A Deep Learning-based Kalman Filter for Speech Enhancement

Sujan Kumar Roy, Aaron Nicolson, and Kuldip K. Paliwal

Signal Processing Laboratory, Griffith School of Engineering,
Griffith University, Brisbane, QLD, Australia, 4111

{sujankumar.roy, aaron.nicolson}@griffithuni.edu.au, k.paliwal@griffith.edu.au

Abstract

The existing Kalman filter (KF) suffers from poor estimates of the noise variance and the linear prediction coefficients (LPCs) in real-world noise conditions. This results in a degraded speech enhancement performance. In this paper, a deep learning approach is used to more accurately estimate the noise variance and LPCs, enabling the KF to enhance speech in various noise conditions. Specifically, a deep learning approach to MMSE-based noise power spectral density (PSD) estimation, called DeepMMSE, is used. The estimated noise PSD is used to compute the noise variance. We also construct a whitening filter with its coefficients computed from the estimated noise PSD. It is then applied to the noisy speech, yielding pre-whitened speech for computing the LPCs. The improved noise variance and LPC estimates enable the KF to minimise the *residual* noise and *distortion* in the enhanced speech. Experimental results show that the proposed method exhibits higher quality and intelligibility in the enhanced speech than the benchmark methods in various noise conditions for a wide-range of SNR levels.

Index Terms: Speech enhancement, Kalman filter, DeepMMSE, Deep Xi, noise PSD, LPC.

1. Introduction

The objective of a speech enhancement algorithm (SEA) is to eliminate the embedded noise from a noisy speech signal. It can be used as a front-end tool for many applications, such as voice communication systems, hearing-aid devices, and speech recognition. Various SEAs, namely spectral subtraction (SS) [1, 2], MMSE [3, 4], Wiener Filter (WF) [5, 6], and Kalman filter (KF) [7] have been introduced over the decades.

The SS method heavily depends on the accuracy of the noise PSD estimate [8]. The MMSE and WF-based SEAs completely rely upon the accurate estimation of the *a priori* SNR [9]. In [3], a decision-directed (DD) approach was proposed to estimate the *a priori* SNR. Since this approach uses the speech and noise power estimates from the previous frame, it is difficult to estimate the *a priori* SNR accurately for the current frame.

The efficiency of KF-based SEA depends on how accurately the noise variance and the LPCs are estimated. In [7], the LPCs are computed from the clean speech, which is unavailable in practice. It is also limited to enhancing speech corrupted with additive white Gaussian noise (AWGN). A sub-band iterative KF for enhancing speech in different noise conditions was proposed in [10]. The noisy speech is first decomposed into 16 sub-bands (SBs). An iterative KF is then employed to enhance the partially reconstructed high-frequency (HF) SBs. It is assumed that the low-frequency (LF) SBs are less affected by noise and are left unprocessed. The noise variance for the sub-band iterative KF is estimated using a derivative-based method.

Nowadays, deep neural networks (DNNs) are widely used for speech enhancement [11]. DNN-based SEAs typically give an estimate of a time-frequency mask, which is used to compute the spectrum of clean speech [11, 12]. A comparative study on six different masks has also been performed in [13] to identify an optimal mask for speech enhancement. However, the masking technique usually introduces residual and musical noise in the enhanced speech [13].

In [14], a fully convolutional neural network (FCNN) based SEA was introduced. This method is particularly designed to enhance babble noise corrupted speech. In [15], a raw waveform-based SEA using FCNN was proposed. Since the input/output of [15] is a raw waveform, the enhanced speech is not affected by the phase issues that are characteristic of magnitude spectrum-based SEAs [11, 13, 14]. Zheng et al. introduced a phase-aware DNN for speech enhancement [16]. Here, the phase information (converted to the instantaneous frequency deviation (IFD)) is jointly used with a time-frequency masks. The enhanced speech is reconstructed 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 deep neural network [17]. However, the noise covariance is estimated during speech pauses, which is not effective in non-stationary noise conditions. In addition, the silence detection process was unspecified.

