Causal Convolutional Neural Network-Based Kalman Filter for Speech Enhancement

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

***Abstract*—Speech enhancement using Kalman filter (KF) suffers from inaccurate estimates of the noise variance and the linear prediction coefficients (LPCs) in real-life noise conditions. This causes a degraded speech enhancement performance. In this paper, a causal convolutional neural network (CCNN) model is used to more accurately estimate the noise variance and LPC parameters of the KF for speech enhancement in real-life noise conditions. Specifically, a CCNN model gives an instantaneous estimate of the noise waveform for each noisy speech frame to compute the noise variance. Each noisy speech frame is pre-whitened by a whitening filter, which is constructed with the coefficients computed from the estimated noise. The LPC parameters are computed from the pre-whitened speech. The improved noise variance and LPCs enables the KF to minimize residual noise as well as distortion in the enhanced speech. Objective and subjective testing on NOIZEUS corpus reveal that the enhanced speech produced by the proposed method exhibits higher quality and intelligibility than some benchmark methods in various noise conditions for a wide range of SNR levels.**

***Index Terms*—Speech enhancement, Kalman filter, convolutional neural network, LPC, whitening filter.**

I. Introduction

A speech enhancement algorithm (SEA) aims to eliminate the embedded noises from the noisy speech signal so that the quality and intelligibility of speech are improved. The SEAs can be used as a pre-processor for many applications, such as voice communication systems, hearing-aid devices, speech recognition. Various SEAs, such as spectral subtraction (SS) [1], MMSE [2], [3], Wiener Filter (WF) [4], Kalman filter (KF) [5], deep neural network (DNN) [6] have been introduced over the decades. This paper focuses on single-channel speech enhancement using deep learning-based KF.

Paliwal and Basu for the first time used KF for speech enhancement in stationary noise condition [5]. In KF, the clean speech signal is represented by an autoregressive (AR) model and incorporated in the Kalman recursive equations. KF gives a linear MMSE estimate of the clean speech given the noisy speech for each sample within a frame. Thus, the performance of KF-based SEA somehow depends on how accurately the key parameters, LPCs are estimated in noisy conditions. In [5], it was demonstrated that the KF showed excellent performance when the LPC parameters were computed from the clean speech. On the other hand, the LPC parameters computed from the noisy speech are inaccurate and degrades KF performance significantly. In [7], Roy *et al.* proposed a sub-band iterative KF-based SEA. Due to processing the high-frequency sub-bands (SBs) among the 16 decomposed SBs for a given noise corrupted utterance, some noise components may still remain in the low-frequency SBs. The enhanced speech also suffers from distortion. In [8], So *et al.* introduced a robustness metric-based tuning of the KF for enhancing stationary noise corrupted speech. However, the robustness metric-based tuning of the KF gain causes distortion in the enhanced speech. To cope with this problem, a sensitivity metric-based tuning of the KF has been proposed [9]. Although it minimizes distortion in the enhanced speech, however, not applicable in real-life noise conditions.

The deep neural network (DNN) has been used as a forefront method for speech enhancement [6]. In [6], the DNN gives an estimate of the ideal binary mask (IBM), which is used to estimate the clean speech spectrum from the noisy speech spectrum. Later on, the ideal ratio mask (IRM) [10] shows better speech quality than the IBM. However, the enhanced speech produced by IBM and IRM-based methods [6], [10] are affected by phase. To address this, Williamson et al. introduced a complex ideal ratio mask (cIRM), which is capable to recover the amplitude and the phase spectrum of the clean speech [11]. In general, the masking technique introduces musical noise in the enhanced speech [10].

In [12], Fu *et al.* proposed a fully convolutional neural network (FCNN)-based SEA. It processes the noisy speech waveform, yielding an estimate of the clean speech waveform. Therefore, the enhanced speech does not affected by phase, unlike the acoustic-domain SEAs [6], [10]. In [13], Zheng et al. introduced a phase-ware SEA using DNN. In this method, the phase information (converted to the instantaneous frequency deviation (IFD)) is jointly used with different masks, namely ideal amplitude mask (IAM) as a training target. The clean speech spectrum is estimated with the estimated mask and the phase information (extracted from the IFD). Yu *et al.* introduced a KF-based SEA, where the LPCs are estimated using a traditional DNN [14]. However, the noise covariance is estimated during speech pauses of the noisy speech, which is irrespective in non-stationary noise conditions.

