ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414868

COMBINING ADAPTIVE FILTERING AND COMPLEX-VALUED DEEP POSTFILTERING FOR ACOUSTIC ECHO CANCELLATION

Mhd Modar Halimeh, Thomas Haubner*, Annika Briegleb*, Alexander Schmidt*, Walter Kellermann*

Multimedia Communications and Signal Processing,
Friedrich-Alexander University Erlangen-Nürnberg, Cauerstr. 7, 91058 Erlangen, Germany,
mhd.m.halimeh@fau.de

ABSTRACT

In this contribution, we introduce a novel approach to noise-robust acoustic echo cancellation employing a complex-valued Deep Neural Network (DNN) for postfiltering. In a first step, early linear echo components are removed using a double-talk robust adaptive filter. The residual signal is subsequently processed by the proposed postfilter (PF). Due to its complex-valued nature, the PF allows to suppress unwanted signal components without introducing distortions to the near-end speaker. For training and evaluation, we exclusively use data from the ICASSP 2021 AEC challenge. Exploiting only a moderate amount of training data, we demonstrate the efficacy of the proposed method. Specifically, we show that the PF (i) benefits significantly from a preceding linear adaptive filter and (ii) significantly outperforms a conventional real-valued DNN-based PF.

Index Terms— Acoustic echo cancellation, Postfiltering, Complex neural network

1. INTRODUCTION

Driven by the ever increasing use of hands-free communication in our daily life, Acoustic Echo Cancellation (AEC) has been a highly active research field for several decades [1, 2, 3, 4, 5]. More specifically and as a by-product of modern applications, new problems are becoming increasingly common for AEC, e.g., signal alterations resulting from codecs and other signal processing incurring during transmission over networks, when AEC is carried out remotely rather than at the end-user's device.

For conventional linear echo cancellation, usually an adaptive FIR filter is employed to model the room acoustics such that the echo components are estimated and then removed from the microphone signal, c.f., [1, 2]. However, the adaptation of the FIR filter is often challenged by local disturbances such as near-end speakers. While these challenges have traditionally been addressed using double talk detection-based techniques [6], more recent approaches were able to achieve noise- and double talk-robust adaptation using, e.g., Bayesian filtering [7], error enhancement approaches [8], or semi-blind source separation methods [9]. In addition to local disturbances, conventional AEC algorithms are also impeded by nonlinearities in the echo path due to nonlinear components such as miniaturized loudspeakers. This can be addressed by modelling the echo path using nonlinear structures such as Hammerstein models [10, 11, 12]. Moreover, linear echo cancellation is often limited due to the use of insufficiently long finite impulse response (FIR) filters. Hence, these and other echoes that are not modelled by the adaptive filter remain in the residual signal in addition to potential background noises. Therefore, a postfilter (PF), also referred to as a residual echo suppressor, is often applied to the residual signal to achieve improved echo and noise suppression. Traditional approaches employ linear Minimum Mean Square Error (MMSE)-based PFs [13, 14, 15] for which a reliable estimation of the residual echo power spectral density (PSD) is of fundamental importance [16, 17]. More recent approaches for PFs are dominated by data-driven techniques such as deep learning. One of the first approaches employing a neural network [18] proposed a feed-forward neural network with only three hidden units per frequency bin, showing a significant improvement of residual echo suppression. Several further ideas emerged from there, which exploited more sophisticated network architectures, such as a restricted Boltzmann machine [19], an encoder-decoder structure with attention [20] or larger dense layers with multiple input signals [21]. In [21], it is also emphasized that using the echo estimate generated using the linear adaptive filter or using the far-end signal together with the residual signal leads to an improvement compared to using only one of these signals. Other approaches tackle AEC with neural networks exclusively, e.g., in [22, 23, 24] where AEC is treated as a source separation problem.

The proposed system comprises two stages: first, a double talk-robust adaptive filter is applied to remove the linear echo components. Second, and as the main contribution of this paper, we propose a deep neural network for postfiltering which estimates a complex-valued mask. It provides a phase-aware processing of its input signal allowing for echo and noise suppression with minimal distortion of the desired near-end speech signal. Note that, while AEC classically refers to the estimation and subsequent subtraction of echo signals, we will refer to the entirety of the proposed system, including the postfilter, as an AEC system, similarly to [25]. For evaluation, we use data from the ICASSP 2021 AEC challenge [25], which includes a wide variety of effects and scenarios, e.g., nonlinear distortions, large network delays, diverse background noises, and microphone clipping effects. We first demonstrate that the Deep Neural Network (DNN)-based PF benefits significantly from a preceding linear echo cancellation stage. Furthermore, we show that the proposed PF outperforms a DNN-based PF which estimates a real-valued mask only. The proposed system is trained with only a moderate amount of training data exclusively from the provided challenge data corpus. We provide a detailed description of the network architecture including all parameters for reproducibility of the results.