In this paper, a deep learning technique is used to resolve the noise variance and the LPC estimation issues of the KF, leading to the capability of performing speech enhancement in various noise conditions. Firstly, the noise PSD is estimated using DeepMMSE [18], which is then used to compute the noise variance. We also construct a whitening filter with its coefficients computed from the estimated noise PSD. The LPCs are then computed from the pre-whitened speech, which is obtained by employing the whitening filter to the noisy speech signal. With the improved noise variance and LPCs, the KF is found to be effective at minimising the *residual* noise as well as *distortion* in the enhanced speech. The efficiency of the proposed method is evaluated against benchmark methods using objective and subjective testing.

2. KF for speech enhancement

At discrete-time sample n, the noisy speech, $y(n)$, can be represented as:

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

where $s(n)$ and $v(n)$ denote the clean speech, and uncorrelated additive noise, respectively. The clean speech can be modeled using a p^{th} order linear predictor, as in [19, Chapter 8]:

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

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 [7]:

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

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

The SSM is comprised of the following:

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

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

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

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

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

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

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 [7]:

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

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

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

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

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

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 [20]:

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

where $K_0(n)$ is the 1^{st} component of $\boldsymbol{K}(n)$ given by [20]:

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

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* mean squared error from the previous sample, $n-1$, to the total *a priori* mean prediction squared error [20].

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

3. Proposed speech enhancement system

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

The noisy speech, $y(n)$ is also analyzed frame-wise using the short-time Fourier transform (STFT):

$$Y(l,m) = S(l,m) + V(l,m), \tag{14}$$

where $Y(l,m)$, $S(l,m)$, and $V(l,m)$ denote the complex-valued STFT coefficients of the noisy speech, the clean speech, and the noise signal, respectively, for time-frame index l and discrete-frequency bin m.

It is assumed that $S(l,m)$ and $V(l,m)$ follow a Gaussian distribution with zero-mean and variances $E\{|S(l,m)|^2\} = \lambda_s(l,m)$, and $E\{|V(l,m)|^2\} = \lambda_v(l,m)$, where $E\{\cdot\}$ represents the statistical expectation operator.

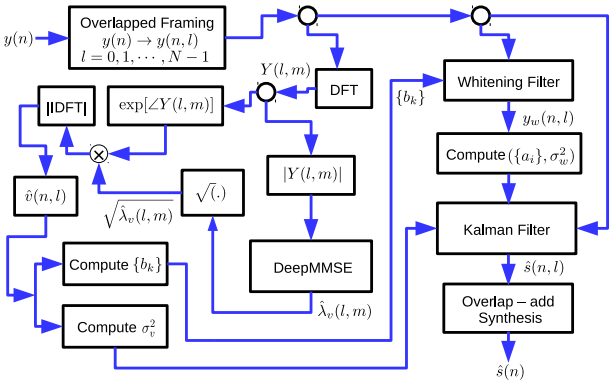


Figure 1: *Block diagram of the proposed SEA.*

3.1. Proposed σ_v^2 and ($\{a_i\}, \sigma_w^2$) estimation method

Firstly, the frame-wise noise PSD, $\hat{\lambda}_v(l,m)$ is estimated using DeepMMSE [18]. DeepMMSE is described in the following subsection. An estimate of noise, $\hat{v}(n,l)$ is given by taking the |IDFT| of $\sqrt{\hat{\lambda}_v(l,m)}\exp[\angle Y(l,m)]$. The noise variance, σ_v^2 is then computed from $\hat{v}(n,l)$ frame-wise as:

$$\sigma_v^2 = \frac{1}{M}\sum_{n=0}^{M-1} \hat{v}^2(n,l). \tag{15}$$

The LPC parameters, ($\{a_i\}$, σ_w^2) are sensitive to noise. We compute ($\{a_i\}$, σ_w^2) ($p = 10$) frame-wise from pre-whitened speech, $y_w(n,l)$, using the autocorrelation method [19]. $y_w(n,l)$ is used to reduce bias in ($\{a_i\}$, σ_w^2). Then, $y_w(n,l)$ is obtained by employing a whitening filter, $H_w(z)$ to $y(n,l)$. $H_w(z)$ is found, as in [19, section 8.1.7]:

