol{a}_{\text{soi},f}$ and $\boldsymbol{h}_f$, by maximizing the statistical independence of the broadband SOI and background signal estimates $\underline{\hat{\boldsymbol{s}}}_\tau^{\text{T}} = \begin{pmatrix}\hat{s}_{1,\tau} & \dots & \hat{s}_{F,\tau}\end{pmatrix}$ and $\underline{\hat{\boldsymbol{z}}}_\tau^{\text{T}} = \begin{pmatrix}\hat{\boldsymbol{z}}_{1,\tau}^{\text{T}} & \dots & \hat{\boldsymbol{z}}_{F,\tau}^{\text{T}}\end{pmatrix}$. The according log-likelihood function is given by [22]

$$\mathcal{L}(\boldsymbol{\theta}) \stackrel{\text{c}}{=} \sum_{\tau=1}^{T} \left(\log p(\underline{\hat{\boldsymbol{s}}}_\tau) + \log p(\underline{\hat{\boldsymbol{z}}}_\tau)\right) + 2T \sum_{f=1}^{F} \log \left|\det \boldsymbol{W}_f\right| \tag{14}$$

with the number of frames T, the number of frequency bins F, and the true signal models $\underline{\boldsymbol{s}}_\tau \sim p(\underline{\boldsymbol{s}}_\tau)$ and $\underline{\boldsymbol{z}}_\tau \sim p(\underline{\boldsymbol{z}}_\tau)$, respectively. Dividing (14) by $-T$ and assuming the background signal $\underline{\boldsymbol{z}}_\tau$ to follow a stationary zero-mean circular complex Gaussian PDF with mutually uncorrelated frequency bins, i.e., $p(\underline{\boldsymbol{z}}_\tau) = \prod_{f=1}^{F} \mathcal{N}_c(\boldsymbol{0}_{M-1\times 1}, \boldsymbol{C}_{zz,f})$, we obtain the cost function:

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}) \stackrel{\text{c}}{=} \widehat{\mathbb{E}}\left[-\log p(\underline{\hat{\boldsymbol{s}}}_\tau) + \sum_{f=1}^{F} \boldsymbol{e}_{f,\tau}^{\text{H}} \boldsymbol{R}_f \, \boldsymbol{e}_{f,\tau}\right] \\ - (M-2) \sum_{f=1}^{F} \log \left|\gamma_f\right|^2\end{aligned} \tag{15}$$

with the AEC error signal $\boldsymbol{e}_{f,\tau} = \boldsymbol{x}_{f,\tau} - \boldsymbol{h}_f \, u_{f,\tau}$, the first component of the SOI ATF $\gamma_f = \left[\boldsymbol{a}_{\text{soi},f}\right]_1$ and $\boldsymbol{R}_f = \boldsymbol{B}_f^{\text{H}} \boldsymbol{C}_{zz,f}^{-1} \boldsymbol{B}_f$ [22]. It has been observed that $\boldsymbol{w}_{\text{bse},f}$ and $\boldsymbol{a}_{\text{soi},f}$ are often not sufficiently strongly coupled by the cost function (15) [22]. This shortcoming is addressed by the orthogonality constraint (OC) $\widehat{\mathbb{E}}\left[\hat{\boldsymbol{z}}_{f,\tau} \, \hat{s}_{f,\tau}^*\right] = \boldsymbol{0}_{M-1\times 1}$ [24], which leads to the relation [22]

$$\boldsymbol{a}_{\text{soi},f} = \begin{pmatrix}\gamma_f \\ \boldsymbol{g}_f\end{pmatrix} = \frac{\widehat{\boldsymbol{C}}_{ee,f} \, \boldsymbol{w}_{\text{bse},f}}{\boldsymbol{w}_{\text{bse},f}^{\text{H}} \widehat{\boldsymbol{C}}_{ee,f} \, \boldsymbol{w}_{\text{bse},f}} \tag{16}$$

