Align-ULCNet: Towards Low-Complexity and Robust Acoustic Echo and Noise Reduction

Shrishti Saha Shetu[1], Naveen Kumar Desiraju[2], Wolfgang Mack[1†], Emanuël A. P. Habets[1]
[1] *International Audio Laboratories Erlangen*, Am Wolfsmantel 33, 91058 Erlangen, Germany*
[2] *Fraunhofer IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany*
{*shrishti.saha.shetu, naveen.kumar.desiraju, emanuel.habets*}*@iis.fraunhofer.de*, *wolfgang.mack@fau.de*

***Abstract*—The successful deployment of deep learning-based acoustic echo and noise reduction (AENR) methods in consumer devices has intensified interest in low-complexity solutions that achieve robust performance in real-world scenarios. In this work, we propose a Kalman filter - deep neural network hybrid method for AENR, which employs a novel channel-wise sampling-based feature reorientation method and time alignment in the latent space for complexity reduction and robust performance. Experimental results show that the proposed method achieves better echo reduction and comparable noise reduction performance to the considered baseline methods with improved generalization capability in many acoustically adverse scenarios, such as nonlinear distortions, low-pass filtering, and large acoustical delays.**

***Index Terms*—echo reduction, noise reduction, low-complexity, time alignment, channel-wise feature reorientation**

I. Introduction

AENR technology is crucial for modern communication devices, such as smartphones, video conferencing systems, and smart speakers, which require robust AENR solutions that can consistently suppress echo and noise, ensuring clear communication in challenging acoustic scenarios. Although classical adaptive filter-based algorithms [1]–[6] are effective in controlled scenarios, their performance degrades in the presence of nonlinear distortions and time-varying acoustic conditions [7]–[9]. Recently, deep neural network (DNN)-based approaches have been widely adopted, offering improved AENR performance, either as part of hybrid systems with adaptive filters [7]–[13] or in end-to-end systems [14]–[18]. However, most high-performance solutions are resource-intensive, making them impractical for embedded devices.

In hybrid systems, adaptive filter-based algorithms are typically employed for linear acoustic echo cancellation, while DNNs suppress residual and nonlinear echo [8], [13], [19], [20]. This division of tasks allows the development of low-complexity DNN post-filters. In [20], we integrated the ULCNet model [21], designed initially for noise reduction (NR), into a hybrid AENR system, achieving performance comparable to other evaluated methods. The obtained $\text{ULCNet}_{\text{AENR}}$ model demonstrated low computational and memory requirements, making it highly suitable for real-time processing on embedded devices. Despite these advantages, it exhibits sub-optimal performance in certain scenarios due to its high dependence on the convergence state of the Kalman filter (KF). Additionally, the employed channel-wise subband feature reorientation method (C-SubFR) [22] can render the ULCNet model vulnerable in certain adverse conditions, such as lowpass or bandpass-filtered input signals.

To address these limitations, i) we propose an improved version of the $\text{ULCNet}_{\text{AENR}}$ model [20], called Align-ULCNet, which integrates two parallel streams for independent encoding of input signals and introduces a cross-attention-based time alignment (TA) block in the latent space. The integration of the TA block improves the system's performance, especially in cases where the preceding adaptive filter has not fully converged. ii) We introduce a novel channel-wise sampling-based feature reorientation (C-SamFR) method to improve robustness against different input signal characteristics, particularly for band-limited input signals. iii) We investigate suitable input feature combinations for the proposed model by evaluating their performance in adverse scenarios.

The remainder of this paper is structured as follows. In Section II, we present the signal model and the proposed Align-ULCNet model, including the proposed C-SamFR method and the TA block. Section III-A contains details regarding model training. Section III-B provides performance evaluations for acoustic echo reduction (AER) and NR tasks, comparing our proposed approach with existing methods. Section III-C includes ablation studies analyzing the impact of key design choices, followed by a discussion of the results.

II. Signal Model and Proposed Method

We consider a duplex-communication scenario, where the far-end signal y is reproduced by a loudspeaker and the microphone signal x is composed of the desired near-end (NE) speech s, the acoustic echo e and the background noise v:

$$x(n) = s(n) + e(n) + v(n), \tag{1}$$

where n denotes the discrete-time sample index. We consider a hybrid system in which the microphone signal is first processed with a KF [23], followed by a DNN-based post-filter. The KF computes an estimate for the linear acoustic

*A joint institution of Fraunhofer IIS and Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany.
† Wolfgang Mack is now with Cisco.

