IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 18, 2021, accepted April 16, 2021, date of publication April 23, 2021, date of current version May 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3075209

DeepLPC: A Deep Learning Approach to Augmented Kalman Filter-Based Single-Channel Speech Enhancement

SUJAN KUMAR ROY[1], (Graduate Student Member, IEEE), AARON NICOLSON[2], AND KULDIP K. PALIWAL[1]

[1]Signal Processing Laboratory, Griffith University, Nathan Campus, Brisbane, QLD 4111, Australia
[2]Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston, QLD 4006, Australia

Corresponding author: Sujan Kumar Roy (sujankumar.roy@griffithuni.edu.au)

ABSTRACT Current deep learning approaches to linear prediction coefficient (LPC) estimation for the augmented Kalman filter (AKF) produce bias estimates, due to the use of a whitening filter. This severely degrades the perceived quality and intelligibility of enhanced speech produced by the AKF. In this paper, we propose a deep learning framework that produces clean speech and noise LPC estimates with significantly less bias than previous methods, by avoiding the use of a whitening filter. The proposed framework, called DeepLPC, jointly estimates the clean speech and noise LPC power spectra. The estimated clean speech and noise LPC power spectra are passed through the inverse Fourier transform to form autocorrelation matrices, which are then solved by the Levinson-Durbin recursion to form the LPCs and prediction error variances of the speech and noise for the AKF. The performance of DeepLPC is evaluated on the NOIZEUS and DEMAND Voice Bank datasets using subjective AB listening tests, as well as seven different objective measures (CSIG, CBAK, COVL, PESQ, STOI, SegSNR, and SI-SDR). DeepLPC is compared to six existing deep learning-based methods. Compared to other deep learning approaches to clean speech LPC estimation, DeepLPC produces a lower spectral distortion (SD) level than existing methods, confirming that it exhibits less bias. DeepLPC also produced higher objective scores than any of the competing methods (with an improvement of 0.11 for CSIG, 0.15 for CBAK, 0.14 for COVL, 0.13 for PESQ, 2.66% for STOI, 1.11 dB for SegSNR, and 1.05 dB for SI-SDR over the next best method). The enhanced speech produced by DeepLPC was also the most preferred by 10 listeners. By producing less biased clean speech and noise LPC estimates, DeepLPC enables the AKF to produce enhanced speech at a higher quality and intelligibility.

INDEX TERMS Speech enhancement, Kalman filter, augmented Kalman filter, deep neural networks, temporal convolutional network, LPC.

I. INTRODUCTION

The main objective of a speech enhancement algorithm (SEA) is to improve the quality and intelligibility of noise corrupted speech (or noisy speech) [1]. This can be achieved by eliminating the embedded noise from a noisy speech signal without distorting the speech. Many applications, such as speech communication systems, hearing aid devices, and speech recognition systems typically rely upon speech enhancement algorithms for robustness. Various SEAs, including spectral subtraction (SS) [2]–[5], the Wiener filter (WF) [6], [7], minimum mean-square error (MMSE) estimators [8]–[11], the Kalman filter (KF) [12], the augmented KF (AKF) [13], computational auditory scene analysis (CASA) [14], and deep learning approaches [15] have been introduced over the decades. This paper focuses on deep learning for the AKF.

The KF is an unbiased linear MMSE estimator, which was first introduced as a SEA by Paliwal and Basu [12]. In this seminal work, each frame of the uncorrupted speech signal (i.e., clean speech) is represented by an auto-regressive (AR) process, whose parameters comprise the linear prediction coefficients (LPCs) and the prediction error variance. The LPC parameters as well as the additive noise variance are inherent to the KF recursive equations. For simplicity, the background noise was assumed to be stationary and white.

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

Given a frame of noisy speech samples, the recursive equations of the KF estimate the clean speech samples. Therefore, the performance of the KF depends on how accurately the LPC parameters and additive noise variance are estimated. However, estimating the LPC parameters and additive noise variance from the noisy speech is difficult in practice, with poor estimates degrading the quality and intelligibility of the enhanced speech produced by the KF. In [12], it was demonstrated that the KF performs well for stationary white noise when the LPC parameters were computed from the clean speech.