The direct estimation of speech spectrum using deep learning methods reported in literature may suffer from musical noise and distortion. We observe that 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 KF-based SEA suffering from inaccurate estimates of the noise variance in practice. In this paper, a causal convolutional neural network (CCNN) model addresses the noise variance and LPC parameter estimates of the KF for speech enhancement. Specifically, the CCNN model gives an instantaneous estimate of the noise waveform to compute the noise variance for each noisy speech frame. A whitening filter is then constructed with the coefficients computed from the estimated noise to pre-whiten each noisy speech frame prior to estimate the LPC parameters. With the improvement of noise variance and LPC parameters, the KF is found to be effective in minimizing residual noise as well as distortion in the enhanced speech. The efficiency of the proposed SEA is compared against some benchmark SEAs using objective and subjective testing on NOIZEUS corpus.

II. KF for Speech Enhancement

Assuming the 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:

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

$s(n)$ can be represented with p^{th} order AR model as [15]:

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

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

Eqs. (1)-(2) can be used to form the following state-space model (SSM) of the KF, as in [5]:

$$\boldsymbol{s}(n) = \boldsymbol{\Phi}\boldsymbol{s}(n-1) + \boldsymbol{d}w(n), \tag{3}$$

$$y(n) = \boldsymbol{c}^\top \boldsymbol{s}(n) + v(n). \tag{4}$$

The SSM is comprised of the following:

1) $\boldsymbol{s}(n)$ is a $p \times 1$ state vector at sample n, represented as:

$$\boldsymbol{s}(n) = [s(n) \quad s(n-1) \quad \ldots \quad s(n-p+1)]^\top, \tag{5}$$

2) $\boldsymbol{\Phi}$ is a $p \times p$ state transition matrix that relates the process states at sample n and $n-1$, represented as:

$$\boldsymbol{\Phi} = \begin{bmatrix} -a_1 & -a_2 & \ldots & a_{p-1} & a_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \tag{6}$$

3) $\boldsymbol{d}$ and $\boldsymbol{c}$ are the $p \times 1$ measurement vectors for the excitation noise and observation, represented as:

$$\boldsymbol{d} = \boldsymbol{c} = \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}^T,$$

4) $y(n)$ represents the noisy observation at sample n.

Firstly, $y(n)$ is windowed into non-overlapped and short (e.g., 20 ms) frames. For a particular frame, the KF computes an unbiased linear MMSE estimate $\hat{\boldsymbol{s}}(n|n)$ at sample n, given $y(n)$, by using the following recursive equations [5]:

$$\hat{\boldsymbol{s}}(n|n-1) = \boldsymbol{\Phi}\hat{\boldsymbol{s}}(n-1|n-1), \tag{7}$$

$$\boldsymbol{\Psi}(n|n-1) = \boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^\top + \sigma_w^2 \boldsymbol{d}\boldsymbol{d}^\top, \tag{8}$$

$$\boldsymbol{K}(n) = \boldsymbol{\Psi}(n|n-1)\boldsymbol{c}(\boldsymbol{c}^\top \boldsymbol{\Psi}(n|n-1)\boldsymbol{c} + \sigma_v^2)^{-1}, \tag{9}$$

$$\hat{\boldsymbol{s}}(n|n) = \hat{\boldsymbol{s}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^\top \hat{\boldsymbol{s}}(n|n-1)], \tag{10}$$

$$\boldsymbol{\Psi}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^\top]\boldsymbol{\Psi}(n|n-1). \tag{11}$$

For a noisy speech frame, the error covariances ($\boldsymbol{\Psi}(n|n-1)$ and $\boldsymbol{\Psi}(n|n)$ corresponding to $\hat{\boldsymbol{s}}(n|n-1)$ and $\hat{\boldsymbol{s}}(n|n)$) and the Kalman gain $\boldsymbol{K}(n)$ are continually updated on a samplewise basis, while σ_v^2 and ($\{a_i\}$, σ_w^2) remain constant. At sample n, $\boldsymbol{c}^\top \hat{\boldsymbol{s}}(n|n)$ gives the estimated speech, $\hat{s}(n|n)$, as in [9]:

$$\hat{s}(n|n) = [1 - K_0(n)]\hat{s}(n|n-1) + K_0(n)y(n), \tag{12}$$

where $K_0(n)$ is the 1^{st} component of $\boldsymbol{K}(n)$ given by [9]:

$$K_0(n) = \frac{\alpha^2(n) + \sigma_w^2}{\alpha^2(n) + \sigma_w^2 + \sigma_v^2}, \tag{13}$$

where $\alpha^2(n) = \boldsymbol{c}^\top \boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^\top \boldsymbol{c}$ is the transmission of *a posteriori* error variance by the speech AR model from the previous sample, $n-1$ [9].