$$H_w(z) = 1 + \sum_{k=1}^{q} b_k z^{-k}, \tag{16}$$

where the whitening filter coefficients, ($\{b_k\}; q = 40$) are computed from $\hat{v}(n,l)$ using the autocorrelation method [19].

3.2. DeepMMSE

DeepMMSE is an MMSE-based noise PSD estimator that employs the Deep Xi framework for *a priori* SNR estimation [18]. DeepMMSE does not exploit any underlying assumptions about the speech or noise, and produces a noise PSD estimate with negligible bias, unlike other MMSE-based noise PSD estimators [21, 22]. DeepMMSE includes the following four stages:

1. The *a priori* SNR estimate, $\xi(l,m)$, of $|Y(l,m)|$, is first found using Deep Xi-ResNet. Deep Xi-ResNet is described in the following subsection. The *a priori* SNR is defined as $\xi(l,m) = \frac{\lambda_s(l,m)}{\lambda_v(l,m)}$.
2. Next, the maximum-likelihood (ML) *a posteriori* SNR estimate is computed using the *a priori* SNR [23]: $\hat{\gamma}(l,m) = \xi(l,m) + 1$.
3. Using $\hat{\xi}$ and $\hat{\gamma}$, the noise periodogram estimate is found using the MMSE estimator [21, 22]: $|\hat{V}(l,m)|^2 = \left[\frac{1}{(1+\xi(l,m))^2} + \frac{\xi(l,m)}{(1+\xi(l,m))\gamma(l,m)}\right]|Y(l,m)|^2$.
4. The final noise PSD estimate, $\hat{\lambda}_v(l,m)$, is found by applying a first-order temporal recursive smoothing operation: $\hat{\lambda}_v(l,m) = \alpha\hat{\lambda}_d[l-1,k] + (1-\alpha)|\hat{V}(l,m)|^2$, where α is the smoothing factor. In this work, $\alpha = 0$ was used, i.e. the instantaneous noise power spectrum estimate from DeepMMSE was used.

3.3. Deep Xi-ResNet for $\xi(l,m)$ estimation

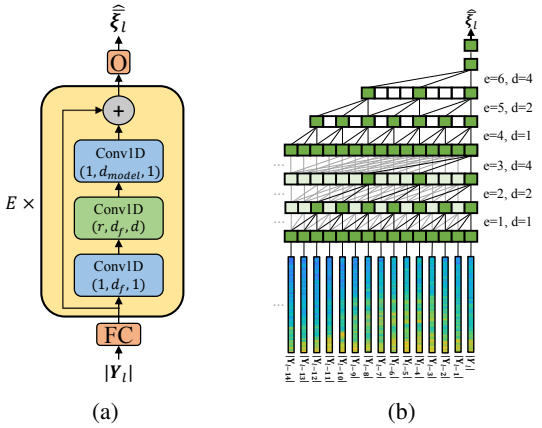
Deep Xi-ResNet is used to estimate $\xi(l,m)$ for DeepMMSE (available at https://github.com/anicolson/DeepXi). Deep Xi is a deep learning approach to *a priori* SNR estimation [24]. During training, the clean speech and noise of the noisy speech are available. This allows the instantaneous case of the *a priori* SNR to be used as the training target. To compute the instantaneous *a priori* SNR, $\lambda_s(l,m)$ and $\lambda_v(l,m)$ are replaced with the squared magnitude of the clean-speech and noise spectral components, respectively.

