ABSTRACT
The Kalman filter has been adopted in acoustic echo cancellation due to its robustness to double-talk, fast convergence, and good steady-state performance. The performance of Kalman filter is closely related to the estimation accuracy of the state noise covariance and the observation noise covariance. The estimation error may lead to unacceptable results, especially when the echo path suffers abrupt changes, the tracking performance of the Kalman filter could be degraded significantly. In this paper, we propose the neural Kalman filtering (NKF), which uses neural networks to implicitly model the covariance of the state noise and observation noise and to output the Kalman gain in real-time. Experimental results on both synthetic test sets and real-recorded test sets show that the proposed NKF has superior convergence and re-convergence performance while ensuring low near-end speech degradation compared with the state-of-the-art model-based methods. Moreover, the model size of the proposed NKF is merely 5.3 K and the RTF is as low as 0.09, which indicates that it can be deployed in low-resource platforms.

Index Terms—Acoustic echo cancellation, Kalman filter, neural networks

1. INTRODUCTION
Acoustic echo cancellation (AEC) plays an important role in the speech processing field. It aims to remove the echoes introduced by the far-end signals in the speech communication scenario, according to the far-end reference signals and microphone signals. Model-based (MB) methods (adaptive filters) such as least-mean square (LMS), normalized least-mean square (NLMS), and recursive least-squares (RLS) are widely used in this field. The state-space models such as Kalman filtering have also been successfully applied in echo cancellation [1, 2, 3, 4]. Compared with other adaptive filters, Kalman filter is more robust to double-talk and has better convergence performance. Because it assumes that the echo path is time-varying, which is modeled by the state equation. The convergence and tracking ability of Kalman filtering are controlled by the covariance of the state noise and observation noise, which can be dynamically estimated according to some assumptions. However, the noise covariance estimation error may lead to unacceptable results. For instance, Yang et al. [4] pointed out that the overestimation of the observation noise covariance can degrade the reconvergence speed when the echo path changes abruptly.

In order to solve this problem, we propose to combine the MB method and data-driven (DD) method. More specifically, we propose to use a trainable recurrent neural network (RNN) to implicitly model the covariance of the state noise and observation noise and output the adaptive Kalman gains in real-time. This can result in better convergence and tracking ability than the Kalman filter [5, 6]. Unlike other neural network aided adaptive filtering methods, e.g., deep neural network-controlled frequency-domain adaptive filter (DNN-FDAF) [7] and deep variable step-size NLMS (DVSS-NLMS) [8] that respectively filtering in the frequency domain and in the time domain, we formulate the neural Kalman filtering operations in the time-frequency domain, resulting in a much smaller model size (with merely 5.3 K compared with 2.4 M and 1 M parameters). Experimental results on both synthetic test sets and real-recorded test sets show that, the proposed NKF achieves much better AEC performance than conventional adaptive filters and the state-of-the-art hybrid MB/DD method Meta-AF [9]. Moreover, compared with the deep-learning based AEC methods that are fully data-driven [10, 11, 12, 13], the proposed method has much smaller model size, and introduces less near-end signal distortion according to the experimental results. The inference code and the pre-trained model are released online 1.

2. PROPOSED METHOD
2.1. AEC Signal Models
In acoustic echo cancellation, the microphone signal $y[n]$ is assumed to be the summation of the near-end signal $s[n]$ and the echo of the far-end signal $x[n]$, which depends on the room impulse response $h[n]$:

$$y[n] = s[n] + \sum_{\tau=0}^{T-1} h[\tau] x[n-\tau], \quad (1)$$

1Code available at https://github.com/fjiang9/NKF-AEC.
where $T$ is the length of the room impulse response (RIR) $h[n]$. The second term in Eq. (1) is the linear acoustic echo. Note that there are also some non-linear components in realistic scenarios, but we only consider linear echo paths in this paper. The objective of AEC is to estimate $h[n]$ so that the echo-cancelled signal can be obtained by

$$
\hat{s}[n] = y[n] - \sum_{\tau=0}^{T-1} \hat{h}[\tau]x[n-\tau].
$$

Similar to Ref. [14], we approximately reformulate Eq. (1) into the time-frequency (TF) domain by applying short-time Fourier transform (STFT) as

$$
Y[m,k] = S[m,k] + \sum_{l=0}^{L-1} H^*[l,k]X[m-l,k],
$$

where $S[m,k]$, $Y[m,k]$, and $X[m,k]$ are respectively the near-end signal, microphone signal, and far-end signal at the frame $m$ and frequency $k$, $L$ is the number of taps, and $*$ refers to the complex conjugate operation. $H[l,k]$ is the convolutive transfer function (CTF) of the echo path. To simplify the expression, we define

$$
x_{m,k} = [X[m,k], \cdots, X[m-L+1,k]]^T,
$$

$$
h_k = [H[0,k], \cdots, H[L-1,k]]^T.
$$

Then the echo-cancelled signal can be expressed as

$$
\hat{S}[m,k] = Y[m,k] - \hat{h}_k^Hx_{m,k},
$$

where $^H$ stands for the Hermitian transpose.

