Deep Residual Network-Based Augmented Kalman Filter for Speech Enhancement

Sujan Kumar Roy and Kuldip K. Paliwal
Signal Processing Laboratory, School of Engineering and Built Environment
Griffith University, Brisbane, QLD, Australia, 4111
E-mail: sujankumar.roy@griffithuni.edu.au, k.paliwal@griffith.edu.au

***Abstract*—Speech enhancement using augmented Kalman filter (AKF) suffers from the inaccurate estimates of the key parameters, linear prediction coefficients (LPCs) of speech and noise signal in noisy conditions. The existing AKF particularly enhances speech in colored noise conditions. In this paper, a deep residual network (ResNet)-based method utilizes the LPC estimates of the AKF for speech enhancement in various noise conditions. Specifically, a ResNet20 (constructed with 20 layers) gives an estimate of the noise waveform for each noisy speech frame to compute the noise LPC parameters. Each noisy speech frame is pre-whitened by a whitening filter, which is constructed with the corresponding noise LPCs. The speech LPC parameters are computed from the pre-whitened speech. The improved speech and noise LPC parameters enable the AKF to minimize residual noise as well as distortion in the enhanced speech. Objective and subjective testing on NOIZEUS corpus reveal that the proposed method exhibits higher quality and intelligibility in the enhanced speech than some benchmark methods in various noise conditions for a wide range of SNR levels.**

***Index Terms*—Speech enhancement, augmented Kalman filter, residual network, LPC, whitening filter.**

I. Introduction

The main objective of a speech enhancement algorithm (SEA) is to eliminate the embedded noise from the noisy speech signal. The SEAs can be used as a pre-processing tool for many signal processing systems, such as voice communication systems, hearing-aid devices, voice operated autonomous systems. Various SEAs, such as spectral subtraction (SS) [1], [2], minimum mean square error (MMSE) [3], [4], Wiener Filter (WF) [5], [6], Kalman filter (KF) [7], augmented KF (AKF) [8], deep neural network (DNN) [9], and machine learning-based KF/AKF [10], [11], [12] have been introduced over the decades. This paper integrates a deep residual network with AKF for single-channel speech enhancement.

Paliwal and Basu for the first time introduced KF for speech enhancement in white noise condition [7]. In KF, a speech signal is represented by an autoregressive (model) and represented in the Kalman recursion equations. KF gives a linear MMSE estimate of the clean speech given the observed noisy speech for each sample within each a frame. Therefore, the performance of KF-based SEA depends on how accurately the key parameters, such as LPCs are estimated in noisy conditions. It is demonstrated in [7] that the LPC parameters estimated from the clean speech shows excellent performance. On the contrary, the LPC parameters computed from the noisy speech are inaccurate and degrades the KF performance for speech enhancement. In [8], Gibson *et al.* introduced an augmented KF (AKF) for enhancing colored noise corrupted speech. In this method, the LPC parameters for the current noisy speech frame are computed from the filtered signal of the previous iteration by AKF. Although the enhanced speech (after 2-3 iterations) shows SNR improvement, however, suffering from spectral distortion as well as musical noise. In [13], Roy *et al.* proposed a sub-band iterative KF-based SEA. Since it only enhances the high-frequency sub-bands (SBs) among the 16 decomposed SBs of noisy speech, some noise components may still remain in the low-frequency SBs. The enhanced speech also suffers from distortion. In [14], George *et al.* introduced a robustness metric-based tuning of the AKF for enhancing colored noise corrupted speech. However, it is shown that the robustness metric-based tuning of the bias in the AKG gain is particularly applicable in colored noise conditions. Also, the tuning process of the AKF gain causes distortion in the enhanced speech. To address this problem, a sensitivity metric-based tuning of the AKF has been proposed [15]. Although it produces less distorted speech, however, performance becomes degraded in real-life noise conditions.

The deep neural network (DNN) has been used widely for speech enhancement over the decades. It shows a noticeable improvement over the traditional SEAs [1], [3], [5], [7]. Motivated by the time-frequency (T-F) masking technique in computational auditory scene analysis [16], the early DNN-based SEAs focus on the mask estimation, which is used to reconstruct the clean speech spectrum. In [9], Wand and Wang introduced a multi-layer perceptron (MLP)-based ideal binary mask (IBM) estimation method. An estimate of the clean speech spectrum is given by multiplying the estimated IBM with the noisy speech spectrum [17]. In [18], it was shown that the ideal ratio mask (IRM) exhibits better speech enhancement accuracy over the IBM. Usually, the masking-based SEAs [9], [17], [18] keep the phase spectrum unprocessed in the sense that it is less affected by noise. However, in [19], Paliwal et al. showed that the improvement of the phase spectrum also improves the perceptual quality of the enhanced speech. In this circumstance, Williamson et al. introduced a complex ideal ratio mask (cIRM)-based SEA for further improving the speech enhancement accuracy [20]. The cIRM is capable to recover both the amplitude and the phase spectrum of the clean speech. In general, it was observed that the masking-based SEAs introduce residual noise in the enhanced speech [18].