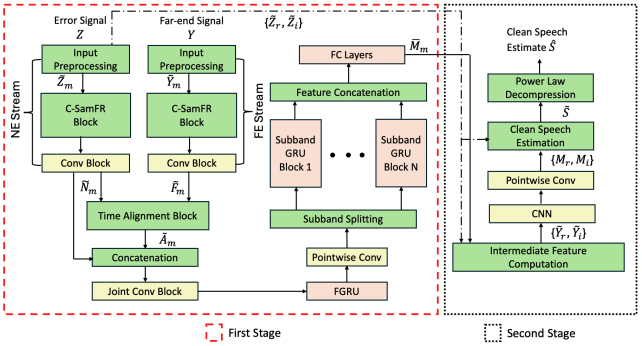


Fig. 1: Proposed low-complexity Align-ULCNet model

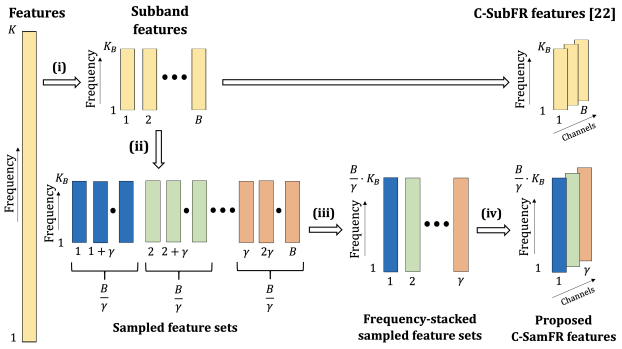


Fig. 2: Proposed C-SamFR method

echo $\hat{e}(n) = \mathrm{KF}\big(x(n), y(n)\big)$, which is then subtracted from the microphone signal to obtain the error signal z, i.e.,

$$z(n) = x(n) - \hat{e}(n). \tag{2}$$

The error signal z contains residual echo r and background noise v, which are subsequently suppressed using the proposed Align-ULCNet model.

Input Preprocessing: The input to the proposed Align-ULCNet model is the short-time Fourier transform (STFT) features of the error signal z and far-end signal y, as shown in Fig. 1. Subsequently, the power-law compressed magnitude features are obtained in the input preprocessing block, as described in [21].

Align-ULCNet: The proposed model has two key modifications to the two-stage ULCNet$_{\mathrm{AENR}}$ model [20], i) the convolutional encoder in the first stage has been redesigned by introducing two parallel convolutional streams — NE and FE — to process the two input signals separately, obtaining the encoded NE and FE features $\widetilde{N}_{\mathrm{m}}$ and $\widetilde{F}_{\mathrm{m}}$, respectively. This parallel structure ensures independent extraction of near-end and far-end features. ii) A cross-attentional TA block is then integrated between the NE and FE streams to align their features in the latent space. The TA block jointly processes $\widetilde{N}_{\mathrm{m}}$ and $\widetilde{F}_{\mathrm{m}}$ to generate time-aligned FE features, represented as $\widetilde{A}_{\mathrm{m}}$. These aligned features are concatenated with NE features $\widetilde{N}_{\mathrm{m}}$ along the channel dimension and further processed in the joint Conv block. The final complex mask M is obtained by processing the convolutional features through FGRU, subband GRUs, FC layers, and CNN blocks. This complex mask M is then used to reconstruct the power-law compressed [21] near-end speech estimate $\tilde{S}$ as follows:

$$\widetilde{S} = \widetilde{Z}_{\mathrm{m}} \cdot M_{\mathrm{m}} \cdot e^{j(\widetilde{Z}_{\mathrm{p}} + M_{\mathrm{p}})}, \tag{3}$$

where $j = \sqrt{-1}$, M_{m} and M_{p} represent the magnitude and phase components of the complex-valued mask M, and $\widetilde{Z}_{\mathrm{m}}$ and $\widetilde{Z}_{\mathrm{p}}$ represent the power-law compressed magnitude and phase components of the error signal Z, respectively.