2.2. Kalman Filtering in the TF domain

Most previous Kalman filtering adopted in AEC are processed either in the frequency domain [1, 15, 4] or in the time domain [16]. In this paper, we introduce Kalman filtering to the TF domain for echo cancellation and denote it as TF-domain Kalman filter (TFDKF).

The echo path $h_k \in \mathbb{C}^L$ can be assumed as a zero-mean random vector, which follows a simplified first-order Markov model, i.e.,

$$
h_{m,k} = Ah_{m-1,k} + w_{m,k},
$$

where $A$ is a transition parameter [1] that is close to $1$, $w_{m,k} \in \mathbb{C}^L$ is a noise vector, and $w_{m,k} \sim \mathcal{N}(0, Q_k)$. Eq. (7) is called the state equation. The covariance matrix $Q_k$ can determine the tracking ability of the Kalman filter. We denote the correlation matrix of the prior misalignment $h_{m,k} - \hat{h}_{m,k}$ as $P_{m,k}$ i.e., $P_{m,k} = E[(h_{m,k} - \hat{h}_{m,k})(h_{m,k} - \hat{h}_{m,k})^H]$, then

$$
P_{m,k} = A^2P_{m-1,k} + Q_k,
$$

where $P_{m,k}^+$ represents the correlation matrix of posterior misalignment $h_{m,k} - \hat{h}_{m,k}$. The observation function can be expressed as

$$
Y[m,k] = h_{m,k}^Hx_{m,k} + S[m,k].
$$

The Kalman filter solution can be expressed as

$$
\hat{h}_{m,k} = A\hat{h}_{m-1,k} + H^*x_{m,k}
$$

$$
P_{m,k} = AP_{m-1,k}A^H + Q_k
$$

$$
e_{m,k} = Y[m,k] - \hat{h}_{m,k}^Hx_{m,k},
$$

$$
k_{m,k} = P_{m,k}x_{m,k} (x_{m,k}^HP_{m,k}x_{m,k} + |S[m,k]|^2)^{-1},
$$

$$
\hat{h}_{m,k}^+ = \hat{h}_{m,k} + k_{m,k}e_{m,k},
$$

$$
P_{m,k}^+ = [I_L - k_{m,k}x_{m,k}^H]P_{m,k},
$$

where $k_{m,k} \in \mathbb{C}^L$ is the Kalman gain. In practice, $Q_k$ can be estimated using different schemes [2, 16, 17], we estimate it by

$$
\hat{Q}_k = (1 - A^2)E[\hat{h}_{m-1,k}\hat{h}_{m-1,k}^H].
$$

The observation noise $S[m,k]$ can be estimated by the prior error signal $e_{m,k}$ according to Eq. (12).

2.3. Neural Kalman Filtering for AEC

Because the estimation of $|S[m,k]|^2$ and $Q_k$ are generally not accurate, the Kalman gain is therefore not optimal. For instance, when an abrupt echo path change occurs, the power of the prior error signal $e_{m,k}$ could be quite large, thus leading to overestimation of the observation noise covariance. As a result, the reconvergence speed will be very slow according to the analysis in [4].

To address the problem of noise statistics estimation in Kalman filter, we propose to use a RNN model to implicitly model the noise statistics and directly estimate the Kalman gain in real-time. From Eqs. (10)-(15) we can see that, the Kalman gain $k_{m,k}$ is mainly related to the far-end signal $x_{m,k}$, the noise covariance $Q_k$, and the near-end signal power $|S[m,k]|^2$. $Q_k$ and $|S[m,k]|^2$ can be inferred from the innovation of $h$ at the last frame and the prior echo-cancelled signal at the current frame, respectively. Therefore, we concatenate $x_{m,k}$, $\Delta h_{m-1,k} = k_{m-1,k}e_{m-1,k}$, and $e_{m,k}$ as the input features $z_{m,k} \in \mathbb{R}^D$ to output the Kalman gain $k_{m,k}$. Obviously, the input feature dimension $D$ is related to the taps $L$, i.e., $D = 2L + 1$.

