Deep Learning-Based Joint Control of Acoustic Echo Cancellation, Beamforming and Postfiltering

Thomas Haubner and Walter Kellermann

Multimedia Communications and Signal Processing, Friedrich-Alexander-University Erlangen-Nürnberg,
Cauerstr. 7, D-91058 Erlangen, Germany, e-mail: {thomas.haubner,walter.kellermann}@fau.de

Abstract—We introduce a novel method for controlling the functionality of a hands-free speech communication device which comprises a model-based acoustic echo canceller (AEC), minimum variance distortionless response (MVDR) beamformer (BF) and spectral postfilter (PF). While the AEC removes the early echo component, the MVDR BF and PF suppress the residual echo and background noise. As key innovation, we suggest to use a single deep neural network (DNN) to jointly control the adaptation of the various algorithmic components. This allows for rapid convergence and high steady-state performance in the presence of high-level interfering double-talk. End-to-end training of the DNN using a time-domain speech extraction loss function avoids the design of individual control strategies.

Index Terms—Acoustic Echo Cancellation, Noise Suppression, Beamformer, Postfilter, Adaptation Control, Deep Learning

I. INTRODUCTION

Driven by the increased usage of hands-free voice communication devices, acoustic echo control has recently become again a highly active research field [1], [2], [3]. Most modern acoustic echo control algorithms can be classified into three classes: Model-based system identification algorithms (traditionally referred to as AECs) [4], [5], pure deep learning-based spectral PFs [6] and a combination of both [1], [2], [7]. While traditional AECs generalize well to unknown acoustic environments and introduce no distortion to the desired near-end speech signal, they are inherently limited by their model assumptions, e.g., the filter length and non-linearity model [8]. In contrast, pure deep learning-based PF approaches avoid this limitation by treating acoustic echo control as an interference suppression problem. Yet, this comes at the cost of requiring large models to minimize the distorting effect of spectrally masking the near-end speaker [6]. Note that echo suppression is in this sense vastly more challenging in comparison to typical acoustic noise suppression tasks, due to the much smaller signal-to-echo ratio (SER) in comparison to signal-to-noise ratio (SNR). This problem can be mitigated by enhancing the SER by a model-based AEC before applying a DNN-based PF for residual echo suppression. This fusion of traditional AECs with deep learning-based PFs has recently been shown to be superior to their individual counterparts for a variety of challenging scenarios [3]. Yet, this benefit is only obtained if the AEC is equipped with a sophisticated adaptation control which ensures a rapid convergence and robust steady-state performance in the presence of double-talk. For this, machine learning-based approaches were shown to be superior to their traditional counterparts [9], [10], [11], [12].

Besides spectral PF-based approaches, BFs have shown great potential for achieving high echo and noise attenuation [13], [14], [15], [16]. They exploit the spatial diversity of the interfering signal components, i.e., echo and noise, and the desired near-end speaker by linear spatiotemporal processing of the microphone or the echo-reduced AEC error signals. Yet, they require precise parameter estimates of, e.g. the relative transfer functions and/or cross power spectral density (CPSD) of the interference, to minimize the distortion of the near-end speech signal. This problem has been addressed by jointly estimating the AEC and BF weights by optimizing a least squares criterion [13], online expectation-maximization algorithms [15], [16] and recently also DNN-supported approaches [17], [18]. However, whereas [18] completely omits the model-based AEC and suggests a DNN-supported MVDR BF [19] for echo and noise reduction, [17] considers a convolutive narrowband AEC and is limited to offline applications. Furthermore, there have also been investigations of multi-microphone DNN-only echo and noise control algorithms [20] which however require again large networks and are trained for specific microphone array topologies.

In this paper we introduce a novel method for jointly controlling a model-based broadband AEC, MVDR BF and spectral PF for online acoustic echo and noise reduction. While the AEC improves the SER by cancelling the early echo, the MVDR BF and PF exploit the spatial and spectrotemporal variability of the signal components to suppress the residual echo and noise. To achieve fast convergence and robust steady-state performance in the presence of double-talk, we suggest to use a single DNN to jointly control the parameter adaptation of the algorithmic components. The DNN is trained end-to-end w.r.t. a component-based time-domain loss function which allows to trade interference suppression against speech distortion. This avoids the development of individual control strategies and allows the algorithmic components to synergistically interact.

We use bold uppercase letters for matrices and bold lowercase letters for vectors with time-domain quantities being indicated by an underline. The $M \times M$-dimensional identity matrix, all-zero matrix and discrete fourier transform (DFT) matrix are denoted by $\boldsymbol{I}_M$, $\boldsymbol{0}_M$ and $\boldsymbol{F}_M$, respectively. Furthermore, we introduce the element-wise product operator $\odot$, the linear convolution operator $*$ and the Euclidean norm $||\cdot||$. The transposition and Hermitian transposition are denoted by $(\cdot)^{\mathrm{T}}$ and $(\cdot)^{\mathrm{H}}$, respectively. Finally, we indicate the mth element of a vector by $[\cdot]_m$.

