Convergence and Performance Analysis of Classical, Hybrid, and Deep Acoustic Echo Control

Ernst Seidel, *Member, IEEE*, Pejman Mowlaee, *Senior Member, IEEE*, and Tim Fingscheidt, *Senior Member, IEEE*

***Abstract*—Acoustic echo cancellation (AEC) and suppression (AES) are widely researched topics. However, only few papers about hybrid or deep acoustic echo control provide a solid comparative analysis of their methods as it was common with classical signal processing approaches. There can be distinct differences in the behaviour of an AEC/AES model which cannot be fully represented by a single metric or test condition, especially when comparing classical signal processing and machine-learned approaches. These characteristics include convergence behaviour, reliability under varying speech levels or far-end signal types, as well as robustness to adverse conditions such as harsh nonlinearities, room impulse response switches or continuous changes, or delayed echo. A first contribution of this article is to present an extended set of *test conditions* and *metrics* that yields a proper characterization of an AEC/AES model and provides researchers with a useful toolbox to benchmark their systems. Second, we evaluate multiple AEC/AES models, each representing a classical, machine-learned, or hybrid paradigm, in various test conditions. We provide an analysis and new insights into their strengths and weaknesses and identify limitations of common metrics in some cases. Our entire toolbox of evaluation metrics and testing conditions is available on GitHub[1].**

***Index Terms*—Acoustic echo control, filter convergence, machine learning, performance analysis.**

I. Introduction

In TODAY'S world, speech communication via smartphones, personal computers, and other electronic devices is omnipresent. One important aspect hereby is acoustic echo control: If one of the devices in a communication system is a hands-free device or in any other way allows its loudspeaker signal to be picked up by its microphone, the speaker on the other end would hear their own voice as an echo. As this would degrade the perceived quality of the conversation, acoustic echo cancellation (AEC) or suppression (AES) systems, jointly referred to as echo control (EC) systems, aim at removing the echo with as few degradation of the near-end signal as possible.

Manuscript received 15 December 2023; revised 11 April 2024; accepted 3 May 2024. Date of publication 20 May 2024; date of current version 31 May 2024. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jan Skoglund. *(Corresponding author: Tim Fingscheidt.)*

Ernst Seidel and Tim Fingscheidt are with the Institute for communications Technology, Technische Universität Braunschweig, Schleinitzstraße 22, 38106 Braunschweig, Germany (e-mail: e.seidel@tu-bs.de; t.fingscheidt@tu-bs.de).

Pejman Mowlaee is with the GN Audio A/S, Lautrupbjerg 7, 2750 Ballerup, Denmark (e-mail: pmowlaee@jabra.com).

[1]https://github.com/ifnspaml/EC-Evaluation-Toolbox.

Digital Object Identifier 10.1109/TASLP.2024.3402552

For decades, classical methods based on adaptive filters have been successfully used for AEC. The most common approaches are the normalized least means square (NLMS) filter [1], [2], [3], [4] and the frequency-domain Kalman filter (FDKF) [5], which have been adapted and extended in many different ways over the past years [6], [7], [8], [9].

Recently, there has also been a great number of approaches employing deep neural networks (DNNs), either as a fully learned EC [10], [11], [12], [13], [14], as a residual (deep) echo suppression postfilter after a linear AEC [8], [15], [16], or as a hybrid approach combining parts of classical approaches with deep learning components [9], [17]. The most prominent recently published work is likely Microsoft's DeepVQE model [14], performing AES, noise suppression, and speech dereverberation in a single network, thereby presenting a solution to challenging tasks that previously required a multi-stage network for good near-end speech preservation [11], [12]. In our work, however, we focus on distinct approaches for echo control to solely enable convergence and performance analysis of the classic and DNN-based EC methods.

The capabilities and behavior of classical approaches are usually well documented, with metrics such as echo return loss enhancement (ERLE) [18] and system distance [4] plotted over time, allowing a detailed analysis of characteristics such as convergence and reconvergence time. Note that recently, tools for (classical) residual echo suppression have been released [19]. Although the evaluation of DNN-based approaches also allows to plot ERLE over time, its reporting is often limited to mean metric scores, which are sufficient to prove a method's good performance under the given conditions, but rarely can provide enough information to achieve a deeper understanding of the DNN's temporal behavior. Beyond that, however, there are many more useful metrics that are worth to remember.

The contribution of our work is not any *algorithmic* novelty, but in the first place, it is a comprehensive analysis of methods representing a wide range of EC paradigms: classical, hybrid, and deep EC. Rather than focusing on a comparison of sophisticated state-of-the-art models, our approach to in-depth analysis of EC systems is demonstrated on representative models with highly diverse methodology and behavior. We create a set of test conditions along with metrics, which analyzes each aspect of an EC system's performance in detail, whereby all metrics are also applicable to DNN or hybrid approaches. Among the test conditions, we also propose the rarely reported continuously

changing room impulse response condition. As a further example, two systems might show similar mean echo suppression scores, but differ in convergence time or sensitivity to speech and noise levels. Choosing the right metrics and test conditions, we show how to gain a better distinction between the exhibited behavior of different methods that might not be detectable in simpler evaluation setups. Last but not least, our synoptic view onto the three paradigms allows a deeper understanding of their pros and cons.

The article is structured as follows: In Section II, we define the task of acoustic echo control and our tested EC methods. Section III defines datasets, the DNN training process, and evaluation metrics. In Section IV, we present our proposed test conditions and discuss the performance of the evaluated methods. Section V provides conclusions.

II. Acoustic Echo Control (EC) Framework

A. Framework Overview and Notations

The task of echo control systems is depicted in Fig. 1. We differentiate two categories: Acoustic echo cancellation (AEC) describes the estimation and subtraction of an echo signal, while acoustic echo suppression (AES) models a gradual attenuation of parts of the signal containing echo via an estimated mask, typically in the frequency domain. As a far-end (FE) speaker talks, the reference signal $x(n)$ is transmitted to the near-end (NE) and played from a loudspeaker. This loudspeaker might also show nonlinear characteristics which lead to a distorted emitted signal $x'(n) = f_{\text{NL}}(x(n))$. The loudspeaker signal propagates in the NE room and is picked up at the microphone as echo $d(n) = h(n) * x'(n)$, with $h(n)$ being the room impulse response (RIR) and $*$ being the convolution operator. The microphone signal $y(n) = s(n) + d(n) + n(n)$ also contains the near-end speech signal $s(n)$ and background noise $n(n)$. In some cases, the system might introduce delay between the reference signal and its corresponding echo signal, e.g., when using a wireless loudspeaker. Such delay, if not compensated, can cause major issues for EC approaches—in classical and hybrid methods due to the limited filter length, but also in deep AEC models if the training data does not contain conditions with delay.

Without any alteration of $y(n)$, this would mean that the FE speaker would receive their own voice as echo, which can be very irritating and significantly lowers the perceived quality of the conversation [20]. The task of AES therefore is to suppress $d(n)$, while the task of AEC is to estimate $d(n)$ and to subtract the estimate according to $e(n) = y(n) - \hat{d}(n)$. By definition, this means that we explicitly exclude the task of noise suppression and models employing it alongside EC—be it in the form of a postfilter or as a joint training target—from the scope of this article. Typically, for an EC system, the only available signals are $x(n)$ and $y(n)$.

B. Echo Control (EC) Methods Under Test

Evaluating a large number of different approaches would clutter our tables and figures. Therefore, we analyze five hand-picked EC methods, each being representative for a broad class

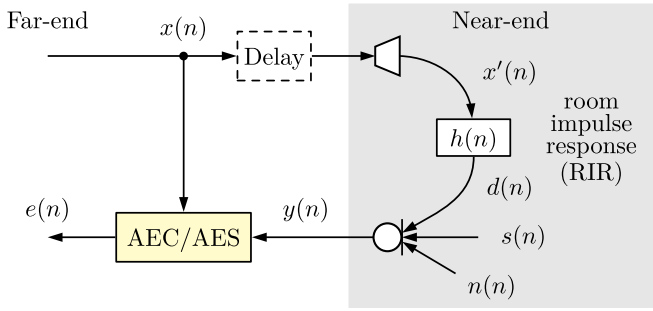

Fig. 1. Generalized overview of our EC framework.

of approaches, including classical adaptive filters, machine-learned deep neural networks (DNNs), and hybrid models. These methods were neither chosen to necessarily represent current state of the art nor to answer the question of whether classical or learned EC is supreme, but rather as greatly different approaches to the given task. This diversity allows us to demonstrate the differences in behaviour observable through our extensive evaluation setup. In line with this idea and to keep our study easier to follow, we opted to not include any auxiliary functions, e.g., double-talk detectors or long-term delay estimation algorithms, even though they might be common in practical applications, especially when using classical algorithms.

The first approach will be a basic normalized least means square (`NLMS`) filter after [3], as it has been widely used in AEC. Variations of this method are still frequently deployed in modern systems and subject to research [7], [8]. A further description of the specific model used here is given in Appendix A.

A more sophisticated AEC approach is the frequency-domain Kalman filter (`FDKF`) [5], [21]. This system is designed for high preservation of NE speech at the trade-off of a less aggressive echo cancellation. A detailed description of our implementation is given in Appendix B.

Recently, several approaches to AEC have substituted parts of classical systems with deep neural networks (DNNs). One such approach is the so-called Neural Kalman Filtering (`NKF`) [9], which replaces the filter update routine of the `FDKF` with a lightweight DNN. Details can be found in Appendix C.

A second approach to such a hybrid EC scheme is the `DeepAdaptive` [22] model. Similar to `NKF`, it utilizes a DNN to estimate the step size of a classical AEC's filter update algorithm, but also further enhances performance by modelling nonlinearities of the echo path before the actual echo prediction algorithm. Details can be found in Appendix D.

As deep AES model in this work, we use the convolutional grouped GRU network with 16 layers (`CGGN16`) [13]. The model is a variation of the convolutional recurrent network archetype and operates in the frequency domain. A detailed description of the network architecture can be found in [13]. All DNN-based methods have been trained from scratch for this work as described in Section III-B, with window and DFT length $K = 512$ and frame shift $R = 128$. With this, all methods evaluated in this article use the same (per-tap) filter size and window length (cf. Appendices A–D). Note that due to `NKF`

and `DeepAdaptive` being multi-tap filters, the `CGGN16`'s inherent potential to model infinitely long time dependencies (due to recurrent layers), and the differences in overlap-add/overlap-save usage between the frequency-domain approaches, the effective filter length varies. This means that some models are by design more capable of modelling longer RIRs or delays.

III. Evaluation Framework

A. Training and Evaluation Datasets

While delivering an ideal training environment for our neural networks (i.e., one that lets them perform well in all possible evaluation conditions) is not a focus of this article, a careful choice of training and evaluation datasets is still paramount. Especially, the evaluation data should be distinct from training conditions to avoid misleading results due to overfitting on seen patterns. To ensure a fair evaluation between all methods, we choose disjoint speakers, background noises, nonlinear distortions, and RIRs for all our dataset splits, as discussed in detail below.