2. PROPOSED SYSTEM WITH COMPLEX-VALUED DNN FOR POSTFILTERING

An overview of the proposed system is given in Fig. 1. The choice of this generic system is motivated by the large variety of different recording setups within the challenge dataset. We therefore re-

* These authors contributed equally to this work.

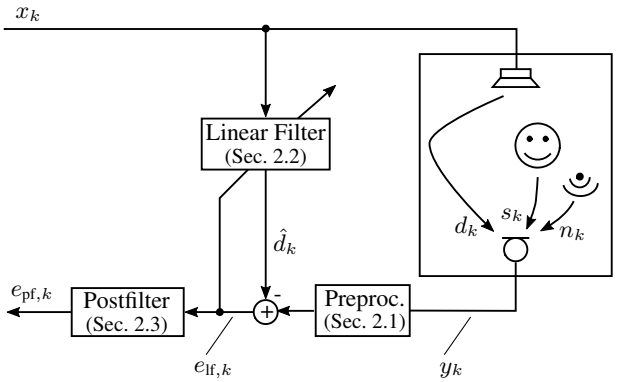

Fig. 1. Overview of proposed system.

nounced on incorporating more sophisticated models for exploiting further prior knowledge on the echo path, e.g., nonlinearities [10, 11, 12]. For the following, note that all frequency-domain quantities are underlined to distinguish them from their time-domain counterparts.

2.1. Signal Model

We model the discrete time-domain microphone signal y_k at sample index k as a linear superposition of the echo signal d_k, the near-end interferer s_k, background noise n_k and a time-varying bias $y_{\text{bias},k}$, which results from potential hardware imperfections:

$$y_k = d_k + s_k + n_k + y_{\text{bias},k} \tag{1}$$

The echo signal d_k is composed of the linear and non-linear components $d_{\text{lin},k}$ and $d_{\text{nl},k}$:

$$d_k = d_{\text{lin},k} + d_{\text{nl},k} = \boldsymbol{h}_k^{\text{T}} \boldsymbol{x}_k + d_{\text{nl},k}, \tag{2}$$

where, the linear component is modelled by a linear convolution of the far-end signal $\boldsymbol{x}_k = \begin{pmatrix} x_{k-L_{\text{RIR}}+1} & \dots & x_k \end{pmatrix}^{\text{T}} \in \mathbb{R}^{L_{\text{RIR}}}$ with the FIR filter $\boldsymbol{h}_k \in \mathbb{R}^{L_{\text{RIR}}}$. For the sake of computational efficiency, the proposed acoustic echo cancellation system (cf. Fig. 1) jointly processes blocks of R samples

$$\boldsymbol{y}_\tau = \boldsymbol{d}_{\text{lin},\tau} + \boldsymbol{d}_{\text{nl},\tau} + \boldsymbol{s}_\tau + \boldsymbol{n}_\tau + \boldsymbol{y}_{\text{bias},\tau} \tag{3}$$

with the vector $\boldsymbol{y}_\tau = \begin{pmatrix} y_{\tau R-R+1}, \dots, y_{\tau R} \end{pmatrix}^{\text{T}}$ representing a length-R block of the microphone signal. To remove the time-varying bias component $y_{\text{bias},k}$ from the microphone signal, the preprocessing block (cf. Fig. 1) subtracts a moving average bias estimate $\hat{y}_{\text{bias},k}$ from y_k, which is obtained as the arithmetic average of the previous L_{br} samples. Note that in the following all other signal quantities in Eq. (3) are defined analogously and their respective bias-reduced counterparts are denoted by $()_{\text{br}}$.