between the SOI ATF $\boldsymbol{a}_{\text{soi},f}$ and the beamforming vector $\boldsymbol{w}_{\text{bse},f}$ with $\widehat{\boldsymbol{C}}_{ee,f} = \widehat{\mathbb{E}}\left[\boldsymbol{e}_{f,\tau} \, \boldsymbol{e}_{f,\tau}^{\text{H}}\right]$ being the sample error covariance matrix. Note that analogously to [22], the OC (16) is equivalent to computing $\boldsymbol{w}_{\text{bse},f}$ as the minimizer of $\boldsymbol{w}_{\text{bse},f}^{\text{H}} \widehat{\boldsymbol{C}}_{ee,f} \, \boldsymbol{w}_{\text{bse},f}$ subject to $\boldsymbol{w}_{\text{bse},f}^{\text{H}} \, \boldsymbol{a}_{\text{soi},f} = 1$, i.e., a minimum *error* power beamformer which is distortionless w.r.t. $\boldsymbol{a}_{\text{soi},f}$ [25]. By enforcing the OC (16), the demixing matrix $\boldsymbol{W}_f$ is entirely parametrized by $\boldsymbol{w}_{\text{bse},f}$ and $\boldsymbol{h}_f$ which need to be estimated by minimizing (15).

3.3. Parameter Update

For rapid convergence and optimum steady-state performance, we derive a Newton-type update (see, e.g., [26]) with a block-diagonal Hessian, i.e., $\frac{\partial^2 \mathcal{J}(\boldsymbol{\theta})}{\partial \boldsymbol{w}_{\text{bse},f} \partial \boldsymbol{h}_f^{\text{H}}} = \frac{\partial^2 \mathcal{J}(\boldsymbol{\theta})}{\partial \boldsymbol{w}_{\text{bse},f} \partial \boldsymbol{h}_f^{\text{T}}} = \boldsymbol{0}_{M\times M}$, to estimate $\boldsymbol{w}_{\text{bse},f}$ and $\boldsymbol{h}_f$. The first-order partial derivatives are given by

$$\frac{\partial \mathcal{J}(\boldsymbol{\theta})}{\partial \boldsymbol{h}_f^*} = -\widehat{\mathbb{E}}\left[\left(\phi_f^*(\underline{\hat{\boldsymbol{s}}}_\tau) \, \boldsymbol{w}_{\text{bse},f} + \boldsymbol{R}_f \, \boldsymbol{e}_{f,\tau}\right) u_{f,\tau}^*\right] \tag{17}$$

$$\frac{\partial \mathcal{J}(\boldsymbol{\theta})}{\partial \boldsymbol{w}_{\text{bse},f}^*} = \widehat{\mathbb{E}}\left[\boldsymbol{e}_{f,\tau} \, \phi_f(\underline{\hat{\boldsymbol{s}}}_\tau)\right] - \boldsymbol{a}_{\text{soi},f} \tag{18}$$

with the score function $\phi_f(\underline{\hat{\boldsymbol{s}}}_\tau) = -\frac{\partial \log p(\underline{\hat{\boldsymbol{s}}}_\tau)}{\partial \hat{s}_{f,\tau}}$ and the SOI ATF $\boldsymbol{a}_{\text{soi},f}$ given by (16). We conclude that (18) corresponds to the BSE filter gradient computed in [22, 27] when exchanging the AEC error signal $\boldsymbol{e}_{f,\tau}$ by the microphone signal $\boldsymbol{x}_{f,\tau}$ (see also sequential demixing model (13)). In [22, 27] it is shown that the score function must fulfill $\mathbb{E}\left[s_{f,\tau} \, \phi_f(\underline{\boldsymbol{s}}_{f,\tau})\right] = 1$ to obtain a valid update, which can be ensured by normalizing $\phi_f(\underline{\hat{\boldsymbol{s}}}_\tau)$ in Eqs. (17) and (18) with $\hat{\nu}_f = \widehat{\mathbb{E}}\left[\hat{s}_{f,\tau} \, \phi_f(\underline{\hat{\boldsymbol{s}}}_\tau)\right]$. By following analogous steps as in [27] we obtain the Newton-type BSE filter update