The *training dataset* $\mathcal{D}_{\text{train}}$ for the methods employing a neural network uses speakers from the publicly available CSTR-VCTK corpus [23] to generate NE and FE utterances. Each speaker can appear as NE or FE to discourage the trained model from overfitting to speakers. To generate the echo signal, first a scaled error function (SEF) [24], [25], defined as

$$x'(n) = f_{\text{SEF}}(x(n)) = \int_0^{x(n)} \exp\left(\frac{z^2}{2\beta^2}\right) dz, \qquad (1)$$

is applied to the FE reference signal $x(n)$, with β randomly chosen from $\{0.5, 1, 10, 999\}$. This simulates nonlinear distortions caused by the loudspeaker. The resulting distorted signal $x'(n)$ is convolved with a room impulse response $h(n)$ (RIR). For the training data, RIRs are modelled as white Gaussian noise modulated with an exponential decay [26] so that a reverberation time (RT60) randomly sampled from a continous range of $[50, 600]$ ms is achieved. The signal-to-echo ratio (SER) of each audio file is randomly sampled from a continuous range of $[-12.4, 22.4]$ dB. For background noise, we use random cuts from the publicly available DEMAND [27] and QUT-NOISE [28] databases at a signal-to-noise ratio (SNR) in the range of $[-2.4, 32.4]$ dB, with a 10% chance of a file to be noise-free. We generate $8,500$ files each of 10 s length. Between training epochs, the speech, noise, and echo components of $\mathcal{D}_{\text{train}}$ are reshuffled with new SER and SNR values. A separate small control split $\mathcal{D}_{\text{con}}$ of $1,000$ files is generated along the lines of the training data generation (no reshuffling, however), that exclusively serves for learning rate control and early stopping during training.

For preliminary evaluations of all EC methods, we use a *development dataset* $\mathcal{D}_{\text{dev}}$ of 200 files per condition. Speakers are drawn from a speaker subset of the publicly available CSTR-VCTK corpus disjoint to the training speakers. Nonlinear distortions are again simulated via (1), but with distinct β values sampled from $\{0.2, 0.4, 1.5, 12, 999\}$. Modulated noise is again used for RIR generation with the same parameter setup as in $\mathcal{D}_{\text{train}}$, but different random seeding. SER values are sampled from discrete values in $\{-10, -5, \ldots, 20\}$ dB. Noise is cut from the publicly available ETSI noise database [29]. The SNR is chosen from $\{0, 5, \ldots, 30\}$ dB.

The *test dataset* $\mathcal{D}_{\text{test}}$ for all EC methods is created from speakers of the publicly available and widely used TIMIT speech corpus [30]. To be distinct from the previously described datasets, we use the arctan nonlinearity function [31], [32], defined as

$$x'(n) = f_{\text{arctan}}(x(n)) = \frac{\arctan(\alpha \cdot x(n))}{\alpha}, \qquad (2)$$

to simulate loudspeaker nonlinearities in this dataset, with $\alpha = 10^{-4}$. Our test dataset $\mathcal{D}_{\text{test}}$ uses real-world RIRs from the Aachen Impulse Response Database [33], which are modified to exclude initial delay. Note that initial delay is introduced for a specific experiment only. Noise is cut from the remaining ETSI files of environments so far unseen both in training and in validation. The SER is chosen from $\{-9, -6, \ldots, 9\}$ dB, the SNR from $\{5, 8, \ldots, 20\}$ dB. We generate 200 files per condition.

The speech material for validation and test has to be further preprocessed, considering the test conditions far-end single-talk (STFE), near-end single-talk (STNE), and double-talk (DT). To support a fair comparison between models, the evaluated sections of each condition (8 s long) are preceded by a few seconds of audio to allow for convergence. In case of DT, this entails first an STFE and then an STNE section before the actual DT section follows. For evaluation of STFE and STNE, only one preceding section also of STFE and STNE is used, respectively. Each preceding section is of 8 s to 12 s length and does not contribute to evaluation metrics.

B. DNN Training Process

All DNN-based models in this work have been trained from scratch to ensure fairness in comparison. The training is conducted on $\mathcal{D}_{\text{train}}$ (cf. III-A) using the Adam optimizer [34] in its standard configuration and a logarithmic MSE loss function defined as

$$J^{\text{logMSE}} = 10 \cdot \log\left(\sum_{n \in \mathcal{N}} \left(e(n) - s(n) - n(n)\right)^2\right), \qquad (3)$$

computed over the entire time sequence $(n \in \mathcal{N} \subset \mathbb{N}_0)$. The batch size is set to 16 with a backpropagation-through-time (BPTT) unrolling sequence length of 200 frames. The initial learning rate (LR) is set to 10^{-4}, which is halved after 4 epochs without loss improvement on the control split $\mathcal{D}_{\text{con}}$. The training is stopped after 100 epochs, if the loss on $\mathcal{D}_{\text{con}}$ does not improve for 10 consecutive epochs, or if the LR drops below 10^{-5}. Between epochs, the microphone signal components $s(n), d(n)$, and $n(n)$ of $\mathcal{D}_{\text{train}}$ are reshuffled with new SER and SNR values to generate more diversified training data.

All models are trained in `PyTorch2` [35], using a `GTX 1080 Ti` GPU. Training runs are implemented in a deterministic fashion to avoid performance variations, e.g., from inconsistent data randomization or non-deterministic CUDA operations.

C. Evaluation Metrics

The DT, STFE, and STNE sections in $\mathcal{D}_{\text{dev}}$ and $\mathcal{D}_{\text{test}}$ are evaluated independently on their own subset of metrics. For evaluation of echo cancellation effectiveness, we use the AECMOS metrics DT Echo (abbreviated DT E) and ST Echo (ST E) [36], describing the suppression of echo during double-talk and (far-end) single-talk, respectively. The NE speech preservation is measured through the Perceptual Evaluation of Speech Quality (PESQ) metric [37] and the AECMOS metrics DT Other (DT O) and ST Other (ST O), describing non-echo-related degradations in DT and STNE, respectively.

For a more precise evaluation, especially in the DT condition, we also employ the black-box component metrics PESQ_{BB} and ERLE_{BB} according to ITU-T Recommendations P.1100 [38] and P.1110 [39], with more details in [40], [41]. The black-box components are derived from the enhanced signal $e(n)$ by calculating a frequency-domain gain $G_\ell(k')$ between microphone and enhanced signal as

$$G_\ell(k') = \min\left[\frac{|E_\ell(k')|}{|Y_\ell(k')|}, 1\right] \cdot \frac{e^{j\phi_{E_\ell}(k')}}{e^{j\phi_{Y_\ell}(k')}}, \tag{4}$$

which can then be applied to individual signal components of the microphone signal such as $\tilde{S}_\ell(k') = G_\ell(k') \cdot S_\ell(k')$, with ℓ being the frame index and k' being the frequency bin index. The domain transformation for the black-box approach always uses a Blackman window with DFT size $K' = 512$ and frame shift $R' = 64$ to obtain its frequency-domain signals [40]. As such, the black-box approach is independent from the employed EC model, as any enhanced signal can be mapped to a spectral gain factor in identical fashion. The only requirement is the availability of the separate signal components of the microphone, as it is generally the case for synthetic datasets. The resulting components $e(n) = \tilde{s}(n) + \tilde{d}(n) + \tilde{n}(n)$ allow us to disentangle the evaluation of NE speech preservation and echo suppression. Here, $\text{PESQ}_{\text{BB}} = \text{PESQ}(s(n), \tilde{s}(n))$ yields the speech *component* quality of the enhanced signal. The black-box variant of the echo return loss enhancement (ERLE_{BB}) after [18] is defined as

$$\text{ERLE}_{\text{BB}}(n) = 10 \cdot \log_{10}\left(\frac{(g(n) * d(n))^2}{(g(n) * \tilde{d}(n))^2}\right), \tag{5}$$

for which a first-order IIR smoothing filter with impulse response $g(n) = \alpha^n$ and coefficient $\alpha = 0.99$ is applied to the signals $d(n)$ and $\tilde{d}(n)$. The operator $*$ denotes convolution. The final ERLE_{BB} is computed as mean over the entire evaluated sequence.[1] Black-box metrics can be advantageously employed also for double-talk conditions, where standard PESQ and ERLE are of limited usefulness.

Another metric to aid in evaluation of DT scenarios is the log-spectral distance (LSD) metric [42] defined as

$$\text{LSD}_\ell = \sqrt{\frac{1}{|\mathcal{K}''|}\sum\nolimits_{k''\in\mathcal{K}''}\left[10\log_{10}\left(\frac{|S_\ell(k'')|^2}{|\hat{S}_\ell(k'')|^2}\right)\right]^2}, \tag{6}$$

calculated over all frequency bins $k'' \in \mathcal{K}''$. Similar to the black-box approach, the LSD metric uses independent DFT parameters $K'' = 512$ and $R'' = 256$. The final LSD results are calculated by averaging over all frames ℓ of a file *that contain near-end speech.* This metric, much like standard PESQ, reports overall quality (both NE speech preservation and echo suppression effectiveness affect the score), but is a straightforward distance metric. As such, it is less prone to performance differences getting masked by noise, but in return is less descriptive on the perceptual impact of residual echo and NE speech degradation. The respective black-box version LSD_{BB} replaces $S_\ell(k'')$ with $\tilde{S}_\ell(k'')$ for a result more focused on the disentangled speech component. To keep the tables more readable, LSD is only reported from Table III on.

Note that the score limit of all metrics (referenced as oracle in the following sections) is dependent on various test set characteristics. For example, perfect echo cancellation on $\mathcal{D}_{\text{test}}$ only leads to a DT PESQ score of 2.70 points since the background noise is not suppressed. In the same condition, the black-box speech *component* metric PESQ_{BB} reaches a score of 3.96, which better reflects the non-distorted speech component.

IV. Evaluation and Discussion

The following subsections now present our extensive comparative evaluation study in various conditions. The goal is to analyze as many aspects of the EC task as possible and to gain detailed insights into the strengths and limitations of the various presented EC models. In Section IV-A, we start with a performance evaluation of all the models in all three conditions DT, STFE, and STNE. Next, in Sections IV-B and IV-C, the convergence and reconvergence behaviour of the systems is investigated. Section IV-D investigates the reconvergence performance under a continuously changing RIR. Lastly, in Sections IV-E, IV-F, and IV-G, we analyze the robustness towards various aspects, namely nonlinearities, varying SER and SNR levels, and delay.

A. Comparative Evaluation of Methods

In this subsection, we start with a comparative performance evaluation in DT, STFE, and STNE. The results are given in Tables I ($\mathcal{D}_{\text{dev}}$) and II ($\mathcal{D}_{\text{test}}$). As a reference, the scores for the unprocessed microphone signal $y(n)$ and perfectly removed echo (oracle) $y(n) - d(n) = s(n) + n(n)$ are given as well.

