Deep Learning with Augmented Kalman Filter for Single-Channel Speech Enhancement

Sujan Kumar Roy, Aaron Nicolson, Kuldip K. Paliwal
School of Engineering, Griffith University, Brisbane, QLD, 4111, Australia
Emails: {sujankumar.roy, aaron.nicolson}@griffithuni.edu.au, k.paliwal@griffith.edu.au

***Abstract*—The existing augmented Kalman filter (AKF) suffers from poor LPC estimates in real-world noise conditions, which degrades the speech enhancement performance. In this paper, a deep learning technique exploits the LPC estimates for the AKF to enhance speech in various noise conditions. Specifically, a deep residual network is used to estimate the noise PSD for computing noise LPCs. A whitening filter is also implemented with the noise LPCs to pre-whiten the noisy speech signal prior to estimating the speech LPCs. It is shown that the improved speech and noise LPCs enable the AKF to minimize the *residual* noise as well as *distortion* in the enhanced speech. Experimental results show that the enhanced speech produced by the proposed method exhibits higher quality and intelligibility than the benchmark methods in various noise conditions for a wide-range of SNR levels.**

***Index Terms*—Speech enhancement, augmented Kalman filter, Deep Xi, noise PSD, LPC.**

I. Introduction

The aim of a speech enhancement algorithm (SEA) is to eliminate embedded noises from a noisy speech signal. SEAs are used in 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 in the literature.

The SS method heavily depends on the accuracy of the noise estimate [8]. The MMSE and WF-based SEAs rely upon the accurate estimation of the *a priori* SNR [9]. In [3], the decision-directed (DD) approach was proposed to estimate the *a priori* SNR. However, the use of speech and noise power estimates from the previous frame makes it inefficient at computing the *a priori* SNR for the current frame.

In KF-based SEA [7], Paliwal and Basu computed the LPCs from clean speech for enhancing white noise corrupted speech. Gibson *et al.* introduced an augmented KF (AKF) to iteratively suppress the colored noise [10]. The LPCs for the AKF of the current iteration are estimated from the filtered signal of the previous iteration. The enhanced speech (after 3-4 iterations) suffers from *musical* noise and *distortion*. Roy *et al.* proposed a sub-band iterative KF-based SEA. Due to enhancing the high-frequency sub-bands (SBs) only, the low-frequency SBs may still get affected by noise.

In [11], a robustness metric-based tuning *offsets* the bias of the KF gain caused by poor LPC estimates. In [12], it was shown that the robustness metric gives an under-estimated Kalman gain, resulting in *distorted* speech, which can be resolved by sensitivity tuning of the KF gain. Both [11], [12] operate in stationary noise conditions. George *et al.* introduced robustness metric-based tuning of the AKF for colored noise suppression [13]. The robustness metric still gives *distorted* speech. Yu *et al.* introduced a KF-based SEA, where the LPCs are estimated using a deep neural network [14]. However, the noise covariance estimated during speech pauses makes the KF ineffective at dealing with non-stationary noise conditions. The silence detection process was also unspecified.

In this paper, a deep learning technique is used to resolve the LPC estimation issues of the AKF, leading to the capability of performing speech enhancement in various noise conditions. Firstly, the noise PSD is estimated using a deep residual network (ResNet) [15], from where the noise LPCs are computed. The noise LPCs are then used to implement a whitening filter to pre-whiten the noisy speech signal prior to computing the speech LPCs. With the improved speech and noise LPCs, the AKF is found to be effective at minimizing the *residual* noise as well as *distortion* in the enhanced speech. The efficiency of the proposed method is evaluated against the benchmark methods using objective and subjective testing.

II. AKF for Speech Enhancement

Assuming that the colored noise $v(n)$ is additive and uncorrelated with speech $s(n)$, the noisy speech $y(n)$ at sample $n\epsilon\{0,1,2,\dots,M-1\}$ can be represented as:

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

Both $s(n)$ and $v(n)$ can be modeled using p^{th} and q^{th} order linear predictors, as in [16]:

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

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

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

Eqs. (1)-(3) can be used to form the following augmented state-space model (ASSM) of the AKF, as [13]:

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

$$y(n) = \boldsymbol{c}^{\top}\boldsymbol{x}(n), \tag{5}$$

where $\boldsymbol{x}(n) = [s(n) \dots s(n-p+1) \; v(n) \dots v(n-q+1)]^\top$ is a $(p+q) \times 1$ state-vector, $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_s & 0 \\ 0 & \boldsymbol{\Phi}_v \end{bmatrix}$ is a $(p+q) \times (p+q)$ state-transition matrix constructed with the $\{a_i\}$ and $\{b_j\}$, $\boldsymbol{d} = \begin{bmatrix} \boldsymbol{d}_s & 0 \\ 0 & \boldsymbol{d}_v \end{bmatrix}$, $\boldsymbol{d}_s = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^\top$, $\boldsymbol{d}_v = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^\top$, $\boldsymbol{z}(n) = \begin{bmatrix} w(n) \\ u(n) \end{bmatrix}$, and $\boldsymbol{c} = \begin{bmatrix} 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}^\top$ is a $(p+q) \times 1$ vector [13].

Firstly, $y(n)$ is windowed into non-overlapped, short (e.g., 20 ms) frames. For a particular frame, the AKF computes an unbiased and linear MMSE estimate $\hat{\boldsymbol{x}}(n|n)$ at sample n, given $y(n)$ by using the following recursive equations [10]:

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

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

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

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

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

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

For a 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 [13], $\hat{s}(n|n)$ 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 [13]:

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

where $\alpha^2(n)$ and $\beta^2(n)$ are the transmission of *a posteriori* error variances (of the speech and noise) by the augmented dynamic model from the previous sample, $n-1$ [13].

Eq. (11) implies that $K_0(n)$ has a significant impact on the $\hat{s}(n|n)$ estimates, which is the output of the AKF. In practice, poor estimates of $(\{a_i\}, \sigma_w^2)$ and $(\{b_k\}, \sigma_u^2)$ introduce bias in $K_0(n)$, which affects the $\hat{s}(n|n)$ estimates. In the proposed SEA, a deep learning technique is used to estimate the LPCs for the AKF, leading to an improved $\hat{s}(n|n)$ estimate.

III. Proposed Speech Enhancement System

Fig. 2 shows the block diagram of the proposed SEA. Firstly, 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 and N is the total number of frames. The DFT coefficients $Y(l,m)$, $S(l,m)$, and $V(l,m)$ are found using the square-root-Hann window and correspond to $y(n)$, $s(n)$ and $v(n)$. These can also be represented as:

$$Y(l,m) = S(l,m) + V(l,m), \tag{13}$$

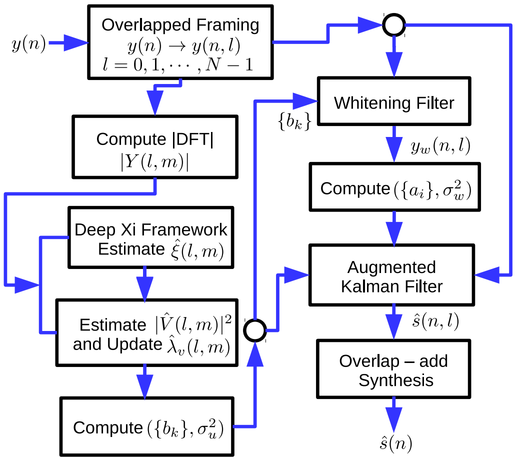


Fig. 1. Block diagram of the proposed deep learning AKF-based SEA.

where m is the discrete-frequency index.

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.

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

The $(\{b_k\}, \sigma_u^2)$ estimates from the initial speech pauses used by the existing AKF [13] makes it limited to suppressing only colored noises. In the proposed SEA, the noise PSD estimate, $\widehat{\lambda}_v(l,m)$, is used to compute $(\{b_k\}, \sigma_u^2)$. Specifically, the noise power estimate, $|\widehat{V}(l,m)|^2$ is obtained through a simplified version[1] of the MMSE method as described in [17], [18]:

$$|\widehat{V}(l,m)|^2 = \left(\frac{1}{1+\xi(l,m)}\right)|Y(l,m)|^2, \tag{14}$$

$$\xi(l,m) = \frac{\lambda_s(l,m)}{\lambda_v(l,m)}, \tag{15}$$

where $\xi(l,m)$ is the *a priori* SNR.

