2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 979-8-3503-0689-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASRU57964.2023.10389780

NEURALKALMAN: A LEARNABLE KALMAN FILTER FOR ACOUSTIC ECHO CANCELLATION

Yixuan Zhang[1*], *Meng Yu*[2], *Hao Zhang*[2], *Dong Yu*[2], *DeLiang Wang*[1]

[1]The Ohio State University, Columbus, OH, USA
[2]Tencent AI Lab, Bellevue, WA, USA

ABSTRACT

The robustness of the Kalman filter to double talk and its rapid convergence make it a popular approach for addressing acoustic echo cancellation (AEC) challenges. However, the inability to model nonlinearity and the need to tune control parameters cast limitations on such adaptive filtering algorithms. In this paper, we integrate the frequency domain Kalman filter (FDKF) and deep neural networks (DNNs) into a hybrid method, called NeuralKalman, to leverage the advantages of deep learning and adaptive filtering algorithms. Specifically, we employ a DNN to estimate nonlinearly distorted far-end signals, a transition factor, and the nonlinear transition function in the state equation of the FDKF algorithm. Experimental results show that the proposed NeuralKalman improves the performance of FDKF significantly and outperforms strong baseline methods.

Index Terms: Acoustic echo cancellation, Kalman filter, deep learning, NeuralKalman

1. INTRODUCTION

Acoustic echo cancellation (AEC), as an active and challenging research problem in the domain of speech processing, has been studied for decades and is widely used in mobile communication and teleconferencing systems. The goal of AEC is to eliminate the far-end signal from the near-end microphone signal so as to remove the echo of the far-end signal (back to the far end). In conventional digital signal processing (DSP)-based adaptive filtering algorithms [1, 2, 3, 4, 5] including normalized least mean square (NLMS) and affine projection, RLS, echo removal is achieved by constantly estimating the linear transfer function between the loudspeaker playing the far-end signal and the near-end microphone, known as the echo path. However, in such AEC algorithms, control parameters need to be tuned to ensure fast convergence, and nonlinearity modeling (i.e. nonlinearity introduced by a loudspeaker) is missing.

With recent advances in deep neural networks, deep learning-based methods [6, 7, 8] have been utilized for AEC, and their ability to model nonlinear relations leads to promising results, even in challenging noisy or double-talk scenarios. Such methods usually treat AEC as a source separation problem and directly estimate the near-end signal based on the microphone and far-end reference signal. In recent AEC challenges [9], two-stage hybrid systems [10, 11, 12, 13] that use DNN as a nonlinear post-processor of a DSP-based adaptive filtering algorithm have shown promising results. In such hybrid systems, DNNs perform nonlinear residual echo suppression, which compensates for the drawbacks of adaptive filtering algorithms. To further leverage the advantages of DNN and adaptive filtering algorithms, method such as Deep Adaptive AEC [14] trains a hybrid model where a linear algorithm is embedded as differentiable layers, which has been proven to be highly effective in modeling a time-varying echo path.

As an adaptive filtering algorithm for AEC, the frequency domain Kalman filter (FDKF) [15, 16] shows robustness in double-talk scenarios and better convergence rates. Hybrid methods based on the Kalman filter algorithm [17, 18, 19] have been used in research fields such as pose estimation, and speech filtering, but have not been well explored in the domain of AEC. The most related study is the Neural Kalman Filtering proposed in [20], where a DNN is trained to estimate a Kalman gain. Directly estimating the Kalman gain, however, omits crucial steps in the Kalman filter and leads to a hybrid model that resembles estimating a step size in NLMS algorithms, such as the Deep Adaptive AEC approach proposed in [14]. Therefore, determining the optimal approach to leverage the Kalman filter and utilize DNNs to enhance the hybrid model remains an uncertain problem that is worth further investigation.