We can observe a very similar behaviour of our models on $\mathcal{D}_{\text{dev}}$ and $\mathcal{D}_{\text{test}}$, with some rare exceptions such as the generalization capabilities of the approaches containing DNNs, where the `CGGN16` shows much better PESQ_{BB} and DT O scores on the test set, while `DeepAdaptive` shows lower PESQ scores. For STNE, all models show excellent pass-through of echo-free signals. The slightly increased ST O and PESQ for `CGGN16` and `DeepAdaptive` can be explained by a light noise suppression, which also causes the PESQ_{BB} to report scores below the ideal value of 4.64 due to imperfect separation of the black-box components. The `CGGN16` model removes the echo almost completely (ERLE_{BB}) at the cost of some near-end (NE) speech quality degradation in DT (PESQ_{BB}), with some aggressive over-suppression observed on $\mathcal{D}_{\text{dev}}$. In contrast, `FDKF` shows

[1] The software for the metrics PESQ_{BB}, ERLE_{BB}, and further previously unpublished measures is available at https://github.com/ifnspaml/EC-Evaluation-Toolbox.

TABLE I
PERFORMANCE ON THE **DEV SET** $\mathcal{D}_{\text{dev}}$ **USING SYNTHETIC RIRS**. ERLE_{BB} IN (dB)

Model	#par.	#FLOPS	Double-Talk (DT)					STFE		STNE		
			PESQ	PESQ_{BB}	ERLE_{BB}	DT O	DT E	ERLE_{BB}	ST E	PESQ	PESQ_{BB}	ST O
Unprocessed	–	–	1.77	4.58	–	3.51	2.46	–	2.37	2.92	4.58	3.09
Oracle (echo-free)	–	–	3.07	3.87	18.78	3.52	4.51	25.36	4.64	2.92	4.58	3.09
Classic `NLMS` [3]	–	0.26 M	1.71	3.49	10.69	2.86	3.44	11.25	3.92	2.92	**4.58**	3.09
`FDKF` [21]	–	0.70 M	1.82	**4.27**	5.63	3.13	2.78	6.05	3.03	2.92	**4.58**	3.09
`NKF` [9]	50 K	0.33 G	1.84	3.78	9.42	**3.20**	3.32	10.39	3.52	2.92	**4.58**	3.09
`DeepAdaptive` [22]	7.7 M	1.92 G	2.02	3.28	16.64	3.09	4.30	19.18	4.40	2.92	4.52	3.11
`CGGN16` [13]	1.3 M	3.21 G	**2.14**	3.39	**19.52**	2.97	**4.35**	**25.10**	**4.55**	**2.97**	4.40	**3.17**

Results of best methods are bold, second-best underlined.

TABLE II
PERFORMANCE ON THE **TEST SET** $\mathcal{D}_{\text{test}}$ **USING REAL-WORLD RIRS**. ERLE_{BB} IN (dB)

Model	#par.	#FLOPS	Double-Talk (DT)					STFE		STNE		
			PESQ	PESQ_{BB}	ERLE_{BB}	DT O	DT E	ERLE_{BB}	ST E	PESQ	PESQ_{BB}	ST O
Unprocessed	–	–	1.68	4.62	–	3.76	2.20	–	1.70	2.50	4.64	3.28
Oracle (echo-free)	–	–	2.70	3.96	15.95	3.95	4.33	21.03	4.57	2.50	4.64	3.28
Classic `NLMS` [3]	–	0.26 M	1.70	3.40	12.06	3.16	3.54	13.64	3.40	2.50	**4.64**	3.28
`FDKF` [21]	–	0.70 M	1.91	**4.20**	7.49	3.37	2.96	8.43	3.37	2.50	**4.64**	3.28
`NKF` [9]	50 K	0.33 G	1.84	3.43	11.46	3.09	3.81	11.30	4.29	2.50	**4.64**	3.28
`DeepAdaptive` [22]	7.7 M	1.92 G	1.81	2.90	**17.08**	3.52	**4.24**	**21.23**	**4.43**	2.50	4.56	3.31
`CGGN16` [13]	1.3 M	3.21 G	**2.03**	3.55	14.53	**3.65**	4.17	20.66	4.40	**2.53**	4.42	**3.33**

Results of best methods are bold, second-best underlined.

great preservation of NE speech at the cost of less aggressive echo suppression, while the `NKF` model shows a performance in-between those two extremes. Please note that `FDKF`'s higher-than-oracle score for DT PESQ_{BB} is due to the fact that less echo suppression allows for a higher maximum score of this metric. Generally, `CGGN16` scores higher than the other models in most categories on $\mathcal{D}_{\text{dev}}$. On $\mathcal{D}_{\text{test}}$, the `DeepAdaptive` model yields the highest echo suppression, but at the cost of more degraded NE speech quality, as seen in PESQ and PESQ_{BB}. On the contrary, the DT O score is quite high, which can be attributed to the high echo suppression and its side-effect of also partly removing disturbing background noise—we found this especially noticeable with babble noise. The simple `NLMS` performs worst, showing mediocre echo cancellation paired with significant NE speech degradation.

We conclude: First, in a simple evaluation, the classical `FDKF` preserves near-end speech best, while the `CGGN16` and `DeepAdaptive` methods are most effective at suppressing the echo. Second, the STNE condition does not very much differentiate the methods.

B. Convergence Behavior for Various Far-End Excitations

An important characteristic of EC algorithms is the convergence behavior. Classical approaches are known to require varying amounts of convergence time to reach their steady state performance on a time-invariant RIR. Hybrid models such as the `NKF` have been shown to improve on convergence time [9], while the convergence behavior of fully learned EC methods has only occasionally been reported.

Firstly, we analyze the performance of the models under two different FE excitations in terms of their mean metric scores. Apart from the previously described excitation by FE *speech* ($\mathcal{D}_{\text{test}}$), the models are also evaluated with *white Gaussian noise* (WGN) as FE excitation ($\mathcal{D}_{\text{test}}^{\text{WGN}}$). The performance of the models on these two datasets is reported in the upper and lower segment of Table III, respectively. We see that for $\mathcal{D}_{\text{test}}^{\text{WGN}}$, compared to the already discussed condition of speech FE excitation, `NKF` surpasses `CGGN16` in terms of echo suppression (ERLE_{BB}) at the cost of even more NE degradation in DT (PESQ_{BB}). Note that ERLE_{BB} for the `NKF` is also higher than the oracle score for the DT condition, which is another clear sign of over-aggressive suppression. `CGGN16` maintains a high echo suppression performance and even improves its NE speech preservation, likely due to the easier distinction between speech and noise-like echo. The evaluated classical methods improve in their echo cancellation, as it is easier for them to converge on a spectrally white excitation. Interestingly, the `DeepAdaptive` model seems to greatly struggle with WGN FE excitation and only displays weak echo suppression. Further inspection of the audio revealed an incoherent suppression of different frequency areas, whereas other models tend to suppress the spectrally white signal more evenly. We attribute this behavior to the sole use of fully connected layers in `DeepAdaptive`, which requires a more broader range of training material to achieve good generalization on various FE excitations.

We also observe that the AECMOS metrics do not seem to be suited for evaluation of the WGN FE excitation condition, as proven by the too high scores of DT E and ST E for unprocessed

TABLE III
PERFORMANCE FOR **SPEECH** ($\mathcal{D}_{\text{test}}$, **UPPER TABLE SEGMENT**) **AND WHITE GAUSSIAN NOISE** ($\mathcal{D}_{\text{test}}^{\text{WGN}}$, **LOWER TABLE SEGMENT**) **AS FE EXCITATION**. ERLE_{BB} AND LSD IN (dB)

Test Set (FE excitation)	**Model**	**Double-Talk (DT)**							**STFE**	
		PESQ	PESQ_{BB}	LSD	LSD_{BB}	ERLE_{BB}	DT O	DT E	ERLE_{BB}	ST E
$\mathcal{D}_{\text{test}}$ (speech)	Unprocessed	1.68	4.62	11.30	2.76	–	3.76	2.20	–	1.70
	Oracle (echo-free)	2.70	3.96	6.46	7.84	15.95	3.95	4.33	21.03	4.57
	Classic `NLMS` [3]	1.70	3.40	9.73	7.85	12.06	3.16	3.54	13.64	3.40
	`FDKF` [21]	1.91	**4.20**	**9.12**	**6.39**	7.49	3.37	2.96	8.43	3.37
	`NKF` [9]	1.84	3.43	10.31	8.09	11.46	3.09	3.81	11.30	4.29
	`DeepAdaptive` [22]	1.81	2.90	11.49	11.41	**17.08**	3.52	**4.24**	**21.23**	**4.43**
	`CGGN16` [13]	**2.03**	3.55	9.89	9.48	14.53	**3.65**	4.17	20.66	4.40
$\mathcal{D}_{\text{test}}^{\text{WGN}}$ (white Gaussian noise)	Unprocessed	1.26	4.64	22.92	2.76	–	2.06	3.79	–	3.66
	Oracle (echo-free)	2.62	3.71	6.64	14.70	17.72	3.22	4.57	23.94	3.90
	Classic `NLMS` [3]	1.46	3.69	15.42	13.50	13.43	2.01	4.47	14.85	**4.02**
	`FDKF` [21]	1.55	4.03	14.57	12.43	10.50	2.30	**4.61**	11.66	3.96
	`NKF` [9]	1.66	3.20	**14.01**	18.21	**18.38**	**2.42**	4.52	**20.98**	3.73
	`DeepAdaptive` [22]	1.59	**4.24**	**14.01**	**7.85**	8.58	2.20	4.25	9.64	3.69
	`CGGN16` [13]	**1.89**	3.88	15.08	20.01	16.35	2.27	4.52	20.45	3.77

Results of best methods per test set are bold, second-best underlined.

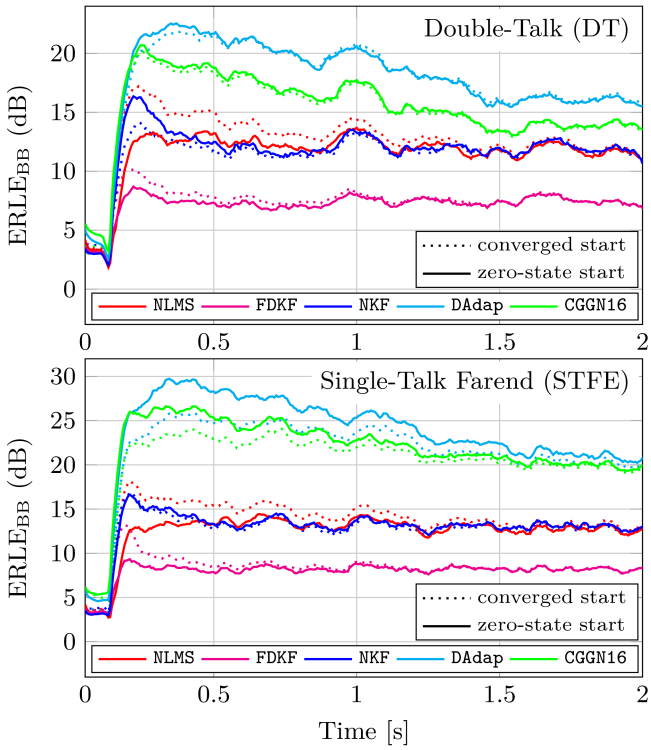

Fig. 2. **Convergence behavior** of the observed approaches during the first 2 seconds of the **FE speech excitation test set** $\mathcal{D}_{\text{test}}$ for DT (top) and STFE condition (bottom). Values represent the mean over the dataset. Dotted lines indicate performance of initially already converged models, solid lines indicate performance starting from a zero state.