Also, in speech enhancement context, the traditional MLP and DNN-based methods [9], [18], are not able to learn the long-term dependencies inherent in the noisy speech.

In [21], a fully-convolutional network (FCNN)-based SEA has been proposed. The FCNN processes the raw-waveform of the noisy speech, giving an estimate of the clean speech waveform. Thus, the enhanced speech does not affected by the phase, unlike other acoustic-domain SEAs [17], [18]. In [22], Zheng *et al.* introduced a phase-ware SEA using DNN. Here, the phase information (converted to the instantaneous frequency deviation (IFD)) is jointly used with different masks, namely the ideal amplitude mask (IAM) as a training target. The clean speech spectrum is reconstructed with the estimated mask and the phase information (extracted from the IFD).

Yu *et al.* introduced a KF-based SEA, where the LPC parameters are estimated using a traditional DNN [23]. However, the noise covariance is estimated during speech pauses of the noisy speech, which is irrespective in conditions having time varying amplitude. Recently, some advance deep learning-based KF/AKF methods [10], [11], [12] have been introduced to estimate the LPC parameters for improving speech enhancement performance. These methods were also found to enhance speech in various noise conditions.

The direct estimation of speech from the noisy speech using the benchmark deep learning methods may suffer from residual noise and distortion. Our investigation reveals that the noise estimation using deep learning technique would be more beneficial, as it is a crucial parameter for most of the SEAs in literature. For example, the AKF-based SEA suffering from the inaccurate estimates of noise LPC parameters in practice. This paper introduces a ResNet20 to accurately estimate the noise LPC parameters of the AKF. Specifically, the ResNet20 gives an estimate of the noise waveform to compute the noise LPC parameters for each noisy speech frame. A whitening filter is also constructed with the noise LPCs to pre-whiten each noisy speech frame. The speech LPC parameters are computed from the pre-whitened speech. The AKF constructed with the improved speech and noise LPC parameters leading to the capability of speech enhancement in various noise conditions. The performance of the proposed method is compared against some benchmark methods using objective and subjective testing on NOIZEUS corpus.

II. AKF FOR COLORED NOISE SUPPRESSION

Assuming the colored noise, $v(n)$ to be additive with the clean speech, $s(n)$ and uncorrelated each other, at sample n, the noisy speech, $y(n)$ is given by:

$$y(n) = s(n) + v(n). \tag{1}$$

$s(n)$ and $v(n)$ in (1) can be modeled with p^{th} and q^{th} order AR models as [24]:

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + w(n), \tag{2}$$

$$v(n) = -\sum_{j=1}^{q} b_j v(n-j) + u(n), \tag{3}$$

where $\{a_i; i = 1, 2, \ldots, p\}$ and $\{b_j; j = 1, 2, \ldots, q\}$ are the LPCs, $w(n)$ and $u(n)$ are assumed to be white noise with zero mean and variance σ_w^2 and σ_u^2, respectively.

Equations (1)-(3) can be used to form the following augmented state-space model (ASSM) of AKF as [14]:

$$\boldsymbol{x}(n) = \boldsymbol{\Phi}\boldsymbol{x}(n-1) + \boldsymbol{d}\boldsymbol{z}(n), \tag{4}$$

$$y(n) = \boldsymbol{c}^T\boldsymbol{x}(n). \tag{5}$$

In the above ASSM,

1) $\boldsymbol{x}(n) = [s(n) \ \ldots s(n-p+1) \ v(n) \ \ldots \ v(n-q+1)]^T$ is a $(p+q)\times 1$ state-vector,
2) $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_s & 0 \\ 0 & \boldsymbol{\Phi}_v \end{bmatrix}$ is a $(p+q)\times(p+q)$ state-transition matrix with:

$$\boldsymbol{\Phi}_s = \begin{bmatrix} -a_1 & -a_2 & \ldots & a_{p-1} & a_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix},$$

$$\boldsymbol{\Phi}_v = \begin{bmatrix} -b_1 & -b_2 & \ldots & b_{q-1} & b_q \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix},$$

3) $\boldsymbol{d} = \begin{bmatrix} \boldsymbol{d}_s & 0 \\ 0 & \boldsymbol{d}_v \end{bmatrix}$, where $\boldsymbol{d}_s = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^\top$, $\boldsymbol{d}_v = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^\top$,
4) $\boldsymbol{z}(n) = \begin{bmatrix} w(n) \\ u(n) \end{bmatrix}$,
5) $\boldsymbol{c}^T = \begin{bmatrix} \boldsymbol{c}_s^T & \boldsymbol{c}_v^T \end{bmatrix}$, where $\boldsymbol{c}_s = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^T$ and $\boldsymbol{c}_v = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^T$ are $p\times 1$ and $q\times 1$ vectors,
6) $y(n)$ is the observed noisy speech at sample n.