Our objective in this study is to develop a hybrid model that maximizes the benefits of both the frequency-domain Kalman filtering algorithm and DNNs. Our findings suggest that solely estimating components in the Kalman filter with DNNs does not necessarily result in improved performance. However, using DNNs to estimate missing or approximated components in the Kalman filter can lead to significant improvements. Specifically, we utilize DNN to estimate the nonlinearly distorted far-end signal, the transition factor and a nonlinear transition function in the state equation of

*This work was done during an internship at Tencent AI Lab.

the frequency-domain Kalman filter. Experimental results show that modeling the nonlinear distortion in far-end signals yields substantial improvements to the NeuralKalman. The transition factor shows adaptations to abrupt echo path changes and introducing a nonlinear transition function in the state equation accelerates training. Compared to modeling the covariance of the state noise and observation noise, we observe that injecting a nonlinear transition function in the state equation achieves similar improvement with less computation. The results show that the proposed hybrid NeuralKalman model suppresses echo well and outperforms the recent NLMS-based Deep Adaptive AEC [14].

2. PROPOSED METHOD: NEURALKALMAN

In a typical acoustic echo scenario, the far-end signal $x(t)$ is transmitted to the near end via a loudspeaker and received by a microphone as acoustic echo $d(t)$:

$$d(t) = h(t) * NL(x(t)) \tag{1}$$

where $h(t)$ represents the echo path, $NL(\cdot)$ represents the nonlinear distortion from the loudspeaker, $*$ denotes convolution. The microphone signal $y(t)$ is composed of echo $d(t)$, near-end speech $s(t)$ and noise $n(t)$:

$$y(t) = s(t) + n(t) + d(t) \tag{2}$$

and it is usually processed, with the far-end signal $x(t)$ as a reference, for echo removal before being sent to the far end.

2.1. Frequency-domain Kalman Filter

Frequency-domain Kalman filter for AEC [15, 16] estimates echo signal by modeling the echo path with an adaptive filter $\mathbf{W}(k)$ where k denotes the frame index, as shown in Fig. 1(a). In this study, we focus on AEC in clean condition and aim at estimating the near-end speech $s(t)$. FDKF can be interpreted as a two-step procedure and the updating of filter weights is achieved through the iterative feedback from the two steps. Following the notation in [15], in the prediction step, the frequency-domain near-end signal vector $\mathbf{S}(k)$ is estimated by the measurement equation,

$$\hat{\mathbf{S}}(k) = \mathbf{Y}(k) - \mathbf{G}^{01}\mathbf{X}(k)\hat{\mathbf{W}}(k), \tag{3}$$

where $\mathbf{X}(k)$ is the frequency-domain far-end signal matrix, $\mathbf{Y}(k)$ corresponds to the frequency-domain microphone signal vector. $\mathbf{G}^{01}$ is the overlap-save projection matrix. $\hat{\mathbf{W}}(k)$ denotes the estimated echo path in the frequency domain. In the update step, the state equation for updating echo path $\hat{\mathbf{W}}(k)$ is defined as,

$$\hat{\mathbf{W}}(k+1) = A[\hat{\mathbf{W}}(k) + \mathbf{G}^{01}\mathbf{K}(k)\hat{\mathbf{S}}(k)], \tag{4}$$