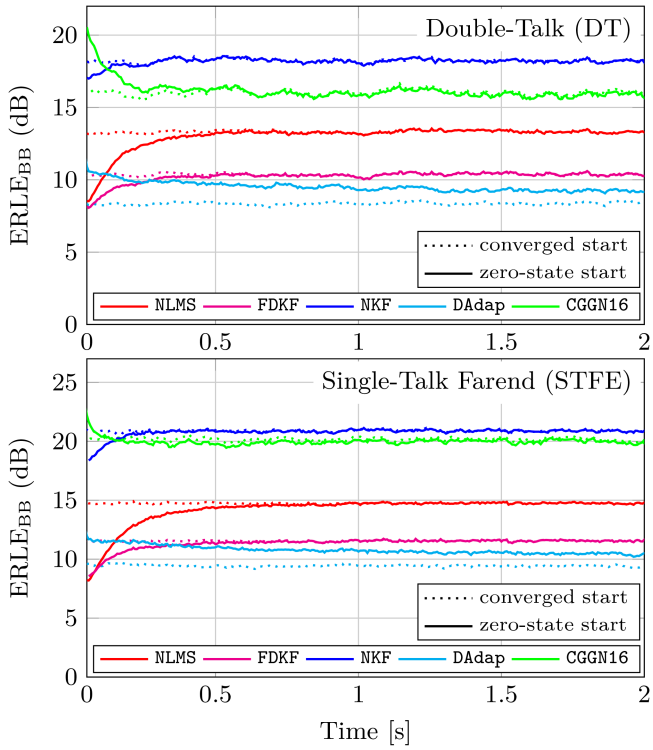

Fig. 3. **Convergence behavior** of the observed approaches during the first 2 seconds of the **FE WGN excitation test set** $\mathcal{D}_{\text{test}}^{\text{WGN}}$ for DT (top) and STFE condition (bottom). Values represent the mean over the dataset. Dotted lines indicate performance of initially already converged models, solid lines indicate performance starting from a zero state.

audio and the too low score of DT O for the echo-free oracle experiment. Moreover, the noise-like FE echo and its residual seem to have a higher impact on overall PESQ scores in DT as well.

As for the now included LSD and LSD_{BB} metrics, we largely see a correlation with our PESQ_{BB} metric for speech FE excitation. The main outlier is the `NLMS` model, which scores comparatively well even though the performance is clearly more degraded than, e.g., with the `CGGN16` model, as found from informal listening tests and supported by all other metrics. We attribute this to the fact that the `NLMS` is the only model to operate directly in the time domain, avoiding a potential impact of the DFT/IDFT and OLA algorithms on LSD. For WGN excitation, we found the reported LSD scores to be less informative and to only weakly correlate with subjective listening impressions.

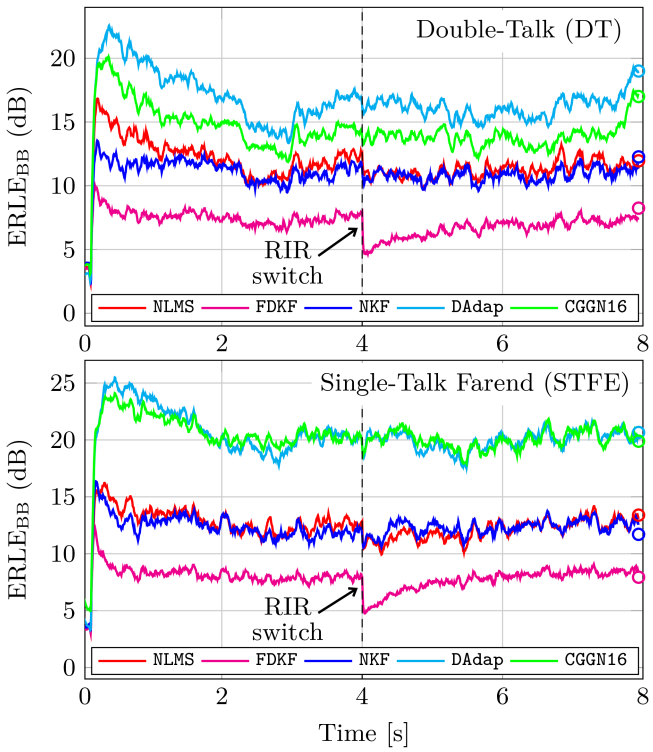

Fig. 4. **Reconvergence behavior** of the observed approaches on the **modified test set** $\mathcal{D}_{\text{test}}^{\text{switch}}$ for DT (top) and STFE condition (bottom). The RIR is abruptly switched after 4 seconds (indicated by dashed line). As the RIR for $t = 4\text{s}\ldots8\text{s}$ is the same as in $\mathcal{D}_{\text{test}}$, the performance of this section can be directly compared to the steady state of Fig. 2. Colored circle markers on the right-hand side mark the Fig. 2 converged start performance after 8 s. Values represent means over the dataset.

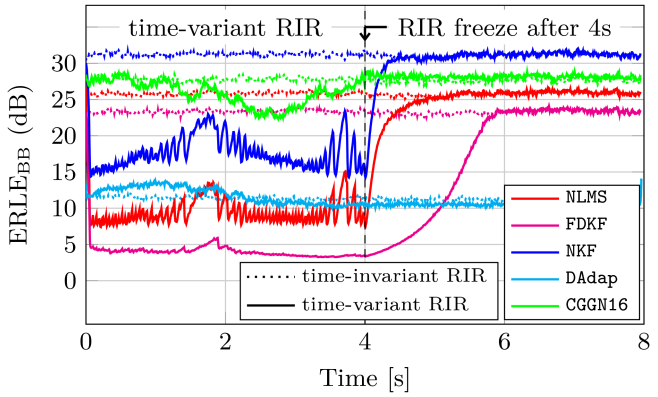

Fig. 5. **Reconvergence/tracking behavior** of the observed approaches with **continuously changing RIR test set** $\mathcal{D}_{\text{test}}^{\text{WGN,dyn}}$ for the **STFE WGN condition**, starting from converged state. Values represent means over the dataset. Dotted lines indicate performance of models on the initial time-invariant RIR, solid lines indicate performance on the here investigated time-variant RIR (0-4 s), while after 4 s the RIR is frozen.

We conclude: For white Gaussian far-end excitation, the DNN approach (`CGGN16`) shows the highest robustness, while `DeepAdaptive` shows severe issues due to the use of fully connected layers. AECMOS and LSD metrics, however, are less suitable for this test condition.

The convergence behavior of the models is evaluated in DT and STFE conditions by comparing the performance between the converged states (using the preceding audio sections as described in Section III-A; in case of the WGN FE excitation experiment, the preceding audio sections contain FE WGN audio as well) and the zero state (running the models only on the evaluated audio section). Once both graphs align, the model has reached its steady state. The convergence behavior of our models in DT and STFE is depicted in Fig. 2 for FE speech excitation and in Fig. 3 for FE WGN excitation.

We see that the convergence time of the `NLMS` filter (red) is the longest, followed by `DeepAdaptive` (cyan) and `FDKF` (magenta). For FE speech excitation, `CGGN16` (green) and `NKF` (blue) show either almost immediate convergence or a tendency towards starting out with very aggressive suppression. The observed better steady-state performance of the hybrid `NKF` method over the purely algorithmic `FDKF`—with both sharing parts of the same core structure—likely stems from multiple factors: the use of overlap-add in `NKF`, its multi-tap filter, and potentially better robustness against background noise. Interestingly, `CGGN16` seems more prone to over-suppression when no NE speech needs to be preserved (STFE condition). For WGN FE excitation, we see mostly identical behaviour, with generally faster convergence times. Again, `DeepAdaptive` displays issues in dealing with WGN FE excitation.

We conclude: Compared to classical approaches, the neural network-based methods show often faster convergence for both speech and white Gaussian far-end excitation.

C. Reconvergence Behavior: RIR Switch

Apart from the initial convergence behavior, another important factor of AEC systems is the *re*convergence behavior. This describes the system's ability to adapt towards a new RIR from an already converged state.

To analyze the models' reconvergence behavior, we modify our test set towards abrupt RIR switches after 4 s of audio. Fig. 4 displays the resulting performance of our models on this modified test set $\mathcal{D}_{\text{test}}^{\text{switch}}$. We see that `FDKF` requires longest to reconverge, while a drop in ERLE$_{BB}$ at the RIR switch seems almost unnoticeable for `CGGN16`. `DeepAdaptive`, `NLMS`, and `NKF` show small drops in performance, which are almost masked by the natural fluctuations in the curves. `NLMS`, however, takes longer to reconverge. When comparing the final performance to the one achieved on $\mathcal{D}_{\text{test}}$ (circles at 8 s), we can see that the models return to about the same steady-state performance after reconvergence.

We conclude: While all models return to their steady-state performance after an RIR change, DNN-based methods show fastest reconvergence.

D. Reconvergence Behavior: Continuously Changing RIR

While reconvergence towards a completely new RIR might happen once in a while, it is far more common that small fluctuations of the RIR occur continuously over time, e.g., due to movement of objects or speakers in the room. Simpler setups with an abrupt RIR change as in Section IV-C, while insightful in their own respect, might not fully inform about a

TABLE IV
PERFORMANCE FOR VARIOUS LOUDSPEAKER NONLINEARITIES ERLE AND LSD IN (DB)

