

# AEC IN A NETSHELL: ON TARGET AND TOPOLOGY CHOICES FOR FCRN ACOUSTIC ECHO CANCELLATION

*Jan Franzen, Ernst Seidel, Tim Fingscheidt*

Institute for Communications Technology, Technische Universität Braunschweig,  
Schleinitzstr. 22, 38106 Braunschweig, Germany

{j.franzen, ernst.seidel, t.fingscheidt}@tu-bs.de

## ABSTRACT

Acoustic echo cancellation (AEC) algorithms have a long-term steady role in signal processing, with approaches improving the performance of applications such as automotive hands-free systems, smart home and loudspeaker devices, or web conference systems. Just recently, very first deep neural network (DNN)-based approaches were proposed with a DNN for joint AEC and residual echo suppression (RES)/noise reduction, showing significant improvements in terms of echo suppression performance. Noise reduction algorithms, on the other hand, have enjoyed already a lot of attention with regard to DNN approaches, with the fully convolutional recurrent network (FCRN) architecture being among state of the art topologies. The recently published impressive echo cancellation performance of joint AEC/RES DNNs, however, so far came along with an undeniable impairment of speech quality. In this work we will heal this issue and significantly improve the near-end speech component quality over existing approaches. Also, we propose for the first time—to the best of our knowledge—a pure DNN AEC in the form of an echo estimator, that is based on a competitive FCRN structure and delivers a quality useful for practical applications.

**Index Terms**— acoustic echo cancellation, echo suppression, convolutional neural network, ConvLSTM

## 1. INTRODUCTION

Applications such as automotive hands-free systems, smart home and loudspeaker devices, web conference systems, and many more share a similar underlying challenge: The microphone signal picks up an undesired echo component stemming from the system's own loudspeakers. With acoustic echo cancellation (AEC) algorithms having a steady role in signal processing over the past decades, these algorithms typically deploy an adaptive filter to estimate the impulse response (IR) of the loudspeaker-enclosure-microphone (LEM) system. The echo component is then estimated and subsequently subtracted from the microphone signal to obtain a widely echo-free enhanced near-end speech signal.

Traditional AEC algorithms [1, 2, 3] have a long-term role in signal processing, with the approaches steadily evolving, resulting in renowned algorithms using normalized least mean squares (NLMS) algorithm [4] or the Kalman filter [5, 6], and including residual echo suppression (RES) approaches [7, 8]. In the recent past, neural networks—especially convolutional neural networks—have shown significant performance in speech enhancement in general, e.g., the work of Strake et al. [9] for noise reduction. However, AEC has only seen very few data-driven approaches so far. Initially, only networks for RES were among them [10, 11].

Just recently, a fully learned AEC was proposed by Zhang et al. [12, 13], revealing an impressive echo cancellation performance. An interesting aspect of these works is the way the AEC problem is tackled. It is treated as source separation approach with the networks being trained to directly output the estimated enhanced signal.