The observation and target of a training example for a deep neural network (DNN) in the Deep Xi framework is $|\mathbf{Y}_l|$ and the mapped *a priori* SNR, $\bar{\boldsymbol{\xi}}_l$, respectively. The mapped *a priori* SNR is a mapped version of the instantaneous *a priori* SNR. The instantaneous *a priori* SNR is mapped to the interval $[0,1]$ in order to improve the rate of convergence of the used stochastic gradient descent algorithm. The cumulative distribution function (CDF) of $\xi_{\mathrm{dB}}(l,m) = 10\log_{10}(\xi(l,m))$ is used as the map. As shown in [24, Fig. 2 (top)], the distribution of $\xi_{\mathrm{dB}}(l,m)$ for the k^{th} frequency component follows a normal distribution. It is thus assumed that $\xi_{\mathrm{dB}}(l,m)$ is distributed normally with mean μ_k and variance σ_k^2: $\xi_{\mathrm{dB}}(l,m) \sim \mathcal{N}(\mu_k, \sigma_k^2)$. Thus, the mapped *a priori* SNR is found by applying the normal CDF to $\xi_{\mathrm{dB}}(l,m)$:

$$\bar{\xi}(l,m) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\xi_{\mathrm{dB}}(l,m) - \mu_k}{\sigma_k\sqrt{2}}\right)\right], \qquad (17)$$

where μ_k and σ_k^2 found in [24] are used in this work. During inference, the *a priori* SNR estimate, $\hat{\xi}(l,m)$, is found from $\hat{\bar{\xi}}(l,m)$ using $\hat{\xi}(l,m) = 10^{\left(\left(\sigma_k\sqrt{2}\mathrm{erf}^{-1}(2\hat{\bar{\xi}}(l,m)-1)+\mu_k\right)/10\right)}$.

Deep Xi-ResNet utilises a residual network consisting of 1-D causal dilated convolutional units withing the Deep Xi framework [18], as shown in Fig. 2 (a). It consists of $E = 40$ bottleneck residual blocks, where $e\epsilon\{1,2,\ldots,E\}$ is the block index.



(a) (b)

Figure 2: *(a) Deep Xi-ResNet and (b) example of the contextual field of Deep Xi-ResNet with $D = 4$, $E = 6$, and $r = 3$.*

Each block contains three convolutional units (CUs) [25], where each CU is pre-activated by layer normalisation [26] followed by the ReLU activation function [27]. The 1^{st} and 3^{rd} CUs have a kernel size of $r = 1$ to that of $r = 3$ for the 2^{nd} CU. The 2^{nd} CU employs a dilation rate (DR) of d, providing a contextual field over previous time steps. As in [18], d is cycled as the block index e increases: $d = 2^{(e-1 \bmod (\log_2(D)+1)}$, where mod is the modulo operation, and D is the maximum DR. An example of how the DR is cycled is shown in Fig. 2 (b), with $D = 4$, and $E = 6$. It can be seen that the DR is reset after block three. This also demonstrates the contextual field gained by the use of causal dilated CUs. For Deep Xi-ResNet, D is set to 16. The 1^{st} and 2^{nd} CUs have an output size of $d_f = 64$ to that of $d_{model} = 256$ for the 3^{rd} CU. **FC** is a fully-connected layer with an output size of d_{model}, where layer normalisation is applied to the output of **FC**, followed by the ReLU activation function. The output layer **O** is a fully-connected layer with sigmoidal units.

4. Speech enhancement experiment

4.1. Training set

For training the ResNet, a total of 74250 clean speech recordings belonging to the *train-clean-100* set from the Librispeech corpus [28] (28539), the CSTR VCTK corpus [29] (42015), and the si^* and sx^* training sets from the TIMIT corpus [30] (3696) are used. 5% of the clean speech recordings are randomly selected and used as a validation set. Thus, 70537 clean speech recordings are used in the training set and 3713 in the validation set. The 2382 noise recordings adopted in [24] are used as the noise training set. All clean speech and noise recordings are single-channel, with a sampling frequency of 16 kHz.