Firstly, $y(n)$ is windowed into non-overlapped and short (e.g., 20 ms) frames. For a particular frame, the AKF computes an unbiased linear MMSE estimate, $\hat{\boldsymbol{x}}(n|n)$ at sample n, given $y(n)$ by using the following recursive equations [14]:

$$\hat{\boldsymbol{x}}(n|n-1) = \boldsymbol{\Phi}\hat{\boldsymbol{x}}(n-1|n-1), \tag{6}$$

$$\boldsymbol{\Psi}(n|n-1) = \boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^T + \boldsymbol{d}\boldsymbol{Q}\boldsymbol{d}^T, \tag{7}$$

$$\boldsymbol{K}(n) = \boldsymbol{\Psi}(n|n-1)\boldsymbol{c}(\boldsymbol{c}^T\boldsymbol{\Psi}(n|n-1)\boldsymbol{c})^{-1}, \tag{8}$$

$$\hat{\boldsymbol{x}}(n|n) = \hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^T\hat{\boldsymbol{x}}(n|n-1)], \tag{9}$$

$$\boldsymbol{\Psi}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^T]\boldsymbol{\Psi}(n|n-1), \tag{10}$$

where $\boldsymbol{Q} = \begin{bmatrix} \sigma_w^2 & 0 \\ 0 & \sigma_u^2 \end{bmatrix}$ is the process noise covariance.

For each noisy speech frame, the error covariances, $\boldsymbol{\Psi}(n|n-1)$ and $\boldsymbol{\Psi}(n|n)$ corresponding to $\hat{\boldsymbol{x}}(n|n-1)$ and $\hat{\boldsymbol{x}}(n|n)$, and the Kalman gain $\boldsymbol{K}(n)$ are continually updated on a samplewise basis, while $(\{a_i\}, \sigma_w^2)$ and $(\{b_k\}, \sigma_u^2)$ remain constant. At sample n, $\boldsymbol{g}^\top\hat{\boldsymbol{x}}(n|n)$ gives the estimated speech,

$\hat{s}(n|n)$, where $\boldsymbol{g} = \begin{bmatrix}1 & 0 & 0 & \dots & 0\end{bmatrix}^\top$ is a $(p+q)\times 1$ column vector. As in [14], $\hat{s}(n|n)$ is given by:

$$\hat{s}(n|n) = [1 - K_0(n)]\hat{s}(n|n-1) + K_0(n)[y(n) - \hat{v}(n|n-1)], \tag{11}$$

where $K_0(n)$ is the 1^{st} component of $\boldsymbol{K}(n)$, given by [14]:

$$K_0(n) = \frac{\alpha^2(n) + \sigma_w^2}{\alpha^2(n) + \sigma_w^2 + \beta^2(n) + \sigma_u^2}, \tag{12}$$

where $\alpha^2(n)$ and $\beta^2(n)$ are the transmission of *a posteriori* error variances by the speech and noise models from the previous time sample, $n-1$ [14].

Equation (11) reveals that $K_0(n)$ has a significant impact on $\hat{s}(n|n)$ estimates (the output of the AKF). In practice, the inaccurate estimates of $(\{a_i\},\ \sigma_w^2)$ and $(\{b_k\},\ \sigma_u^2)$ introduce bias in $K_0(n)$, which affects the estimates of $\hat{s}(n|n)$. In the proposed SEA, a ResNet20 is used to utilize the LPC estimates for the AKF, leading to an improved $\hat{s}(n|n)$ estimate.

III. Proposed Speech Enhancement System

Fig. 1 shows the block diagram of the proposed SEA. Unlike the AKF method in Section II, a 32 ms rectangular window with 50% overlap was considered for converting $y(n)$ (1) into frames, $y(n,l)$, i.e., $y(n,l) = s(n,l) + v(n,l)$, where $l\epsilon\{0,1,2,\dots,N-1\}$ is the frame index with N being the total number of frames in an utterance, and M is the total number of samples within each frame, i.e., $n\epsilon\{0,1,2,\dots,M-1\}$.

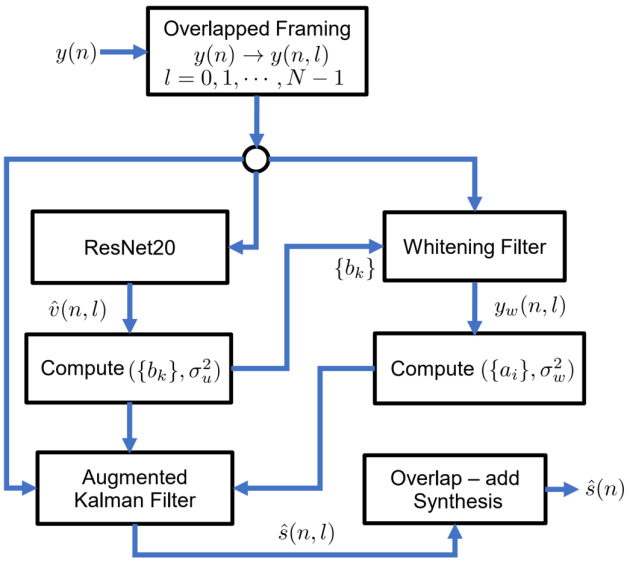

Fig. 1. Block diagram of the proposed SEA.