2.2. Linear Adaptive Filter Unit

The linear adaptive filter unit aims at cancelling the early reflections $d_{\text{early},k} = \left(\begin{bmatrix} \boldsymbol{I}_L & \boldsymbol{0}_{L\times(L_{\text{RIR}}-L)} \end{bmatrix} \boldsymbol{h}_k \right)^{\text{T}} \boldsymbol{x}_k$ in the linear echo component $d_{\text{lin},k}$, i.e., the echo resulting from the first $L \leq L_{\text{RIR}}$ taps of the linear propagation path model $\boldsymbol{h}_k$, by subtracting an echo estimate $\hat{\boldsymbol{d}}_\tau$ from the bias-reduced microphone signal block $\boldsymbol{y}_{\text{br},\tau}$

$$\boldsymbol{e}_{\text{lf},\tau} = \boldsymbol{y}_{\text{br},\tau} - \hat{\boldsymbol{d}}_\tau. \tag{4}$$

Due to its computational efficiency and low algorithmic delay, we use a Discrete Fourier Transform (DFT)-domain partitioned block convolution model as a linear echo estimator

$$\hat{\boldsymbol{d}}_\tau = \sum_{b=0}^{B-1} \boldsymbol{Q}_1^{\text{T}} \boldsymbol{F}_M^{-1} \underline{\boldsymbol{X}}_{\tau-b} \underline{\hat{\boldsymbol{h}}}_{b,\tau-1} \tag{5}$$

with the DFT domain adaptive filter vector of the b-th partition $\underline{\hat{\boldsymbol{h}}}_{b,\tau-1} \in \mathbb{C}^M$ and the DFT-domain far-end signal matrix $\underline{\boldsymbol{X}}_\tau$. The latter one is computed by $\underline{\boldsymbol{X}}_\tau = \text{diag}\left(\boldsymbol{F}_M \boldsymbol{x}_\tau\right) \in \mathbb{C}^{M\times M}$ with $\boldsymbol{x}_\tau = \begin{pmatrix} x_{\tau R-M+1}, \dots, x_{\tau R} \end{pmatrix}^{\text{T}} \in \mathbb{R}^M$ as the respective time-domain far-end signal block, $\boldsymbol{F}_M$ as the M-dimensional DFT matrix and $\text{diag}(\cdot)$ as the diagonalization operator. Furthermore, the constraint matrix $\boldsymbol{Q}_1^{\text{T}} = \begin{pmatrix} \boldsymbol{0}_{R\times M-R} & \boldsymbol{I}_R \end{pmatrix}$, with $\boldsymbol{I}_R$ and $\boldsymbol{0}_{R\times S}$ denoting the identity matrix and all-zero matrix respectively, ensures a linear convolution of the far-end signal with the estimated FIR filter.

Due to its double talk-robust adaptation performance, the adaptive filter partitions $\underline{\hat{\boldsymbol{h}}}_{b,\tau}$ with $b = 0, \dots, B-1$ are updated by the gradient-constrained version of the diagonalized Partitioned-Block Kalman Filter (PBKF) proposed in [26]. Note that, instead of the microphone signal, the bias-reduced version $\boldsymbol{y}_{\text{br},\tau}$ is used as noisy observation (cf. Fig. 1). The noise-robustness of the Kalman filter-based adaptation decisively depends on a precise estimation of the observation noise covariance matrix $\underline{\boldsymbol{\Psi}}_\tau^{SS} = \mathbb{E}\left[\underline{\tilde{\boldsymbol{n}}}_\tau \underline{\tilde{\boldsymbol{n}}}_\tau^*\right]$ with $\mathbb{E}[\cdot]$ as the expectation operator. It represents the covariance matrix of the DFT-domain signal vector $\underline{\tilde{\boldsymbol{n}}}_\tau = \boldsymbol{F}_M \boldsymbol{Q}_1 \left(\boldsymbol{y}_{\text{br},\tau} - \boldsymbol{d}_{\text{early},\tau} \right)$ comprising all signal components which cannot be explained by the linear adaptive filter model. By assuming spectral uncorrelatedness, the observation noise covariance matrix $\underline{\boldsymbol{\Psi}}_\tau^{SS}$ is modelled to be diagonal and thus allows for a computationally efficient implementation.

For modelling the observation noise, we use an Expectation-Maximization (EM)-inspired optimization scheme [27] in which the Kalman filter update acts as E-step and the estimation of the observation and process noise covariance matrices represent the M-step. In contrast to [27], we suggest to estimate both the observation noise covariance matrix $\underline{\boldsymbol{\Psi}}_\tau^{SS}$ and the process noise covariance matrix $\underline{\boldsymbol{\Psi}}_{b,\tau}^{\Delta\Delta}$ of the b-th partition by

$$\left[\underline{\boldsymbol{\Psi}}_\tau^{SS} \right]_{mm} = \hat{\mathbb{E}} \left[\left(\underline{\boldsymbol{e}}_{\text{post},\tau} \underline{\boldsymbol{e}}_{\text{post},\tau}^{\text{H}} \right) \right]_{mm} \tag{6}$$