In practice, the existing *decision-directed* approach [17], [18] gives a biased estimate of $\hat{\xi}(l,m)$, which affects the $|\widehat{V}(l,m)|^2$ estimate. To resolve this, we employ a ResNet [15] within the Deep Xi framework (Deep Xi-ResNet) [19] to estimate $\hat{\xi}(l,m)$, as described in section III-B. The smoothed noise PSD estimate, $\widehat{\lambda}_v(l,m)$ is obtained as:

$$\widehat{\lambda}_v(l,m) = \eta\widehat{\lambda}_v(l-1,m) + (1-\eta)|\widehat{V}(l,m)|^2. \tag{16}$$

where η is a smoothing constant and set to 0.9.

The |IDFT| of $\widehat{\lambda}_v(l,m)$ yields an estimate of the noise autocorrelation, $\widehat{R}_{vv}(\tau)$, where τ is the autocorrelation lag. By solving $\widehat{R}_{vv}(\tau)$ using the Levinson-Durbin recursion [16], the $(\{b_k\}, \sigma_u^2)$ $(q = 40)$ estimates are obtained. Then $\{b_k\}$'s are used to design the whitening filter, $H_w(z)$ as [16]:

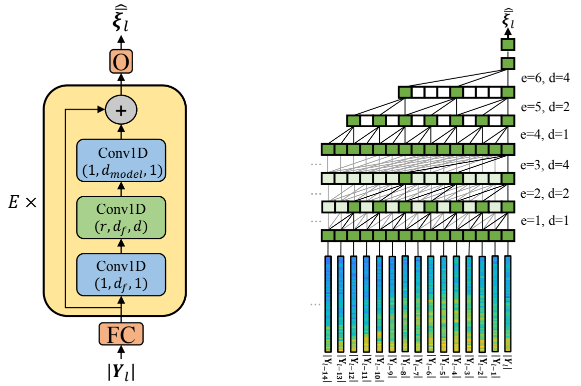
$$H_w(z) = 1 + \sum_{k=1}^{q} b_k z^{-k}. \tag{17}$$

[1] The simplification is a result of setting the *a posteriori* SNR to $\hat{\gamma}(l,m) = \hat{\xi}(l,m) + 1$, which is the maximum-likelihood estimate.

Employing $H_w(z)$ to $y(n,l)$ gives the whitened speech, $y_w(n,l)$ for computing the $(\{a_i\}, \sigma_w^2)$ $(p=10)$ [16].

B. Deep Xi-ResNet for $\hat{\xi}(l,m)$ Estimation

Deep Xi-ResNet is used to estimate $\hat{\xi}(l,m)$ (model **3e** from https://github.com/anicolson/DeepXi). Specifically, it takes $|\boldsymbol{Y}_l|$ (which contains all frequency components for l^{th} frame) as its input and gives an estimate of the mapped *a priori* SNR, $\hat{\bar{\boldsymbol{\xi}}}_l$, as described in section III-C. Deep Xi-ResNet is shown in Fig. 2 (a). It consists of $E=40$ bottleneck residual blocks, where $e\epsilon\{1,2,\dots,E\}$ is the block index. Each block contains three one-dimensional causal dilated convolutional units (CDCUs) [20], where each convolutional unit (CU) is pre-activated by layer normalisation [21] followed by the ReLU activation function [22]. 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 [23], 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 CDCUs. 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 [24]. **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.



(a) (b)

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

C. Mapped a priori SNR Training Target

The training target for the ResNet is a mapped version of the instantaneous *a priori* SNR. For the instantaneous case, $|S(l,m)|$ and $|V(l,m)|$ in Eq. (15) are known to compute $\lambda_s(l,m)$ and $\lambda_v(l,m)$. In [19], $\xi_{\text{dB}}(l,m) = 10\log_{10}[\xi(l,m)]$ was 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 of $\xi_{\text{dB}}(l,m)$ was used as the map. It can be seen from [19, Fig. 2 (top)] that the distribution of ξ_{dB} for a given frequency component, m follows a normal distribution. Thus, it was assumed that $\xi_{\text{dB}}(l,m)$ is distributed normally with mean μ_m and variance σ_m^2: $\xi_{\text{dB}}(l,m) \sim \mathcal{N}(\mu_m, \sigma_m^2)$. The mapped *a priori* SNR $\bar{\xi}(l,m)$ is given by:

$$\bar{\xi}(l,m) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\xi_{\text{dB}}(l,m) - \mu_m}{\sigma_m\sqrt{2}}\right)\right]. \tag{18}$$