A. Proposed $(\{b_k\},\ \sigma_u^2)$ and $(\{a_i\},\ \sigma_w^2)$ Estimation Method

The speech LPC parameters, $(\{a_i\},\ \sigma_w^2)$ are very sensitive to noise. Since the clean speech, $s(n,l)$ is unavailable in practice, it is difficult to estimate these parameters accurately. In the existing AKF-based SEA, an estimate of the noise waveform, $\hat{v}(n,l)$ is obtained from some initial noisy speech frames by considering that there remains no speech [14]. Then compute $(\{b_k\},\ \sigma_u^2)$ from $\hat{v}(n,l)$, which remains constant during processing all noisy speech frames for a given utterance. This concept is only applicable to enhancing colored noise corrupted speech to some extent. However, due to the real-world noise may contain time varying amplitudes, it requires to update $(\{b_k\},\ \sigma_u^2)$ for each noisy speech frame. Therefore, $(\{b_k\},\ \sigma_u^2)$ estimation process in [14] becomes irrespective with the noise conditions having time varying amplitudes.

In this paper, we introduce a ResNet20 (described in section III-B) to estimate the noise waveform, $\hat{v}(n,l)$ corresponding to each $y(n,l)$. Then $(\{b_k\},\ \sigma_u^2)$ $(q=20)$ are computed from $\hat{v}(n,l)$ using the autocorrelation method [24]. To reduce bias in the estimated $(\{a_i\},\ \sigma_w^2)$ for each noisy speech frame, we compute them from the corresponding pre-whitened speech, $y_w(n,k)$ using the autocorrelation method [24]. The framewise $y_w(n,k)$ is obtained by employing a whitening filter, $H_w(z)$ to $y(n,k)$. With estimated $\{b_k\}$, $H_w(z)$ is constructed as [24]:

$$H_w(z) = 1 + \sum_{k=1}^{q} b_k z^{-k}. \tag{13}$$

B. ResNet20 for Noise Waveform Estimation

Fig. 2 shows the architecture of the proposed ResNet20 for noise waveform estimation. Motivated by the Resnet50 (containing 50 layers) [25], we propose a reduced version, namely the ResNet20 (containing 20 layers) model. It is due to the ResNet50 [25] was introduced for image recognition, where a stack of 50 2-dimensional convolutional (Conv2D) layers-based deep learning technique improved the accuracy of recognition. However, the deep architecture of a network varies over applications. We investigate and find that the reduced ResNet model, i.e., ResNet20 to be effective in estimating the noise waveform from the noisy speech waveform on a framewise basis. Instead of Conv2D layer in ResNet50 [25], the proposed ResNet20 is constructed with the 1-dimensional convolution (Conv1D) layer, since the target is to process the 1D speech signal. It reduces the number of training parameters, which minimizes the training time accordingly. In addition, we have used the causal Conv1D layer [26]. Fig. 3 demonstrates the operating principle of the standard and causal Conv1D layers. The standard Conv1D layers (Fig. 3 (a)) are comprised of filters that capture the local correlation of nearby data points, thus leaking the future information into the current data during operating. Conversely, in the causal Conv1D layer (Fig. 3 (b)), the output at any time step t only uses the information from the previous time steps, i.e., 0 to $t-1$ [26]. It allows the ResNet20 for real-time noise waveform estimation.

The proposed ResNet20-based method takes the noisy speech waveform, $\boldsymbol{y}_l = \{y(0,l), y(1,l), \dots, y(M-1,l)\}$ as input, yielding an estimate of the noise waveform, $\hat{\boldsymbol{v}}_l = \{\hat{v}(0,l), \hat{v}(1,l), \dots, \hat{v}(M-1,l)\}$. Specifically, $\boldsymbol{y}_l$ is passed through the input layer, which is a fully-connected layer of size 512, followed by the layer normalization (LN) [27] and SELU activation [28] layer. Reason of using SELU activation is that it has less impact on vanishing gradients than that of ReLU

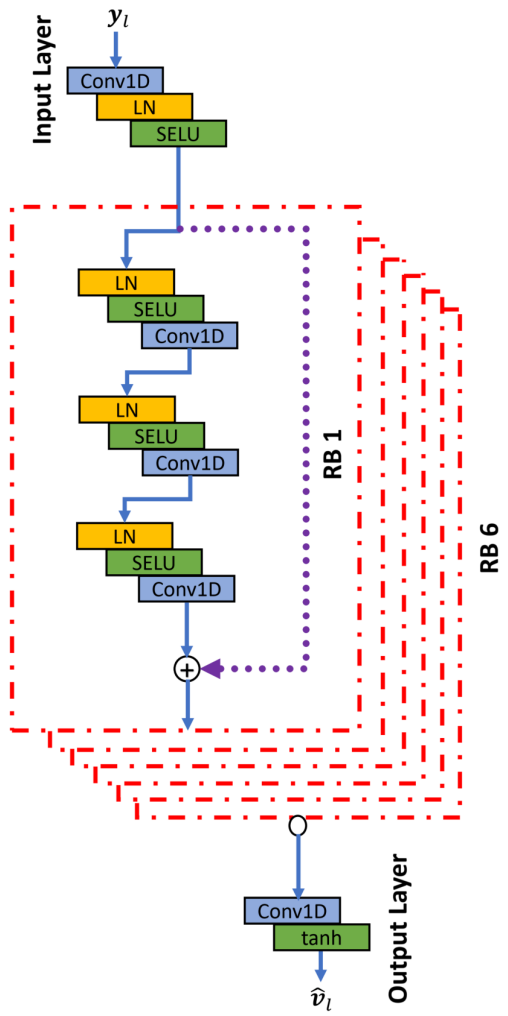

Fig. 2. Architecture of the proposed ResNet20 for noise waveform estimation.