C-SamFR method: The C-SamFR method rearranges input features in a non-sequential manner along the frequency dimension as shown in Fig. 2. The C-SamFR method is applied separately to the K-element power-law compressed magnitude features $\widetilde{Z}_{\mathrm{m}}$ and $\widetilde{Y}_{\mathrm{m}}$, involving four different steps, (i) for each input signal in each frame, K features are divided into B subbands, each containing K_B contiguous frequency bins, with an overlap factor of $0 \leq \beta < 1$ [22]. (ii) These B subbands are then rearranged using a sampling factor γ, forming sets containing $\frac{B}{\gamma}$ subbands each. (iii) The subbands in each set are then stacked along the frequency dimension to generate a total of γ sampled feature sets. (iv) Finally, these feature sets are stacked along the channel dimension to obtain the final C-SamFR features. The resulting C-SamFR features enhance feature representation compared to the C-SubFR features (upper part of Fig. 2) by reducing the likelihood of subbands containing only zeros in the channel dimension.

Time Alignment: To reduce computational complexity and leverage convolutional features for soft alignment, we implement a convolutional cross-attentional TA block in the latent space. This processing was inspired by the approach in [16]. The NE features $\widetilde{N}_{\mathrm{m}} \in \mathbb{R}^{L \times T \times P}$ and FE features $\widetilde{F}_{\mathrm{m}} \in \mathbb{R}^{L \times T \times P}$ (where L denotes the number of channels, T denotes the number of frames in the time dimension, and P denotes the number of features in the feature dimension) are first transformed by point-wise convolution layers into $N \in \mathbb{R}^{H \times T \times P}$ and $F \in \mathbb{R}^{H \times T \times P}$, respectively, where H is the number of similarity channels. Then, F is unfolded along the time axis to produce delayed features $F_u \in \mathbb{R}^{H \times T \times D_{\max} \times P}$, where $D_{\max}$ is the maximum echo delay. A dot product between N and F_u along the feature axis yields $C \in \mathbb{R}^{H \times T \times D_{\max}}$:

$$C(h,t,d) = \sum_{p=1}^{P} N(h,t,p) \cdot F_u(h,t,d,p), \tag{4}$$

which is further processed by a convolutional layer to obtain the delay probability distribution $D \in \mathbb{R}^{T \times D_{\max}}$ using a softmax along the delay dimension d. Finally, the aligned FE features $\widetilde{A}_{\mathrm{m}} \in \mathbb{R}^{H \times T \times P}$ are computed as a weighted sum of F_u with the delay probabilities from D along the time axis:

$$\widetilde{A}_{\mathrm{m}}(h,t,p) = \sum_{d=1}^{D_{\max}} D(t,d) \cdot F_u(h,t,d,p). \tag{5}$$

III. Experiments and Results

A. Experimental Design

Model Parameters: The Conv blocks in the NE and FE streams comprised two separable convolution layers, each with $L = 32$ filters and stride of (1×1), and with kernel sizes of (1×5) and (1×3), respectively. Downsampling by a factor of 2 was performed through max-pooling along the

TABLE I: AECMOS [24] results on AEC Challenge blind test sets.

	Computational Complexity		Interspeech 2021 [25]				ICASSP 2023 [26]			
			DT		FST	NST	DT		FST	NST
Processing[1]	Params [M]	GMACS	EMOS	DMOS	EMOS	DMOS	EMOS	DMOS	EMOS	DMOS
Peng et al.* [9]	10.20	2.52	4.36	4.23	4.34	4.26	-	-	-	-
Braun et al. [14]	-	-	4.55	**4.25**	4.35	4.18	-	-	-	-
Zhang et al.* [11]	9.56	-	-	-	-	-	**4.72**	4.16	4.70	-
Deep-VQE* [16]	7.50	4.02	-	-	-	-	4.70	**4.29**	4.69	-
Align-CRUSE [15]	0.74	-	4.45	4.07	4.67	-	4.60	3.95	4.56	-
ULCNet$_{\text{AENR}}$(C-SubFR) [20]	**0.69**	**0.10**	4.61	3.79	4.64	4.28	4.54	3.58	4.73	4.15
ULCNet$_{\text{AENR}}$(C-SamFR)	**0.69**	**0.10**	4.58	3.93	4.53	**4.32**	4.58	3.74	4.71	4.15
Proposed Align-ULCNet	**0.69**	**0.10**	**4.66**	3.95	**4.75**	4.29	4.60	3.80	**4.77**	**4.28**

frequency dimension. In the TA block, we employed point-wise convolutions with $H = 32$ filters and used a kernel size of (5×3) for the convolutional operation on the dot product features C. The Joint Conv block included two convolution layers with 64 and 96 filters, a kernel size of (1×3), and a stride of (1×2). The remaining blocks in the Align-ULCNet model follow the same parameterization as described in [21].