II. SIGNAL MODEL

We consider a hands-free voice communication scenario with P microphones as shown in Fig. 1. The multichannel time-domain microphone signal $\underline{\boldsymbol{y}}_\kappa$ at sample index κ is modelled as a linear superposition of an echo component $\underline{\boldsymbol{d}}_\kappa$, a speech component $\underline{\boldsymbol{s}}_\kappa$ and a noise component $\underline{\boldsymbol{n}}_\kappa$ as follows:

$$\underline{\boldsymbol{y}}_\kappa = \underline{\boldsymbol{d}}_\kappa + \underline{\boldsymbol{s}}_\kappa + \underline{\boldsymbol{n}}_\kappa \in \mathbb{R}^P. \tag{1}$$

Furthermore, the echo and the speech images at the microphones are modelled by a linear convolution of the loudspeaker signal $\underline{x}_\kappa$ and the dry, i.e., non-reverberant, speech signal $\underline{s}_\kappa^{\mathrm{dr}}$ with according room impulse responses (RIRs) $\underline{h}_{p,\kappa}$ and $\underline{g}_{p,\kappa}$

$$[\underline{\boldsymbol{d}}_\kappa]_p = \underline{d}_{p,\kappa} = \underline{h}_{p,\kappa} * \underline{x}_\kappa \tag{2}$$

$$[\underline{\boldsymbol{s}}_\kappa]_p = \underline{s}_{p,\kappa} = \underline{g}_{p,\kappa} * \underline{s}_\kappa^{\mathrm{dr}}. \tag{3}$$

As we consider in the following a block-based online processing of the signals, we introduce the time-domain loudspeaker signal block of length $M = 2R$

$$\underline{\boldsymbol{x}}_\tau^{\mathrm{bl}} = \begin{pmatrix} \underline{\boldsymbol{x}}_{\tau-1}^{\mathrm{in}} \\ \underline{\boldsymbol{x}}_\tau^{\mathrm{in}} \end{pmatrix} \in \mathbb{R}^M \tag{4}$$

which is composed of two innovation blocks of the form

$$\underline{\boldsymbol{x}}_\tau^{\mathrm{in}} = \left(\underline{x}_{\tau R - R + 1}, \underline{x}_{\tau R - R + 2}, \ldots, \underline{x}_{\tau R} \right)^{\mathrm{T}} \in \mathbb{R}^R \tag{5}$$

with R being the frameshift and τ the block index. Analogously, we introduce the microphone innovation block $\underline{\boldsymbol{y}}_{p,\tau}^{\mathrm{in}} \in \mathbb{R}^R$ at the pth microphone.

III. PROPOSED ONLINE ACOUSTIC ECHO AND NOISE CONTROL ALGORITHM

We first introduce the individual speech enhancement components, i.e., linear AEC, MVDR BF and spectral PF, before we turn to the deep learning-based method for joint control.

A. Acoustic Echo Cancellation

As AEC we consider a straightforward extension of the popular frequeny-domain adaptive filter (FDAF) algorithm [21], [22] to multiple microphones. Here, an echo estimate at the pth microphone and block index τ is obtained by a linear convolution of the finite impulse response (FIR) filter $\underline{\hat{\boldsymbol{h}}}_{p,\tau}$ of length R with the loudspeaker signal block $\underline{\boldsymbol{x}}_\tau^{\mathrm{bl}}$ (cf Eq. (4)). The linear convolution can efficiently be implemented by overlap-save processing in the DFT domain

$$\widehat{\underline{\boldsymbol{d}}}_{p,\tau}^{\mathrm{in}} = \boldsymbol{Q}_1^{\mathrm{T}} \boldsymbol{F}_M^{-1} \left(\left(\boldsymbol{F}_M \underline{\boldsymbol{x}}_\tau^{\mathrm{bl}} \right) \odot \hat{\boldsymbol{h}}_{p,\tau} \right) \in \mathbb{R}^R \tag{6}$$

with the DFT-domain FIR filter $\hat{\boldsymbol{h}}_{p,\tau} = \boldsymbol{F}_M \boldsymbol{Q}_2 \underline{\hat{\boldsymbol{h}}}_{p,\tau} \in \mathbb{C}^M$ and the zero-padding matrix $\boldsymbol{Q}_2^{\mathrm{T}} = \begin{pmatrix} \boldsymbol{I}_R & \boldsymbol{0}_R \end{pmatrix}$. Note that the constraint matrix $\boldsymbol{Q}_1^{\mathrm{T}} = \begin{pmatrix} \boldsymbol{0}_R & \boldsymbol{I}_R \end{pmatrix}$ ensures a linear convolution of the DFT-domain product in brackets by discarding the elements corresponding to a circular convolution [21]. Subsequently, the time-domain error innovation block is computed by subtracting the echo estimate $\widehat{\underline{\boldsymbol{d}}}_{p,\tau}^{\mathrm{in}}$ (cf. Eq. (6)) from the respective microphone observations $\underline{\boldsymbol{y}}_{p,\tau}^{\mathrm{in}}$