The model consists of one complex fully-connected (FC) layer, one complex GRU layer, and another two complex FC layers. The PReLU activation function [18] is used after each FC layers except the last layer. The numbers of units in the three FC layers are $2D$, $2D$, and $L$ respectively. The number of units in the GRU layer is $L^2 + 2$. Interestingly, we
Empirically find that increasing the number of hidden units in the GRU layers could only marginally improve the AEC results. This is reasonable because the dimension $L^2 + 2$ is sufficient for modeling both $Q_k$ and $|S[m, k]|^2$. It is noted that the RNN model independently runs at each frequency bin, which means that the feature dimension of the input features can be viewed as the batch dimension. This design can greatly decrease the number of parameters of the model as the feature dimension is only at the $L$ level and the networks can be shared among different frequencies.

The forward pass of neural Kalman filtering at the $k$-th frequency bin is shown in Algorithm 1. The initialization of echo path $h_{0,k}$, RNN hidden state $g_{0,k}$, and $\Delta h_{0,k}$ can be either all-zero vectors or white Gaussian noises. Using an all-zero vector to initialize the echo path $h_{0,k}$ is quite common in most previous research. The reason we use white Gaussian noise as the initialization is to simulate the system state when an abrupt echo path change occurs. We empirically set the probability of zero initialization and noise initialization as both 50% during training, while using only zero initialization in the inference stage. After one forward pass, the estimated echo $\hat{D}[m, k]$ and echo-cancelled signal $\hat{S}[m, k]$ can be obtained. The loss function can then be calculated by measuring the distance between the estimated echo $\hat{D}[m, k]$ and ground-truth echo $D[m, k]$, e.g.,

$$\mathcal{L}(D, \hat{D}(\Theta)) = \sum_{k=1}^{F} \sum_{m} |D[m, k] - \hat{D}[m, k]|^2. \quad (17)$$

Finally, The RNN model $\mathcal{F}_\Theta(\cdot, \cdot)$ is optimized using the gradient descent algorithm. In the inference time, the time-domain echo-cancelled signal $\hat{s}[n]$ can be reconstructed according to $\hat{S}[m, k]$ using inverse STFT.

### 3. EXPERIMENTS

#### 3.1. Experimental Setup

We use the far-end speech in the synthetic fold of the AEC challenge [19] as the training corpus, which contains 10000 speech clips and each clip is 10 seconds long. During training, both the far-end signal and near-end signal are randomly sampled from this speech corpus. We use randomly generated white Gaussian noise as the RIR to generate the far-end echo signal. The signal-to-echo ratio (SER) is uniformly sampled from -5 dB to 5 dB. The duration of the far-end signal is randomly clipped to 1 s, while the near-end signal is randomly segmented to 0.5~1 s. The sampling rate of the audio is 16 kHz in this paper. As for the STFT parameters, the FFT size window length, and hop size are 1024, 1024, and 256, respectively. In this paper, the number of taps of the NKF is set to be 4 (i.e., $L = 4$), which results in only 5.3 K real-valued trainable parameters.

Two test sets are used to evaluate the AEC performance of the proposed method. One is a synthetic test set, which uses the speech data from the Librispeech [20] dev-clean and test-clean folds as the far-end and near-end signals respectively, and uses simulated RIRs in [21] to generate echo signals. Note that these data are not seen during training. The echo paths are cut off after 64 ms. The SER is uniformly sampled from -10 dB to 10 dB. The duration of each samples is 8 seconds. Four subsets, namely far-end single-talk (FST), far-end single-talk with echo-path change (FST-EPC), double-talk (DT), and double-talk with echo-path change (DT-EPC) are generated. Each subset contains 500 samples. The echo path change is implemented by abruptly switching the RIR from one simulated room to another at a certain moment (randomly from 3.5 s to 4.5 s). The other is a real recorded test set, i.e., the clean blind test set from the ICASSP 2021 AEC Challenge [19]. We use the double-talk partition for testing, which contains 200 samples and half of them suffers echo path change. Note that some samples in this test set have large time delay, we therefore performs time-delay compensation using the GCC-PHAT algorithm [22] like in Ref. [23] before linear filtering.

#### 3.2. Results

Three fully model-based methods, i.e., proportionate normalized least-mean-squares (PNLMS) [24], partitioned-block-based frequency-domain Kalman filter (PFDKF) [15], and TFDKF, as well as one hybrid MB/DD method Meta-AF [9] are used as the baseline methods to compare with the proposed method. We directly use the pre-trained Meta-AF model released online by the authors. The hyper-parameters of the model-based baseline methods are tuned to be optimal upon the synthetic test set. For a fair comparison, the filter lengths of these linear AEC systems are set to be identical (i.e., 64 ms) in our experiments.