Experimental Parameters: The KF used a recursive Kalman gain of 0.8 and 10 partition blocks, as mentioned in [10], [23]. We chose $N_{\text{FFT}} = 512$, a window length of 32 ms, and a hop size of 16 ms such that $K = 257$. We used a power-law compression factor $\alpha = 0.3$ for the input preprocessing step. For the C-SamFR method, we used $K_B = 2$ with an overlap factor of $\beta = 0$, such that $B = 130$, and a sampling factor of $\gamma = 5$, while for the TA block, we used $D_{\text{max}} = 64$, which roughly corresponded to a maximum echo delay of 1 s. For training, we used the Adam optimizer with an initial learning rate of 0.004, which decayed by a factor of 10 when the validation loss did not improve with a patience of one epoch. Each training sample was of 3 s duration, a batch size of 64 was chosen, and the model was trained for 20k steps per epoch.

Training Dataset: We used the measured echo and far-end signals provided in [26], as well as clean and noisy speech signals from [27] to create a training dataset of 1100 hours at 16 kHz sampling rate using the methodology mentioned in [20]. Different scenarios for the microphone signal, e.g., near-end single-talk (NST), far-end single-talk (FST), and double-talk (DT), were simulated following the methods proposed in [12], with SNR $\in [-5, 30]$ and SER $\in [-20, 20]$ dB.

Evaluation Dataset and Metrics: The Interspeech 2021 [25] and ICASSP 2023 [26] AEC Challenge datasets are used for evaluating the AER performance, and the DNS challenge 2020 dataset [27] is used for evaluating the NR performance. We use AECMOS [24], SI-SDR [28], BAKMOS, and SIGMOS [29] as the evaluation metrics.

B. Results

AER Performance: Table I presents a comparison of the echo reduction performance of the proposed Align-ULCNet model against six recent baseline methods from literature, as well as the ULCNet$_{\text{AENR}}$ model combined with the C-SamFR methods. The results demonstrate that the proposed model achieves substantial improvements over both versions of ULCNet$_{\text{AENR}}$ across all objective metrics. The enhancement in EMOS metrics can be attributed to the integration of the TA block, while the improvement in DMOS metrics primarily results from using the C-SamFR method. Additionally, our approach surpasses the baseline methods in EMOS metrics while delivering slightly lower DT DMOS scores, but at a significantly reduced computational cost and memory footprint.

[1]Metrics reported for SOTA methods are taken directly from their respective publications, with an asterisk (*) indicating results reported as part of the AEC challenge.

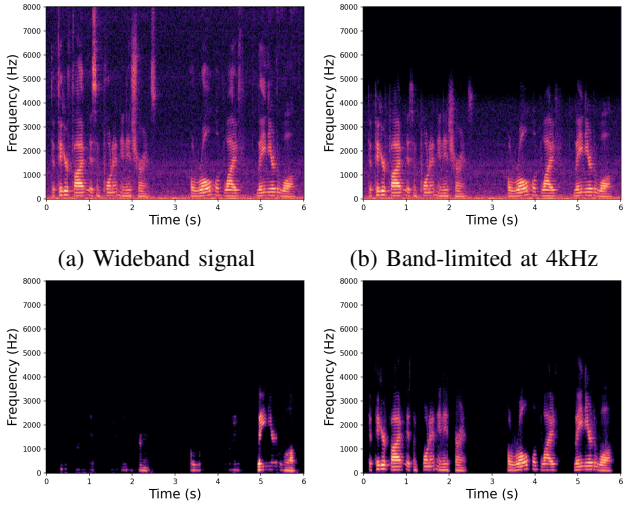
TABLE II: Objective results for NR on DNS Challenge [27] non-reverb test set

Processing	SI-SDR	SIGMOS	BAKMOS
Noisy	9.06	3.39	2.62
DeepFilterNet [30]	16.17	3.49	4.03
DeepFilterNet2 [31]	16.60	**3.51**	**4.12**
ULCNet$_{\text{MS}}$ [21]	16.34	3.46	4.06
ULCNet$_{\text{Freq}}$ [21]	**16.67**	3.38	4.09
ULCNet$_{\text{AENR}}$ [20]	15.58	3.30	4.05
Proposed Align-ULCNet	16.12	3.33	4.08