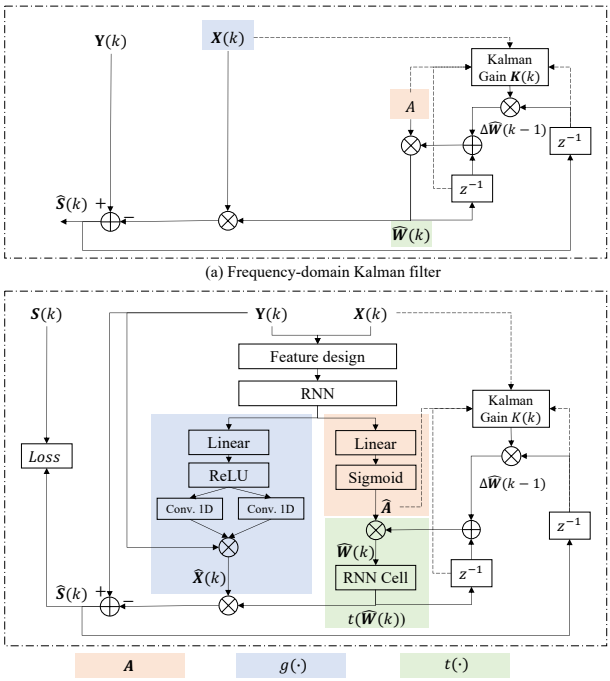

Fig. 1. Diagrams of (a) Frequency-domain Kalman filter and (b) proposed NeuralKalman, where z^{-1} denotes the unit delay.

where A is the transition factor. $\mathbf{K}(k)$ denotes the Kalman gain. As shown in Fig. 1(a), $\mathbf{K}(k)$ is related to far-end signal $\mathbf{X}(k)$, echo path $\hat{\mathbf{W}}(k-1)$ and estimated near-end signal $\hat{\mathbf{S}}(k-1)$. The dash line indicates the relations not expressed directly in the equations, e.g., $\hat{\mathbf{S}}(k)$ in $\mathbf{\Psi}_{vv}(k)$ [15]. The calculation of $\mathbf{K}(k)$ is defined as,

$$\mathbf{K}(k) = \mathbf{P}(k)\mathbf{X}^H(k)[\mathbf{X}(k)\mathbf{P}(k)\mathbf{X}^H(k) + 2\mathbf{\Psi}_{vv}(k)]^{-1}, \tag{5}$$

$$\mathbf{P}(k+1) = A^2[\mathbf{I} - \frac{1}{2}\mathbf{K}(k)\mathbf{X}(k)]\mathbf{P}(k) + \mathbf{\Psi}_{\Delta\Delta}(k), \tag{6}$$

where $\mathbf{P}(k)$ is the state estimation error covariance. $\mathbf{\Psi}_{vv}(k)$ and $\mathbf{\Psi}_{\Delta\Delta}(k)$ are observation noise covariance and process noise covariance respectively and are approximated by the covariance of the estimated near-end signal $\mathbf{\Psi}_{\hat{s}\hat{s}}(k)$ and the echo-path $\mathbf{\Psi}_{\hat{W}\hat{W}}(k)$, respectively. More details can be found in [15].

2.2. NeuralKalman Framework

While being robust to double-talk and achieving a better convergence rate, the FDKF algorithm still faces several challenges. First, the echo is modeled as a linear transform of far-end signal $\mathbf{X}(k)$ while neglecting the nonlinear distortion caused by amplifiers. Second, in FDKF algorithm, $\mathbf{\Psi}_{vv}(k)$ and $\mathbf{\Psi}_{\Delta\Delta}(k)$ are approximated. An inaccurate estimate of covariance will degrade the performance of FDKF algorithm

[15]. Third, in the FDKF algorithm, the transition factor A in the state equation (Eq. 4) is typically assigned a fixed value that is manually adjusted based on the non-stationarity of the echo path. However, a fixed A is less likely to adapt well to the changing environment.

To address these problems, we propose NeuralKalman. Unlike [20] which directly estimates Kalman gain from concatenated input features comprising estimated far-end, near-end signals, and innovation of $\hat{W}$, the proposed framework leverages DNNs to estimate the transition factor A, the far-end nonlinear distortion $g(\cdot)$, and a nonlinear transition function $t(\cdot)$, as shown in Figure 1(b). Since the echo path and nonlinear distortion information can be retrieved from the microphone and far-end signals, the input feature computed from the complex STFT of the microphone signal and far-end signal is shared for estimating transition factor A and nonlinear distortion $g(\cdot)$. Similar to NeuralEcho [8], the employed input feature is a concatenation of temporal correlation, frequency correlation, channel covariance, and normalized log power spectrum of microphone and far-end signal. As shown in Fig. 1(b), an recurrent neural network (RNN) takes the computed input feature and is followed by two branches for estimating nonlinearly distorted far-end signal $\hat{\mathbf{X}}(k)$ and transition factor $\hat{A}(k)$ respectively. The shared RNN is a 4-layer long short-term memory (LSTM) network where each layer has 257 hidden units.