---

**Algorithm 1: NKF forward pass at frequency $k$**

**Input:** mic signal $Y$, far-end signal $x$, model $\mathcal{F}_\Theta$

**Output:** echo-cancelled signal $\hat{S}$

1. Initialize $h_{0,k}$, $g_{0,k}$, and $\Delta h_{0,k}$;
2. for $m \leftarrow 1$ to $T$ do
   3. $e_{m,k} = Y[m, k] - \hat{h}^H_{m-1,k} x_{m,k}$;
   4. $z_{m,k} = [x_{m,k}^H, \Delta \hat{h}^H_{m-1,k}, e_{m,k}]^T$;
   5. $k_{m,k}, g_{m,k} = \mathcal{F}_\Theta(z_{m,k}, g_{m-1,k})$;
   6. $\Delta \hat{h}_{m,k} = k_{m,k} e_{m,k}$;
   7. $\hat{h}_{m,k} = \hat{h}_{m-1,k} + \Delta \hat{h}_{m,k}$;
   8. $\hat{S}[m, k] = Y[m, k] - \hat{h}^H_{m,k} x_{m,k}$;
9. end
Table 1. Comparison of AEC methods with ERLE (dB) and PESQ across four subsets of the synthetic test set, and with AECMOS score for the real-recorded test set. The ERLE reported in this table is the mean of the segmental ERLEs. Model size (MS) refers to the number of real-valued parameters. Grey cells indicate that the method is not data-driven.

| Method        | Lin. | MS     | Synthetic | Real-recorded |
|---------------|------|--------|-----------|---------------|
|               |      |        | FST       | FST-EPC       | DT            | DT-EPC        | DT (AECMOS)  |
|               |      |        | ERLE      | ERLE          | ERLE          | PESQ          | ERLE         | PESQ         | Echo | Other |
| PNLMS [24]    | ✓    | –      | 10.86     | 8.75          | 7.06          | 1.50          | 5.52         | 1.43          | 2.64 | 3.94  |
| PFDKF [15]    | ✓    | –      | 25.87     | 17.56         | 7.46          | 1.46          | 5.88         | 1.41          | 2.75 | 3.78  |
| TFDKF         | ✓    | –      | 24.50     | 18.62         | 15.11         | 2.29          | 10.99        | 1.77          | 2.89 | 3.96  |
| DTLN [12]     | –    | 10.4 M | –         | –             | –             | –             | –            | –             | –    | 4.31  |
| Meta-AF [9]   | ✓    | 28 K   | 19.45     | 14.60         | 12.59         | 1.78          | 9.09         | 1.56          | 2.64 | 3.85  |
| NKF           | ✓    | 5.3 K  | 28.41     | 24.75         | 15.99         | 2.77          | 13.75        | 2.37          | 4.02 | 4.02  |

Fig. 1. ERLE curves on the synthetic DT-EPC subset. Abrupt echo path change occurs at the shaded region.

We use the segmental echo return loss enhancement (ERLE) [25] and wide-band perceptual evaluation of speech quality (PESQ) [26] to evaluate the AEC performance on the synthetic test set. The results are shown in Table 1. We can see that the ERLE improvements from the best baseline results on the FST, FST-EPC, DT, and DT-EPC subsets are 2.54 dB, 6.13 dB, 0.88 dB, and 2.76 dB, respectively. The PESQ improvements from the best baseline results on the DT and DT-EPC subsets reach 0.48 and 0.60 respectively, which means that the NKF can better suppress the echo without introducing extra near-end speech distortion. The improvements on the FST-EPC and DT-EPC subsets are more significant, indicating that the proposed NKF can estimate the Kalman gain more accurately than its model-based counterpart TFDKF especially when the system suffers abrupt echo path changes. Similar observations can also be obtained from Fig. 1, which shows the ERLE curves (averaged on 500 test samples) of different linear AEC methods on the DT-EPC subset. We can clearly see that, the proposed NKF has much faster convergence and re-convergence speed compared with other methods. In particular, it is interesting to notice that the re-convergence speed of the TFDKF is much slower than its convergence speed, whereas the re-convergence speed of the NKF is still fast enough. Such excellent tracking ability of the proposed NKF may be due to the dynamic estimations of the noise statistics by the hidden states of the RNN layer. In addition, the model size of the proposed NKF is only 5.3 K, which is the smallest published neural AEC model to the best of our knowledge. Its real-time factor (RTF), which is tested on a single CPU core, is as low as 0.09.