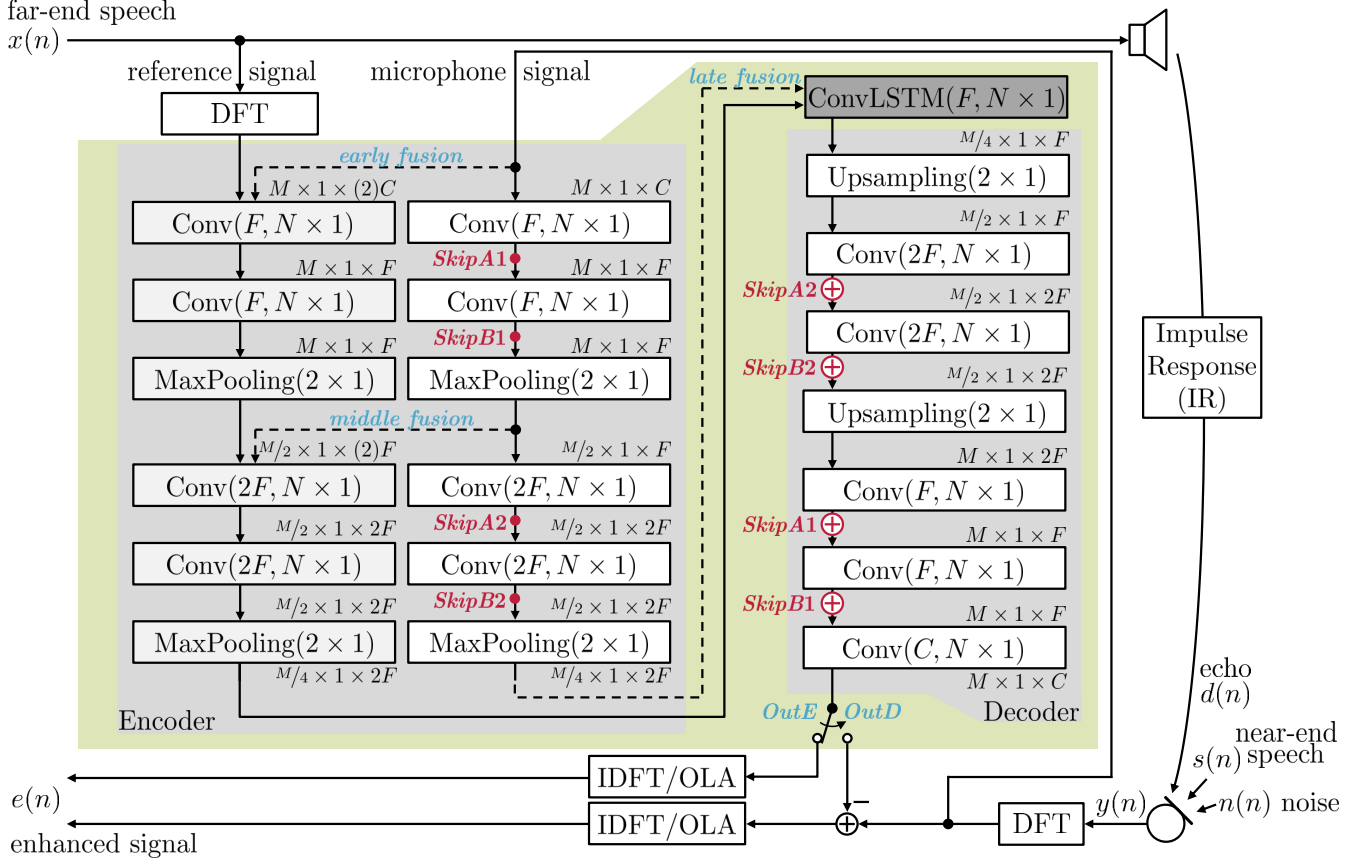
The difficulty with AEC DNNs is, however, that they so far come along with an undeniable impairment of the near-end speech component quality. In this work, we will investigate this issue with a set of experiments to show and reveal trade-offs between different performance aspects in terms of echo suppression, noise reduction, and near-end speech quality. With the fully convolutional recurrent network (FCRN) [9, 14] and its proven capability of autoencoding speech at high fidelity as a basis, we will introduce several DNN AEC architectures overcoming earlier problems, thereby significantly improving over existing approaches. We will provide useful insights into network design choices, giving the reader guidance on the not yet widely explored field of DNN AEC.

The remainder of this paper is structured as follows: In Section 2, a system overview including the framework and general network topology is given. The training and different experimental variants including novel network topology choices are described in Section 3. In Section 4, the experimental validation and discussion of all approaches is given. Section 5 provides conclusions.

## 2. NETWORK TOPOLOGY, SIMULATION FRAMEWORK, AND DATA

### 2.1. Novel FCRN Network Topology

In contrast to traditional adaptive filters, the AEC itself is performed by a neural network in this work. Basis for our experiments is the well-performing fully convolutional recurrent network (FCRN) encoder-decoder structure proposed for noise reduction in [9]. However, we introduce important AEC specifics into the network topology. Our proposed network is depicted within the green box of Figure 1, operates on discrete Fourier transform (DFT) inputs  $X_\ell(k)$  with frame index  $\ell$  and frequency bin  $k$ , and contains a couple of novelties: Originally consisting of only one encoder (i.e., here, most likely comparable to performing early fusion with the microphone signal DFT  $Y_\ell(k)$  and following only the respective dashed signal path), we investigate a parallel second encoder (portion) consisting of up to two times two convolutional layers, followed by maximum pooling over the feature dimension with stride 2. The first two convolutional layers use  $F$  filter kernels of size  $N \times 1$  (convolution over the feature axis), whereas the latter two use  $2F$  filter kernels of the same size. Leaky ReLU activations [15] are used for these layers.