$$\underline{\boldsymbol{e}}_{p,\tau}^{\mathrm{in}} = \underline{\boldsymbol{y}}_{p,\tau}^{\mathrm{in}} - \widehat{\underline{\boldsymbol{d}}}_{p,\tau}^{\mathrm{in}}. \tag{7}$$
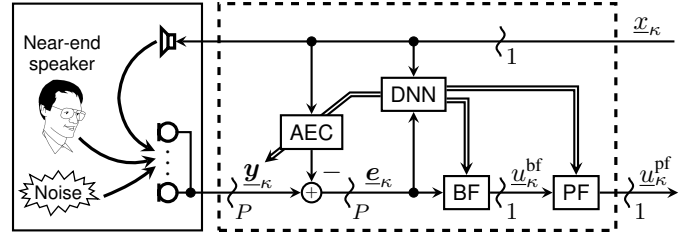

Fig. 1. Block diagram of the proposed DNN-controlled speech enhancement algorithm with the loudspeaker signal $\underline{x}_\kappa$, the microphone signal $\underline{\boldsymbol{y}}_\kappa$, the multichannel AEC error signal $\underline{\boldsymbol{e}}_\kappa$, the BF output $\underline{u}_\kappa^{\mathrm{bf}}$ and the PF output $\underline{u}_\kappa^{\mathrm{pf}}$. The numbers below the curved lines correspond to the respective channel dimensions.

The filter coefficients $\hat{\boldsymbol{h}}_{p,\tau}$ are adapted by the gradient-based update rule [21], [22]

$$\Delta \hat{\boldsymbol{h}}_{p,\tau} = \left(\boldsymbol{F}_M \underline{\boldsymbol{x}}_\tau^{\mathrm{bl}} \right)^* \odot \left(\boldsymbol{F}_M \boldsymbol{Q}_1 \underline{\boldsymbol{e}}_{p,\tau}^{\mathrm{in}} \right) \tag{8}$$

$$\hat{\boldsymbol{h}}_{p,\tau} = \hat{\boldsymbol{h}}_{p,\tau-1} + \boldsymbol{Q}_3 \left(\boldsymbol{\mu}_{p,\tau} \odot \Delta \hat{\boldsymbol{h}}_{p,\tau} \right) \tag{9}$$

with the frequency- and microphone-dependent step-size vector $\boldsymbol{\mu}_{p,\tau}$. Furthermore, the gradient-constraint projection matrix $\boldsymbol{Q}_3 = \boldsymbol{F}_M \boldsymbol{Q}_2 \boldsymbol{Q}_2^{\mathrm{T}} \boldsymbol{F}_M^{-1}$ ensures that the estimate $\hat{\boldsymbol{h}}_{p,\tau}$ corresponds to a zero-padded FIR filter $\underline{\hat{\boldsymbol{h}}}_{p,\tau} = \boldsymbol{Q}_2^{\mathrm{T}} \boldsymbol{F}_M^{-1} \hat{\boldsymbol{h}}_{p,\tau}$ [22]. The performance of this update decisively depends on a proper choice of the step-size vector $\boldsymbol{\mu}_{p,\tau}$. Due to its rapid convergence speed and double-talk robustness, we use a straightforward extension of the DNN-FDAF approach [11] to multiple microphones. In particular, we suggest the time-varying step-size vector

$$\left[\boldsymbol{\mu}_{p,\tau} \right]_f = \frac{m_{f,\tau}^\mu}{\Psi_{f,\tau}^{\mathrm{XX}} + \frac{M}{R} \left| m_{f,\tau}^{\mathrm{aec}} \left[\boldsymbol{F}_M \boldsymbol{Q}_1 \underline{\boldsymbol{e}}_{p,\tau}^{\mathrm{in}} \right]_f \right|^2} \tag{10}$$

with $\Psi_{f,\tau}^{\mathrm{XX}}$ being the loudspeaker signal PSD and $m_{f,\tau}^\mu \in [0,1]$ and $m_{f,\tau}^{\mathrm{aec}} \in [0,1]$ being auxiliary variables which are provided for each frequency bin f and time τ by a DNN (cf. Sec. III-D). The step-size vector (10) allows the DNN to set entries entirely to zero, while in addition being robust to varying signal powers due to the frequency-selective loudspeaker signal power and error power normalization. In addition, the DNN can eliminate the error power normalization to address the different reasons for large errors, i.e., interfering signal activity (double-talk) or system misalignment. The loudspeaker signal PSD $\Psi_{f,\tau}^{\mathrm{XX}}$ is estimated by recursive averaging

$$\Psi_{f,\tau}^{\mathrm{XX}} = \lambda_{\mathrm{X}} \, \Psi_{f,\tau-1}^{\mathrm{XX}} + (1 - \lambda_{\mathrm{X}}) \, \left| \left[\boldsymbol{F}_M \underline{\boldsymbol{x}}_\tau^{\mathrm{bl}} \right]_f \right|^2 \tag{11}$$

with $\lambda_{\mathrm{X}} \in (0,1)$ being an easy-to-choose hyperparameter.