$$\left[\underline{\boldsymbol{\Psi}}_{b,\tau}^{\Delta\Delta} \right]_{mm} = (1 - A^2)\, \hat{\mathbb{E}} \left[\left(\underline{\hat{\boldsymbol{h}}}_{b,\tau} \underline{\hat{\boldsymbol{h}}}_{b,\tau}^{\text{H}} \right) \right]_{mm} \tag{7}$$

with $\hat{\mathbb{E}}[\cdot]$ denoting recursive averaging. Here, the posterior error is calculated by $\underline{\boldsymbol{e}}_{\text{post},\tau} = \underline{\boldsymbol{y}}_{\text{br},\tau} - \sum_{b=0}^{B-1} \boldsymbol{C}_{\tau-b} \underline{\hat{\boldsymbol{h}}}_{b,\tau}$ with $\boldsymbol{C}_{\tau-b} = \boldsymbol{F}_M \boldsymbol{Q}_1 \boldsymbol{Q}_1^{\text{T}} \boldsymbol{F}_M^{-1} \underline{\boldsymbol{X}}_{\tau-b}$ being the overlap-save constrained far-end signal block. As a sequential application of a single E-step, i.e., updating the adaptive filter partitions $\underline{\hat{\boldsymbol{h}}}_{b,\tau-1} \in \mathbb{C}^M$ with $b = 1, \dots, B$, followed by a single M-step, would not take into account the current observation $\boldsymbol{y}_{\text{br},\tau}$ for the estimated observation noise covariance matrix $\underline{\boldsymbol{\Psi}}_\tau^{SS}$ that is used in the Kalman stepsize ([26] Eq. (26)), at least two EM steps are required.

2.3. Complex-valued Deep Neural Network for Postfiltering

2.3.1. Network architecture

The proposed PF uses a complex-valued DNN architecture that is a modified version of the complex U-net introduced in [28] for speech enhancement. The proposed architecture is depicted in Fig. 2 and it comprises a complex-valued autoencoder. Furthermore, to enable the network to model time dependencies beyond

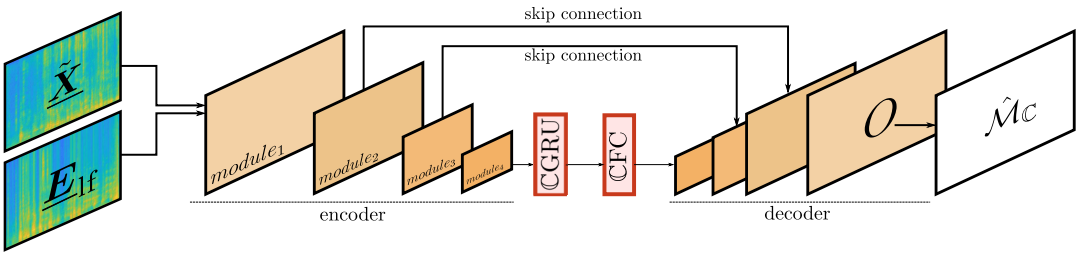

Fig. 2. The proposed PF complex-valued DNN.

the input frames' temporal range, we introduce a complex Gated Recurrent Unit (GRU), denoted by ℂGRU, and a complex Fully Connected (FC) layer, denoted by ℂFC, between the encoder and the decoder. This is inspired by [29], where a complex-valued Long Short-Term Memory (LSTM) layer is used instead of the complex-valued GRU for speech enhancement.

Both the encoder and the decoder consist of four complex-valued modules constructed as a complex two-dimensional convolutional layer followed by a complex batch normalization [30] and a complex-valued leaky Rectified Linear Unit (ReLU) [28]. Moreover, skip connections were employed between the intermediate modules as shown in Fig. 2. The complex GRU after the encoder comprises two traditional real-valued GRUs denoted GRU_r and GRU_i. The complex output of ℂGRU for a complex input $z = a + ib$ is then computed from

$$\mathbb{C}\text{GRU}(z) = (\text{GRU}_r(a) - \text{GRU}_i(b)) + i\,(\text{GRU}_r(b) + \text{GRU}_i(a))\,. \tag{8}$$

It should be noted that Eq. (8) suggests that the complex GRU is a linear function, which, naturally, is not the case. Nevertheless, Eq.(8) allows for an easy implementation of backpropagation algorithms and it remains of interest for future work to find better approximations.