4.2. Training strategy

The ResNet is trained using cross-entropy as the loss function and the *Adam* algorithm [31] with default hyper-parameters. The gradients are also clipped between $[-1,1]$. The selection order for the clean speech recordings is randomised for each epoch. 175 epochs are used to train the ResNet, where a mini-batch size of 10 noisy speech signals is used. The noisy signals are created as follows: each clean speech recording selected for the mini-batch is mixed with a random section of a randomly

selected noise recording at a randomly selected SNR level (-10 to 20 dB, in 1 dB increments).

4.3. Test set

For objective experiments, 30 utterances belonging to six speakers are taken from the NOIZEUS corpus and are sampled at 16 kHz [9, Chapter 12]. We generate a noisy data set that has been corrupted by the *passing car* and *café babble* noise recordings that were adopted in [24] at SNR levels from -5dB to 15dB, in 5 dB increments. Note that these clean speech and noise recordings are not used during training.

4.4. Evaluation metrics

The objective quality and intelligibility evaluation was carried out using the perceptual evaluation of speech quality (PESQ) [32] and quasi-stationary speech transmission index (QSTI) [33] measures. We also analyse the enhanced speech spectrograms of the SEAs. The subjective evaluation was carried out through blind AB listening tests [34, Section 3.3.4]. Five English speaking listeners participated in the tests, where the utterance sp05 ("*Wipe the grease off his dirty face*") was corrupted with 5 dB *passing car* noise and used as the stimulus.

The proposed method is compared with benchmark methods, such as raw waveform processing using FCNN (RWF-FCN) method [15], phase-aware DNN (IAM+IFD) method [16], deep learning KF (DNN-KF) method [17], KF-Ideal method (where $(\{a_i\}, \sigma_w^2)$ and σ_v^2 are computed from the clean speech and noise signal) and Noisy (noise corrupted speech).

5. Results and discussion

Fig. 3 (a)-(b) demonstrates that the proposed method consistently shows improved PESQ scores over the benchmark methods, except the KF-Ideal method for all noise conditions and SNR levels. The IAM+IFD method [16] attained the highest PESQ scores amongst the benchmark methods. The Noisy speech shows the worse PESQ score for all conditions.

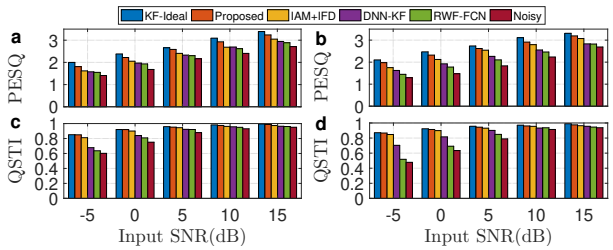


Figure 3: *Performance of each SEA in terms of: (a) PESQ for passing car, (b) PESQ for café babble, (c) QSTI for passing car, and (d) QSTI for café babble.*

Fig. 3 (c)-(d) also shows that the proposed method demonstrates a consistent QSTI improvement across the noise experiments as well as the SNR levels, apart from the KF-Ideal method. The existing IAM+IFD method [16] is found to be competitive with the proposed method in terms of QSTI, typically at low SNR levels. However, the QSTI for each method at high SNR levels is competitive.

It can be seen that the enhanced speech produced by the proposed method (Fig. 4 (f)) exhibits significantly less *residual* noise than that of the benchmark methods (Fig. 4 (c)-(e)) and is similar to the KF-Ideal method (Fig. 4 (g)). The informal