**Fig. 1.** System model and network with various options for skip connections, encoder fusion and training targets. The parameters of the convolutional layers are  $\text{Conv}(\# \text{ filters}, \text{kernel dimensions})$ , maximum pooling  $\text{MaxPooling}(\text{pool dimensions})$  over the feature axis, and likewise for upsampling. Skip connection variants (*none*, *SkipA*, *SkipB*) are indicated by the respective start and end positions with the same naming. Only a single dashed path (---) is involved at a time, determining either *early*, *middle*, or *late fusion* of the reference and microphone signals in the encoder. Training target options are indicated by the switch positions *OutE* (speech target) or *OutD* (echo target).

For easier readability, feature dimensions can be seen at the in- and output of each layer, denoted as *feature axis*  $\times$  *time axis*  $\times$  *number of feature maps*. During inference, the network is subsequently processing single input frames, which is indicated by the time axis value being set to 1.

At the bottleneck, right after the encoder, where the feature axis reaches its maximum compression of  $M/4$ , a convolutional LSTM [16] with  $F$  filter kernels of size  $N \times 1$  is placed, enabling the network to model temporal context. The decoder is set up exactly as inverse to the encoder and followed by a final convolutional layer with linear activations to yield the final output of dimension  $M \times 1 \times C$ . To extract input features and training targets for the structures given in Figure 1, at a sampling rate of  $f_s = 16$  kHz a frame length of  $K = 512$  samples is used and the frame shift is set to 256 samples. By applying a square root Hann window and 512-point DFT, complex spectra are obtained. Separated into real and imaginary parts, and zero-padded to feature maps of height  $M = 260$ , this leads to  $C = 2$  channels for the reference, the microphone, and the estimated echo or (clean) speech signal.

## 2.2. Simulation Framework and Data

To model the acoustic setup shown in Figure 1, we adopt the procedure described in [13] with some modifications. Thus, to model typical single- and double-talk scenarios, far-end speech  $x(n)$  and

near-end speech  $s(n)$  are set up using the TIMIT dataset [17]. Background noises  $n(n)$  are taken from the QUT dataset [18] for training and validation, while babble, white noise, and the operations room noises are used from the NOISEX-92 dataset [19] for the test set. Noise  $n(n)$  is superimposed with near-end speech  $s(n)$  at the microphone, and echo signals  $d(n)$  are generated by imposing loudspeaker nonlinearities [13] on far-end signals  $x(n)$  and convolving them with impulse responses (IRs) of 512 samples length. The IRs are created using the image method [20] with reverberation times  $T_{60} \in \{0.2, 0.3, 0.4\}$  s for training and validation, and 0.2 s for test mixtures, thereby following [13]. A test with additional real IRs is omitted here for space reasons, since it was impressively shown in [13] that apparently for DNN AECs comparable results are obtained for both, real and simulated IRs. For a broad variety of simulations, signal-to-echo ratios (SER) are selected randomly between  $\{-6, -3, 0, 3, 6, \infty\}$  dB per mixture and signal-to-noise ratios (SNR) are selected randomly between  $\{8, 10, 12, 14, \infty\}$  dB per mixture. Note, that we included  $\infty$  dB to the SER and SNR values, since for a practical application it is absolutely mandatory that the absence of echo or noise can be handled by the network as well. In our setup this leads to a total of 3000 training, 500 evaluation, and 280 test mixtures, whereas the latter—differing from [13]—consist of *unseen* speakers from the CSTR VCTK database [21] with unseen utterances, impulse responses, and noise sequences. SER and SNR for the test mixtures are set to 0 dB and 10 dB, respectively. For

deeper insights into the network performance, we additionally evaluate the test files but consisting either of echo only, or of near-end noise or near-end speech only.