The AECMOS results on the ICASSP 2021 AEC challenge blind test set double-talk partitions are shown in the real-recorded section of Table 1. In addition to the linear AEC baselines, we also report the results of a fully data-driven model DTLN [12] which ranked in the 3-rd place in the AEC challenge for comparison. We directly use its official pre-trained model for testing. Since there is no ground-truth signal provided in the AEC challenge blind test set, a non-intrusive perceptual objective metric named AECMOS [27] is used to evaluate the AEC performance on the real recorded test set. We can see that, the proposed NKF works surprisingly well on this real-recorded test set. It surpasses all the other linear AEC baselines by a large margin, especially in terms of echo suppression (with at least 1.13 AECMOS-echo improvement). This indicates that conventional linear filtering methods may need to carefully tune the algorithm parameters in different scenarios. By contrast, The proposed NKF does not require tuning after training and has excellent generalization ability. Although the echo suppression performance of NKF is not as good as that of DTLN, the near-end speech quality of NKF is significantly higher. Considering that the NKF only suppresses part of the linear echo, the results may be further improved with longer filter length and a residual echo suppression network.

4. CONCLUSIONS

In this paper, we propose a new linear AEC framework based on Kalman filtering and neural networks. It significantly improves the performance of the Kalman filters while keeps a very small model size and fast inference speed, which is feasible to be adopted in the low-resource applications. For future work, we will further investigate better architecture for modeling the Kalman gain and extend the proposed method to multi-channel and non-linear echo cancellation.
5. REFERENCES

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

[2] G. Enzner, “Bayesian inference model for applications of time-varying acoustic system identification,” in *2010 18th European Signal Processing Conference*, 2010, pp. 2126–2130.

[3] S. Malik and G. Enzner, “State-Space Frequency-Domain Adaptive Filtering for Nonlinear Acoustic Echo Cancellation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2065–2079, 2012.

[4] F. Yang, G. Enzner, and J. Yang, “Frequency-Domain Adaptive Kalman Filter With Fast Recovery of Abrupt Echo-Path Changes,” *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1778–1782, 2017.

[5] G. Revach, N. Shlezinger, R. J. G. van Sloun, and Y. C. Eldar, “Kalmannet: Data-Driven Kalman Filtering,” in *ICASSP*, 2021, pp. 3905–3909.

[6] G. Revach, N. Shlezinger, X. Ni, A. L. Escoriza, R. J. G. van Sloun, and Y. C. Eldar, “Kalmannet: Neural Network Aided Kalman Filtering for Partially Known Dynamics,” *IEEE Transactions on Signal Processing*, vol. 70, pp. 1532–1547, 2022.

[7] T. Haubner, A. Brendel, and W. Kellermann, “End-To-End Deep Learning-Based Adaptation Control for Frequency-Domain Adaptive System Identification,” in *ICASSP*, 2022, pp. 766–770.

[8] A. Irevy, I. Cohen, and B. Berdugo, “Deep Adaptation Control for Acoustic Echo Cancellation,” in *ICASSP*, 2022, pp. 741–745.

[9] J. Casebeer, N. J. Bryan, and P. Smaragdis, “Meta-AF: Meta-Learning for Adaptive Filters,” *arXiv preprint arXiv:2204.11942*, 2022.

[10] H. Zhang and D. Wang, “Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios,” in *Proc. Interspeech 2018*, 2018, pp. 3239–3243.

[11] A. Fazel, M. El-Khamy, and J. Lee, “CAD-AEC: Context-Aware Deep Acoustic Echo Cancellation,” in *ICASSP*, 2020, pp. 6919–6923.

[12] N. L. Westhausen and B. T. Meyer, “Acoustic EchoCancellation with the Dual-Signal Transformation LSTM Network,” in *ICASSP*, 2021, pp. 7138–7142.

[13] S. Zhang, Y. Kong, S. Lv, Y. Hu, and L. Xie, “F-T-LSTM Based Complex Network for Joint Acoustic Echo Cancellation and Speech Enhancement,” in *Proc. Interspeech 2021*, 2021, pp. 4758–4762.

[14] B. J. Cho and H.-M. Park, “Stereo Acoustic Echo Cancellation Based on Maximum Likelihood Estimation With Inter-Channel-Correlated Echo Compensation,” *IEEE Transactions on Signal Processing*, vol. 68, pp. 5188–5203, 2020.

[15] F. Kuech, E. Mabande, and G. Enzner, “State-space architecture of the partitioned-block-based acoustic echo controller,” in *ICASSP*, 2014, pp. 1295–1299.