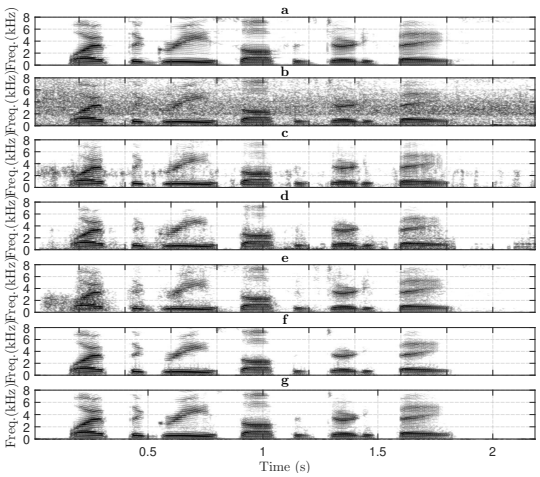


Figure 4: *(a) Clean speech, (b) noisy speech (sp05 is corrupted with 5 dB passing car noise), and the enhanced speech spectrograms produced by the: (c) RWF-FCN, (d) DNN-KF, (e) IAM+IFD, (f) proposed, and (g) KF-Ideal methods.*

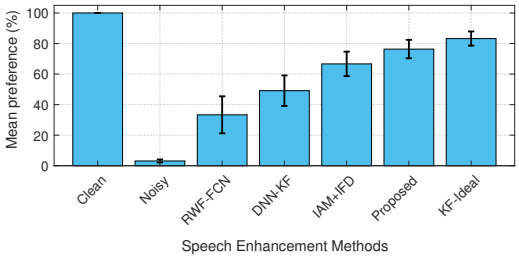


Figure 5: *The mean preference score (%) for each SEA on sp05 corrupted with 5 dB passing car noise.*

listening tests also confirm that the benchmark methods produce enhanced speech with significantly more disturbances than the proposes method.

Fig. 5 shows that the enhanced speech produced by the proposed method is widely preferred by the listeners (76.33%) than the benchmark methods, apart from the KF-Ideal (83.22%) and clean speech. The IAM+IFD method [16] is found to be the best preferred (66.67%) amongst the benchmark methods.

6. Conclusions

This paper introduced a deep learning and Kalman filter-based speech enhancement algorithm. Specifically, DeepMMSE is used to estimate the noise PSD for computing the noise variance. A whitening filter is also constructed using coefficients estimated from the noise PSD. It is employed to the noisy speech signal, yielding a pre-whitened speech. The LPCs are computed from the pre-whitened signal. The large training set of DeepMMSE yields more accurate estimates of the noise variance and the LPCs in various noise conditions. As a result, the KF constructed with the improved parameters minimises the *residual* noise as well as the *distortion* in the resultant enhanced speech. Extensive objective and subjective testing implies that the proposed method outperforms the benchmark methods in various noise conditions for a wide range of SNR levels.

7. 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] Y. Ephraim and D. Malah, "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] N. Upadhyay and A. Karmakar, "Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study," *Procedia Computer Science*, vol. 54, pp. 574 – 584, 2015.

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

[10] 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.

[11] 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.

[12] A. Nicolson and K. K. Paliwal, "Bidirectional long-short term memory network-based estimation of reliable spectral component locations," in *Proc. Interspeech 2018*, 2018, pp. 1606–1610. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1134

[13] 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.

[14] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *Proceedings of Interspeech*, p. 1993–1997, 2017.

[15] 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.

[16] 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.

[17] 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.

[18] Q. Zhang, A. M. Nicolson, M. Wang, K. Paliwal, and C. Wang, "Deepmmse: A deep learning approach to mmse-based noise power spectral density estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

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

[20] S. So, A. E. W. George, R. Ghosh, and K. K. Paliwal, "Kalman filter with sensitivity tuning for improved noise reduction in speech," *Circuits, Systems, and Signal Processing*, vol. 36, no. 4, pp. 1476–1492, April 2017.

[21] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4266–4269, March 2010.

[22] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, December 2012.

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

[24] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44–55, August 2019.

[25] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *ArXiv*, vol. abs/1803.01271, 2018. [Online]. Available: http://arxiv.org/abs/1803.01271

[26] J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *ArXiv*, vol. abs/1607.06450, 2016.

[27] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *27th International Conference on Machine Learning*, pp. 807–814, June 2010.

[28] 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.

[29] 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.

[30] 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.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ArXiv*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[32] 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.

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

[34] 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.