$$\boldsymbol{w}_{\text{bse},f} \leftarrow \boldsymbol{w}_{\text{bse},f} + \frac{\hat{\nu}_f^*}{\hat{\rho}_f^* - \hat{\nu}_f^*} \widehat{\boldsymbol{C}}_{ee,f}^{-1} \left(\widehat{\mathbb{E}}\left[\boldsymbol{e}_{f,\tau} \, \frac{\phi_f(\underline{\hat{\boldsymbol{s}}}_\tau)}{\hat{\nu}_f}\right] - \boldsymbol{a}_{\text{soi},f}\right) \tag{19}$$

with $\hat{\rho}_f = \widehat{\mathbb{E}}\left[\frac{\partial \, \phi_f(\underline{\hat{\boldsymbol{s}}}_\tau)}{\partial \hat{s}_{f,\tau}^*}\right]$. Note that the update (19) corresponds to the well-known one-unit FastICA/FastIVA algorithms [27,28] applied to the AEC error signal $\boldsymbol{e}_{f,\tau}$. We continue by computing the second-order derivatives for the AEC filter update

$$\frac{\partial^2 \mathcal{J}(\boldsymbol{\theta})}{\partial \boldsymbol{h}_f \, \partial \boldsymbol{h}_f^{\text{H}}} = \left(\boldsymbol{R}_f^{\text{T}} + \hat{\rho}_f \boldsymbol{w}_{\text{bse},f}^* \boldsymbol{w}_{\text{bse},f}^{\text{T}}\right) \widehat{\mathbb{E}}\left[\left|u_{f,\tau}\right|^2\right] \tag{20}$$

$$\frac{\partial^2 \mathcal{J}(\boldsymbol{\theta})}{\partial \boldsymbol{h}_f \, \partial \boldsymbol{h}_f^{\text{T}}} = \hat{\xi}_f \boldsymbol{w}_{\text{bse},f}^* \boldsymbol{w}_{\text{bse},f}^{\text{H}} \widehat{\mathbb{E}}\left[u_{f,\tau}^2\right] \tag{21}$$

with $\hat{\xi}_f = \widehat{\mathbb{E}}\left[\frac{\partial \, \phi_f(\underline{\hat{\boldsymbol{s}}}_\tau)}{\partial \hat{s}_{f,\tau}}\right]$. As most common STFT-domain loudspeaker signals, e.g., speech and music, can be well-modeled as circular random variables, $\widehat{\mathbb{E}}\left[u_{f,\tau}^2\right]$ will be close to zero, and thus Eq. (21) is approximately equal to the all-zero matrix $\boldsymbol{0}_{M\times M}$. This simplifies the Newton step to [26]

$$\boldsymbol{h}_f \leftarrow \boldsymbol{h}_f - \left(\left(\frac{\partial^2 \mathcal{J}(\boldsymbol{\theta})}{\partial \boldsymbol{h}_f \, \partial \boldsymbol{h}_f^{\text{H}}}\right)^*\right)^{-1} \frac{\partial \mathcal{J}(\boldsymbol{\theta})}{\partial \boldsymbol{h}_f^*}. \tag{22}$$