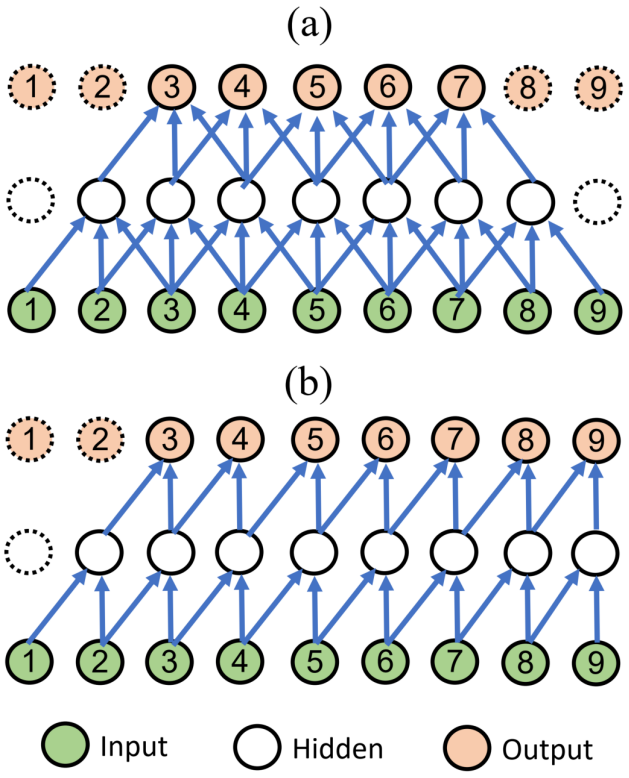

Fig. 3. One-dimensional CNN structure with (a) standard convolution and (b) causal convolution.

[29] and ELU [30]. Also, SELUs itself learn faster and better than ReLU and ELU even if they are combined with layer normalization [28]. The input layer is followed by 6 bottleneck residual blocks (RBs). Each RB contains 3 Conv1D layers. Each of the Conv1D layers is pre-activated by LN followed by SELU activation function. The output size of the first and the second Conv1D layer is 64, while the third one is 512. In addition, the first and third Conv1D layer has the kernel size of 1, whilst the second Conv1D layer has the kernel size of 3. Therefore, the first Conv1D layer in each RB compresses the input to a lower-dimensional embedding. The last RB (6^{th}) is followed by the output layer, which is a fully-connected layer (output size 512) with tanh units [31].

The stack of six RBs containing 18 Conv1D layers in the proposed ResNet20 exhibits a deep architecture. It is observed that the Conv1D layers in the lower RBs (close to the input layer), the gradients calculated from the backpropagated error signals of the Conv1D layers in the higher RBs, become progressively smaller or vanishing. It is referred to as the vanishing gradient problem [32]. Due to the vanishing gradients, connection weights at Conv1D layers in the lower RBs are not modified much, which reduces the learning capability during training. As long as the ResNet20 goes deeper, its performance gets saturated or even starts degrading rapidly. To alleviate this problem, a skip connection mechanism has been introduced in [25]. To improve the flow of information and gradients throughout the proposed ResNet20, we also utilize skip connections between the input and out layers of the RBs. The skip connection is represented by dotted line (Fig. 2). It can be seen that the skip connection bypass the output of each RB and added to the output of the next RB. To facilitate the skip-connection, the output size of the third Conv1D layer in each RB is set to 512. The skip-connection does not add any extra parameter or computational complexity. Rather, it acts as an identity mapping of the ResNet20 model, which ensures that the Conv1D layers in the higher RBs will perform as good as the Conv1D layers in the lower RBs.

IV. Speech Enhancement Experiment

A. Training Set

For training the proposed ResNet20, a total of $30,000$ clean speech recordings are randomly selected belonging to the *train-clean-100* set of the Librispeech corpus [33], the CSTR VCTK corpus [34], and the si^* and sx^* training sets of

the TIMIT corpus [35]. Among the 5% of $30,000$, i.e., 1500 speech recordings are randomly selected for cross-validation of the ResNet20 accuracy during training. That means, $28,500$ speech recordings are used for training of the ResNet20. On the other hand, a total of 500 noise recordings are randomly selected from the QUT-NOISE dataset [36], the Nonspeech dataset [37], the Environmental Background Noise dataset [38], [39], the noise set from the MUSAN corpus [40]. In addition, the 5% of 500, i.e., 25 noise recordings are selected for cross-validation purposes, while the remaining 475 of them are used for training. All the clean speech and noise recordings are single-channel with a sampling frequency of 16 kHz.