TABLE III: EMOS results for the proposed Align-ULCNet model for different input combinations

	Input Signals		EMOS	
TA block	NE stream	FE stream	DT	FST
KF	-	-	1.73	2.19
No	Z	Y	3.73	3.91
Yes	$\boldsymbol{Z}$	$\boldsymbol{Y}$	**4.20**	**4.69**
Yes	Z and $\widehat{E}$	Y	4.04	4.18
Yes	Z	Y and $\widehat{E}$	3.99	4.39

NR Performance: In Table II, we compare the NR performance of the proposed Align-ULCNet model against five different low-complexity dedicated NR methods. While DeepFilterNet2 [31] achieves slightly higher SIGMOS, our method maintains comparable NR performance (BAKMOS) with significantly lower computational complexity [21]. This suggests that Align-ULCNet effectively balances echo suppression and noise reduction without excessive parameter overhead. In informal listening tests, we observed that the perceptual quality



(a) Wideband signal (b) Band-limited at 4kHz

(c) ULCNet$_{\text{AENR}}$(C-SubFR) (d) ULCNet$_{\text{AENR}}$(C-SamFR)

Fig. 3: Example of (a) a noisy NST wideband signal, (b) band-limited at 4 kHz frequency, output of ULCNet$_{\text{AENR}}$ model [20] using (c) C-SubFR [22] and (d) proposed C-SamFR method.

of the Align-ULCNet is comparable to the other methods under test. The processed samples can be found here: https://fhgainr.github.io/alignulcnet-aenr/.

C. Ablation Study and Discussion

Effect of C-SamFR Method: The C-SubFR method [22], as integrated in the ULCNet$_{\text{AENR}}$ [20] model, exhibits performance issues when the input signal is band-limited. This is due to the sequential subband splitting process, which can result in one or more subbands being completely filled with zeros in the channel dimension, thereby negatively impacting the model's overall performance. Our proposed C-SamFR method addresses this issue by ensuring that no subbands contain only zeros in the channel dimension. As shown in Fig. 3, in the NST scenario with the microphone signal band-limited at 4kHz, the (d) ULCNet$_{\text{AENR}}$(C-SamFR) method maintains satisfactory performance, whereas the (c) ULCNet$_{\text{AENR}}$(C-SubFR) model tends to suppress the NE signal excessively. The proposed C-SamFR method also significantly enhances speech quality compared to the C-SubFR method, as reflected in the DMOS metrics in Table I, while maintaining similar echo reduction performance, as indicated by the EMOS metrics.

Different Input Combinations: In this work, we also examine the impact of various input signal combinations (the STFT-domain error signal Z, echo estimate $\widehat{E}$ and far-end signal Y) on the Align-ULCNet model performance for the scenarios where the KF fails to converge. For this experiment, we used the same AEC challenge test datasets as mentioned in Table I, and selected the samples where the KF failed to perform satisfactorily ($\approx$ 60 samples). We can observe the results in Table III, where the first row contains the scores obtained after KF processing. The results clearly demonstrate that our proposed Align-ULCNet model, which takes the error signal Z in the NE stream and the far-end signal Y in the FE stream, outperforms all other input signal combinations.

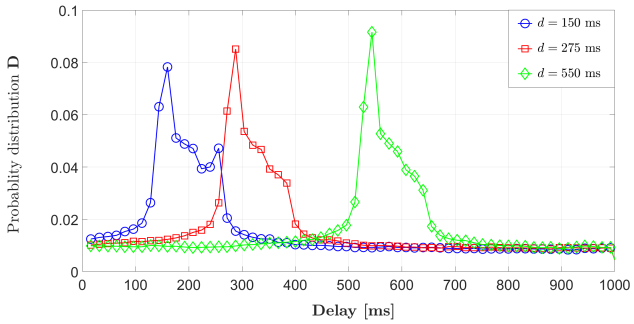


Fig. 4: Estimated delay probability distribution $\mathbf{D}$ for different amounts of acoustic delay d

Analysis of Time Alignment: The experiment with different input combinations further highlights the effectiveness of the TA block. As shown in Table III, omitting the TA block with the same input combination (Z and Y) significantly reduces performance, with EMOS dropping by -0.47 and -0.78 for DT and FST scenarios, respectively. This supports the hypothesis that the TA block reduces the system's reliance on the KF. Figure 4 shows the accurate delay probability distribution $\mathbf{D}$ computed by the TA block for varying acoustic delays. These probabilities are used in the weighted sum computation of time-aligned FE features $\widetilde{A}_{\text{m}}$, highlighting the need for a sufficiently long delay buffer in real-time streaming applications.