Eq. (12) implies that $K_0(n)$ has a significant impact on $\hat{s}(n|n)$ estimates, which is the output of the KF. In practice, the inaccurate estimates of σ_u^2 and ($\{a_i\}$, σ_w^2) introduce bias in $K_0(n)$, which degrades $\hat{s}(n|n)$ estimates. In the proposed SEA, CCNN model is used to accurately estimate σ_u^2 and ($\{a_i\}$, σ_w^2), leading to a more accurate $\hat{s}(n|n)$.

III. Proposed Speech Enhancement System

Fig. 1 shows the block diagram of the proposed SEA. Unlike the KF method in section II, a 32 ms rectangular window with 50% overlap was considered for converting $y(n)$ (eq. (1)) into frames $y(n, l)$, i.e., $y(n, l) = s(n, l) + v(n, l)$, where $l \epsilon \{0, 1, 2, \ldots, N-1\}$ 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., $n \epsilon \{0, 1, 2, \ldots, M-1\}$.

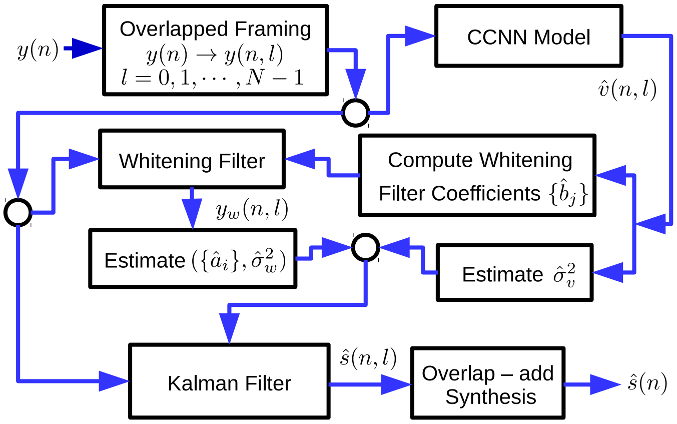

Fig. 1. Block diagram of the proposed KF-based SEA.