2.2.1. Nonlinear Distortion $g(\cdot)$

To address the nonlinear distortion introduced by the loudspeaker, we estimate the far-end nonlinear distortion with DNNs and use the nonlinear far-end signal as a reference for updating the Kalman filter. As illustrated in Figure 1(b), the sub-network responsible for estimating $g(\cdot)$ consists of a linear layer with Rectified Linear Unit (ReLU) activation, followed by two one-dimensional convolution layers (Conv. 1D). The nonlinearly distorted far-end signal $\hat{\mathbf{X}}(k)$ is obtained by applying the complex-valued ratio filters cRF [21, 8] to the microphone signal $\mathbf{Y}(k)$.

2.2.2. Nonlinear Transition Function $t(\cdot)$

In [20], the paper illustrates the effectiveness of implicitly modeling the covariance of state noise and observation noise during the Kalman gain estimation process. Our observation, as discussed in Sec. 4, also demonstrates that incorporating covariance modeling enhances performance. However, we find that introducing a nonlinear transition function in the state equation leads to a comparable performance improvement with less computation. More specifically, we replace the linear transition function in Eq. 4 with a nonlinear one:

$$\hat{\mathbf{W}}(k+1) = t(A[\hat{\mathbf{W}}(k) + \mathbf{G^{01}K}(k)\hat{\mathbf{S}}(k)]), \qquad (7)$$

where the nonlinear transition function $t(\cdot)$ is estimated from an LSTM cell which has 256 hidden units. The input to the LSTM cell consists of the estimated $\hat{\mathbf{W}}(k+1)$ from Eq. 4 and previous state h_{k-1}. Then two linear layers take h_k as input and output the real and imaginary parts of the processed $t(\hat{\mathbf{W}}(k+1))$.

$$\begin{aligned} h_k &= \mathbf{RNN}(\hat{\mathbf{W}}(k+1), h_{k-1}), \\ t(\hat{\mathbf{W}}(k+1)) &= \mathbf{FNN}(h_k), \end{aligned} \qquad (8)$$

where $\mathbf{RNN}$ and $\mathbf{FNN}$ denote the LSTM cell and linear layers respectively.

2.2.3. Transition Factor A

Transition factor A in the range of [0,1] depicts the variation of the Kalman filter and it is often manually tuned to a value that is close to 1. To incorporate the influence of possible changes in the echo path on the transition factor, instead of using a fixed value, we employ DNN to estimate a time-varying transition factor for Eq. 4. The branch for estimating frame-based transition factor $A(k)$ is composed of a linear layer followed by a sigmoidal activation function.

2.2.4. Loss Function

The loss function is defined to jointly optimize SI-SDR [22] in the time domain and mean absolute error (MAE) of magnitude spectrogram between the target and estimated near-end signal.

$$L = -\mathbf{SI\text{-}SDR}(s, \hat{s}) + \alpha\mathbf{MAE}(|\mathbf{S}|, |\hat{\mathbf{S}}|), \qquad (9)$$

where $\hat{s}$ and $\hat{\mathbf{S}}$ are the estimated time-domain and frequency-domain near-end signal, respectively. And α is set to $10,000$ in our implementation to balance the value range of the two losses.