Following [19], the statistics of $\xi_{\text{dB}}(l,m)$ for each noisy speech spectral component are found over a sample of $1,000$ noisy speech files from the training set. During inference, $\hat{\xi}(l,m)$ is found from $\hat{\xi}_{\text{dB}}(l,m)$ as follows:

$$\hat{\xi}(l,m) = 10^{(\hat{\xi}_{\text{dB}}(l,m)/10)}, \tag{19}$$

where the $\hat{\xi}_{\text{dB}}(l,m)$ is computed from $\hat{\bar{\xi}}(l,m)$ as follows:

$$\hat{\xi}_{\text{dB}}(l,m) = \sigma_m\sqrt{2}\text{erf}^{-1}\big(2\hat{\bar{\xi}}(l,m) - 1\big) + \mu_m. \tag{20}$$

IV. Speech Enhancement Experiment

A. Training Set

For training Deep Xi-ResNet, a total of $74,250$ clean speech recordings belonging to the *train-clean-100* set from the Librispeech corpus [25] $(28,539)$, the CSTR VCTK corpus [26] $(42,015)$, and the si^* and sx^* training sets from the TIMIT corpus [27] $(3,696)$ are used. 5% of the clean speech recordings are randomly selected and used as a validation set. Thus, $70,537$ clean speech recordings are used in the training set and $3,713$ in the validation set. The $2,382$ noise recordings adopted in [19] are used as the noise training set. All clean speech and noise recordings are single-channel, with a sampling frequency of 16 kHz.

B. Training Strategy

The following strategy was employed to train the ResNet:

- Cross-entropy as the loss function.
- The *Adam* algorithm [28] with default hyper-parameters is used for gradient descent optimisation.
- Gradients are 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.
- A mini-batch size of 10 noisy speech signals.
- 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).

C. 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 non-stationary (*babble*) and colored (*factory2*) noises [29] at SNR levels from -5dB to 15dB, in 5 dB increments.

D. Evaluation Metrics

The objective quality and intelligibility evaluation is carried out using the perceptual evaluation of speech quality (PESQ) [30] and quasi-stationary speech transmission index (QSTI) [31] measures. We also analyze the enhanced speech spectrograms of the SEAs. The subjective evaluation was carried out through blind AB listening tests [32, 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 *babble* noise and used as the stimulus.

The proposed method is compared with benchmark methods, such as MMSE-STSA [3], AKF-IT [10], robustness-metrics-based tuning of AKF (AKF-RMBT) [13], AKF-Ideal (where ($\{a_i\}$, σ_w^2) and ($\{b_k\}$, σ_u^2) are computed from the clean speech and noise signal) and Noisy (noise corrupted speech).

V. Results and Discussion

Fig. 3 (a)-(b) demonstrates that the proposed method consistently shows improved PESQ scores over the benchmark methods, except for AKF-Ideal. The AKF-RMBT method [13] exhibits competitive PESQ scores with the proposed method for *babble* noise (Fig. 3 (a)), however, for *factory2* noise, its efficiency is reduced and is only competitive with the other benchmark methods (Fig. 3 (b)).

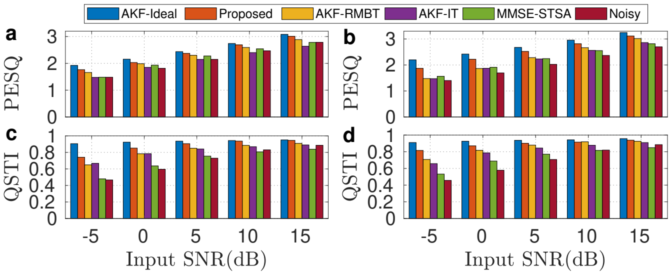


Fig. 3. Performance comparison of the SEAs in terms of average: PESQ; (a) *babble*, (b) *factory2* and QSTI; (c) *babble*, (d) *factory2* noise conditions.

Fig. 3 (c)-(d) shows that the proposed method demonstrates a consistent QSTI improvement across the noise experiments, apart from AKF-Ideal. The existing AKF-RMBT method [13] is also competitive with the proposed method. The QSTI scores of MMSE-STSA [3] and Noisy methods are significantly lower than the AKF-IT method [10] at low SNR levels.