By incorporating the score function normalization into (17) and (20) we obtain the proposed AEC filter update

$$\begin{aligned}\boldsymbol{h}_f \leftarrow \boldsymbol{h}_f + \left(\left(\boldsymbol{R}_f + \frac{\hat{\rho}_f^*}{\hat{\nu}_f^*} \boldsymbol{w}_{\text{bse},f} \boldsymbol{w}_{\text{bse},f}^{\text{H}}\right) \widehat{\mathbb{E}}\left[\left|u_{f,\tau}\right|^2\right]\right)^{-1} \\ \widehat{\mathbb{E}}\left[\left(\frac{\phi_f^*(\underline{\hat{\boldsymbol{s}}}_\tau)}{\hat{\nu}_f^*} \, \boldsymbol{w}_{\text{bse},f} + \boldsymbol{R}_f \, \boldsymbol{e}_{f,\tau}\right) u_{f,\tau}^*\right].\end{aligned} \tag{23}$$

The AEC filter update (23) can be interpreted as an SOI- and interference-aware multichannel BNLMS. The traditional single-channel BNLMS is obtained by setting the number of microphones $M = 1$, which simplifies the beamforming vector to $w_{\text{bse},f} = 1$, i.e., $\hat{s}_{f,\tau} = e_{f,\tau}$, and discards $\boldsymbol{R}_f$, and choosing a stationary Gaussian SOI source model with score function $\phi_f(\underline{\hat{\boldsymbol{s}}}_\tau) = \hat{s}_{f,\tau}^*$ which leads to

$$\frac{\partial \mathcal{J}(\boldsymbol{\theta})}{\partial h_f^*} = -\widehat{\mathbb{E}}\left[e_{f,\tau} u_{f,\tau}^*\right] \quad \text{and} \quad \frac{\partial^2 \mathcal{J}(\boldsymbol{\theta})}{\partial h_f \partial h_f^*} = \widehat{\mathbb{E}}\left[\left|u_{f,\tau}\right|^2\right]. \tag{24}$$

The proposed AEC update extends the BNLMS (cf. Eqs. (22) and (24)) by incorporating knowledge about the spatial and spectral signal characteristics of the SOI and interference and thus acts as an inherent adaptation control.

Algorithm 1 Algorithmic description of the proposed joint AEC and BSE algorithm[1].

1: **for** each iteration **do**
2: Update AEC filter $\boldsymbol{h}_f$ by Eq. (23)
3: Update BSE filter: $\boldsymbol{w}_{\text{bse},f}$ by Eq. (19)
4: Normalize: $\boldsymbol{w}_{\text{bse},f} \leftarrow \boldsymbol{w}_{\text{bse},f} / (\boldsymbol{w}_{\text{bse},f}^{\text{H}} \widehat{\boldsymbol{C}}_{ee,f} \boldsymbol{w}_{\text{bse},f})$
5: **end for**
6: Backprojection: $\hat{s}_{f,\tau} \leftarrow \widehat{\mathbb{E}}[\hat{s}_{f,\tau}^* [\boldsymbol{e}_{f,\tau}]_r] / \widehat{\mathbb{E}}[| \hat{s}_{f,\tau} |^2] \, \hat{s}_{f,\tau}$

3.4. Algorithmic Description

The proposed joint AEC and BSE algorithm is summarized in Alg. 1. Thereby we approximate the covariance matrix $\boldsymbol{C}_{zz,f}$ in $\boldsymbol{R}_f = \boldsymbol{B}_f^{\text{H}} \boldsymbol{C}_{zz,f}^{-1} \boldsymbol{B}_f$ by the estimate $\widehat{\boldsymbol{C}}_{zz,f} = \widehat{\mathbb{E}}\left[\hat{\boldsymbol{z}}_{f,\tau} \hat{\boldsymbol{z}}_{f,\tau}^{\text{H}}\right]$ and normalize the BSE filter $\boldsymbol{w}_{\text{bse},f}$ after each iteration by $\boldsymbol{w}_{\text{bse},f}^{\text{H}} \widehat{\boldsymbol{C}}_{ee,f} \boldsymbol{w}_{\text{bse},f}$ (cf. line 4) to ensure that the SOI estimate $\hat{s}_{f,\tau}$ has unit scale, i.e., $\widehat{\mathbb{E}}\left[\left|\hat{s}_{f,\tau}\right|^2\right] = 1$ [22]. Finally, to address the scale ambiguity, which is inherent to any ICA approach [14], we multiply the SOI estimate $\hat{s}_{f,\tau}$ after the optmization by $\min_{\alpha_f} \widehat{\mathbb{E}}\left[\left|\alpha_f \, \hat{s}_{f,\tau} - [\boldsymbol{e}_{f,\tau}]_r\right|^2\right]$ [29] which corresponds to a projection of $\hat{s}_{f,\tau}$ onto the rth error signal $[\boldsymbol{e}_{f,\tau}]_r$. Note that, because of the reduced echo power, we use the error signal as reference instead of the commonly used microphone signal.