Test Set (LS nonlinearity)	Model	Double-Talk (DT)							STFE	
		PESQ	PESQ_{BB}	LSD	LSD_{BB}	ERLE_{BB}	DT O	DT E	ERLE_{BB}	ST E
$\mathcal{D}_{\text{test}}^{\text{linear}}$ (no nonlinearity)	Unprocessed	1.70	4.63	11.05	2.76	–	3.87	1.95	–	1.70
	Oracle (echo-free)	2.70	3.97	6.46	7.62	15.71	3.94	4.33	20.70	4.57
	Classic NLMS [3]	1.73	3.40	9.57	7.85	12.02	3.20	3.53	13.77	4.34
	FDKF [21]	1.94	**4.20**	**8.77**	**6.43**	7.49	3.42	3.00	8.51	3.43
	NKF [9]	1.85	3.43	10.21	8.08	11.38	3.08	3.82	12.73	4.28
	DeepAdaptive [22]	1.82	2.87	11.56	11.54	**17.40**	3.54	**4.26**	**21.87**	4.42
	CGGN16 [13]	**2.07**	3.55	9.82	9.45	14.69	**3.64**	4.18	21.25	**4.45**
$\mathcal{D}_{\text{test}}$ (mild nonlinearity)	Unprocessed	1.68	4.62	11.30	2.76	–	3.76	2.20	–	1.70
	Oracle (echo-free)	2.70	3.96	6.46	7.84	15.95	3.95	4.33	21.03	4.57
	Classic NLMS [3]	1.70	3.40	9.73	7.85	12.06	3.16	3.54	13.64	3.40
	FDKF [21]	1.91	**4.20**	**9.12**	**6.39**	7.49	3.37	2.96	8.43	3.37
	NKF [9]	1.84	3.43	10.31	8.09	11.46	3.09	3.81	11.30	4.29
	DeepAdaptive [22]	1.81	2.90	11.49	11.41	**17.08**	3.52	**4.24**	**21.23**	**4.43**
	CGGN16 [13]	**2.03**	3.55	9.89	9.48	14.53	**3.65**	4.17	20.66	4.40
$\mathcal{D}_{\text{test}}^{\text{sigmoid}}$ (strong nonlinearity)	Unprocessed	1.62	4.63	12.11	2.76	–	3.72	2.00	–	1.73
	Oracle (echo-free)	2.70	3.91	6.46	8.55	15.52	3.94	4.34	20.44	4.59
	Classic NLMS [3]	1.58	3.43	10.82	7.36	10.53	2.98	3.47	11.51	4.27
	FDKF [21]	1.76	**4.23**	10.81	**5.62**	6.42	3.25	2.78	6.95	3.03
	NKF [9]	1.71	3.47	12.11	7.61	9.89	2.93	3.78	10.87	4.27
	DeepAdaptive [22]	1.72	3.07	11.40	10.60	**13.87**	3.51	**4.14**	**16.71**	**4.39**
	CGGN16 [13]	**1.81**	3.81	**10.28**	8.00	10.39	**3.61**	3.94	13.00	4.04

Best results per test set are bold, second-best underlined.

model's behaviour under such conditions. Therefore, it is crucial to include an experiment with highly dynamic RIR conditions in one's test suite for a comprehensive analysis. We evaluate the models' behaviour in such a condition for the new dataset $\mathcal{D}_{\text{test}}^{\text{WGN,dyn}}$. This dataset uses a single, constantly changing RIR $h^{\text{dyn}}(n)$, which was recorded as described by Jung et al. [31], see also Jung's contribution to ITU-T P. 1130 [43]. The changing RIR is simulated by rotating a reflective plate at a speed of $\omega \approx 360^\circ/4$ s placed in the recording environment. For more details on the RIR data of this condition, the reader is referred to Appendix E. A total of 60 audio files were generated for $\mathcal{D}_{\text{test}}^{\text{WGN,dyn}}$, using WGN as FE excitation.[2]

Fig. 5 presents the performance of our models on $\mathcal{D}_{\text{test}}^{\text{WGN,dyn}}$. The models start from a converged state on the initial RIR, followed by 4 s of audio with continuously changing RIR (equals roughly one full rotation of the reflective plate) before it is frozen again at $t = 4$ s. To better analyze the effects of a time-variant RIR on the reconvergence behaviour (solid lines), the figure also contains graphs for a condition with time-invariant RIR (dotted lines), generated from the final state of $h^{\text{dyn}}(n)$ at $t = 4$ s.

We can see that several methods are greatly affected by the continuous RIR change, with an initial ERLE drop by 15 to 20 dB. Particularly FDKF struggles to achieve any meaningful echo cancellation during this experiment, only slowly recovering after the RIR is frozen again. NLMS and NKF show a similar oscillating pattern as observed by Jung [31], with slightly better performance at the 180° (∼1.8 s) and 360° (∼3.7 s) marks, where the RIR resembles the original one. After the RIR freeze, both models quickly return to their steady-state performance. The DNN CGGN16 presents the best tracking capabilities, only showing a performance drop after 2 s and even recovering back to steady-state performance *before* the RIR freeze. With all models returning to their initial steady-state performance within 2 s after the RIR freeze and the NKF again reaching the highest score, it becomes clear that the performance of a model after convergence/reconvergence gives no information about a model's tracking capabilities. DeepAdaptive continues to exhibit weak performance for WGN FE excitation, although staying quite consistent for the time-variant RIR.

We conclude: The DNN method (CGGN16) shows great tracking behaviour under a continuously changing RIR. Robustness against time-variant RIRs is not reflected by steady-state performance (where the hybrid NKF shows the highest scores for white Gaussian noise far-end excitation).

E. Nonlinearity Robustness

While classical methods of sufficient filter length are in theory able to converge towards a close representation of the RIR for an ideal system, the addition of loudspeaker nonlinearities poses a problem. Since these methods (as well as the NKF) estimate filter coefficients to be applied to the reference signal, nonlinearities cannot be directly modelled, therefore hindering the convergence process and putting a limit on the estimation precision. Other models are by design (CGGN16) or through explicit distortion modelling (DeepAdaptive) more capable of taking nonlinearities into account. However, their effectiveness in doing so is not guaranteed on unseen types of nonlinearities.

In this experiment, we substitute the nonlinear function (2) employed in our test set to analyze the effects on the different

[2]The code for generating dynamic RIRs with this method can be found at https://github.com/ifnspaml/EC-Evaluation-Toolbox.

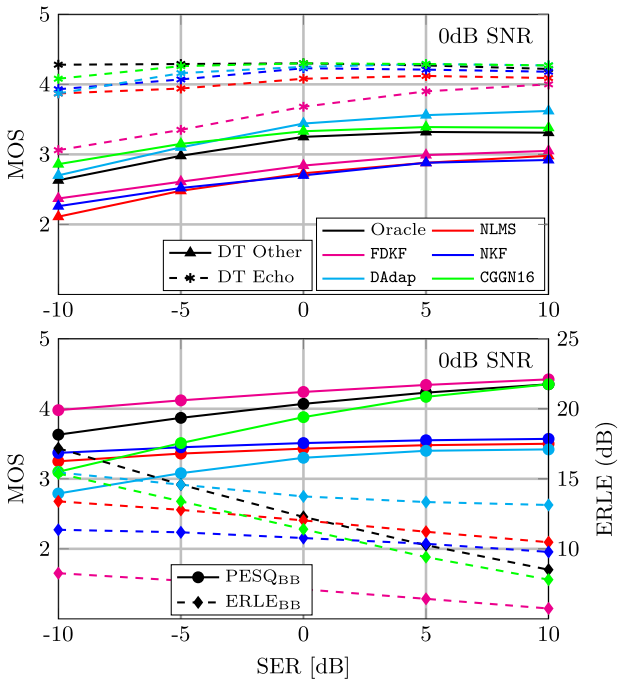

Fig. 6. **Impact of SER level** on the models' performance in DT condition of $\mathcal{D}_{\text{test}}$ for **SNR = 0 dB**.

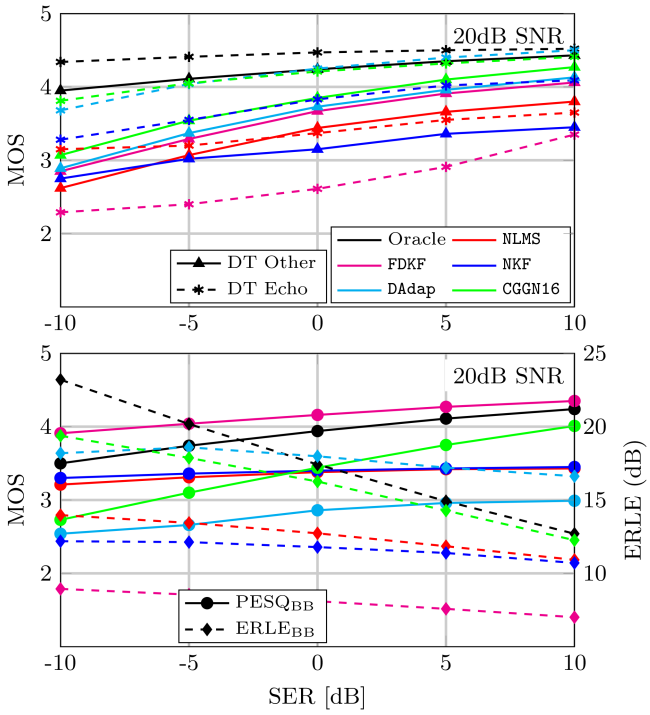

Fig. 7. **Impact of SER level** on the models' performance in DT condition of $\mathcal{D}_{\text{test}}$ for **SNR = 20 dB**.

models. The test set $\mathcal{D}_{\text{test}}^{\text{linear}}$ contains no loudspeaker nonlinearities. The nonlinearities in the original $\mathcal{D}_{\text{test}}$ can be considered fairly mild, while $\mathcal{D}_{\text{test}}^{\text{sigmoid}}$ contains harsh loudspeaker distortions modelled by the memoryless sigmoidal function [44]. The nonlinearities of $\mathcal{D}_{\text{test}}$ and $\mathcal{D}_{\text{test}}^{\text{sigmoid}}$ are both disjoint from the ones used in training.

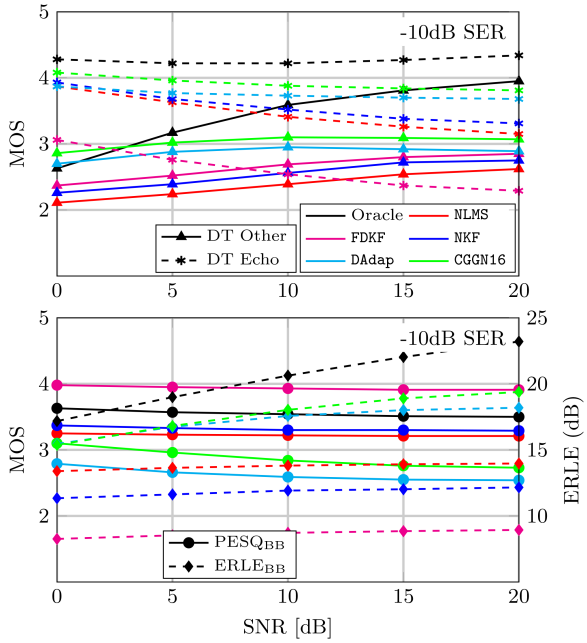

Fig. 8. **Impact of SNR level** on the models' performance in DT condition of $\mathcal{D}_{\text{test}}$ for **SER = −10 dB**.