### 3. EXPERIMENTAL VARIANTS AND TRAINING

#### 3.1. Training Target Variants

One major question we investigate is rather significant and concerns the choice of the training targets. Here, [13] differs from the traditional concept for AEC, where an estimated echo  $\hat{d}(n)$  is generated, which is then subtracted from the microphone signal to obtain an (ideally) echo-free enhanced signal  $e(n)$ . In [12, 13], however, the echo problem is tackled by a source separation approach trained to directly output the estimated enhanced signal  $E_\ell(k)$ , thereby enabling two meaningful possibilities for the regression training targets  $\bar{E}_\ell(k)$  in the DFT domain: The complex-valued target can either be chosen as  $\bar{E}_\ell(k) = S_\ell(k) + N_\ell(k)$  (i.e., only echo cancellation is performed by the network), or just as  $\bar{E}_\ell(k) = S_\ell(k)$  (i.e., echo and noise cancellation are performed). This leads to the question which of the mentioned targets are best suited and if there are any tradeoffs to be dealt with.

Indicated by the network output switch in Figure 1, we investigate the two different variants of training targets with an MSE loss  $J_\ell = \frac{1}{K} \sum_{k \in K} |\hat{E}_\ell(k) - \bar{E}_\ell(k)|^2$  in the frequency domain (switch position **OutE**), with  $\hat{E}_\ell(k)$  being the respective network outputs. As the third variant, the MSE loss  $J_\ell = \frac{1}{K} \sum_{k \in K} |\hat{D}_\ell(k) - \bar{D}_\ell(k)|^2$  is applied using the echo component training targets  $\bar{D}_\ell(k) = D_\ell(k)$  directly with subsequent subtraction from the microphone signal (switch position **OutD**).

#### 3.2. Skip Connection Variants

Throughout this work we will experiment with different positions for skip connections reaching from encoder to decoder. The original model has a skip connection placed between the red marked points *SkipB1* and another one between the points *SkipB2* [9]. Hereinafter, this setup will be denoted as **SkipB**. With the varying dimensions of the feature maps, a second possibility is given by placing the skip connections in a symmetric manner, i.e., one between the points *SkipA1* and another one between the points *SkipA2*. This setup will be denoted as **SkipA**. The last variant is to use no skip connections at all, which will be denoted as **NoSkips** (—).

#### 3.3. Encoder Fusion Variants

A traditional AEC algorithm uses the reference signal  $x(n)$  as input and replicates the IR with an adaptive filter. The microphone signal  $y(n)$  (or rather a thereon based error signal) serves as control input for the adaptive filter. In contrast, for the network of [13], as for our network when a combined encoder is used (early fusion), the feature maps of reference and microphone signal are directly concatenated at the network input, denoted as **EarlyF**.

However, the original idea of using an encoder-decoder structure is to allow the network to preprocess and find a suitable representation of its input signals throughout the encoder, which can then be well-handled by its bottleneck layer. At this point it is important to notice that convolutional layers with  $N \times 1$  kernels along the frequency axis are not able to model delay, which we consider of crucial importance for handling the temporal shift between reference and microphone signal. Since our main processing unit in the bottleneck layer is a convolutional LSTM that can indeed model delay, we experiment with performing fusion of the reference and microphone

signals at different positions in the encoder. This shall allow the network to process microphone and reference signals separately to a certain degree, before the respective feature maps are concatenated and processed together throughout the remaining network.

Two further variants for the encoder fusion will be considered: The first one is the middle fusion, in the following denoted as **MidF**, where only the respective dashed signal path is involved. The second variant duplicates the entire encoder and performs feature map concatenation at the input to the convolutional LSTM. Here, only the last respective dashed signal path is used. The method is denoted as **LateF** in the experiments.

If middle or late fusion is performed in combination with skip connections, the skip connections are branched off of the microphone signal path as indicated in Figure 1. We also considered placing their starting points in the reference signal path, but as can be expected this does not lead to any meaningful results. When early fusion is performed, the skip connections are branched off of the respective positions of the common encoder.