Discussion: Our proposed Align-ULCNet model, while being computationally lightweight, delivers competitive AER and on-par NR results as compared to the baseline AER and dedicated low-complexity NR methods. We demonstrate that the proposed C-SamFR method ensures robust performance, particularly when the input signal is band-limited - a scenario likely in mass consumer device deployments. Including the TA block in the Align-ULCNet model enhances its independence from the KF, further improving overall performance under adverse conditions. Although objective metrics such as DMOS and SIGMOS indicate slightly inferior results with respect to speech quality, previous studies [20], [21] have shown that the use of power-law compression can negatively impact these metrics. However, in informal listening tests, the results of the proposed model remain preferable. When implemented efficiently, our model requires only a small buffer for storing the far-end features from previous frames, achieving a real-time factor of $\approx$16% on a Cortex-A53 1.43 GHz processor.

IV. Conclusion

We proposed a low-complexity hybrid approach for robust AENR. Our proposed Align-ULCNet model achieves objective results that are superior or comparable to other evaluated methods for both echo and noise reduction tasks. With a small model size of only 0.69M parameters and a computational complexity of 0.10 GMACS, the proposed method is well-suited for deployment in resource-constrained consumer devices. Additionally, we provide insights into the various processing blocks within the model and their contributions to the overall performance of the proposed model.

V. Acknowledgment

This work has been supported by the Free State of Bavaria in the DSAI project.

References

[1] Jacob Benesty, Tomas Gänsler, Dennis R Morgan, M Mohan Sondhi, Steven L Gay, et al., "Advances in network and acoustic echo cancellation," *Springer*, 2001.

[2] Stefan Gustafsson, Rainer Martin, Peter Jax, and Peter Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Transactions on Speech and Audio Processing*, pp. 245–256, 2002.

[3] María Luis Valero, Edwin Mabande, and Emanuël A.P. Habets, "A state-space partitioned-block adaptive filter for echo cancellation using inter-band correlations in the Kalman gain computation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2015, pp. 599–603.

[4] María Luis Valero, Edwin Mabande, and Emanuël A.P. Habets, "A low-complexity state-space architecture for multi-microphone acoustic echo control," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 1–5.

[5] Fabian Kuech, Andreas Mitnacht, and Walter Kellermann, "Nonlinear acoustic echo cancellation using adaptive orthogonalized power filters," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2005.

[6] Mhd Modar Halimeh and Walter Kellermann, "Efficient multichannel nonlinear acoustic echo cancellation based on a cooperative strategy," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2020, pp. 461–465.

[7] Jan Franzen and Tim Fingscheidt, "Deep residual echo suppression and noise reduction: A multi-input FCRN approach in a hybrid speech enhancement system," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2022, pp. 666–670.

[8] Mhd Modar Halimeh, Thomas Haubner, Annika Briegleb, Alexander Schmidt, and Walter Kellermann, "Combining adaptive filtering and complex-valued deep postfiltering for acoustic echo cancellation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2021, pp. 121–125.

[9] Renhua Peng, Linjuan Cheng, Chengshi Zheng, and Xiaodong Li, "Acoustic echo cancellation using deep complex neural network with nonlinear magnitude compression and phase information," in *INTERSPEECH*, 2021, pp. 4768–4772.

[10] Thomas Haubner, Mhd Modar Halimeh, Andreas Brendel, and Walter Kellermann, "A synergistic Kalman-and deep postfiltering approach to acoustic echo cancellation," in *European Signal Processing Conference (EUSIPCO)*, 2021, pp. 990–994.

[11] Zihan Zhang, Shimin Zhang, Mingshuai Liu, Yanhong Leng, Zhe Han, Li Chen, and Lei Xie, "Two-step band-split neural network approach for full-band residual echo suppression," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2023, pp. 1–2.