B. Training Strategy

The following training strategy was employed to train the proposed ResNet20 for noise waveform estimation:

- The widely used 'mean square error' is used as the loss function during training.
- The *Adam* algorithm [41] with default hyperparameters is also adopted for the gradient descent optimisation.
- Gradients are clipped between $[-1, 1]$.
- A total of 120 epochs are used to train the ResNet20.
- The number of training examples in an epoch is equal to the number of clean speech recordings used in the training set, i.e., $28,500$.
- A mini-batch size of 1 noisy speech signal is used.
- The noisy speech signals are generated as follows: each randomly selected clean speech recording (without replacement) is corrupted with a randomly selected noise recording (without replacement) at a randomly selected SNR level (-10 to +20 dB, in 1 dB increments).

C. Test Set

For objective experiments, 30 clean speech utterances belonging to six speakers (3 male and 3 female) are taken from the NOIZEUS corpus. The speech recordings are sampled at 16 kHz [42, Chapter 12]. We generate a noisy speech data set by corrupting the speech recordings with (*traffic*) and (*restaurant*) noise recordings selected from the noise database used in [38], [39] at multiple SNR levels varying from -5dB to +15 dB, in 5 dB increments. It is also important to note that the speech and the noise recordings are unseen and not used in training the proposed ResNet20 method.

D. Evaluation Metrics

The objective quality and intelligibility evaluation are carried out through the perceptual evaluation of speech quality (PESQ) [43] and quasi-stationary speech transmission index (QSTI) [44] measures. We also analyze the spectrograms of the enhanced speech produced by the proposed and benchmark SEAs to quantify the level of residual noise and distortion.

The subjective evaluation was carried out through blind AB listening tests [45, Section 3.3.4]. It is conducted on the utterance sp05 ("*Wipe the grease off his dirty face*") corrupted with 5 dB *traffic* noise. The enhanced speech produced by five SEAs as well as the corresponding clean and noisy speech recordings, a total of 42 stimuli pairs played in a random order to each listener, excluding the comparisons between the same method. For each stimuli pair, the listener prefers the first or second stimuli which is perceptually better, or a third response indicating no difference is found between them. A 100% award is given to the preferred method, 0% to the other, and 50% to each method for the similar preference response. Participants could re-listen to stimuli if required. Five English speaking listeners participate in the AB listening tests. The average of the preference scores given by the listeners, termed as the mean preference score (%).

The performance of the proposed method is carried out by comparing it with the benchmark methods, such as raw-waveform processing using FCNN (RWF-FCN) method [21], phase-aware DNN (IAM+IFD) method [22], deep learning-based KF (DNN-KF) method [23], AKF-Oracle method (where $(\{a_i\}, \sigma_w^2)$ and $(\{b_k\}, \sigma_u^2)$ are computed from the clean speech and noise signal) and no-enhancement (Noisy).

E. Results and Discussion

Fig. 4 (a)-(b) demonstrates that the proposed SEA consistently shows improved PESQ score over the benchmark SEAs, except the AKF-Oracle method for all test noise conditions as well as the SNR levels. The IAM+IFD method [22] relatively exhibits better PESQ score among the benchmark methods across the noise experiments. The no-enhancement (Noisy) shows the worse PESQ score in any condition.

Fig. 4 (c)-(d) also shows that the proposed method demonstrates a consistent QSTI score improvement across the noise experiments as well as the SNR levels, apart from the AKF-Oracle method. The existing IAM+IFD method [22] is found to be competitive with the proposed method typically at low SNR levels. However, at high SNR levels, all SEAs, even the no-enhancement (Noisy) case relatively shows competitive QSTI scores across the noise conditions.

It can be seen that the enhanced speech produced by the proposed SEA (Fig. 5 (f)) exhibits significantly less residual noise than that of the benchmark SEAs (Fig. 5 (c)-(e)) and is closely similar to the AKF-Oracle method (Fig. 5 (g)). When going from RWF-FCN method [21] to IAM+IFD method [22] (Fig. 5 (c)-(e)), noise-flooring is seen decreasing. The informal listening tests conducted on the enhanced speech also confirm that the benchmark SEAs relatively produce annoying sound as compared to negligible audio artifacts by the proposed method.

The outcome of AB listening tests in terms of mean preference score (%) is shown in Fig. 6. It can be seen that the enhanced speech produced by the proposed SEA is widely preferred by the listeners (around 72%) than the benchmark methods, apart from the AKF-Oracle method (around 81%) and clean speech signal (100%). The IAM+IFD method [22] is found to be the best preferred (60%) amongst the benchmark methods, followed by the DNN-KF method [23] (48%), and RWF-FCN method [21] (31%).