As an input to the network an *image* of two channels is used. Two windowed Short-Time Fourier Transform (STFT) frames of the far-end signal x_k are arranged as the input's first channel $\underline{\boldsymbol{X}}_\tau = \left[\underline{\boldsymbol{x}}_\tau, \underline{\boldsymbol{x}}_{\tau-1}\right]$, while two STFT frames of the residual signal $e_{\text{lf},k}$ are arranged as the input's second channel $\underline{\boldsymbol{E}}_{\text{lf},\tau} = \left[\underline{\boldsymbol{e}}_{\text{lf},\tau}, \underline{\boldsymbol{e}}_{\text{lf},\tau-1}\right]$. For each time-frequency bin (τ, f), the network outputs the complex-valued unprocessed mask $O_{\tau,f}$ as shown in Fig. 2. This mask is then processed to obtain a bounded complex-valued mask $\hat{\mathcal{M}}_\mathbb{C}$, where $\hat{\mathcal{M}}_{\mathbb{C},\tau,f} = |\hat{\mathcal{M}}_{\mathbb{C},\tau,f}|e^{i\theta_{\tau,f}}$ is calculated as follows [28]

$$|\hat{\mathcal{M}}_{\mathbb{C},\tau,f}| = \tanh(|O_{\tau,f}|), \tag{9}$$

and

$$e^{i\theta_{\tau,f}} = \frac{O_{\tau,f}}{|O_{\tau,f}|}. \tag{10}$$

Using the complex-valued mask $\hat{\mathcal{M}}_\mathbb{C}$, an STFT-domain near-end speech signal's estimate is obtained by

$$\underline{\hat{\boldsymbol{s}}}_\tau = \hat{\mathcal{M}}_\mathbb{C} \odot \underline{\boldsymbol{e}}_{\text{lf},\tau}, \tag{11}$$

where $\odot$ denotes the Hadamard product operation.

2.3.2. *Training target and loss function*

The proposed PF is trained to estimate the clean near-end speech signal s_k. This is done by employing a block-wise variant of the weighted-Signal-to-Distortion Ratio (SDR) loss introduced in [28]

$$\mathcal{J}(\boldsymbol{e}_{\text{lf},\tau}, \boldsymbol{s}_\tau, \hat{\boldsymbol{s}}_\tau) = \alpha_\tau \frac{-\boldsymbol{s}_\tau^\text{T}\hat{\boldsymbol{s}}_\tau}{\|\boldsymbol{s}_\tau\|\,\|\hat{\boldsymbol{s}}_\tau\|} + (1-\alpha_\tau)\frac{-\boldsymbol{n}_\tau^\text{T}\hat{\boldsymbol{n}}_\tau}{\|\boldsymbol{n}_\tau\|\,\|\hat{\boldsymbol{n}}_\tau\|}, \tag{12}$$

where $\|\cdot\|$ denotes the Euclidean norm and $\hat{\boldsymbol{s}}_\tau$ denotes the time-domain estimate of the near-end speech signal obtained by means of the Inverse Short-Time Fourier Transform (ISTFT) of the estimate $\underline{\hat{\boldsymbol{S}}}_\tau$, while α_τ denotes the energy ratio

$$\alpha_\tau = \frac{\|\boldsymbol{s}_\tau\|^2}{\|\boldsymbol{e}_{\text{lf},\tau}\|^2}. \tag{13}$$

Furthermore, $\boldsymbol{n}_\tau = \boldsymbol{e}_{\text{lf},\tau} - \boldsymbol{s}_\tau$ denotes the undesired signal components while $\hat{\boldsymbol{n}}_\tau = \boldsymbol{e}_{\text{lf},\tau} - \hat{\boldsymbol{s}}_\tau$ denotes its estimate. Consequently, the network is trained to suppress not only the residual echoes resulting from late reflections, but also any potential nonlinear distortions as well as background noise.

It is worth noting that using a two-components loss function, consisting of a desired source-related component and a noise-related one, balanced adaptively by the energy ratio given in Eq. (13), allows for continuous learning even for noise-only frames.

3. RESULTS

In this section, we will evaluate the proposed algorithm by using objective and subjective performance measures[1] and discuss in detail the benefits of the individual algorithmic parts, i.e., linear adaptive filter, PF and the combination of both. As objective performance measures we use signal power ratios of the type

$$g(\boldsymbol{a}, \boldsymbol{b}) = 10\log_{10}\left(\frac{||\boldsymbol{a}||^2}{||\boldsymbol{b}||^2}\right) \tag{14}$$