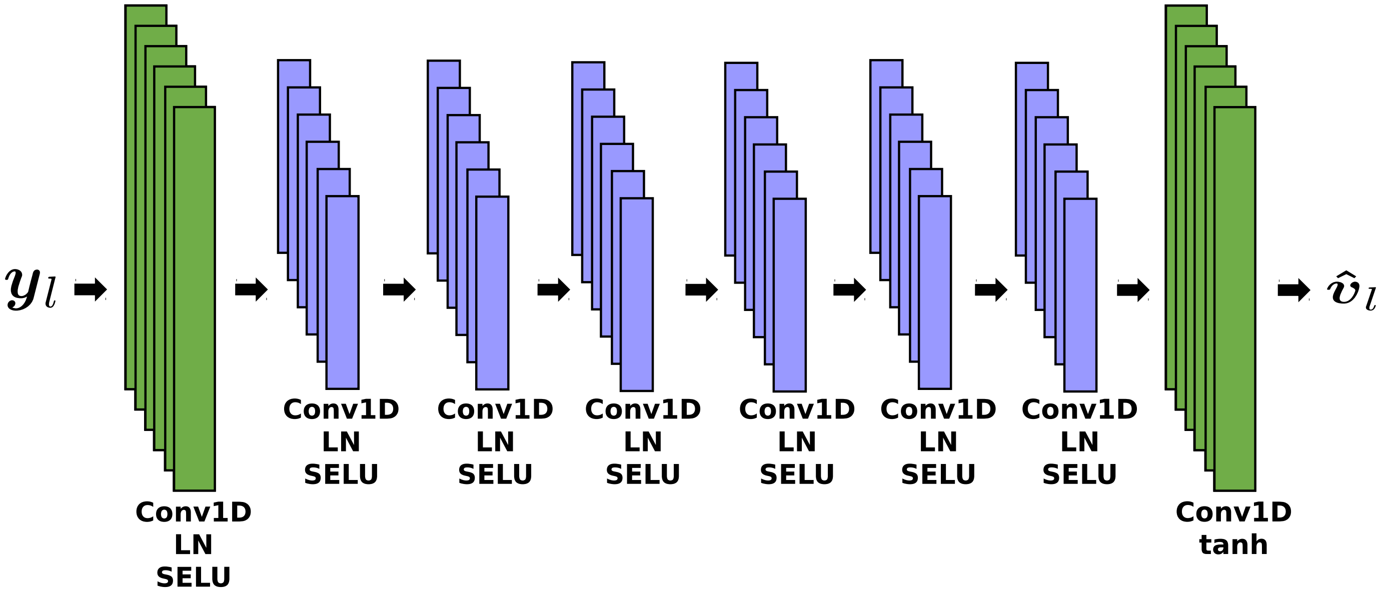

Fig. 2. Architecture of the proposed CCNN model for noise waveform estimation.

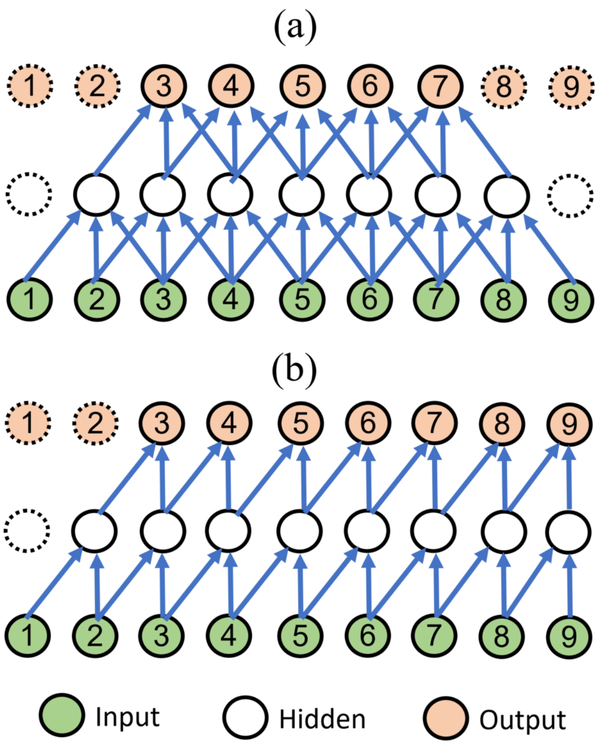

Fig. 3. Working principle of: (a) standard and (b) causal Conv1D layer.

A. Proposed $\hat{\sigma}_v^2$ and ($\{\hat{a}_i\}$, $\hat{\sigma}_w^2$) Estimation Method

The LPC parameters, ($\{a_i\}$, σ_w^2) are very sensitive to noise. Since the clean speech, $s(n, l)$ is not available in practice, it is difficult to estimate these parameters accurately. Therefore, we first focus on noise variance estimation. For this purpose, we introduce a CCNN model (described in section III-B) to estimate the noise waveform, $\hat{v}(l, n)$ for each noisy speech frame, $y(n, l)$. $\hat{\sigma}_v^2$ is then computed from $\hat{v}(n, l)$ as:

$$\hat{\sigma}_v^2 = \frac{1}{M}\sum_{n=0}^{M-1} \hat{v}^2(n, l). \tag{14}$$

To reduce bias in the estimated ($\{\hat{a}_i\}$, $\hat{\sigma}_w^2$) for each noisy speech frame, $y(n, l)$, we compute them from the corresponding pre-whitened speech, $y_w(n, l)$ using the autocorrelation method [15]. $y_w(n, l)$ is obtained by employing a whitening filter, $H_w(z)$ to $y(n, l)$. $H_w(z)$ is constructed as [15]:

$$H_w(z) = 1 + \sum_{k=1}^{q} b_k z^{-k}, \tag{15}$$

where the coefficients, ($\{\hat{b}_k\}; q = 20$) are computed from $\hat{v}(n, l)$ using the autocorrelation method [15].

B. CCNN Model for Noise Waveform Estimation

The proposed CCNN model for $\hat{v}(l, n)$ estimation is shown in Fig. 2. It consists of a stack of eight 1-dimensional convolutional (Conv1D) layers. Unlike 2-dimensional convolutional layers (Conv2D), we have used Conv1D, since it is appropriate to process the 1D speech signal. Due to using Conv1D layer, the proposed CCNN model reduces huge training parameter as well as training time than that of Conv2D layer. In addition, we have used the causal Conv1D layer [16]. 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 future information is leaking into the current data during operating. Conversely, the output of the causal Conv1D layer (Fig. 3 (b)) at any time step t only uses the information from the previous time steps, i.e., 0 to $t-1$ [16]. It allows the CCNN model for real-time noise waveform estimation.