V. Conclusion

This paper introduced a deep residual network-based augmented Kalman filter for speech enhancement in various

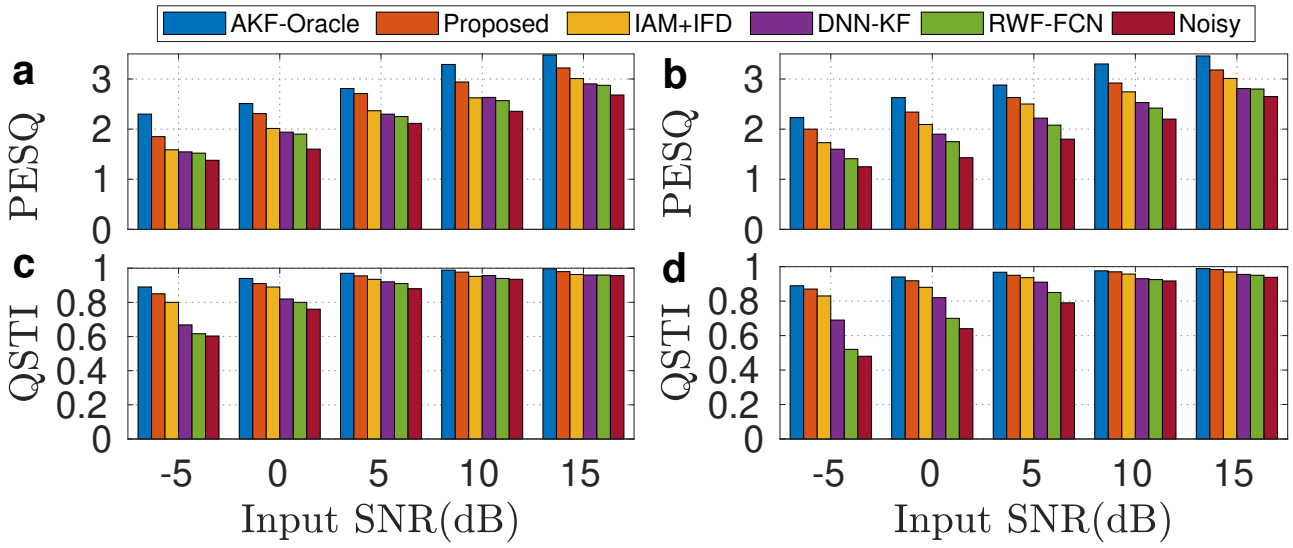

Fig. 4. Performance comparison of the proposed SEA with the benchmark SEAs in terms of the average: PESQ; (a) *traffic*, (b) *restaurant* and QSTI; (c) *traffic*, (d) *restaurant* noise conditions.

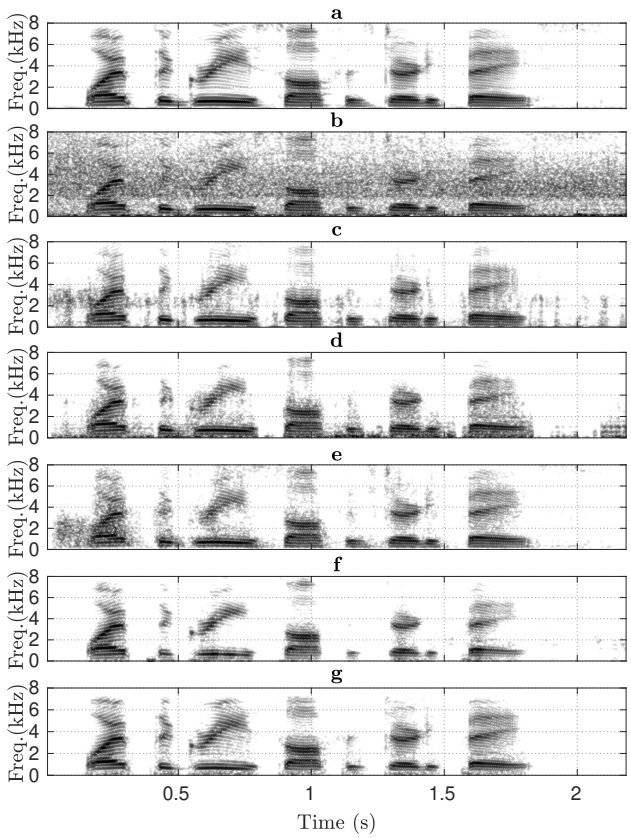

Fig. 5. Comparing the spectrograms of: (a) clean speech, (b) noisy speech (corrupt sp05 with 5 dB *traffic* noise), to that of the enhanced speech produced by the: (c) RWF-FCN [21], (d) DNN-KF [23], (e) IAM+IFD [22], (f) proposed, and (g) AKF-Oracle methods.