B. MVDR Beamforming

As the linear spatial filtering is conducted in the short-time Fourier transform (STFT) domain, we introduce the broadband STFT-domain error signal at the pth microphone (cf. Eq. (7))

$$\boldsymbol{e}_{p,\tau} = \boldsymbol{F}_M \left(\boldsymbol{b} \odot \begin{pmatrix} \underline{\boldsymbol{e}}_{p,\tau-1}^{\mathrm{in}} \\ \underline{\boldsymbol{e}}_{p,\tau}^{\mathrm{in}} \end{pmatrix} \right) \in \mathbb{C}^M \tag{12}$$

with $\boldsymbol{b}$ being a Hamming window. By concatenating the narrowband error signal components (cf. Eq. (12)) at the different microphones, we obtain the multichannel error signal vector

$$\tilde{\boldsymbol{e}}_{f,\tau} = \left([\boldsymbol{e}_{1,\tau}]_f \quad \cdots \quad [\boldsymbol{e}_{P,\tau}]_f\right)^{\mathrm{T}} \in \mathbb{C}^P \tag{13}$$

at frequency bin f. Note that in the following all STFT-domain narrowband signals are indicated by a tilde.

The single-channel STFT-domain output of the MVDR BF

$$\tilde{u}_{f,\tau}^{\mathrm{bf}} = \tilde{\boldsymbol{w}}_{f,\tau}^{\mathrm{H}} \tilde{\boldsymbol{e}}_{f,\tau} \tag{14}$$

is computed as the inner product of the multichannel error signal $\tilde{\boldsymbol{e}}_{f,\tau}$ (cf. Eq. (13)) and the MVDR weight vector [23]

$$\tilde{\boldsymbol{w}}_{f,\tau} = \frac{\left(\tilde{\boldsymbol{\Psi}}_{f,\tau}^{\mathrm{ZZ}} + \delta_1 \boldsymbol{I}_P\right)^{-1} \tilde{\boldsymbol{a}}_{f,\tau}}{\tilde{\boldsymbol{a}}_{f,\tau}^{\mathrm{H}} \left(\tilde{\boldsymbol{\Psi}}_{f,\tau}^{\mathrm{ZZ}} + \delta_1 \boldsymbol{I}_P\right)^{-1} \tilde{\boldsymbol{a}}_{f,\tau} + \delta_2} \tag{15}$$

with δ_1 and δ_2 being regularization constants. The interfering signal CPSD matrix $\tilde{\boldsymbol{\Psi}}_{f,\tau}^{\mathrm{ZZ}} \in \mathbb{C}^{P \times P}$ and near-end relative transfer function (RTF) vector $\tilde{\boldsymbol{a}}_{f,\tau} \in \mathbb{C}^P$ are computed from estimates of the interference, i.e., residual echo and noise, and speech image

$$\left[\hat{\tilde{\boldsymbol{z}}}_{f,\tau}\right]_p = \left(1 - m_{p,f,\tau}^{\mathrm{bf}}\right) [\tilde{\boldsymbol{e}}_{f,\tau}]_p, \tag{16}$$

$$\left[\hat{\tilde{\boldsymbol{s}}}_{f,\tau}\right]_p = m_{p,f,\tau}^{\mathrm{bf}} [\tilde{\boldsymbol{e}}_{f,\tau}]_p, \tag{17}$$

respectively, with $m_{p,f,\tau}^{\mathrm{bf}} \in [0,1]$ being a microphone-, frequency- and time-dependent mask that is provided by a DNN (cf. Sec. III-D). While the interfering signal CPSD matrix $\tilde{\boldsymbol{\Psi}}_{f,\tau}^{\mathrm{ZZ}}$ is directly calculated by recursively averaging the outer products of the interference estimates $\hat{\tilde{\boldsymbol{z}}}_{f,\tau}$ (cf. Eq. (16))

$$\tilde{\boldsymbol{\Psi}}_{f,\tau}^{\mathrm{ZZ}} = \lambda_{\mathrm{Z}} \tilde{\boldsymbol{\Psi}}_{f,\tau-1}^{\mathrm{ZZ}} + (1 - \lambda_{\mathrm{Z}}) \hat{\tilde{\boldsymbol{z}}}_{f,\tau} \hat{\tilde{\boldsymbol{z}}}_{f,\tau}^{\mathrm{H}}, \tag{18}$$