with the signal vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ containing the entire signal, i.e., $\boldsymbol{a} = \begin{pmatrix} a_1 & \dots & a_K \end{pmatrix}^\text{T}$, with a typical duration of 10 s. To quantify the effects of the different algorithmic parts on the microphone signal $\boldsymbol{y}$, we introduce the echo cancellation measure $\mathcal{E}_\text{lf}$ and the echo suppression measure $\mathcal{E}_\text{pf}$, which correspond to the conventional Echo Return Loss Enhancement (ERLE) contributions of the linear filter and the postfilter, respectively. Suppression of local noise and near-end distortion are measured by $\mathcal{N}_\text{pf}$ and the scale-invariant measure $\mathcal{S}_\text{pf}$, respectively:

$$\begin{aligned} \mathcal{E}_\text{lf} &= g(\boldsymbol{d}_\text{br}, \boldsymbol{d}_\text{br} - \hat{\boldsymbol{d}}), & \mathcal{N}_\text{pf} &= g(\boldsymbol{n}, \text{pf}(\boldsymbol{n})), \\ \mathcal{E}_\text{pf} &= g(\boldsymbol{d}_\text{br}, \text{pf}(\boldsymbol{d}_\text{br} - \hat{\boldsymbol{d}})), & \mathcal{S}_\text{pf} &= g(\beta\boldsymbol{s}, \beta\boldsymbol{s} - \text{pf}(\boldsymbol{s})), \end{aligned} \tag{15}$$

Note that $\text{pf}(\cdot)$ describes the processing of the respective argument by the PF and $\beta = \frac{\boldsymbol{s}^\text{T}\text{pf}(\boldsymbol{s})}{||\boldsymbol{s}||^2}$ is a scaling factor [31]. Besides the signal power-based performance measures given in Eq. (15), we also evaluate the PESQ (Perceptual Evaluation of Speech Quality [32]) gain

$$\Delta\text{PESQ} = \text{PESQ}(\boldsymbol{s}, \tilde{\boldsymbol{s}}) - \text{PESQ}(\boldsymbol{s}, \boldsymbol{y}) \tag{16}$$

[1] Implementation and audio samples are provided at `https://github.com/LMSAudio/Complex_PF`

Table 1. Performance of the various AEC approaches, given by the respective mean and standard deviation (in parentheses). Note that PF-only approaches are trained without the linear filter unit, i.e., $\boldsymbol{e}_{\text{lf},\tau}$ is replaced by $\boldsymbol{y}_{\text{br},\tau}$ at the input of the networks.

		Real-valued DNN		Complex-valued DNN	
	linear filter-only	PF-only	linear filter + PF	PF-only	linear filter + PF (Proposed)
t_{pr}[ms] / RTF	**0.7**/**0.05**	4.5/0.34	5.2/0.39	7.6/0.57	8.3/0.63
$\mathcal{E}_{(\cdot)}$	10.3 (5.6)	11.8 (5.6)	**20.2** (8.3)	11.7 (4.0)	18.0 (7.2)
$\mathcal{N}_{\text{pf}}$	0.0 (0.0)	**9.0** (3.9)	8.2 (3.8)	5.4 (2.9)	3.4 (1.9)
$\mathcal{S}_{\text{pf}}$	∞ (0.0)	9.5 (4.7)	14.0 (4.8)	14.5 (4.4)	**24.9** (6.1)
ΔPESQ	0.7 (0.6)	0.0 (0.1)	0.4 (0.4)	0.3 (0.3)	**1.1** (0.7)
MOS	—	—	—	—	3.48

with $\tilde{\boldsymbol{s}}$ being the processed signal at the output of the algorithm.

In all experiments we use a frame length of $M = 424$ taps with a frame shift of $R = \frac{M}{2}$ at a sampling frequency of $f_s = 16$ kHz. Concerning the preprocessing, the length of the mean subtraction filter is set to $L_{\text{br}} = 1024$. The linear adaptive filter (cf. Sec. 2.2) models $B = 9$ partitions which results in an overall filter length of 1908 taps. We conduct 2 EM steps with a state transition factor of $A = 0.9999$ and the recursive averaging factors of both the process and the observation noise estimator being set to 0.9. As for the PF, the convolutional layers in $\{module_1, module_2, module_3, module_4\}$ on the encoder side have $\{32, 32, 64, 32\}$ output channels and use kernels with sizes of $\{(7,2),(7,2),(7,2),(5,2)\}$ and strides of $\{(2,2),(2,1),(2,2),(2,1)\}$. The decoder's modules are configured similarly in a reversed order except for the last module which has a single output channel. As a result, the overall network has approximately 1.8 M adaptive parameters. The network was trained using ten hours of the training data provided by [25] where seven hours were sampled randomly from the synthetic dataset, while three hours were sampled randomly from the real dataset. The training was conducted using the truncated backpropagation through time algorithm where the truncated time sequence was limited to 15 frames. Moreover, the Adam optimizer [33] was used to adapt the network's weights with a learning rate of 10^{-3}.