#### 3.4. Training Parameters

Networks are trained with the Adam optimizer [22] using its standard parameters. The batch size and sequence length are set to 16 and 50, respectively. With an initial learning rate of  $5 \cdot 10^{-5}$ , the learning rate is multiplied with a factor of 0.6 if the loss did not improve for 3 epochs. Training is stopped when the learning rate drops below  $5 \cdot 10^{-6}$  or if the loss did not improve for 10 epochs. The number of parameters varies with the encoder fusion position, resulting in 5.2 M and 5.6 M parameters for **EarlyF** and **MidF**, respectively, and 7.1 M parameters for **LateF**.

## 4. RESULTS AND DISCUSSION

Experiment results for all combinations of our proposed variants are shown in Tables 1 to 3 using three measures to rate the performance in different categories: the recently updated wideband PESQ MOS LQO [23, 24] for speech quality, SNR improvement as deltaSNR (dSNR) in [dB] for noise reduction, and echo return loss enhancement (ERLE) as  $ERLE(n) = 10 \cdot \log \left( d^2(n) / (d(n) - \hat{d}(n))^2 \right)$  for echo suppression. The final ERLE is computed as in [25] using first order IIR smoothing with factor 0.9996 on each of the sample-wise components  $d(n)$  and  $\hat{d}(n)$ , and is averaged over whole files.

Each table is split into two major parts: the three right-most columns provide the network performance when the input files consist either of echo only ( $d(n)$ , rated with ERLE), or of near-end noise ( $n(n)$ , rated with dSNR) or near-end speech only ( $s(n)$ , rated with PESQ). These results allow deep insights into each network model: How does it handle echo or near-end noise if no other signals are present? And most importantly: Can the model 'simply' pass through clean near-end speech?

The four center columns provide results for the normal previously described test sets, i.e., *full mixture* input signals. Here, the PESQ MOS for the full output signal is evaluated, and the so-called black box approach according to ITU-T Recommendation P.1110 [26, sec. 8] and [27, 28, 29] is used to obtain the processed *components* of the enhanced signal  $e(n) = \hat{d}(n) + \hat{n}(n) + \hat{s}(n)$ , thereby allowing to compute ERLE, dSNR, and PESQ on the separated processed components  $\hat{d}(n)$ ,  $\hat{n}(n)$ , and  $\hat{s}(n)$ , respectively. These measures are marked with the index **BB** (for black box).

To allow for a better rating of the results, we additionally provide the performance of a traditional AEC algorithm, the well-known variationally diagonalized state-space frequency domain

**Table 1.** Experiment results: ERLE and deltaSNR given in [dB], and PESQ MOS LQO for all models with **clean speech** training target **OutE**:  $\bar{E}_\ell(k) = S_\ell(k)$  following [13]. For deeper insights, the three columns on the right show the respective performance when only one component is present at the microphone. Best result per measure is marked in **bold** font, second best is underlined.

Model/ Skip	full mixture				$d(n)$	$n(n)$	$s(n)$
	PESQ	ERLE <sub>BB</sub>	dSNR <sub>BB</sub>	PESQ <sub>BB</sub>	ERLE	dSNR	PESQ
Kalman	1.15	3.49	-0.94	<b>4.64</b>	18.36	—	<b>4.64</b>
EarlyF/B	1.52	19.49	11.05	2.82	11.83	32.94	3.24
EarlyF/A	1.44	20.73	10.94	2.56	13.20	<u>33.33</u>	<u>3.65</u>
<b>EarlyF/—</b>	<b>1.57</b>	<b>25.85</b>	<u>11.47</u>	2.64	<b>21.33</b>	32.94	3.44
MidF/B	1.52	20.87	10.24	<u>2.83</u>	15.79	29.65	3.64
MidF/A	1.49	21.00	11.25	2.70	15.45	27.12	3.32
MidF/—	<u>1.56</u>	24.63	11.05	2.81	18.25	32.97	3.33
LateF/B	1.45	23.27	<b>11.68</b>	2.62	17.77	27.87	3.44
LateF/A	1.52	20.40	10.43	2.70	15.23	27.38	3.42
LateF/—	1.53	<u>24.80</u>	10.81	2.66	<u>19.23</u>	<b>33.62</b>	3.28