the near-end RTF vector $\tilde{\boldsymbol{a}}_{f,\tau}$ is computed as normalized eigenvector corresponding to the maximum eigenvalue of the speech CPSD matrix estimate

$$\tilde{\boldsymbol{\Psi}}_{f,\tau}^{\mathrm{SS}} = \lambda_{\mathrm{S}} \tilde{\boldsymbol{\Psi}}_{f,\tau-1}^{\mathrm{SS}} + (1 - \lambda_{\mathrm{S}}) \hat{\tilde{\boldsymbol{s}}}_{f,\tau} \hat{\tilde{\boldsymbol{s}}}_{f,\tau}^{\mathrm{H}}. \tag{19}$$

A computationally efficient solution to the eigenvector problem is given by the power iteration algorithm [24]

$$\tilde{\boldsymbol{a}}_{f,\tau} = \tilde{\boldsymbol{\Psi}}_{f,\tau}^{\mathrm{SS}} \tilde{\boldsymbol{a}}_{f,\tau-1} \tag{20}$$

$$\tilde{\boldsymbol{a}}_{f,\tau} = \frac{\tilde{\boldsymbol{a}}_{f,\tau}}{[\tilde{\boldsymbol{a}}_{f,\tau}]_1} \tag{21}$$

whose simplicity allows for a numerically robust end-to-end training of the DNN parameters (cf. Sec. III-D) [25]. Note that, in Eq. (21), we use the first microphone as reference without loss of generality.

C. Spectral Postfilter

The MVDR BF output $\tilde{u}_{f,\tau}^{\mathrm{bf}}$ (cf. Eq. (14)) is subsequently processed by a scalar spectral PF as follows

$$\tilde{u}_{f,\tau}^{\mathrm{pf}} = m_{f,\tau}^{\mathrm{pf}} \tilde{u}_{f,\tau}^{\mathrm{bf}} \tag{22}$$

with $m_{f,\tau}^{\mathrm{pf}}$ being a frequency-dependent mask which is inferred by a DNN (cf. Sec. III-D).
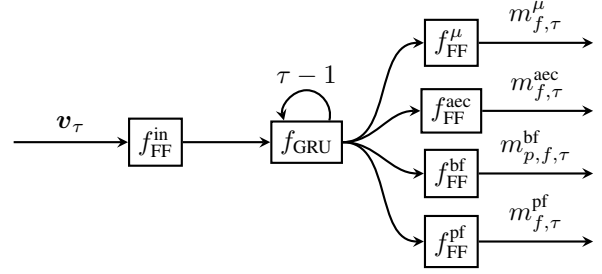

Fig. 2. DNN architecture for jointly controlling the AEC, BF and PF.

D. Deep Learning-Based Adaptation Control

In the following, we will describe how to determine the masks $m_{f,\tau}^{\mu}$ and $m_{f,\tau}^{\mathrm{aec}}$ (cf. (10)), $m_{p,f,\tau}^{\mathrm{bf}}$ (cf. (16) and (17)) and $m_{f,\tau}^{\mathrm{pf}}$ (cf. (22)) which control the adaptation of the AEC, MVDR BF and PF, respectively, by a single DNN. As input feature vector $\boldsymbol{v}_\tau$ for the DNN we use a concatenation of the logarithmic magnitude spectrum of the STFT-domain loudspeaker signal and the AEC error signals at the different microphones $\boldsymbol{e}_{1,\tau}, \ldots, \boldsymbol{e}_{P,\tau}$ (cf. Eq. (12)). The feature elements are normalized by estimating their mean and standard deviation during training. Note that due to the symmetry of the STFT-domain signals we only use the frequency components up to $F = \frac{M}{2} + 1$ with M being the even DFT length.

The DNN is composed of a feed-forward layer, which condenses the input feature vector $\boldsymbol{v}_\tau$ to a lower dimension Q, two stacked gated recurrent unit (GRU) layers, which extract temporal information, and finally four different feed-forward layers with sigmoid activation which map to the various masks as shown in Fig. 2. The architecture is chosen to yield after the GRU layer a condensed representation of the convergence state of the AEC, the noise spectrum and the near-end source activity. The DNN parameters $\boldsymbol{\theta}$ are trained end-to-end w.r.t. the component-based loss function

$$\mathcal{J}(\boldsymbol{\theta}) = \alpha \, ||\mathrm{pf}(\underline{\boldsymbol{d}}_{1:K})|| + \beta \, ||\mathrm{pf}(\underline{\boldsymbol{n}}_{1:K})|| + ||\underline{\boldsymbol{s}}_{1:K}^{\mathrm{ref}} - \mathrm{pf}(\underline{\boldsymbol{s}}_{1:K})|| \tag{23}$$