4. EXPERIMENTAL EVALUATION

The proposed joint AEC and BSE algorithm is evaluated for a variety of challenging multi-microphone acoustic echo and interference cancellation scenarios. We consider a 4-element circular microphone array with random diameter in the range [5cm, 15cm] which is placed randomly in a shoebox room with dimensions in the ranges [4m, 6m], [4m, 6m] and [2.2m, 3.2m], respectively, with random reverberation time $T_{60} \in$ [0.2s, 0.3s]. The loudspeaker, SOI and a single interferer are located orthogonally to the microphone array axis with random angles of arrival and random distances in the ranges [5cm, 20cm], [0.7m, 1.4m] and [1.7m, 2.5m], respectively. All room impulse responses are simulated according to the image method [30] with a minimum filter length of $\max(4000, \lceil f_s T_{60} \rceil)$ and sampling frequency f_s = 16kHz. The loudspeaker, SOI and interferer emit 5s speech signals which are sampled randomly from a subset of the LibriSpeech corpus [31] including 100 speakers. The SOI and interfering signal images (cf. Eq. (1)) are scaled according to a random SOI-to-echo and interference-to-echo ratio in the ranges [5dB, 10dB] and [0dB, 5dB], respectively. Note that we consider a dominant SOI to alleviate the outer permutation problem, i.e., selecting the correct SOI. To simulate more realistic scenarios, we added spatially uncorrelated white Gaussian sensor noise with a random echo-to-noise ratio in the range [25dB, 35dB].

We chose an STFT frame length and frameshift of 2048 and 1024, respectively. The filter vectors $\boldsymbol{h}_f$ and $\boldsymbol{w}_{\text{bse},f}$ were updated by 50 iterations after being initialized with $\boldsymbol{h}_f = \boldsymbol{0}_{M \times 1}$ and $\boldsymbol{w}_{\text{bse},f} = \begin{pmatrix} 1 & \boldsymbol{0}_{1 \times M-1} \end{pmatrix}^{\text{T}}$, respectively. As score function we considered

$$\phi_f(\hat{\underline{\boldsymbol{s}}}_\tau) = \frac{\hat{s}_{f,\tau}^*}{\sqrt{\sum_{f=1}^{F} |\hat{s}_{f,\tau}|^2}} \tag{25}$$

which models joint broadband activity of the SOI. We backprojected the SOI estimates (cf. Sec. 3.4) to the first error signal.

As performance measures we use the logarithmic time-domain SOI-to-interference power ratio (SIR), SOI-to-echo power ratio (SER), SOI-to-interference-plus-echo power ratio (SIER) and the logarithmic echo return loss enhancement (ERLE) [8] after the AEC unit and the beamformer, respectively. All performance measures are averaged over 50 experiments with randomly drawn speech signals and randomly sampled acoustic scenarios, i.e., sampling room geometry and reverberation time, microphone array properties, and positions of SOI, interferer, and loudspeaker, respectively, in the ranges given above.