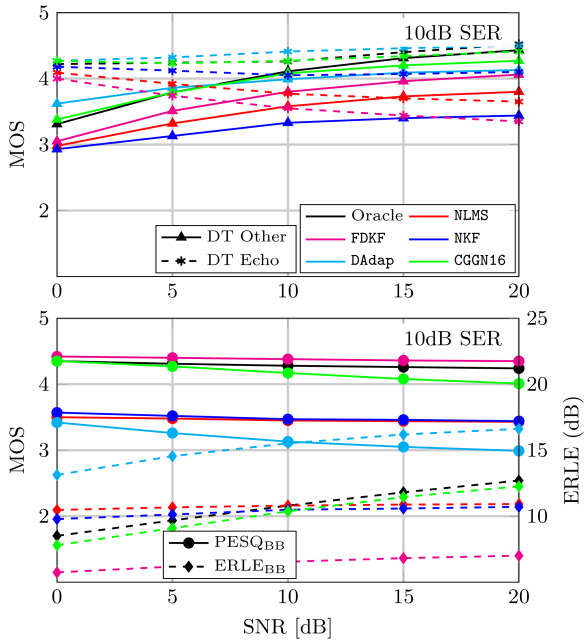

Fig. 9. **Impact of SNR level** on the models' performance in DT condition of $\mathcal{D}_{\text{test}}$ for **SER = 10 dB**.

In Table IV, we can see that the performance differences between no nonlinearity ($\mathcal{D}_{\text{test}}^{\text{linear}}$) and a mild nonlinearity ($\mathcal{D}_{\text{test}}$) are quite small for all methods. The harsh nonlinearities in $\mathcal{D}_{\text{test}}^{\text{sigmoid}}$, however, result in a noticeable performance change for all models. FDKF is again excellent for near-end speech component preservation in DT (PESQ$_{BB}$), but at the price of the lowest echo performance (ERLE$_{BB}$, DT E) in the same condition. The deep AES method CGGN16, on the other hand,

performs strong w.r.t. ERLE_{BB}, DT E, and ST E in almost all cases, and also overall in PESQ during double talk.

With strong nonlinearity, NLMS performs well w.r.t. ERLE_{BB}, while both NLMS and NKF yield high ST E scores. This can be attributed to the models' characteristic of indiscriminately suppressing major parts of all signal components during echo activity, which can be seen in their poor NE speech preservation in DT (PESQ_{BB}). Interestingly, CGGN16 seems to react with the strongest drop in echo suppression on $\mathcal{D}_{\text{test}}^{\text{sigmoid}}$, which is likely due to the fact that such harsh nonlinearities have not been part of the training material. DeepAdaptive, as the highest scoring model in terms of echo suppression, also displays a noticeable decrease in ERLE_{BB}, while DT E and ST E remain on a high level. It is also the only model to maintain a steady LSD score. While this model also did not see harsh nonlinearities in training, the designated nonlinearity modelling by its DNN is likely the cause for this less pronounced performance decrease.

We also observe that the ERLE_{BB} and the DT E metric do not deliver the same rank orders in any of the nonlinearity cases. This is because of their diverging evaluation objective: ERLE_{BB} merely measures residual echo, but DT E reflects a subjective impression about echo removal.

We conclude: Deep echo control models can obtain strongest echo cancellation performance under nonlinearity strengths that have been seen in training. The linear FDKF provides the highest fidelity of the near-end speech component also in harsh nonlinear conditions.

F. Impact of SER and SNR

Another important aspect of echo cancellation is the reliability of a system under varying levels of echo and noise. Ideally, we want an EC system to deliver a good performance independently of SER and SNR within the operational range of the desired use case, but in practice, high noise levels often complicate system identification and loud echoes might result in leaking residual echo.

We evaluate our models under varying combinations of SER and SNR as shown in Figs. 6–9. Figs. 6 and 7 display the performance of the models for different SER levels with fixed SNR $=$ 0 dB and SNR $=$ 20 dB, respectively. Figs. 8 and 9, on the other hand, display the performance of the models for different SNR levels with fixed SER $= -10$ dB and SER $=$ 10 dB, respectively. We also display the scores of the echo-free oracle signal, which not only gives us a better frame of reference, but also reveals the metrics' upper limit dependencies regarding audio levels.

We can see that oracle scores are very consistent across all SER and SNR levels for DT Echo, while DT Other scoring is highly dependent on noise levels in the microphone signal. This is due to the fact that DT Other considers noise as degradation. We also observe a tendency for AECMOS scores to cluster more in low SNR conditions. The most likely cause is that for strong noise components, the differences in performance of the evaluated models are more difficult to distinguish for the underlying neural network of the AECMOS metric. For the black-box metrics, we can see a high dependency of ERLE_{BB} on both SER and SNR, while PESQ_{BB} scoring is only significantly influenced by echo levels. These dependencies are attributed to the fact that the black-box component separation is not always perfect, e.g., in the case of overlapping speakers or between noise and highly reverberated echo. Therefore, the removal of a stronger echo will degrade the *black-box* (but not the actual) NE speech component more, while (unsuppressed) strong noise makes residual echo appear more severe in its black-box component.

We conclude: All shown metrics depend either on SNR, on SER, or on both. It is therefore crucial to mind the oracle scores when comparing a model's performance under different echo/noise levels. High noise levels seem to partially mask model differences in the AECMOS metrics.

For the investigated models, we can see that their characteristics are very distinct. CGGN16 performs well in high SER scenarios and mostly independently of the current SNR, but for low SER we can see a tendency towards aggressive suppression at the cost of degraded NE speech quality (see Figs. 6 and 7). As discussed before, low SNR masks this behaviour in the AECMOS metrics. Again, FDKF shows relatively mild echo cancellation (especially for low SER), but maintains a good NE speech preservation across all SER and SNR conditions. For NLMS and NKF, we can see an increasing gap to the ERLE_{BB} and DT Echo oracle scores with decreasing SER, while strong noise for high SER levels seems to encourage overly aggressive suppression (see Fig. 9). DeepAdaptive exhibits an interesting behaviour, as the echo suppression seems more aggressive in high SER scenarios. The significantly higher ERLE_{BB} scores than oracle hint at an over-suppression (see Figs. 6 and 7).

We conclude: All models show stronger echo leak for lower SER levels. FDKF preserves the NE speech best in all conditions, while the DNN model (CGGN16) gives the most consistent echo suppression across most SER/SNR levels.

G. Delay Robustness

Depending on the deployed hardware and environment, systems can also experience delays between the reference signal and the resulting echo. An example could be a wireless loudspeaker, which introduces additional delay in the loudspeaker path, see the optional delay in the dashed box in Fig. 1. This can be problematic as it makes a precise system identification more difficult or even impossible, especially for classical approaches with a limited filter length.

Fig. 10 shows the effects of additional delay to $\mathcal{D}_{\text{test}}$ for the different models. We apply a fixed added delay between the reference signal $x(n)$ and the resulting echo $d(n)$ to all files of the dataset, for multiple delay values. Note that, while it is common to employ some sort of long-term delay estimation [45], [46], [47], this experiment explicitly focuses on the raw performance of the evaluated models without any prior delay compensation. We see that almost all models loose a significant portion of their performance for delays as low as 50 ms already. For the classical models and NKF, this is to be expected, as their filter lengths are much shorter. The DNN CGGN16 also cannot handle higher delays due to the lack of respective training material.

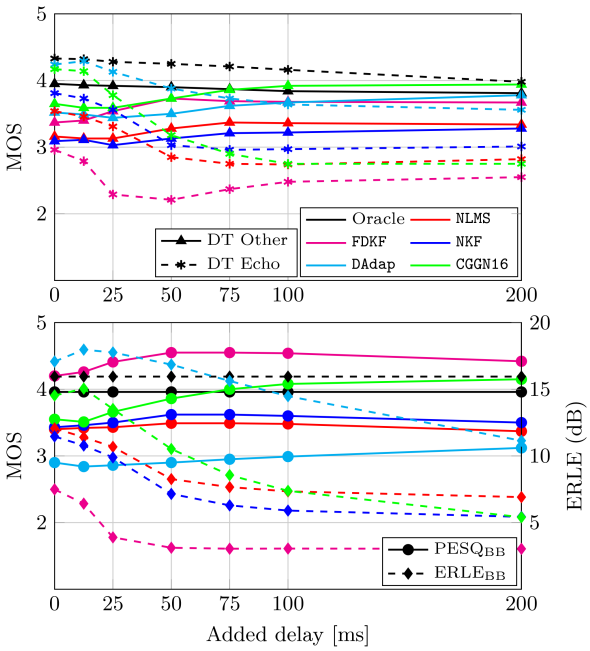

Fig. 10. **Impact of additional initial delay** between the reference input $x(n)$ and corresponding echo component $d(n)$ on the models' performance, measured on the DT condition of $\mathcal{D}_{\text{test}}$. **None** of the DNN-based models has **seen delay in training**.

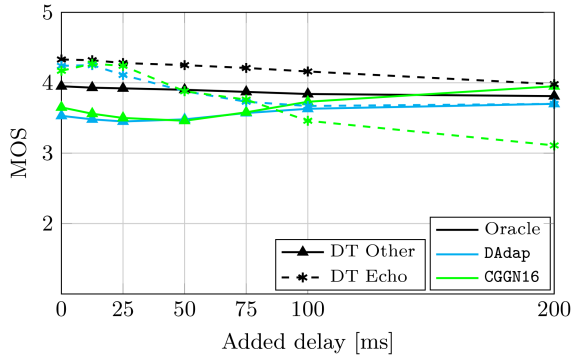

Fig. 11. **Impact of additional initial delay** between the reference input $x(n)$ and corresponding echo component $d(n)$ on the performance of models **trained with delay-augmented data**, measured on the DT condition of $\mathcal{D}_{\text{test}}$.

Interestingly, even though both NKF and NLMS loose most of their echo cancellation capabilities, both models still show significantly reduced NE speech quality (AECMOS and PESQ$_{BB}$) when compared to the oracle. The DeepAdaptive model can deal with delay to a certain extent due to its longer multi-tap filter and shows a slower drop in echo suppression effectiveness (even increasing in ERLE$_{BB}$ for minor added delay).

As a follow-up experiment, we test the hypothesis of a potential benefit in training data with delay augmentation. Therefore, echo signals $d(n)$ in the training dataset were delayed randomly by up to 50 ms. The training of DeepAdaptive and CGGN16 was repeated under this condition. The NKF model was not retrained, as its shorter receptive field already prevented the integration of such long delays in training.

Fig. 11 shows the results for the retrained models in DT AECMOS scores. It can be seen that the echo suppression capabilities of the CGGN16 model is now less impacted by the introduced delay, only starting to drop at the 50 ms mark. Notably, the performance on undelayed test data is not majorly impacted by the training augmentation. DeepAdaptive only marginally deviates from its performance when trained without delay augmentation (compare to Fig. 10).

We conclude: Delay robustness is dependent on filter length / receptive field and—for some neural network-based approaches—on the respective training material.

V. Conclusion

In this article, we provided a comparative analysis of classical, hybrid, and deep learning-based acoustic echo cancellation (AEC) and suppression (AES) approaches, employing a broad assortment of metrics and test conditions. We demonstrated the benefits of our expansive evaluation by comparing five AEC/AES methods as representatives for classical, machine-learned, and hybrid approaches, in terms of their performance regarding convergence behaviour, reliability under different SER/SNR combinations or far-end signal types, as well as robustness to adverse conditions such as harsh nonlinearities, room impulse response switches or continuous changes, or delayed echo. We discussed limitations of the employed metrics in these conditions. This increased scope of evaluation revealed clear differences in the investigated models, which a simpler test condition would not have been able to detect. We observed a faster convergence and better tracking capabilities in the neural-network based model, while the traditional frequency-domain adaptive Kalman filter remains unbeaten in its preservation of near-end speech during double-talk. It was also shown that data-driven models such as NKF, DeepAdaptive, and CGGN16, despite their advantages over classical approaches in many conditions, can still be vulnerable to conditions unaccounted for in their training, such as harsh nonlinearities or highly delayed echo.