with $\mathrm{pf}(\underline{\boldsymbol{d}}_{1:K})$, $\mathrm{pf}(\underline{\boldsymbol{n}}_{1:K})$ and $\mathrm{pf}(\underline{\boldsymbol{s}}_{1:K})$ denoting the length-K time-domain signal of the echo, noise and speech images, respectively, after being processed by the algorithm. While the first two terms in (23) quantify the suppression of interference, i.e., echo and noise, the last term represents the distortion of the near-end speech signal w.r.t. a time-domain reference signal $\underline{\boldsymbol{s}}_\kappa^{\mathrm{ref}}$. The trade-off between interference suppression and near-end distortion can be controlled by the hyperparameters α and β. Eq. (23) is an echo-aware time-domain version of the frequency-domain noise suppression loss proposed in [26]. Note that the end-to-end optimization of the DNN avoids the design of desired oracle target masks and instead directly evaluates their effect on the speech extraction performance.

IV. EXPERIMENTS

We will now evaluate the proposed algorithm for a variety of different multi-microphone acoustic echo and noise control scenarios. Each scenario comprises a circular microphone array with four uniformly-distributed elements and a random

diameter in the range [7cm, 15cm]. The array is placed randomly in a shoebox room with random lengths, widths and heights in the ranges [3m, 8m], [3m, 8m] and [2.0m, 3.5m], respectively, and random reverberation time $T_{60} \in [0.2s, 0.6s]$. The loudspeaker and near-end speaker positions are sampled from the azimuthal angle range $[0°, 360°]$, elevation angle range $[-20°, 20°]$ and distance ranges $[0.1m, 0.5m]$ and $[0.5m, 2.0m]$, respectively. All RIRs are simulated by the image method [27], [28] with a filter length of $\max\{6000, \lfloor f_s T_{60} \rfloor\}$ and sampling frequency $f_s = 16$ kHz. The dry near-end speaker and loudspeaker signals $\underline{s}_\kappa^{dr}$ and $\underline{x}_\kappa$ (cf. Eq. (1)), respectively, were drawn from different subsets of the *LibriSpeech* dataset [29] comprising 143 speakers each. The near-end speaker activity started randomly in the range [1s, 4s] which allows to individually evaluate single-talk and double-talk performance of the algorithm. We consider recorded single-channel background noise signals from cafés, street junctions, public transportation and pedestrian areas [30] which were made spherically diffuse by using [31]. The near-end speech and background noise signals were scaled to create a random echo-to-near-end and echo-to-noise ratio in the ranges $[-10\text{dB}, 10\text{dB}]$ and $[10\text{dB}, 25\text{dB}]$, respectively.

The frame shift R and block length M were chosen as 1024 and 2048, respectively, which results in an AEC filter length of 1024 taps. Furthermore, the recursive PSD averaging factors were set to $\lambda_X = 0.5$, $\lambda_Z = \lambda_S = 0.99$ and the MVDR regularization constants were chosen as $\delta_1 = \delta_2 = 0.01$. The condensed feature dimension Q of the DNN was set to 256 which results in overall 3.9 million parameters. As reference signal s_κ^{ref} in the cost function (23), we use a delay-and-sum BF applied to the near-end speech image $\underline{s}_\kappa$, and the weighting parameters were chosen as $\alpha = \beta = 1$. The network was trained by using the ADAM optimizer with a step-size of 0.001 and 1.4h of training data. For evaluating the algorithms we simulated an additional 8.3min (50 scenarios with 10s each) which were disjoint from the training data, i.e., different signals, arrays and environments.

To evaluate the convergence rate and steady-state echo suppression performance of the various algorithmic components, we introduce the time-dependent logarithmic echo return loss enhancement (ERLE) obtained at the first microphone

$$\mathcal{E}_\kappa = 10 \log_{10} \frac{\hat{\mathbb{E}}\left[|\underline{d}_{1,\kappa}|^2\right]}{\hat{\mathbb{E}}\left[|\text{pr}(\underline{d}_\kappa)|^2\right]} \tag{24}$$