We compare the proposed complex-valued PF to a real-valued one. Its design is adapted from the noise suppression network proposed in [34]. This PF uses one dense layer as a feature extraction layer followed by two stacked GRU layers and a dense output layer. As input features, we use the logarithmic power spectrum of the residual signal block $\boldsymbol{e}_{\text{lf},\tau}$ and of the corresponding far-end signal block $\boldsymbol{x}_\tau$ normalized to zero-mean and unit standard deviation. The resulting real-valued PF has approximately 2.8 M adaptive parameters. As cost function for training, we use a modified version of the two-components loss proposed in [35], weighted time-frequency bin-wise similar to Eq. (13). Furthermore, we normalize each component by the magnitude of the respective clean signal component. This way, the overall cost function is no longer biased towards the component with the higher energy contribution, which might be a problem when one source is significantly stronger. A local SDR-based weight is introduced by our choice of α (cf. Eq. (13)). To increase comparability, we use the same data, optimizer and learning rate as for the complex-valued PF for training.

To allow for more general conclusions, all of the above mentioned performance measures are averaged over 100 experiments which are randomly drawn from the synthetic test set, which is disjoint from the training set, introduced in [25]. Tab. 1 summarizes the respective mean and standard deviation of the experiments. Furthermore, we report the mean MOS score as evaluated for the AEC challenge, which is based on the blind test set of the provided data. We compute the presented evaluation measures after the linear filter (*linear filter-only*) and the PF (*linear filter + PF*). For further comparison, we trained both PFs with far-end and microphone signal as input (*PF-only*), i.e., no linear adaptive filter is used.

As can be concluded from Tab. 1, the combination of a linear adaptive filter unit with a DNN-based PF outperforms the individual approaches, i.e., linear filter-only or PF-only, for both the real-valued PF and the complex-valued PF. More specifically, using an adaptive linear filter only, without any PF, results in limited cancellation of the echoes as it can only model the first part of the echo path. On the other hand, using a PF-only system resulted in a good echo and noise suppression at the expense of significantly distorting the desired near-end source. This can be explained by the fact that the PF is effectively extracting a source signal from a mixture with low Signal to Echo Ratio (SER), resulting in suppression masks that are more aggressive, i.e., sparse, than those of the linear filter + PF case.

Moreover, we can observe that the complex-valued PF, alone and in combination with the linear filter, distinctly outperforms the real-valued PF in terms of speech distortion and ΔPESQ given similar echo suppression performance. This highlights the benefit of estimating a complex-valued mask where both the signal's phase and amplitude are reconstructed. And while the real-valued PF outperforms the proposed complex-valued PF in terms of noise suppression, this gain is achieved via aggressive masking that heavily distorts the desired source signals.

A comparison between the previous results and the moderate MOS score, which is evaluated on the blind test dataset as part of the AEC challenge [25], points to a discrepancy between our training data and the blind test dataset, which was not observed for the synthetic test set. Unfortunately, objective performance measures could not be computed for the blind test dataset due to the absence of references for clean speech and noise signals.

Finally, Tab. 1 also provides the block processing runtime t_{pr}, i.e., the runtime to process one signal block $\boldsymbol{y}_\tau$ of duration 13.25 ms, and the corresponding real time factor RTF of the various algorithmic variants on an *Intel(R) Xeon(R) CPU E3-1275 v6*.

4. CONCLUSION

In this contribution, we introduced a complex-valued DNN for postfiltering and evaluated it on the ICASSP 2021 AEC challenge dataset. The proposed network leads to a significantly lower speech distortion compared to a real-valued DNN-based PF, while achieving similar echo suppression performance. We could furthermore show that, given a limited amount of training data, the proposed approach shows special efficacy if combined with a state-of-the-art adaptive linear echo path model.

5. REFERENCES

[1] M. M. Sondhi, "An adaptive echo canceller," *Bell Syst. Tech. J.*, vol. 46, no. 3, pp. 497–511, Mar. 1967.

[2] C. Breining et al., "Acoustic echo control. An application of very-high-order adaptive filters," *IEEE Signal Processing Mag.*, vol. 16, no. 4, pp. 42–69, July 1999.

[3] E. Hänsler and G. Schmidt, "Hands-free telephones: Joint control of echo cancellation and postfiltering," *Signal Process.*, vol. 80, no. 11, pp. 2295–2305, Nov. 2000.