**Table 2.** Experiment results for all models as in Table 1, but with **noisy speech** training target **OutE**:  $\bar{E}_\ell(k) = S_\ell(k) + N_\ell(k)$ . Best result per measure is marked in **bold** font, second best is underlined. Additional result for best model EarlyF/A with separate subsequent noise reduction from [9], retrained on this work’s data (EarlyF/A+).

Model/ Skip	full mixture				$d(n)$	$n(n)$	$s(n)$
	PESQ	ERLE <sub>BB</sub>	dSNR <sub>BB</sub>	PESQ <sub>BB</sub>	ERLE	dSNR	PESQ
Kalman	1.15	3.49	-0.94	<b>4.64</b>	18.36	—	<b>4.64</b>
EarlyF/B	<b>1.51</b>	8.99	0.63	3.10	6.95	3.20	4.19
<b>EarlyF/A</b>	1.46	9.57	0.76	2.85	15.93	0.66	<u>4.42</u>
EarlyF/—	1.44	<u>11.24</u>	0.79	2.70	15.89	2.62	3.83
MidF/B	1.46	9.41	0.73	2.94	13.39	-0.99	4.05
MidF/A	<u>1.49</u>	9.21	0.68	2.91	16.01	-1.08	4.05
MidF/—	1.47	8.48	0.53	3.00	15.68	0.85	3.60
LateF/B	1.43	<b>11.67</b>	0.76	2.64	<b>18.51</b>	-0.62	4.20
LateF/A	1.46	10.12	0.57	2.82	14.90	2.39	4.20
LateF/—	1.44	9.56	0.41	2.80	<u>17.74</u>	1.19	3.91
EarlyF/A+	1.49	18.78	4.44	2.38	19.96	14.95	3.52

**Table 3.** Experiment results for all models as in Table 2, but with **echo** training target **OutD**:  $\bar{D}_\ell(k) = D_\ell(k)$ , and subsequent subtraction from the microphone signal. Best result per measure is marked in **bold** font, second best is underlined. Additional result for best model LateF/A with separate subsequent noise reduction from [9], retrained on this work’s data (LateF/A+).

Model/ Skip	full mixture				$d(n)$	$n(n)$	$s(n)$
	PESQ	ERLE <sub>BB</sub>	dSNR <sub>BB</sub>	PESQ <sub>BB</sub>	ERLE	dSNR	PESQ
Kalman	1.15	3.49	-0.94	<b>4.64</b>	<b>18.36</b>	—	<b>4.64</b>
EarlyF/B	1.94	8.71	0.68	3.21	7.89	4.40	4.26
EarlyF/A	1.86	<b>9.75</b>	0.47	2.66	7.81	2.02	4.45
EarlyF/—	1.93	6.56	0.72	<u>3.68</u>	6.70	3.50	4.34
MidF/B	<u>1.98</u>	7.90	0.69	3.28	8.50	4.51	4.02
MidF/A	1.89	7.17	0.87	3.14	14.49	4.27	4.46
MidF/—	1.92	6.73	0.74	3.57	7.48	1.81	4.32
LateF/B	1.88	<u>9.17</u>	0.50	3.01	13.66	3.63	4.02
<b>LateF/A</b>	1.86	9.06	0.69	3.17	<u>16.65</u>	0.36	<u>4.50</u>
LateF/—	<b>2.10</b>	5.85	0.36	3.59	2.87	2.06	4.47
LateF/A+	1.58	18.42	5.29	2.66	20.65	21.23	3.55

adaptive Kalman filter including its residual echo suppression post-filter [5, 30, 31, 32], as reference point.