with $\hat{\mathbb{E}}$ denoting recursive averaging. Furthermore, $\text{pr}(\underline{d}_\kappa)$ denotes the processed echo image $\underline{d}_\kappa$ at the output of the various speech enhancement components, i.e., AEC, MVDR BF and PF. Note that $\text{pr}(\cdot)$ provides a scalar output because we select the first channel of the AEC and the outputs of the BF and the PF are anyway scalar. Fig. 3 shows the average time-dependent ERLE $\bar{\mathcal{E}}_\kappa$ over 50 different scenarios. The starting point of the near-end speaker activity is indicated by a black dot and sampled randomly from the shaded range. We conclude from Fig. 3 that the adaptive AEC and MVDR BF converge despite the interfering double-talk rapidly to its steady-state. The PF
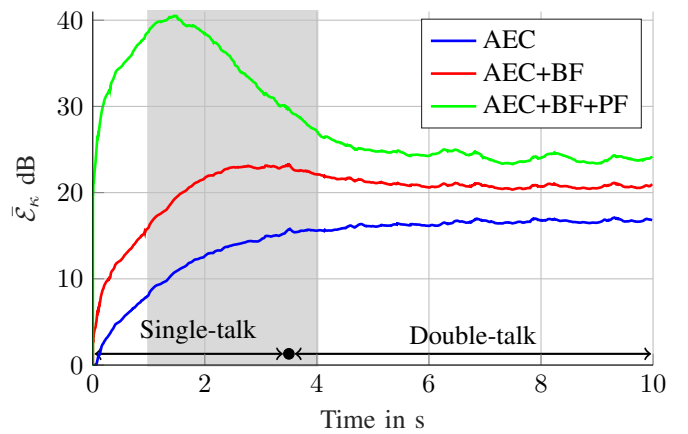

Fig. 3. Average time-dependent ERLE $\bar{\mathcal{E}}_\kappa$ of the various algorithmic components. The starting point of the near-end speaker activity is indicated by a black dot and sampled randomly from the shaded range.

drastically increases the echo suppression during the single-talk period at the beginning. Yet, after the near-end speaker starts talking, the PF attenuation is significantly reduced to minimize the distortion of the desired speech signal.

We will now investigate the echo and noise suppression performance of the algorithmic components during single-talk and double-talk and relate it to the induced distortion of the near-end speech signal. For this we introduce the logarithmic ERLE and the noise suppression factor at the first microphone

$$\mathcal{E} = 10 \log_{10} \frac{\sum_{\kappa=N_1}^{N_2} |\underline{d}_{1,\kappa}|^2}{\sum_{\kappa=N_1}^{N_2} |\text{pr}(\underline{d}_\kappa)|^2} \tag{25}$$

$$\mathcal{N} = 10 \log_{10} \frac{\sum_{\kappa=N_1}^{N_2} |\underline{n}_{1,\kappa}|^2}{\sum_{\kappa=N_1}^{N_2} |\text{pr}(\underline{n}_\kappa)|^2}, \tag{26}$$

respectively, with $\text{pr}(\cdot)$ denoting again the processed signal in brackets at the output of the algorithmic processing chain. Note that choosing the summation bounds N_1 and N_2 allows to assess different time periods, e.g., single- or double-talk. To evaluate the near-end speech distortion during double-talk we use the PESQ (Perceptual Evaluation of Speech Quality [32])

$$\mathcal{S}_{dist} = \text{pesq}(\underline{s}_{N_1:N_2}^{ref}, \text{pr}(\underline{s}_{N_1:N_2})) \tag{27}$$

of the clean reference signal $\underline{s}_{N_1:N_2}^{ref}$ and the processed speech image $\text{pr}(\underline{s}_{N_1:N_2})$. In addition, to assess the general performance during double-talk we introduce the PESQ

$$\mathcal{S} = \text{pesq}(\underline{s}_{N_1:N_2}^{ref}, \text{pr}(\underline{x}_{N_1:N_2})) \tag{28}$$

with $\text{pr}(\underline{x}_{N_1:N_2})$ being the near-end speech estimate after the various speech enhancement components. All performance measures have been averaged over 50 different scenarios (8.3min) which is denoted by an overbar. Tab. I shows the average performance measures during single-talk and double-talk in comparison to an oracle baseline, i.e., using the first R coefficients of the true RIRs $\underline{h}_{p,\kappa}$ as AEC, the oracle signals for computing the CPSD matrices of the MVDR BF and the oracle magnitude ratio mask as spectral PF. We conclude that the cancellation of the early echo by the DNN-controlled AEC

Algorithm		Single-talk		Double-talk			
		$\bar{\mathcal{E}}$	$\bar{\mathcal{N}}$	$\bar{\mathcal{E}}$	$\bar{\mathcal{N}}$	$\bar{\mathcal{S}}_{\text{dist}}$	$\bar{\mathcal{S}}$
Unprocessed		—	—	—	—	—	1.20
Proposed	AEC	9.8	—	16.5	—	—	1.83
	AEC+BF	17.9	8.9	20.8	5.7	3.57	2.12
	AEC+BF+PF	37.9	26.3	23.8	8.0	3.00	2.13
Oracle	AEC	19.9	—	19.6	—	—	1.90
	AEC+BF	29.1	9.7	28.5	6.4	3.48	2.35
	AEC+BF+PF	60.3	36.9	34.2	12.3	3.46	3.22