Appendix A
NLMS Algorithm

The normalized least means squares algorithm (NLMS, after [3]) in this article operates in the time domain and calculates the RIR estimate $\hat{\boldsymbol{h}}(n) = [\hat{h}_0(n), \hat{h}_1(n), \ldots, \hat{h}_{N-1}(n)]^T$ for a filter length of $N = 512$ as

$$\hat{\boldsymbol{h}}(n+1) = \hat{\boldsymbol{h}}(n) + \frac{\mu \cdot e^*(n)\boldsymbol{x}(n)}{||\boldsymbol{x}(n)||^2}, \tag{7}$$

with $\boldsymbol{x}(n) = [x(n), x(n-1), \ldots, x(n-N+1)]^T$, the step size $\mu = 0.7$, and the enhanced signal $e(n)$ calculated as

$$e(n) = y(n) - \hat{\boldsymbol{h}}^T(n) \cdot \boldsymbol{x}(n). \tag{8}$$

Note that $()^T$ is the transpose and $()^*$ is the conjugate complex operator.

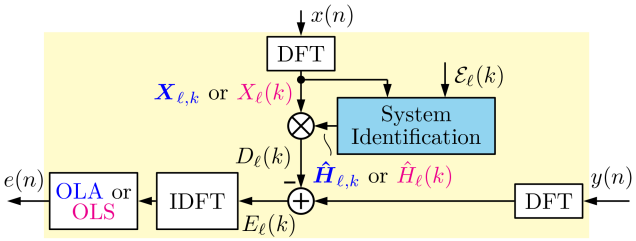

Fig. 12. FDKF and NKF algorithms (differences in respective colors); Details are shown in Figs. 13 and 14.

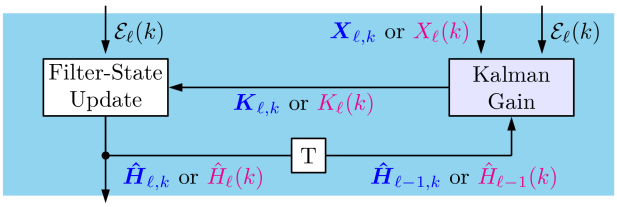

Fig. 13. **System identification** algorithm as being used in Fig. 12 for FDKF and NKF with their respectively colored signals. The block "T" delays incoming signals by one frame.

Appendix B
Frequency-Domain Kalman Filter

The frequency-domain Kalman filter (FDKF, after [5], [21]) used in this article is implemented in diagonalized form (all matrices have only entries on their diagonal axes and can be considered vectors). The overall structure is shown in Fig. 12 (magenta-colored options). The input signals $Y_\ell(k)$ and $X_\ell(k)$ are computed by rectangular windowing of the microphone signal $y(n)$ and reference signal $x(n)$ with window (and in consequence filter) length $K = 512$ and frame shift $R = 128$, and applying a K-point DFT to each frame. Its system identification block is depicted in Fig. 13 and can be broken down into two main modules: The filter-state update and the Kalman gain estimation. To update the filter state, first the current error signal $\mathcal{E}_\ell(k)$ has to be computed as

$$\mathcal{E}_\ell(k) = Y_\ell(k) - A\frac{R}{K}\hat{H}_{\ell-1}(k)X_\ell(k), \tag{9}$$

with the state transition factor $A = 0.998$. Then, the Kalman gain $K_\ell(k)$ is calculated from the step size $\mu_\ell(k)$ through

$$K_\ell(k) = \mu_\ell(k)X_\ell^*(k) \tag{10}$$

$$\mu_\ell(k) = \frac{R}{K}P_\ell^+(k)\cdot\left(\frac{R}{K}\left[|X_\ell(k)|^2P_\ell^+(k)\right] + \Psi_\ell^{\mathrm{S}}(k)\right)^{-1}, \tag{11}$$

whereby the process noise covariance is computed as

$$\Psi_\ell^{\mathrm{S}}(k) = (1-\beta)\left(|\mathcal{E}_\ell(k)|^2 + \frac{R}{K}|X_\ell(k)|^2P_\ell^+(k)\right) + \beta\Psi_{\ell-1}^{\mathrm{S}}(k), \tag{12}$$

with smoothing factor $\beta = 0.5$. The state error covariance $P_\ell^+(k)$ is computed as

$$P_\ell^+(k) = A^2P_\ell(k) + \lambda\Psi_\ell^{\triangle}(k)$$

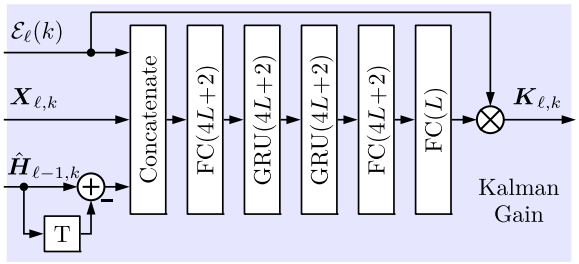

Fig. 14. **NKF Kalman gain estimation** block after [9]. The values in parentheses indicate the number of units per fully connected (FC) or GRU layer based on the filter-tap count L.

$$P_{\ell+1}(k) = P_\ell^+(k)\left(1 - \frac{R}{K}K_\ell(k)X_\ell(k)\right), \tag{13}$$

with the estimation error covariance

$$\Psi_\ell^{\triangle}(k) = \left(P_\ell(k) + |\hat{H}_{\ell-1}(k)|^2\right)\cdot(1-A^2). \tag{14}$$

With the Kalman gain $K_\ell(k)$ calculated for the current step by (10), the filter state $\hat{H}_\ell(k)$ can be updated through

$$\hat{H}_\ell(k) = A\hat{H}_{\ell-1}(k) + K_\ell(k)\mathcal{E}_\ell(k). \tag{15}$$

The enhanced signal $E_\ell(k)$, again according to Fig. 12, is finally calculated as

$$E_\ell(k) = Y_\ell(k) - \frac{R}{K}\hat{H}_\ell(k)X_\ell(k), \tag{16}$$

which is transformed back into the time domain via K-point IDFT and overlap-save.

Appendix C
Neural Kalman Filtering (NKF)

The Neural Kalman Filtering (NKF, after [9]) approach lends the overall structure from the FDKF in Figs. 12 and 13 (blue-colored options), but replaces the Kalman gain estimation algorithm for a *multi-tap* filter $\hat{\boldsymbol{H}}_{\ell,k} = [\hat{H}_\ell(k), \hat{H}_{\ell-1}(k), \ldots, \hat{H}_{\ell-L+1}(k)]^T$ with a DNN as shown in Fig. 14. Following the original implementation [9], $L = 4$ filter taps are predicted for window/DFT length $K = 512$, frame shift $R = 128$, and (contrary to the original FDKF) a square-root Hann windowing function. The effective filter length between all filter-taps accumulates to $M = K + (L-1)\cdot R = 896$. Note that the same DNN is applied to each frequency bin individually, but including context from $L-1$ past frames. The error signal $\mathcal{E}_\ell(k)$ is calculated as

$$\mathcal{E}_\ell(k) = Y_\ell(k) - \hat{\boldsymbol{H}}_{\ell-1,k}^T\boldsymbol{X}_{\ell,k}. \tag{17}$$

The filter state $\hat{\boldsymbol{H}}_{\ell,k}$ is updated through

$$\hat{\boldsymbol{H}}_{\ell,k} = A\hat{\boldsymbol{H}}_{\ell-1,k} + \boldsymbol{K}_{\ell,k}\mathcal{E}_\ell(k), \tag{18}$$

with $A = 1$, and then used to calculate the enhanced signal $E_\ell(k)$ as

$$E_\ell(k) = Y_\ell(k) - \hat{\boldsymbol{H}}_{\ell,k}^T\boldsymbol{X}_{\ell,k}. \tag{19}$$

If the magnitude of all entries in the reference signal $\boldsymbol{X}_{\ell,k} = [X_\ell(k), X_{\ell-1}(k), \ldots, X_{\ell-L+1}(k)]^T$ is below 10^{-5} for all k,

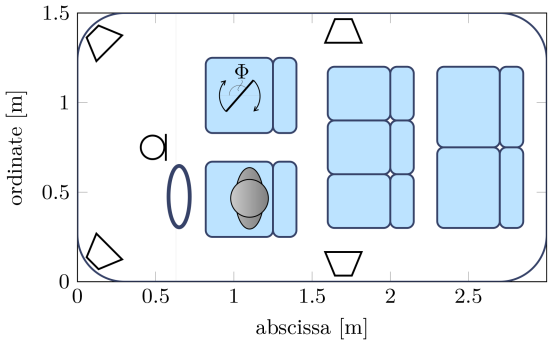

Fig. 15. Car setup for dynamic RIR generation [31], with hands-free microphone at rear-view mirror position and four loudspeakers. Changes in the RIR are simulated by rotating a reflecting surface on the passenger seat along the azimuth angle Φ.

no update is calculated and no echo is compensated ($E_\ell(k) = Y_\ell(k)$). Finally, the enhanced signal is subject to another square-root Hann window and transformed back into the time-domain via K-point IDFT and overlap-add (contrary to the `FDKF` implementation).

Appendix D
DeepAdaptive

Similar to the `NKF` approach, the `DeepAdaptive` model [22] augments a traditional frequency-domain acoustic echo canceller in its filter state update rule:

$$\hat{H}_{\ell+1}(k) = \hat{H}_\ell(k) + \frac{\mu_\ell(k)}{X_\ell^*(k)X_\ell(k)} E_\ell(k)X_\ell(k), \qquad (20)$$

with the complex conjugate operator $*$. This update rule can be seen as the frequency-domain equivalent to (7). A DNN consisting of four LSTM layers with 300 units each and a full-connected output layer is employed instead of the classical step size $\mu_\ell(k)$ calculation. Furthermore, the same DNN is also used (via a second fully connected output layer) to estimate a spectral magnitude mask $M_\ell(k)$, which is applied to the microphone signal to extract the nonlinearly distorted reference signal $X'_\ell(k)$. This is used as substitute for $X_\ell(k)$ in (20), thereby ideally reducing the task complexity back to a linear system identification problem. Analogous to the previous models, we choose the window/DFT length $K = 512$ and frame shift $R = 128$, while following the original authors in using $L = 10$ filter taps. Note that the resulting extended effective filter length $M = 1664$ does give the model an advantage over the other classical and hybrid models in terms of longer RIRs and delay, but might also result in longer convergence times.

Appendix E
Dynamic Impulse Response Recording

The setup for generating a dynamic RIR (as sketched in Fig. 15) follows the description of Jung et al. [31], which is also reflected in ITU-T P.1110 [39] and P.1130 [43].