The proposed CCNN model maps each frame of the given noisy speech waveform, $\boldsymbol{y}_l = \{y(0,l),\, y(1,l),\, \ldots,\, y(M-1,l)\}$ to that of the instantaneous noise waveform, $\hat{\boldsymbol{v}}_l = \{\hat{v}(0,l),\, \hat{v}(1,l),\, \ldots,\, \hat{v}(M-1,l)\}$. Specifically, $\boldsymbol{y}_l$ is passed through the first fully connected Conv1D layer, which is the input layer. The output size and kernel size of the input layer are 512 and 1, respectively. The input layer is followed by the layer normalization (LN) [17] and SELU activation [18] layer. We have used SELU activation function since it has less impact on vanishing gradients than that of ReLU [19] and ELU [20]. Also, it has faster and better learning capability than ReLU and ELU even if it is combined with layer normalization [18]. The input layer is followed by 6 Conv1D layers of output size 64 and kernel size 3. Therefore, the middle six Conv1D layers encode the features into lower dimension. The reduced output size in the Conv1D layers decreases the training time of the CCNN model. Each of these six Conv1D layers is followed by LN and SELU activation function. The last Conv1D layer is also a fully connected layer of output size 512, and kernel size 1. It is the output layer of the proposed CCNN model. The output layer is followed by tanh activation function, since the mapped noise waveform ranged between -1 to +1.

IV. Speech Enhancement Experiment

A. Training Set

For training the proposed CCNN, a total of $30,000$ clean speech recordings are randomly selected belonging to the *train-clean-100* set of the Librispeech corpus [21], the CSTR VCTK corpus [22], and the si^* and sx^* training sets of the TIMIT corpus [23]. 5% of $30,000$, i.e., 1500 speech recordings are randomly selected for validating the training accuracy of the CCNN model. Thus, $28,500$ speech recordings are used for training of the CCNN model. Also, a total of 500 noise recordings are randomly selected from the QUT-NOISE dataset [24], the Nonspeech dataset [25], the Environmental Background Noise dataset [26], [27], the noise set from the MUSAN corpus [28]. 5% of 500, i.e., 25 noise recordings are selected for validation purposes, while the remaining 475 of them are used for training. All the clean speech and the noise recordings are sampled at 16 kHz.

B. Training Strategy

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

- The 'mean square error' is chosen as the loss function.
- The *Adam* algorithm [29] with default hyperparameters is also selected for gradient descent optimisation.
- Gradients are clipped between $[-1, 1]$.
- 120 epochs are used to train the CCNN model.
- The number of training examples in an epoch is equal to the number of clean speech recordings used in the training set $(28,500)$.
- 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 [30, Chapter 12]. We generate a noisy speech data set by corrupting the speech recordings with (*traffic*) and (*restaurant*) noise recordings selected from [26], [27] at multiple SNR levels varying from -5dB to +15 dB, in 5 dB increments. It is important to note that both the speech and the noise recordings are not used in training the CCNN model.

D. Evaluation Metrics

The objective quality and intelligibility evaluation are carried out through the perceptual evaluation of speech quality (PESQ) [31] and quasi-stationary speech transmission index (QSTI) [32] 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 blind AB listening test [33, Section 3.3.4] was adopted from subjective evaluation. This test was 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 the 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 pair, the listeners prefer the first or the second stimuli which is perceptually better, or a third response indicating no difference is found between them. The preferred method is given a 100% award, 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 take part 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 compared with the benchmark methods, such as noisy speech waveform processing using FCNN (RWF-FCN) method [12], phase-aware DNN (IAM+IFD) method [13], deep learning-based KF (DNN-KF) method [14], KF-Oracle method (where $(\{a_i\}, \sigma_w^2)$ and $(\{b_k\}, \sigma_u^2)$ are computed from the clean speech and the noise signal) and no-enhancement (Noisy).

E. Results and Discussion

Fig. 4 (a)-(b) demonstrates that the proposed SEA consistently shows improved PESQ scores over the benchmark SEAs, except the KF-Oracle method for all noise conditions as well as the SNR levels. The IAM+IFD method [13] relatively exhibits better PESQ score among the benchmark methods across the noise experiments. The no-enhancement (Noisy) shows the worse PESQ score in any case.