3. EXPERIMENTAL SETUP

3.1. Dataset

Following [8], we simulate the single-channel AEC dataset using AISHELL-2 [23] and AEC-Challenge [9] datasets. We use clean and nonlinearly distorted far-end signals from AEC-Challenge's synthetic echo set [9]. Nonlinear distortions such as maximum amplitude clipping with a Sigmoidal function [6], learned distortion functions, etc. are included in the far-end signals. To simulate acoustic echo, 10k room impulse responses (RIRs) sets with random room characteristics are generated using the image-source method [24] with reverberation time (RT60) ranging from 0 to 0.6 seconds. Each of the 10k RIRs sets comprises the RIRs from locations of the loudspeaker, near-end speaker. During data generation, the signal-to-echo-ratio (SER) ranges from -10 dB to 10 dB and RIRs set is randomly picked. The training set has 90k utterances, and 10k utterances are randomly selected in each epoch for training. Each network is trained for 90 epochs. We generated

Table 1. Performance of NeuralKalman with various settings in the presence of double-talk.

	PESQ	WER
Unprocessed	1.87	79.85%
Kalman Filter [16]	2.32	32.89%
NeuralKalman-$g(\cdot)/t(\cdot)/A$	**2.67**	**20.41%**
NeuralKalman-$g(\cdot)/A$	2.57	22.37%
NeuralKalman-$g(\cdot)/t(\cdot)$	2.59	21.98%
NeuralKalman-$g(\cdot)/\mathbf{\Psi}/A$	2.65	21.79%
NeuralKalman-$g(\cdot)/\mathbf{\Psi}/t(\cdot)/A$	**2.67**	20.72%

200 utterances for validation and 300 utterances for testing. The test set is generated from utterances and RIRs that are not seen in training process. All input audios are sampled at 16 kHz. STFT is computed with a 32 ms frame length and 50% frame shift.

3.2. Evaluation Metrics

We evaluate the echo cancellation performance of the proposed NeuralKalman using perceptual evaluation of speech quality (PESQ) [25] and word error rate (WER). To evaluate WER, we use a commercial general-purpose speech recognition API [26] to test the automatic speech recognition (ASR) performance.

4. EXPERIMENTAL RESULTS

4.1. NeuralKalman Evaluation

We conduct an ablation study to primarily investigate the impact of modeling the nonlinear transition function $t(\cdot)$ and the transition factor A. The influence of modeling of nonlinear distortion $g(\cdot)$ has been well examined in prior research [14]. NeuralKalman models with different DNN components are built for comparison and the results are shown in Table 1. The NeuralKalman model discussed in Sec. 2.2 and shown in Fig. 1(b) corresponds to NeuralKalman-$g(\cdot)/t(\cdot)/A$ which estimates a nonlinear distortion function $g(\cdot)$, a nonlinear transition function $t(\cdot)$ and the transition factor A. NeuralKalman-$g(\cdot)/A$ uses DNN to jointly estimate a nonlinear distortion function $g(\cdot)$ and the transition factor A. NeuralKalman-$g(\cdot)/t(\cdot)$ estimates a nonlinear distortion function $g(\cdot)$ and a nonlinear transition function $t(\cdot)$.

Since an accurate estimation of covariance matrices in FDKF would contribute to better convergence rate and AEC performance [15], we also train a NeuralKalman-$g(\cdot)/\mathbf{\Psi}/A$ model which involves training two LSTM cells with 256 hidden units to estimate the covariance matrices $\mathbf{\Psi}_{vv}(k)$ and $\mathbf{\Psi}_{\Delta\Delta}(k)$. The inputs to the RNNs for estimating $\mathbf{\Psi}_{vv}(k)$ and $\mathbf{\Psi}_{\Delta\Delta}(k)$ are the estimated near-end speech $\hat{\mathbf{S}}(k)$ and updated $\hat{\mathbf{W}}(k)$, respectively. NeuralKalman-$g(\cdot)/\mathbf{\Psi}/t(\cdot)/A$ which performs both covariance matrices and nonlinear transition function estimation is also trained for comparison.

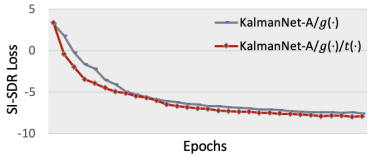

Fig. 2. SI-SDR loss curve of NeuralKalman-$A/g(\cdot)$ and NeuralKalman-$A/g(\cdot)/t(\cdot)$.