Table 1 shows the results for all models with **clean speech** training target **OutE**:  $\bar{E}_\ell(k) = S_\ell(k)$ . The models without skip connections achieve higher echo and noise suppression with up to 21.33 dB ERLE and 33.62 dB dSNR when only the respective component is present at the microphone. A clear preference for the encoder fusion position cannot be seen for this target choice, but the early fusion model *EarlyF/—* shows best overall trade-off results; note that Zhang et al. [12, 13] also perform early fusion with clean speech targets. However, the strong suppression performance of clean speech targets comes at a price: With the PESQ values not going above 3.65 MOS, none of the models is able to pass through clean speech. This can also be seen in the full mixture results, especially when compared with the perfect near-end speech component score PESQ<sub>BB</sub> of the Kalman filter reference.

With the **noisy speech** target choice **OutE**:  $\bar{E}_\ell(k) = S_\ell(k) + N_\ell(k)$  in Table 2, the PESQ scores are slightly increased, whereas the echo suppression performance is slightly decreased, when only the respective components are present at the microphone. It can be seen that skip connections are extremely helpful to pass through clean speech, and—considering the near-end speech quality—the best overall trade-off design for these targets is the early fusion model *EarlyF/A*. However, the PESQ scores on the full mixture remain comparable to those in Table 1. The diversity of the results again shows how important the design choices are in order to find a good tradeoff between suppression performance and near-end speech quality.

Finally, the results for our newly proposed **echo** training target **OutD**:  $\bar{D}_\ell(k) = D_\ell(k)$ , and subsequent subtraction from the microphone signal, are displayed in Table 3. The later fusion positions prove highly beneficial and lead to the best model *LateF/A* for these targets. In contrast to the previous tables, this model does not only achieve a high echo suppression but maintains the best near-end speech quality at the same time. While this specific model also outperforms the best trade-off model from Table 2, the full mixture PESQ scores (leftmost column) for all models are clearly above all other target choices.

For the two best trade-off models in Tables 2 and 3, we considered to perform a subsequent separate noise reduction [9] on output signal  $e(n)$  as postprocessor after AEC, trained on this work’s data (symbol+). The results are displayed in the bottom lines of the tables. As to be expected, they show improved noise and residual echo suppression, but interestingly reveal a degradation of near-end speech again—whereas only our proposed *LateF/A* DNN echo target AEC was able to maintain the near-end speech quality.

## 5. CONCLUSIONS

We presented a deeper investigation of acoustic echo cancellation with fully convolutional neural networks. Along with a newly proposed network structure in the form of an echo estimator that delivers a significantly improved near-end speech quality over existing approaches (model: *LateF/A* DNN, echo target, Table 3), we revealed trade-offs between different performance aspects in terms of echo suppression, noise reduction, and near-end speech quality, thereby giving the reader guidance on crucial design choices for the not yet widely explored field of DNN AEC.