results in a PESQ gain of 0.63 which is close to the oracle baseline. The subsequent MVDR BF increases the PESQ by additional 0.29. This improvement is obtained by the increased echo and noise suppression while introducing only minimal distortion to the near-end speech signal (cf. $\bar{\mathcal{S}}_{\text{dist}}$ for double-talk and MVDR BF). The spectral PF achieves almost no additional PESQ gain during double-talk as the improved interference suppression is traded against increased speech distortion. Yet, the PF obtains remarkable echo and noise suppression during the single-talk period at the beginning. This is of particular importance due to the limited performance of the AEC and MVDR BF during the initial convergence phase. Note that the partly overlapping convergence phase and single-talk period explain also the different interference suppression performance of the AEC and MVDR BF during single-talk and double-talk. Finally, as the average runtime of the proposed algorithm for processing one signal innovation block of duration 64ms on an *IntelXeon CPU-E3-1275 v6@3.80GHz* is 3ms, we confirm real-time capability on such platforms.

V. CONCLUSION

In this paper, we introduced a novel online acoustic echo and noise reduction algorithm with a DNN-based joint control of acoustic echo cancellation, MVDR beamforming and post-filtering as most distinctive feature. The proposed algorithm achieves rapid convergence and robust steady-state performance in the presence of high-level interfering double-talk.

REFERENCES

[1] 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.

[2] J. Valin et al., "Low-Complexity, Real-Time Joint Neural Echo Control and Speech Enhancement Based On Percepnet," in *Int. Conf. Acoust., Speech, Signal Process.*, Toronto, CA, June 2021, pp. 7133–7137.

[3] 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.

[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, June 2006.

[5] S. Malik et al., "Double-talk Robust Multichannel Acoustic Echo Cancellation using Least-Squares MIMO Adaptive Filtering: Transversal, Array, and Lattice Forms," *IEEE Trans. Signal Process.*, 2020.

[6] N. L. Westhausen and B. T. Meyer, "Acoustic echo cancellation with the dual-signal transformation lstm network," in *Int. Conf. Acoust., Speech, Signal Process.*, Toronto, CA, June 2021, pp. 7138–7142.

[7] Mhd. M. 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.

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

[9] 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.

[10] T. Haubner et al., "A Synergistic Kalman- and Deep Postfiltering Approach to Acoustic Echo Cancellation," in *European Signal Process. Conf.*, Dublin, IR, Aug. 2021, pp. 990–994.

[11] 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, pp. 766–770.

[12] A. Ivry et al., "Deep adaptation control for acoustic echo cancellation," in *Int. Conf. Acoust., Speech, Signal Process.*, Singapore, SG, May 2022, pp. 741–745.

[13] W. Herbordt et al., "Joint Optimization of LCMV Beamforming and Acoustic Echo Cancellation for Automatic Speech Recognition," in *Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, USA, March 2005.

[14] 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, Mar. 2019.

[15] 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.

[16] 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, Dec. 2021.

[17] G. Carbajal et al., "Joint NN-Supported Multichannel Reduction of Acoustic Echo, Reverberation and Noise," *IEEE Audio, Speech, Language Process.*, vol. 28, pp. 2158–2173, 2020.

[18] H. Zhang and D. Wang, "Multi-Channel and Multi-Microphone Acoustic Echo Cancellation Using A Deep Learning Based Approach," in *Annual Conf. of the Int. Speech Comm. Assoc.*, Brno, Czechia, Aug. 2021.

[19] S. Chakrabarty and E. A. P. Habets, "Time–Frequency Masking Based Online Multi-Channel Speech Enhancement With Convolutional Recurrent Neural Networks," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 4, pp. 787–799, Aug. 2019.

[20] V. Kothapally et al., "Joint AEC and Beamforming with Double-Talk Detection using RNN-Transformer," *arXiv preprint:2111.04904*, 2021.

[21] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, NJ, USA, 2002.

[22] J. J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Process. Mag.*, vol. 9, no. 1, pp. 14–37, 1992.

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

[24] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, third edition, 1996.

[25] C. Boeddeker et al., "Convolutive Transfer Function Invariant SDR Training Criteria for Multi-Channel Reverberant Speech Separation," in *Int. Conf. Acoust., Speech, Signal Process.*, Toronto, CA, June 2021, pp. 8428–8432.

[26] Y. Xia et al., "Weighted Speech Distortion Losses for Neural-Network-Based Real-Time Speech Enhancement," in *Int. Conf. Acoust., Speech, Signal Process.*, Barcelona, ES, May 2020, pp. 871–875.

[27] 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, 1979.

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

[29] 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.

[30] J. Barker et al., "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *IEEE Autom. Speech Recognit. Underst. Workshop*, Scottsdale, USA, Dec 2016, pp. 504–511.

[31] E. A. P. Habets et al., "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Am.*, vol. 124, pp. 2911–2917, 2008.

[32] ITU-T Recommendation P.862.2, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Recommendation, ITU, Nov. 2007.