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 KF-Oracle method. The existing IAM+IFD method [13] is found

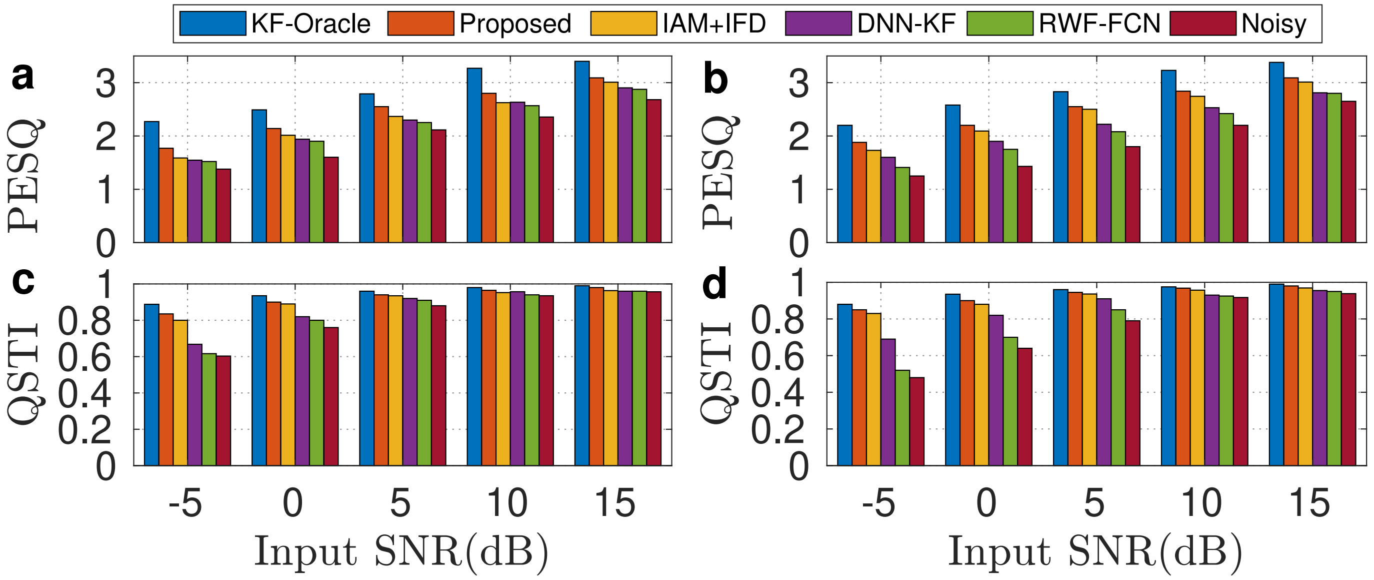

Fig. 4. Performance comparison of the proposed SEA with the benchmark SEAs in terms of the average: PESQ; (a) *traffic*, (b) *restaurant* and QSTI; (c) *traffic*, (d) *restaurant* noise conditions.