4.1.1. *Nonlinear Distortion* $g(\cdot)$

While the influence of modeling $g(\cdot)$ has been discovered in [14], we also observe that the key improvement comes from modeling the far-end nonlinear distortion. We find that by solely estimating A, we achieve a PESQ score of 2.32 and a WER of 31.67% which is slightly better than Kalman filter. Compared to only estimating A, we find that further introducing far-end nonlinear distortion $g(\cdot)$ increases PESQ by 0.25, and reduces WER relatively by 29.4%.

4.1.2. *Nonlinear Transition Function* $t(\cdot)$

With the learned nonlinear transition function $t(\cdot)$, the performance of NeuralKalman is further improved. As shown in Table 1, by comparing the results of NeuralKalman-$g(\cdot)/A$ and NeuralKalman-$g(\cdot)/t(\cdot)/A$, we find that PESQ is improved by 0.1, and WER is relatively reduced by 8.8%. NeuralKalman-$g(\cdot)/\mathbf{\Psi}/A$ which substitute approximated $\mathbf{\Psi}_{vv}(k)$ and $\mathbf{\Psi}_{\Delta\Delta}(k)$ in FDKF with DNN learned covariance, has also shown improved performance in terms of all metrics. Compared to NeuralKalman-$g(\cdot)/\mathbf{\Psi}/A$, we observe that NeuralKalman-$g(\cdot)/t(\cdot)/A$ can achieve slightly better performance with less computational cost. Also, the result of NeuralKalman-$g(\cdot)/\mathbf{\Psi}/t(\cdot)/A$ shows that estimating both covariance and nonlinear transition function does not bring further improvement. In addition, we observe from Fig. 2 that estimating the nonlinear transition function $t(\cdot)$ brings faster training convergence speed. Therefore, we decide to only estimate $t(\cdot)$ instead of $\mathbf{\Psi}$.

4.1.3. *Transition Factor* A

By incorporating the modeling of the transition factor A, the Kalman filter gains enhanced flexibility in controlling the update of the echo path We validate the necessity of estimating transition factor A by eliminating the estimation of A from the best-performing NeuralKalman-$g(\cdot)/t(\cdot)/A$ model, where we find that PESQ reduces by 0.1 and WER increases to 21.98%. To examine the response of A in the scenario of abrupt echo path change, we have also evaluated a model that solely estimates A (named as NeuralKalman-A) on a sample

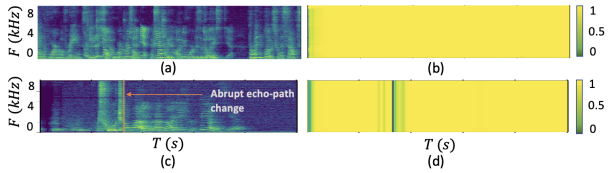

Fig. 3. Estimated transition factor A for signals with no echo path change (a, b) and with abrupt echo path change (c, d): (a)(c) Magnitude STFT of a microphone signal, (b)(d) Estimation of A from NeuralKalman-A model.

Table 2. Performance comparison with baseline methods.

	PESQ	WER
Unprocessed	1.87	79.85%
Kalman Filter [16]	2.32	32.89%
NLMSNet [14][1]	2.52	23.33%
DNN-AEC	2.62	23.14%
NeuralKalman-$g(\cdot)/t(\cdot)/A$	**2.67**	**20.41%**

with abrupt echo path change. The RIR is aburptly switched from one to another at the indicated position in Fig. 3. We plot the value of learned A from the NeuralKalman-A model on audios with and without abrupt echo path change. It is observed that for the audio signal without echo path change, the learned A is close to 1 and relatively stable throughout time, which explains why the performance of NeuralKalman-A is similar to that of FDKF. For the signal with abrupt echo path change, we observe that the value of learned A decreases to nearly 0 when the echo path abruptly changes and gradually increases afterward, which is reasonable to obtain a stable convergence. We believe that it is proper to use a time-varying A in scenarios with echo path changes to make the updating of the algorithm stable and diminish the chances of divergence.