Tab. 1 shows the average performance of the unprocessed microphone signals, a least squares (LS) AEC filter estimate, a fast IVE-controlled beamformer, i.e., discarding the AEC unit, and finally the proposed joint AEC+IVE in comparison to the individual update (BNLMS+IVE), i.e., replacing the AEC filter update (23) by an independent BNLMS in each channel (cf. Sec. 3.3). We conclude from Tab. 1 that the exploitation of both algorithmic parts, i.e., echo canceller and beamformer, significantly outperforms the individual approaches (cf. LS AEC and IVE). When comparing the proposed joint update to the decoupled BNLMS+IVE approach, we observe a significantly improved SER which results from enhanced parameter updates of both algorithmic parts (cf. ERLE_{aec} and ERLE_{bf} in Tab. 1). Besides an improved echo attenuation, the proposed joint parameter update achieves also slightly higher interference cancellation in comparison to the individual approach. We attribute this to the higher echo attenuation provided by the echo canceller, which allows the beamformer to focus on the interferer suppression.

Table 1. Average performance of the proposed joint AEC and BSE algorithm in comparison to various baselines with the best values being typeset bold.

Algorithm	SIR	SER	SIER	ERLE_{aec}	ERLE_{bf}
Unprocessed	4.78	7.39	2.80	0.00	0.00
LS AEC	4.78	17.66	4.65	10.27	10.27
IVE	6.41	10.01	4.25	0.00	6.97
BNLMS+IVE	6.75	17.73	6.25	10.27	14.40
Joint AEC+IVE	**7.02**	**19.78**	**6.73**	**11.39**	**16.36**

5. CONCLUSION

In this paper, we described a computationally efficient and fast-converging joint AEC and BSE algorithm which improves the AEC performance in comparison to the independent application of both algorithmic components. As future research we plan to investigate convolutive AEC models and the incorporation of spatial prior knowledge about the SOI [32].

[1] Source code implementation will be made available at https://github.com/ThomasHaubner/joint_AEC_BSE

6. REFERENCES

[1] K. Sridhar et al., "ICASSP 2021 Acoustic Echo Cancellation Challenge: Datasets, Testing Framework, and Results," in *Int. Conf. Acoust., Speech, Signal Process.*, Toronto, CA, June 2021, pp. 151–155.

[2] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *Int. Conf. Acoust., Speech, Signal Process.*, Munich, DE, April 1997, pp. 219–222.

[3] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A practical Approach*, Wiley-Interscience, NJ, USA, 2004.

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

[5] M. L. Valero and E. A. P. Habets, "Low-Complexity Multi-Microphone Acoustic Echo Control in the Short-Time Fourier Transform Domain," *IEEE Audio, Speech, Language Process.*, vol. 27, no. 3, pp. 595–609, 2019.

[6] Mhd. Halimeh et al., "Combining Adaptive Filtering and Complex-Valued Deep Postfiltering for Acoustic Echo Cancellation," in *Int. Conf. Acoust., Speech, Signal Process.*, Toronto, CA, June 2021, pp. 121–125.

[7] W. Kellermann, "Acoustic Echo Cancellation for Beamforming Microphone Arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds., pp. 281–306. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[8] G. Enzner et al., "Acoustic Echo Control," *Academic Press Library in Signal Process.*, vol. 4, pp. 807–877, 2014.

[9] J. Park and J.-H. Chang, "State-Space Microphone Array Nonlinear Acoustic Echo Cancellation Using Multi-Microphone Near-End Speech Covariance," *IEEE Audio, Speech, Language Process.*, vol. 27, no. 10, pp. 1520–1534, Oct. 2019.

[10] N. Cohen et al., "An online algorithm for echo cancellation, dereverberation and noise reduction based on a Kalman-EM Method," *Eurasip J. Audio Speech Music Process.*, vol. 2021, no. 1, pp. 33, 2021.

[11] T. Haubner et al., "A Synergistic Kalman- and Deep Postfiltering Approach to Acoustic Echo Cancellation," in *European Signal Process. Conf.*, Dublin, IR, Aug. 2021.