[4] G. Schmidt E. Hänsler, *Topics in Acoustic Echo and Noise Control*, Springer-Verlag, Heidelberg, Germany, 2006.

[5] S. Theodoridis and R. Chellappa, *Academic press library in signal processing, volume 4: Image, video processing and analysis, hardware, audio, acoustic and speech processing*, Academic Press, Inc., Orlando, FL, USA, 1st edition, 2014.

[6] J. Benesty et al., "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech, Audio Process.*, vol. 8, no. 2, pp. 168–172, Mar. 2000.

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

[8] T. Wada and B. Juang, "Enhancement of residual echo for robust acoustic echo cancellation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 175–189, Jan. 2012.

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

[10] F. Kuech et al., "Nonlinear acoustic echo cancellation using adaptive orthogonalized power filters," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Philadelphia, PA, Mar. 2005, pp. iii/105–iii/108.

[11] A. Carini et al., "Introducing Legendre nonlinear filters," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 7939–7943.

[12] M. M. Halimeh et al., "A neural network-based nonlinear acoustic echo canceller," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1827–1831, Dec. 2019.

[13] S. Gustafsson et al., "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Process.*, vol. 64, no. 1, pp. 21–32, Jan. 1998.

[14] R. Martin and J. Altenhoner, "Coupled adaptive filters for acoustic echo control and noise reduction," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Detroit, MI, May 1995, vol. 5, pp. 3043–3046.

[15] R. Martin and S. Gustafsson, "The echo shaping approach to acoustic echo control," *Speech Commun.*, vol. 20, no. 3, pp. 181–190, Dec. 1996.

[16] G. Enzner et al., "Partitioned residual echo power estimation for frequency-domain acoustic echo cancellation and postfiltering," *European Trans. Telecommun.*, vol. 13, no. 2, pp. 103–114, 2002.

[17] G. Enzner et al., "Unbiased residual echo power estimation for hands-free telephony," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Orlando, FL, May 2002, vol. 2, pp. 1893–1896.

[18] A. Schwarz et al., "Spectral feature-based nonlinear residual echo suppression," in *Proc. IEEE Workshop Applicat. of Signal Process. to Audio and Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[19] C. M. Lee et al., "DNN-Based Residual Echo Suppression," in *Proc. Interspeech*, Dresden, Germany, Sept. 2015, pp. 1775–1779.

[20] A. Fazel et al., "CAD-AEC: Context-Aware Deep Acoustic Echo Cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 6919–6923.

[21] G. Carbajal et al., "Multiple-Input Neural Network-Based Residual Echo Suppression," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Calgary, AB, Apr. 2018, pp. 231–235.

[22] H. Zhang and D. Wang, "Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios," in *Proc. Interspeech*, Hyderabad, India, Sept. 2018, pp. 3239–3243.

[23] H. Zhang et al., "Deep Learning for Joint Acoustic Echo and Noise Cancellation with Nonlinear Distortions," in *Proc. Interspeech*, Graz, Austria, Sept. 2019, pp. 4255–4259.

[24] A. Fazel et al., "Deep Multitask Acoustic Echo Cancellation," in *Proc. Interspeech*, Graz, Austria, Sept. 2019, pp. 4250–4254.

[25] K. Sridhar et al., "ICASSP 2021 Acoustic Echo Cancellation Challenge: Datasets and Testing Framework," in *arXiv:2009.04972*, Sept. 2020.

[26] F. Kuech et al., "State-space architecture of the partitioned-block-based acoustic echo controller," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 1295–1299.

[27] S. Malik and G. Enzner, "Online maximum-likelihood learning of time-varying dynamical models in block-frequency-domain," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Dallas, TX, Mar. 2010, pp. 3822–3825.

[28] H. Choi et al., "Phase-aware speech enhancement with deep complex U-net," in *arXiv:1903.03107*, Feb. 2019.

[29] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *arXiv:2008.00264*, Aug. 2020.

[30] C. Trabelsi et al., "Deep complex networks," in *Proc. Int. Conf. Learning Representations*, Vancouver, BC, Feb. 2018.

[31] J. LeRoux et al., "SDR – Half-baked or Well Done?," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Brighton, UK, May 2019, pp. 626–630.

[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.

[33] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980v9 [cs.LG]*, Jan. 2017.

[34] Y. Xia et al., "Weighted Speech Distortion Losses for Neural-Network-Based Real-Time Speech Enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 871–875.

[35] Z. Xu et al., "Components Loss for Neural Networks in Mask-Based Speech Enhancement," *arXiv:1908.05087*, Aug. 2019.