4.1.4. Other observations

Other experiments such as using additional LSTM cells to estimate the complex matrices of $\mathbf{K}(k)$ and $\mathbf{P}(k)$ on top of NeuralKalman-$g(\cdot)/t(\cdot)/A$ are performed and achieve similar performance. It is observed that solely estimating Kalman filter components using DNNs did not consistently improve performance. However, estimating missing or approximated components yields significant improvements.

4.2. Comparison with Baselines

We compare the proposed NeuralKalman-$g(\cdot)/t(\cdot)/A$ model with strong baseline methods including frequency-domain Kalman filter [16], NLMSNet which is based on deep adaptive AEC [14] , and a fully DNN-based model DNN-AEC.

[1]NLMSNet is a modified and retrained version of [14] for fair comparison.

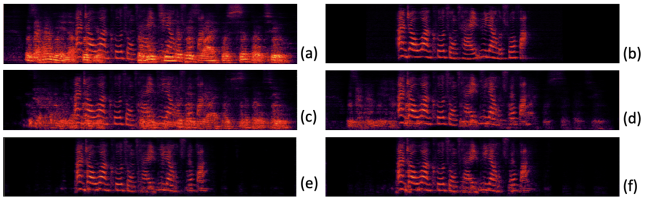

Fig. 4. Spectrograms of (a) microphone signal, (b) target near-end signal, and outputs of (c) Kalman filter, (d) NLMSNet, (e) DNN-AEC, and (f) our proposed NeuralKalman-$g(\cdot)/t(\cdot)/A$.

Following [14], NLMSNet is a hybrid model based on NLMS algorithm and takes microphone and far-end signal as inputs and estimate the step size parameter and the non-linear far-end signal. DNN-AEC is a method based entirely on DNNs that uses the microphone and far-end signal as inputs to directly predict the speech at the near-end. In our experiments, NLMSNet and DNN-AEC adopt the same RNN network, input feature and loss function as the proposed NeuralKalman which are described in Sec. 2.2. From Table 2, we observe that all hybrid methods outperform the frequency-domain Kalman filter algorithm. Among the hybrid methods, NeuralKalman has the best performance. We observe that NLMSNet does not show superiority over DNN-AEC when trained on data with stationary echo path. Compared to NLMSNet, NeuralKalman improves the PESQ by 0.15 and relatively improves WER by 12.5%. NeuralKalman outperforms DNN-AEC in terms of PESQ and WER, with PESQ showing a 0.05 improvement and WER showing a relative improvement of 11.8%. Fig. 4 shows the magnitude STFT of the near-end signal estimated by different methods. The frequency-domain Kalman filter's ability to suppress echoes appears to be limited. Hybrid method such as NLMSNet shows promising echo suppression results in double talk regions; however, echoes remain present in single talk regions. Among the baseline methods, DNN-AEC and NeuralKalman-$g(\cdot)/t(\cdot)/A$ demonstrate superior effectiveness in removing echoes.

5. CONCLUSION

In this paper, we have proposed a learnable Kalman filter for acoustic echo cancellation. The proposed model leverages the advantages of DNN to improve the Kalman filter by estimating the missing or approximated components, including the transition factor, nonlinear distortion of the far-end signal, and nonlinear transition function for the estimated echo path. Systematic evaluations show that the proposed method outperforms recent baseline methods. For future work, with more access to data, we will explore training NeuralKalman on real-recorded signals with echo path changes and explore utilizing it in real-world devices.

6. REFERENCES

[1] Donald L Duttweiler, "Proportionate normalized least-mean-squares adaptation in echo cancelers," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 508–518, 2000.

[2] Steven L Gay, "The fast affine projection algorithm," in *Acoustic Signal processing for Telecommunication*, pp. 23–45. 2000.