[12] Wolfgang Mack and Emanuël A.P. Habets, "A hybrid acoustic echo reduction approach using Kalman filtering and informed source extraction with improved training," in *IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 502–508.

[13] Jean-Marc Valin, Srikanth Tenneti, Karim Helwani, Umut Isik, and Arvindh Krishnaswamy, "Low-complexity, real-time joint neural echo control and speech enhancement based on percepnet," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2021, pp. 7133–7137.

[14] Sebastian Braun and Maria Luis Valero, "Task splitting for DNN-based acoustic echo and noise removal," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[15] Evgenii Indenbom, Nicolae-Cătălin Ristea, Ando Saabas, Tanel Pärnamaa, and Jegor Gužvin, "Deep model with built-in cross-attention alignment for acoustic echo cancellation," in *arXiv preprint arXiv:2208.11308*, 2022.

[16] Evgenii Indenbom, Nicolae-Catalin Ristea, Ando Saabas, Tanel Parnamaa, Jegor Guzvin, and Ross Cutler, "DeepVQE: Real time deep voice quality enhancement for joint acoustic echo cancellation, noise suppression and dereverberation," in *arXiv preprint arXiv:2306.03177*, 2023.

[17] Shimin Zhang, Ziteng Wang, Jiayao Sun, Yihui Fu, Biao Tian, Qiang Fu, and Lei Xie, "Multi-task deep residual echo suppression with echo-aware loss," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2022, pp. 9127–9131.

[18] Nils L Westhausen and Bernd T Meyer, "Acoustic echo cancellation with the dual-signal transformation LSTM network," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2021, pp. 7138–7142.

[19] Hao Zhang, Srivatsan Kandadai, Harsha Rao, Minje Kim, Tarun Pruthi, and Trausti Kristjansson, "Deep adaptive AEC: Hybrid of deep learning and adaptive acoustic echo cancellation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2022, pp. 756–760.

[20] Shrishti Saha Shetu, Naveen Kumar Desiraju, Jose Miguel Martinez Aponte, Emanuël A. P. Habets, and Edwin Mabande, "A hybrid approach for low-complexity joint acoustic echo and noise reduction," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024.

[21] Shrishti Saha Shetu, Soumitro Chakrabarty, Oliver Thiergart, and Edwin Mabande, "Ultra low complexity deep learning based noise suppression," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2024, pp. 466–470.

[22] Haohe Liu, Lei Xie, Jian Wu, and Geng Yang, "Channel-wise subband input for better voice and accompaniment separation on high resolution music," in *INTERSPEECH*, 2020.

[23] Fabian Kuech, Edwin Mabande, and Gerald Enzner, "State-space architecture of the partitioned-block-based acoustic echo controller," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2014, pp. 1295–1299.

[24] Marju Purin, Sten Sootla, Mateja Sponza, Ando Saabas, and Ross Cutler, "AECMOS: A speech quality assessment metric for echo impairment," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2022, pp. 901–905.

[25] Ross Cutler, Ando Saabas, Tanel Pärnamaa, Markus Loide, Sten Sootla, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sørensen, Robert Aichner, et al., "InterSpeech 2021 Acoustic Echo Cancellation Challenge," in *INTERSPEECH*, 2021, pp. 4748–4752.

[26] Ross Cutler, Ando Saabas, Tanel Pärnamaa, Marju Purin, Evgenii Indenbom, Nicolae-Cătălin Ristea, Jegor Gužvin, Hannes Gamper, Sebastian Braun, and Robert Aichner, "ICASSP 2023 Acoustic Echo Cancellation Challenge," *IEEE Open Journal of Signal Processing*, 2024.

[27] Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al., "The InterSpeech 2020 Deep Noise Suppression Challenge: Datasets, subjective testing framework, and challenge results," in *INTERSPEECH*, 2020.

[28] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2019, pp. 626–630.

[29] Chandan KA Reddy, Vishak Gopal, and Ross Cutler, "DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2022, pp. 886–890.

[30] Hendrik Schröter, Alberto N Escalante-B, Tobias Rosenkranz, and Andreas Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2022, pp. 7407–7411.

[31] Hendrik Schröter, A Maier, Alberto N Escalante-B, and Tobias Rosenkranz, "DeepFilterNet2: Towards real-time speech enhancement on embedded devices for full-band audio," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.