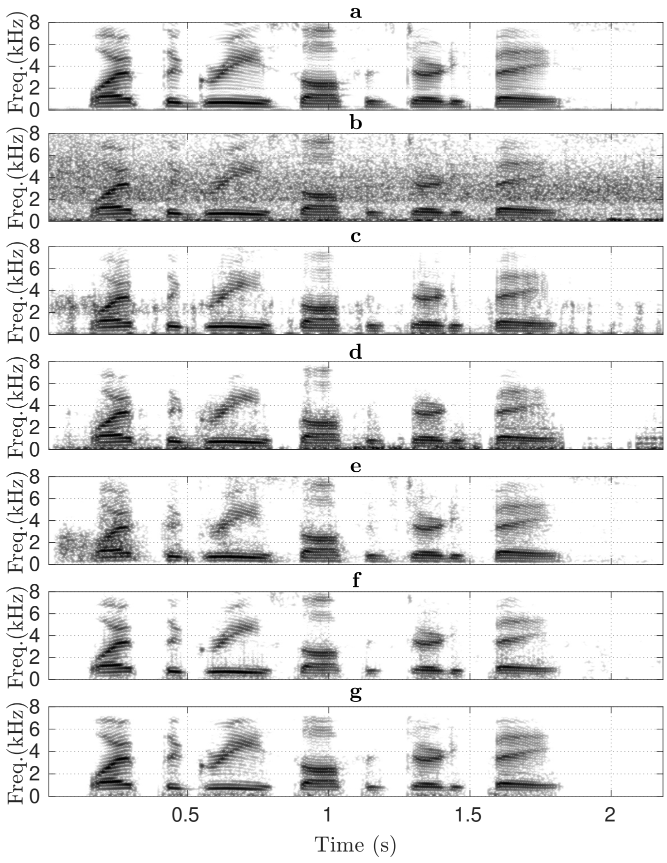

Fig. 5. (a) Comparing the spectrograms of: clean speech, (b) noisy speech (sp05 is corrupted with 5 dB traffic noise), to that of the enhanced speech produced by: (c) RWF-FCN [12], (d) DNN-KF [14], (e) IAM+IFD [13], (f) proposed, and (g) KF-Oracle methods.

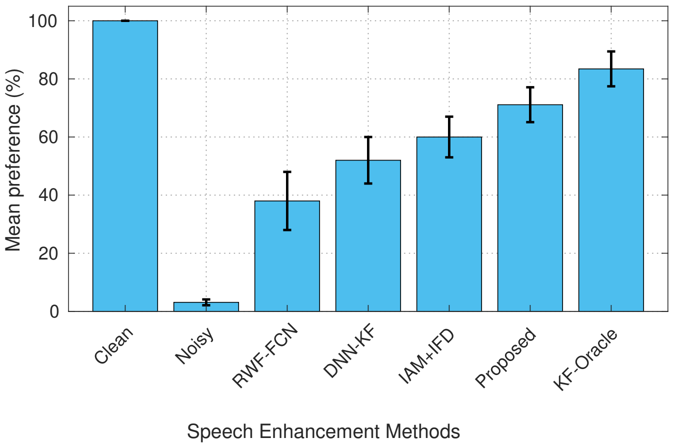

Fig. 6. The mean preference score (%) for each SEA on sp05 corrupted with 5 dB *traffic* noise.

to be competitive with the proposed method in QSTI improvement typically at low SNR levels. However, at high SNR levels, all the SEAs, even the no-enhancement (Noisy) case relatively shows a competitive QSTI score for all conditions.

The proposed SEA is also compared with the benchmark methods in terms of spectrogram analysis. It can be seen that the proposed SEA (Fig. 5 (f)) exhibits significantly less residual noise in the enhanced speech than that of the benchmark SEAs (Fig. 5 (c)-(e)) and is closely similar to the KF-Oracle method (Fig. 5 (g)). When going from RWF-FCN method [12] to IAM+IFD method [13] (Fig. 5 (c)-(e)), noise reduction 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 the blind 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 71%) than the benchmark methods, apart from the KF-Oracle method (around 83%) and clean speech signal (100%). It is due to accurate estimates of the noise variance and LPC parameters by the proposed CCNN model. Among the benchmark methods, the IAM+IFD method [13] is found to be the best preferred (60%), followed by the DNN-KF method [14] (52%), and RWF-FCN method [12] (38%).

V. Conclusion

This paper introduced a causal convolutional neural network-based Kalman filter for speech enhancement. Specifically, the proposed CCNN gives an instantaneous estimate of the noise waveform for each noisy speech frame to compute the noise variance. A whitening filter constructed with the coefficients computed from the estimated noise is employed to each noisy speech frame, yielding a pre-whitened speech. The LPC parameters are computed from the pre-whitened speech. The whitening filter basically reduces bias in the estimated LPC parameters. Since the CCNN is trained with a large training set, it is capable to accurately estimate the noise variance and the LPC parameters in various noise conditions. As a result, the KF constructed with the improved parameters minimizes residual noise as well as distortion in the enhanced speech. Extensive objective and subjective testing 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] 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.

[3] ——, "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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] S. So, A. E. W. George, R. Ghosh, and K. K. Paliwal, "A non-iterative Kalman filtering algorithm with dynamic gain adjustment for single-channel speech enhancement," *International Journal of Signal Processing Systems*, vol. 4, no. 4, pp. 263–268, August 2016.

[9] ——, "Kalman filter wih sensitivity tuning for improved noise reduction in speech," *Circuits, Systems, and Signal Processing*, vol. 36, no. 4, pp. 1476–1492, April 2017.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

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

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

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

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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

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

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

[31] 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.

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

[33] 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.