[3] Andreas Mader, Henning Puder, and Gerhard Uwe Schmidt, "Step-size control for acoustic echo cancellation filters–an overview," *Signal Processing*, vol. 80, pp. 1697–1719, 2000.

[4] Sarmad Malik and Gerald Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 2065–2079, 2012.

[5] Jafar Ramadhan Mohammed and Gurnam Singh, "An efficient RLS algorithm for output-error adaptive IIR filtering and its application to acoustic echo cancellation," in *IEEE Symposium on Computational Intelligence in Image and Signal Processing*, 2007, pp. 139–145.

[6] Hao Zhang and DeLiang Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," in *Proc. Interspeech*, 2018, p. 322.

[7] Hao Zhang and DeLiang Wang, "Neural cascade architecture for multi-channel acoustic echo suppression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2326–2336, 2022.

[8] Meng Yu, Yong Xu, Chunlei Zhang, Shi-Xiong Zhang, and Dong Yu, "NeuralEcho: A self-attentive recurrent neural network for unified acoustic echo suppression and speech enhancement," *arXiv preprint arXiv:2205.10401*, 2022.

[9] Ross Cutler, Ando Saabas, Tanel Parnamaa, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sørensen, and Robert Aichner, "ICASSP 2022 acoustic echo cancellation challenge," in *Proc. ICASSP*, 2022, pp. 9107–9111.

[10] 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 *Proc. ICASSP*, 2021, pp. 7133–7137.

[11] Ziteng Wang, Yueyue Na, Zhang Liu, Biao Tian, and Qiang Fu, "Weighted recursive least square filter and neural network based residual echo suppression for the AEC-challenge," in *Proc. ICASSP*, 2021, pp. 141–145.

[12] 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 *Proc. Interspeech*, 2021, pp. 4768–4772.

[13] Thomas Haubner, Mhd Modar Halimeh, Andreas Brendel, and Walter Kellermann, "A synergistic Kalman and deep postfiltering approach to acoustic echo cancellation," in *Proc. EUSIPCO*, 2021, pp. 990–994.

[14] 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 *Proc. ICASSP*, 2022, pp. 756–760.

[15] Feiran Yang, Gerald Enzner, and Jun Yang, "Frequency-domain adaptive Kalman filter with fast recovery of abrupt echo-path changes," *IEEE Signal Processing Letters*, vol. 24, pp. 1778–1782, 2017.

[16] Gerald Enzner and Peter Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, pp. 1140–1156, 2006.

[17] Beren Millidge, Alexander Tschantz, Anil Seth, and Christopher Buckley, "Neural Kalman filtering," *arXiv preprint arXiv:2102.10021*, 2021.

[18] Guy Revach, Nir Shlezinger, Xiaoyong Ni, Adria Lopez Escoriza, Ruud JG Van Sloun, and Yonina C Eldar, "Kalmannet: Neural network aided kalman filtering for partially known dynamics," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1532–1547, 2022.

[19] Huseyin Coskun, Felix Achilles, Robert DiPietro, Nassir Navab, and Federico Tombari, "Long short-term memory kalman filters: Recurrent neural estimators for pose regularization," in *Proc. of the IEEE International Conference on Computer Vision*, 2017, pp. 5524–5532.

[20] Dong Yang, Fei Jiang, Wei Wu, Xuefei Fang, and Muyong Cao, "Low-complexity acoustic echo cancellation with neural Kalman filtering," *arXiv preprint arXiv:2207.11388*, 2022.

[21] Wolfgang Mack and Emanuël AP Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Processing Letters*, vol. 27, pp. 61–65, 2019.

[22] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *Proc. ICASSP*, 2019, pp. 626–630.

[23] Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu, "AISHELL-2: Transforming mandarin ASR research into industrial scale," *arXiv:1808.10583*, 2018.

[24] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, 1979.

[25] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.

[26] "Tencent ASR," in http://ai.qq.com/product/aaiasr.shtml.