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 that of the AKF-Ideal (Fig. 4 (g)). Some *distortion* and noise-flooring is found for the AKF-RMBT method [13] (Fig. 4 (e)). The enhanced speech of the MMSE-STSA method [3] contains significant *residual* noise (Fig. 4 (c)).

Fig. 5 shows that the enhanced speech produced by the proposed method is widely preferred by the listeners (78%) than the benchmark methods, apart from the AKF-Ideal (81.75%) and clean speech. The AKF-RMBT method [13] is found to be the best preferred (54%) amongst the benchmark methods.

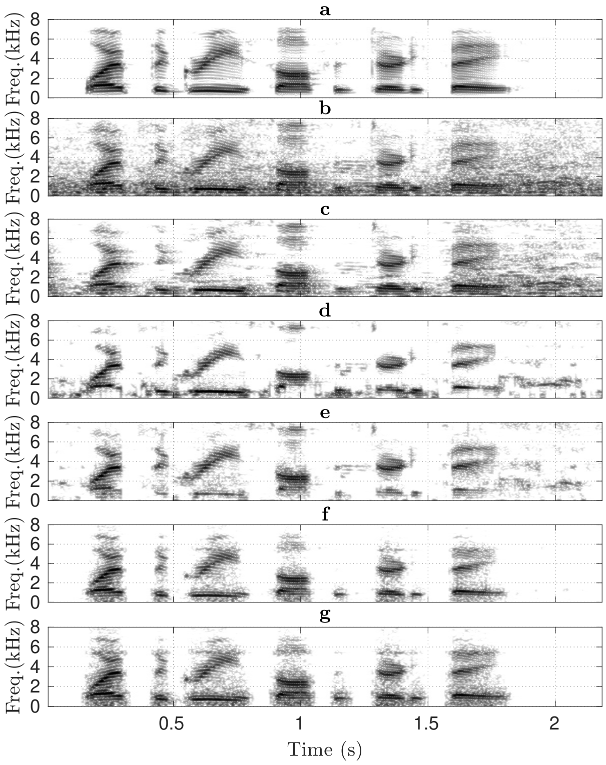


Fig. 4. (a) Clean speech, (b) noisy speech (sp05 is corrupted with 5 dB babble noise), the enhanced speech spectrograms produced by the: (c) MMSE-STSA, (d) AKF-IT, (e) AKF-RMBT, (f) proposed, and (g) AKF-Ideal methods.

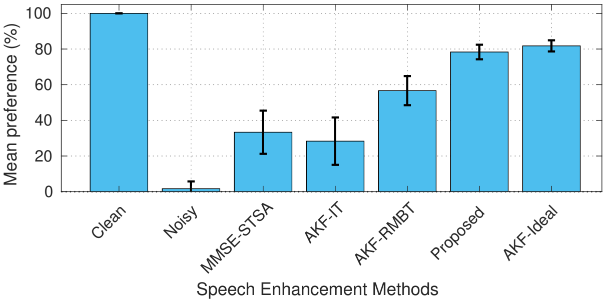


Fig. 5. The mean preference score (%) for each SEA on sp05 corrupted with 5 dB *babble* noise.

VI. Conclusions

This paper introduced a deep learning and augmented Kalman filter-based single channel speech enhancement algorithm. Specifically, Deep Xi-ResNet is used to estimate the noise PSD for computing the noise LPCs. A whitening filter is then constructed with the noise LPCs to pre-whiten the noisy speech signal prior to the speech LPC estimates. The large training set of Deep Xi-ResNet enables the LPC estimates to be effective in various noise conditions. As a result, the improved speech and noise LPCs enable the AKF to minimize the *residual* noise as well as *distortion* in the resultant enhanced speech. Extensive objective and subjective testing imply that the proposed method outperforms the 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] 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*, CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

[10] 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.

[11] 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, pp. 263–268, August 2016.

[12] 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.

[13] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

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

[23] Y. Luo and N. Mesgarani, "Tasnet: Surpassing ideal time-frequency masking for speech separation," *ArXiv*, vol. abs/1809.07454, 2018.

[24] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. V. D. Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," *ArXiv*, vol. abs/1610.10099, 2016.

[25] 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.

[26] 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.

[27] 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.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ArXiv*, vol. abs/1412.6980, 2014.

[29] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

[30] 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.

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

[32] 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.