## 6. REFERENCES

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

- [2] J. Lee and C. Un, "Block Realization of Multirate Adaptive Digital Filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 105–117, Feb. 1986.
- [3] H. Shin, A. H. Sayed, and W. Song, "Variable Step-Size NLMS and Affine Projection Algorithms," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 132–135, Feb. 2004.
- [4] K. Steinert, M. Schönle, C. Beaugéant, and T. Fingscheidt, "Hands-free System with Low-Delay Subband Acoustic Echo Control and Noise Reduction," in *Proc. of ICASSP*, Las Vegas, NV, USA, Apr. 2008, pp. 1521–1524.
- [5] G. Enzner and P. Vary, "Frequency-Domain Adaptive Kalman Filter for Acoustic Echo Control in Hands-Free Telephones," *Signal Processing (Elsevier)*, vol. 86, no. 6, pp. 1140–1156, June 2006.
- [6] J. Franzen and T. Fingscheidt, "A Delay-Flexible Stereo Acoustic Echo Cancellation for DFT-Based In-Car Communication (ICC) Systems," in *Proc. of INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 181–185.
- [7] F. Kuech, E. Mabande, and G. Enzner, "State-Space Architecture of the Partitioned-Block-Based Acoustic Echo Controller," in *Proc. of ICASSP*, Florence, Italy, May 2014, pp. 1295–1299.
- [8] J. Franzen and T. Fingscheidt, "An Efficient Residual Echo Suppression for Multi-Channel Acoustic Echo Cancellation Based on the Frequency-Domain Adaptive Kalman Filter," in *Proc. of ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 226–230.
- [9] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Fully Convolutional Recurrent Networks for Speech Enhancement," in *Proc. of ICASSP*, Barcelona, Spain, May 2020, pp. 6674–6678.
- [10] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral Feature-Based Nonlinear Residual Echo Suppression," in *Proc. of WASPAA*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.
- [11] G. Carbajal, R. Serizel, E. Vincent, and É. Humbert, "Multiple-Input Neural Network-Based Residual Echo Suppression," in *Proc. of ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 231–235.
- [12] H. Zhang and D.L. Wang, "Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios," in *Proc. of INTERSPEECH*, Hyderabad, India, Sept. 2018, pp. 3239–3243.
- [13] H. Zhang, K. Tan, and D.L. Wang, "Deep Learning for Joint Acoustic Echo and Noise Cancellation with Nonlinear Distortions," in *Proc. of INTERSPEECH*, Graz, Austria, Sept. 2019, pp. 4255–4259.
- [14] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional Neural Networks to Enhance Coded Speech," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 663–678, Apr. 2019.
- [15] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier Nonlinearities Improve Neural Network Acoustic Models," in *Proc. of ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*, Atlanta, GA, USA, June 2013, pp. 1–6.
- [16] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in *Proc. of NIPS*, Montreal, QC, Canada, Dec. 2015, pp. 802–810.
- [17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, Philadelphia, PA, USA, 1993.
- [18] 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. of INTERSPEECH*, Makuhari, Japan, Sept. 2010, p. 3110–3113.
- [19] A. Varga and H. J. Steeneken, "Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.
- [20] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.
- [21] 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.
- [22] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of ICLR*, San Diego, CA, USA, May 2015, pp. 1–15.
- [23] "ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," ITU, Nov. 2007.
- [24] "ITU-T Recommendation P.862.2 Corrigendum 1, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," ITU, Oct. 2017.
- [25] M.-A. Jung and T. Fingscheidt, "A Shadow Filter Approach to a Wideband FDAF-Based Automotive Handsfree System," in *5th Biennial Workshop on DSP for In-Vehicle Systems*, Kiel, Germany, Sept. 2011, pp. 60–67.
- [26] "ITU-T Recommendation P.1110, Wideband Hands-Free Communication in Motor Vehicles," ITU, Mar. 2017.
- [27] T. Fingscheidt and S. Suhadi, "Quality Assessment of Speech Enhancement Systems by Separation of Enhanced Speech, Noise, and Echo," in *Proc. of INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 818–821.
- [28] T. Fingscheidt, S. Suhadi, and K. Steinert, "Towards Objective Quality Assessment of Speech Enhancement Systems in a Black Box Approach," in *Proc. of ICASSP*, Las Vegas, NV, USA, Apr. 2008, pp. 273–276.
- [29] K. Steinert, S. Suhadi, T. Fingscheidt, and M. Schönle, "Instrumental Speech Distortion Assessment of Black Box Speech Enhancement Systems," in *Proc. of IWAENC*, Seattle, WA, USA, Sept. 2008, pp. 1–4.
- [30] S. Malik and J. Benesty, "Variationally Diagonalized Multi-channel State-Space Frequency-Domain Adaptive Filtering for Acoustic Echo Cancellation," in *Proc. of ICASSP*, Vancouver, BC, Canada, May 2013, pp. 595–599.
- [31] M. A. Jung, S. Elshamy, and T. Fingscheidt, "An Automotive Wideband Stereo Acoustic Echo Canceller Using Frequency-Domain Adaptive Filtering," in *Proc. of EUSIPCO*, Lisbon, Portugal, Sept. 2014, pp. 1452–1456.
- [32] J. Franzen and T. Fingscheidt, "In Car Communication: From Single- to Four-Channel with the Frequency Domain Adaptive Kalman Filter," in *Vehicles, Drivers, and Safety*, John H. L. Hansen et al., Eds., pp. 213–227. Walter de Gruyter GmbH Berlin/Boston, 2020.