[12] T. Haubner et al., "Noise-Robust Adaptation Control for Supervised Acoustic System Identification Exploiting a Noise Dictionary," in *Int. Conf. Acoust., Speech, Signal Process.*, Toronto, CA, June 2021, pp. 945–949.

[13] T. Haubner et al., "End-To-End Deep Learning-Based Adaptation Control for Frequency-Domain Adaptive System Identification," in *Int. Conf. Acoust., Speech, Signal Process.*, Singapore, SG, May 2022.

[14] P. Comon, "Independent component analysis, a new concept?," *Signal Pproc.*, vol. 36, no. 3, pp. 287–314, 1994.

[15] F. Nesta et al., "Batch-online semi-blind source separation applied to multi-channel acoustic echo cancellation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 3, pp. 583–599, 2011.

[16] J. Gunther, "Learning Echo Paths During Continuous Double-Talk Using Semi-Blind Source Separation," *IEEE Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 646–660, 2012.

[17] Z. Wang et al., "Weighted Recursive Least Square Filter and Neural Network Based Residual Echo Suppression for the AEC-Challenge," in *Int. Conf. Acoust., Speech, Signal Process.*, Toronto, CA, June 2021, pp. 141–145.

[18] G. Cheng et al., "Semi-blind source separation for nonlinear acoustic echo cancellation," *IEEE Signal Process. Lett.*, vol. 28, pp. 474–478, 2021.

[19] H. Buchner and W. Kellermann, "A fundamental relation between blind and supervised adaptive filtering illustrated for blind source separation and acoustic echo cancellation," in *Hands-Free Speech Comm and Microphone Arrays (HSCMA)*, Trento, IT, May 2008, pp. 17–20.

[20] Y. Na et al., "Joint Online Multichannel Acoustic Echo Cancellation, Speech Dereverberation and Source Separation," in *Interspeech*, Brno, CZ, Sept. 2021, pp. 1144–1148.

[21] C. Boeddeker et al., "A comparison and combination of unsupervised blind source separation techniques," in *ITG Conf. on Speech Comm.*, Kiel, DE, Sept. 2021.

[22] Z. Koldovský and P. Tichavský, "Gradient algorithms for complex non-gaussian independent component/vector extraction, question of convergence," *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1050–1064, 2019.

[23] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*, Wiley, 2002.

[24] J.-F. Cardoso, "On the performance of orthogonal source separation algorithms," in *European Signal Process. Conf.*, Edinburgh, UK, 1994, vol. 94, pp. 776–779.

[25] Z. Koldovský et al., "Orthogonally Constrained Independent Component Extraction: Blind MPDR Beamforming," in *European Signal Processing Conf.*, Kos Island, GR, Aug. 2017, pp. 1155–1159.

[26] H. Li and T. Adali, "Complex-valued adaptive signal processing using nonlinear functions," *EURASIP J. Adv. Signal Process.*, vol. 2008, Jan. 2008.

[27] Z. Koldovský et al., "Dynamic independent component/vector analysis: Time-variant linear mixtures separable by time-invariant beamformers," *IEEE Trans. Signal Process.*, vol. 69, pp. 2158–2173, 2021.

[28] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Comput.*, vol. 9, no. 7, pp. 1483–1492, Oct. 1997.

[29] K. Matsuoka, "Minimal distortion principle for blind source separation," in *SICE Annual Conf.*, Osaka, JP, Aug. 2002, pp. 2138–2143.

[30] E. A. P. Habets, "Room Impulse Response Generator," Tech. Rep., Technische Universiteit Eindhoven, Sept. 2010.

[31] V. Panayotov et al., "Librispeech: An ASR corpus based on public domain audio books," in *Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, AU, Apr. 2015, pp. 5206–5210.

[32] A. Brendel et al., "A Unified Probabilistic View on Spatially Informed Source Separation and Extraction Based on Independent Vector Analysis," *IEEE Trans. Signal Process.*, vol. 68, pp. 3545–3558, 2020.