The RIR is measured in a Volkswagen Touran car equipped with a hands-free microphone at the rear-view mirror position and four loudspeakers placed around the driver seat. A perfect sweep excitation signal [48] is played back from all loudspeakers simultaneously at high volume to achieve a good SNR at the microphone. A rotating piece of plywood is placed on the co-driver's seat to enforce changes in the RIR, being manually rotated at a speed of $\omega \approx 360°/4$ s. The driver's seat is occupied during the measurement.

References

[1] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1985.
[2] A. Feuer and E. Weinstein, "Convergence analysis of LMS filters with uncorrelated Gaussian data," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. TASSP-33, no. 1, pp. 222–230, Feb. 1985.
[3] P. Diniz, *Adaptive Filtering: Algorithms and Practical Implementation*. Boston, MA, USA: Kluwer, 1997.
[4] E. Hänsler and G. Schmidt, *Echo Cancellation*. Hoboken, NJ, USA: Wiley, 2004.
[5] 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, Jun. 2006.
[6] J. Franzen, I. Meyer zum Alten Borgloh, and T. Fingscheidt, "On the benefit of a stereo acoustic echo cancellation in an in-car communication system," in *Proc. IEEE 13th ITG Symp. Speech Commun.*, 2018, pp. 41–45.
[7] E. Shachar, I. Cohen, and B. Berdugo, "Acoustic echo cancellation with the normalized sign-error least mean squares algorithm and deep residual echo suppression," *Algorithms*, vol. 16, no. 3, 2023, Art. no. 137.
[8] Z. Chen et al., "A progressive neural network for acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–2.
[9] D. Yang, F. Jiang, W. Wu, X. Fang, and M. Cao, "Low-complexity acoustic echo cancellation with neural Kalman filtering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.
[10] N. L. Westhausen and B. T. Meyer, "Acoustic echo cancellation with the dual-signal transformation LSTM network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7138–7142.
[11] E. Seidel, J. Franzen, M. Strake, and T. Fingscheidt, "Y^2-Net FCRN for acoustic echo and noise suppression," in *Proc. Interspeech*, 2021, pp. 4763–4767.
[12] S. Braun and M. L. Valero, "Task splitting for DNN-based acoustic echo and noise removal," in *Proc. IEEE Int. Workshop Acoustic Signal Enhancement*, 2022, pp. 386–390.
[13] E. Seidel, P. Mowlaee, and T. Fingscheidt, "Efficient deep acoustic echo suppression with condition-aware training," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023, pp. 1–5.
[14] E. Indenbom, N.-C. Ristea, A. Saabas, T. Parnamaa, J. Guzvin, and R. Cutler, "DeepVQE: Real time deep voice quality enhancement for joint acoustic echo cancellation, noise suppression and dereverberation," in *Proc. Interspeech*, 2023, pp. 3819–3823.
[15] J.-M. Valin, S. Tenneti, K. Helwani, U. Isik, and A. Krishnaswamy, "Low-complexity, real-time joint neural echo control and speech enhancement based on percepnet," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7133–7137.
[16] S. Zhang et al., "Multi-task deep residual echo suppression with echo-aware loss," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 9127–9131.
[17] G. Revach, N. Shlezinger, R. J. G. v. Sloun, and Y. C. Eldar, "KalmanNet: Data-driven Kalman filtering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 3905–3909.
[18] P. Vary and R. Martin, *Digital Speech Transmission*. Hoboken, NJ, USA: Wiley, 2006.
[19] A. Ivry and I. Cohen, "Objective metrics to evaluate residual-echo suppression during double-talk in the stereophonic case," in *Proc. Interspeech*, 2022, pp. 5348–5352.
[20] S. L. Gay and J. Benesty, *Acoustic Signal Processing for Telecommunication*. Norwell, MA, USA: Kluwer Academic Publishers, 2000.
[21] J. Franzen and T. Fingscheidt, "An efficient residual echo supression for multi-channel acoustic echo cancellation based on the frequency-domain adaptive Kalman filter," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 226–230.
[22] H. Zhang, S. Kandadai, H. Rao, M. Kim, T. Pruthi, and T. Kristjansson, "Deep adaptive AEC: Hybrid of deep learning and adaptive acoustic echo cancellation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 756–760.

[23] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," University of Edinburgh. The Centre for Speech Technology Research, 2017.
[24] H. Zhang and D. Wang, "A deep learning approach to active noise control," in *Proc. Interspeech*, 2020, pp. 1141–1145.
[25] W. Klippel, "Tutorial: Loudspeaker nonlinearities–causes, parameters, symptoms," *J. Audio Eng. Soc.*, vol. 54, no. 10, pp. 907–939, Oct. 2006.
[26] M.-A. Jung, S. Elshamy, and T. Fingscheidt, "An automotive wideband stereo acoustic echo canceler using frequency-domain adaptive filtering," in *Proc. IEEE 22nd Eur. Signal Process. Conf.*, 2014, pp. 1452–1456.
[27] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 3591–3591, 2013.
[28] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Interspeech*, 2010, pp. 3110–3113.
[29] "Speech Processing, Transmission and quality aspects (STQ); speech quality performance in the presence of background noise; part 1: Background noise simulation technique and background noise database," *Eur. Telecommun. Standard Inst.*, Tech. Rep. ETSI EG 202 396-1, Sep. 2008.
[30] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, *TIMIT Acoustic-Phonetic Continous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.
[31] M.-A. Jung, L. Richter, and T. Fingscheidt, "Towards reproducible evaluation of automotive hands-free systems in dynamic conditions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 8144–8148.
[32] U. Zölzer, *DAFX: Digital Audio Effects*. Hoboken, NJ, USA: Wiley, 2003.
[33] M. Jeub, M. Schäfer, H. Krüger, C. M. Nelke, C. Beaugeant, and P. Vary, "Do we need dereverberation for hand-held telephony?," in *Proc. Int. Congr. Acoust.*, 2010, pp. 1–7.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.
[35] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Annu. Conf. Neural Int. Process. Syst.*, 2019, pp. 8024–8035.
[36] M. Purin, S. Sootla, M. Sponza, A. Saabas, and R. Cutler, "AECMOS: A speech quality assessment metric for echo impairment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 901–905.
[37] ITU, "Rec. P.862.2 corrigendum 1: Wideband extension to Rec. P.862 for the assessment of wideband telephone networks and speech codecs," International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), P.862.2, Oct. 2017.
[38] ITU, "Rec. P.1100: Narrowband hands-free communication in motor vehicles," International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), P.1100, Jan. 2019.
[39] ITU, "Rec. P.1110: Wideband hands-free communication in motor vehicles, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), P.1110, Jan. 2019.
[40] T. Fingscheidt and S. Suhadi, "Quality assessment of speech enhancement systems by separation of enhanced speech, noise, and echo," in *Proc. Interspeech*, 2007, pp. 818–821.
[41] T. Fingscheidt, S. Suhadi, and K. Steinert, "Towards objective quality assessment of speech enhancement systems in a black box approach," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 273–276.
[42] I. Katsir, I. Cohen, and D. Malah, "Speech bandwidth extension based on speech phonetic content and speaker vocal tract shape estimation," in *Proc. Eur. Signal Process. Conf.*, 2011, pp. 461–465.
[43] ITU, Rec., "P.1130: Subsystem requirements for automotive speech services," *Int. Telecommun. Standardization Sector*, P.1130, Jun. 2015.
[44] H. Zhang and D. Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," in *Proc. Interspeech*, 2018, pp. 3239–3243.
[45] S. Zhang et al., "F-T-LSTM based complex network for joint acoustic echo cancellation and speech enhancement," in *Proc. Interspeech*, 2021, pp. 4758–4762.
[46] L. Pfeifenberger, M. Zoehrer, and F. Pernkopf, "Acoustic echo cancellation with cross-domain learning," in *Proc. Interspeech*, 2021, pp. 4753–4757.
[47] R. Peng et al., "Acoustic echo cancellation using deep complex neural network with nonlinear magnitude compression and phase information," in *Proc. Interspeech*, 2021pp. 4768–4772.
[48] C. Antweiler, A. Telle, P. Vary, and G. Enzner, "Perfect-sweep NLMS for time-variant acoustic system identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 517–520.

Ernst Seidel received the M.Sc. degree in electrical engineering in 2021 from Technische Universität Braunschweig, Braunschweig, Germany, where he is currently working toward the Ph.D. degree with the Institute for Communications Technology, in the field of machine-learned acoustic echo cancellation. His research interests include designs and evaluates classical, machine-learned, and hybrid approaches to acoustic echo cancellation and suppression. His work achieved the 4th rank in the Interspeech 2021 Acoustic Echo Cancellation Challenge.

Pejman Mowlaee (Senior Member, IEEE) received the M.Sc. degree in communication systems from the Iran University of Science and Technology, Tehran, Iran, in 2006, and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 2010. From 2011 to 2012, he was a Postdoc Fellow with Marie Curie Programme AUDIS (digital signal processing in audiology), at the Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany. From 2012 to 2017, he was an Assistant Professor with Graz University of Technology, Graz, Austria. In 2017, he was appointed as Privatdozent with Habilitation in speech signal processing and Adjunct Professor. From 2017 to 2020, he was a Signal Processing Specialist with WS Audiology A/S, Denmark. Since 2020, he has been a Lead Research Scientist with GN Audio A/S with main responsibility in scouting and development of new signal processing concepts for speech communication devices. He is coauthor of the book "*Single Channel Phase-Aware Signal Processing in Speech Communication: Theory and Practice*", John Wiley & Sons 2017. His research interests include signal processing and machine learning for audio applications. He is an Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING and an elected Member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing.

Tim Fingscheidt (Senior Member, IEEE) received the Dipl.-Ing. degree in electrical engineering and the Ph.D. degree from RWTH Aachen University, Aachen, Germany, in 1993 and 1998, respectively. He was with AT&T Labs, Florham Park, NJ, USA, in 1998 and Siemens AG (Mobile Devices), Munich, Germany, in 1999. He was leading the speech technology development activities, with Siemens Corporate Technology, Munich, during 2005–2006. Since 2006, he has been a Full Professor with the Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany. His research interests include signal processing and machine learning, with applications in speech processing and computer vision. He has been the Speaker of the Speech Acoustics Committee ITG AT3 since 2015, and is currently the Speaker of the ITG branch Audio Technology. He was an Associate Editor for IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING during 2008–2010, and *EURASIP Journal on Audio, Speech, and Music Processing* during 2013–2018. He was a Member of the IEEE Speech and Language Processing Technical Committee during 2012–2014 and 2016–2018. Since 2023, he has been ITG Fellow and ITG Board Member. He was the recipient of several awards, including the Vodafone Mobile Communications Foundation prize in 1999 and 2002 ITG award of the Association of German Electrical Engineers (VDE ITG). In 2017 and 2020, he coauthored the ITG award-winning publication, and in 2019, 2020, and 2021 he was given the Best Paper Award of a CVPR Workshop.