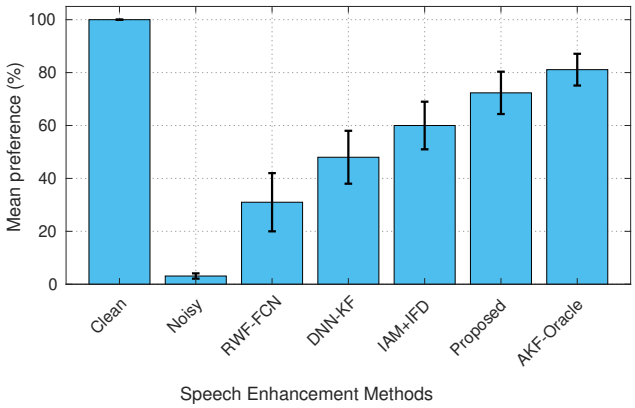

Fig. 6. The mean preference score (%) for each SEA on sp05 corrupted with 5 dB *traffic* noise.

noise conditions. Specifically, the proposed ResNet20 gives an estimate of the instantaneous noise waveform for each noisy speech frame. The noise LPC parameters are computed from the estimated noise. Each noisy speech frame is pre-whitened by a whitening filter, which is constructed with the corresponding noise LPCs. The speech LPC parameters are computed from the pre-whitened speech. Since the ResNet20 is trained with a large training set, it is capable to accurately estimate the speech and the noise LPC parameters in various noise conditions. The AKF constructed with the improved speech and noise LPC parameters is capable to minimize residual noise and distortion in the enhanced speech. Extensive objective and subjective testing on NOIZEUS corpus reveal that the proposed method outperforms some benchmark methods in various noise conditions for a wide range of SNR levels.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, April 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 208–211, April 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, December 1984.

[4] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, April 1985.

[5] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 629–632, May 1996.

[6] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, "A two-step noise reduction technique," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 289–292, May 2004.

[7] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 12, pp. 177–180, April 1987.

[8] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Transactions on Signal Processing*, vol. 39, no. 8, pp. 1732–1742, August 1991.

[9] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[10] S. K. Roy, A. Nicolson, and K. K. Paliwal, "Deep learning with augmented Kalman filter for single-channel speech enhancement," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2020, pp. 1–5.

[11] S. K. Roy, P. Nicolson, and K. K. Paliwal, "A deep learning-based Kalman filter for speech enhancement," *prof. of Interspeech2020*, October 2020.

[12] S. K. Roy and K. K. Paliwal, "Causal convolutional encoder decoder-based augmented Kalman filter for speech enhancement," *14th International Conference on Signal Processing and Communication Systems (ICSPCS) 2020*, December 2020.

[13] S. K. Roy, W. P. Zhu, and B. Champagne, "Single channel speech enhancement using subband iterative Kalman filter," *IEEE International Symposium on Circuits and Systems*, pp. 762–765, May 2016.

[14] A. E. W. George, S. So, R. Ghosh, and K. K. Paliwal, "Robustness metric-based tuning of the augmented Kalman filter for the enhancement of speech corrupted with coloured noise," *Speech Communication*, vol. 105, pp. 62 – 76, December 2018.

[15] S. K. Roy and K. K. Paliwal, "Sensitivity metric-based tuning of the augmented Kalman filter for speech enhancement," *14th International Conference on Signal Processing and Communication Systems (ICSPCS) 2020*, December 2020.

[16] J. Rouat, "Computational auditory scene analysis: Principles, algorithms, and applications (wang, d. and brown, g.j., eds.; 2006) [book review]," *IEEE Transactions on Neural Networks*, vol. 19, no. 1, pp. 199–199, 2008.

[17] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[18] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[19] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, p. 465–494, Apr. 2011. [Online]. Available: https://doi.org/10.1016/j.specom.2010.12.003

[20] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[21] S. Fu, T. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[22] N. Zheng and X. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 63–76, 2019.

[23] H. Yu, Z. Ouyang, W. Zhu, B. Champagne, and Y. Ji, "A deep neural network based Kalman filter for time domain speech enhancement," *IEEE International Symposium on Circuits and Systems*, pp. 1–5, May 2019.

[24] S. V. Vaseghi, "Linear prediction models," in *Advanced Digital Signal Processing and Noise Reduction*. John Wiley & Sons, 2009, ch. 8, pp. 227–262.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[26] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. van den Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," 2016.

[27] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[28] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," 2017.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015.

[30] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," 2015.

[31] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," 2018.

[32] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, 2018.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, April 2015.

[34] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[35] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, Feb. 1993.

[36] 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 *Proceedings Interspeech 2010*, 2010, pp. 3110–3113.

[37] G. Hu, "100 nonspeech environmental sounds," *The Ohio State University, Department of Computer Science and Engineering*, 2004.

[38] F. Saki, A. Sehgal, I. Panahi, and N. Kehtarnavaz, "Smartphone-based real-time classification of noise signals using subband features and random forest classifier," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 2204–2208.

[39] F. Saki and N. Kehtarnavaz, "Automatic switching between noise classification and speech enhancement for hearing aid devices," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2016, pp. 736–739.

[40] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015. [Online]. Available: http://arxiv.org/abs/1510.08484

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[42] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[43] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 749–752, May 2001.

[44] B. Schwerin and K. K. Paliwal, "An improved speech transmission index for intelligibility prediction," *Speech Communication*, vol. 65, pp. 9–19, December 2014.

[45] K. K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450–475, May 2010.