In [13], Gibson *et al.* introduced the augmented KF (AKF) for speech enhancement in coloured noise conditions. For the AKF, both the clean speech and additive background noise are represented by AR processes. The clean speech and noise LPC parameters form an augmented matrix, which is used to construct the recursive equations of the AKF. In [13], the AKF processes the noisy speech iteratively (usually three to four iterations) to suppress the coloured background noise, yielding the enhanced speech. During this, the clean speech and noise LPC parameters for the current frame are estimated from the corresponding filtered speech frame of the previous iteration. Although this method demonstrated the ability to improve the signal-to-noise ratio (SNR) of noisy speech, the resultant enhanced speech suffered from *musical noise* and *speech distortion*. This is because the AKF is not robust to inaccurate LPC estimates [16], [17].

In [18], Roy *et al.* introduced a sub-band (SB) iterative KF (SBIT-KF) for SEA. With the assumption that the impact of noise in low-frequency SBs is negligible, SBIT-KF enhances only the high-frequency sub-bands (SBs) of the noisy speech using two KF iterations. However, low-frequency SBs can also be affected by noise—typically when operating in real-life noise conditions. Moreover, the iterative processing employed by SBIT-KF produces *speech distortion* [13]. George *et al.* used a robustness metric to tune the AKF for coloured noise [16]. The authors demonstrated that inaccurate estimates of the clean speech and noise LPC parameters introduce bias in the AKF gain, leading to a degradation in speech enhancement performance. Typically, the adjusted AKF gain is under-estimated in speech regions, resulting in distorted speech.

As of late, deep learning has been used widely for speech enhancement. Motivated by the time-frequency (T-F) masking in CASA [14], Wand and Wang proposed the use of multilayer perceptrons (MLPs) to estimate the ideal binary mask (IBM) [19]. The estimated IBM is used to reconstruct the clean speech spectrum from the noisy speech spectrum. Subsequently, it was demonstrated that the ideal ratio mask (IRM) is able to attain better speech quality than the IBM [20]. In [21], post-processing was applied after masking with the IBM, IRM, or ideal amplitude mask (IAM) [22], resulting in an improvement in objective quality and intelligibility. In [23], Williamson *et al.* introduced a complex ideal ratio mask (cIRM), which is able to estimate both the amplitude and phase spectra of the clean speech. In [24], Zheng *et al.* combine the instantaneous frequency deviation (IFD) with the IAM to form the phase sensitive mask (PSM). The clean speech spectrum is then reconstructed using the estimated mask and the phase information (extracted from the IFD). Different from masking-based methods, mapping-based methods employ a DNN to extract the spectral features of the clean speech from that of the noisy speech. In [15], Xu *et al.* proposed a DNN to map the noisy speech log power spectra (LPS) to the clean speech LPS. In [25], Han *et al.* trained a DNN to learn a spectral mapping from the magnitude spectrum of noisy speech to that of clean speech.

Deep learning methods have also been proposed to improve the performance of statistical model-based SEAs, such as the MMSE short-time spectral amplitude (MMSE-STSA) estimator [8], MMSE log-spectral amplitude (MMSE-LSA) estimator [9], WF [1], and square-root WF (SRWF) [1]. The performance of these SEAs relies upon the accurate estimation of the *a priori* SNR. Recently, a deep learning framework was proposed to estimate the *a priori* SNR directly from the noisy speech spectral magnitude, called Deep Xi [26]. Deep Xi significantly improved the performance of the statistical model-based SEAs, and demonstrates the ability to produce higher quality enhanced speech than the IRM. In [27], Zhang *et al.* proposed a DeepMMSE framework, which employed a ResNet temporal convolutional network (ResNet-TCN) for MMSE-based noise power spectral density (PSD) estimation. DeepMMSE demonstrates a significant improvement in noise PSD tracking over previous methods.

Deep learning has also been employed for time-domain speech enhancement. In [28], end-to-end utterance enhancement using a fully-convolutional neural network (EEUE-FCNN) has been proposed. The authors claimed that the discontinuities present at the boundaries of framed speech are detrimental to the enhancement process. In this SEA, an FCNN facilitates a direct mapping of the noisy speech waveform to the clean speech waveform. The FCNN model is constructed with ten one-dimensional convolutional layers; each comprises of 30 filters with a filter size of 55. The authors claim that the processing of the whole noisy speech waveform results in enhanced speech with an improvement in intelligibility.

A. RELATED WORK

In this section, we briefly review existing deep learning approaches to LPC parameter estimation for the KF and AKF. In [29], Pickersgill *et al.* employed a similar DNN to that used in [15] for LPC estimation. They evaluate the LPC estimation performance in terms of the spectral distortion (SD) level. However, the performance of LPC estimation at low SNR levels was unspecified. In [30], Yu *et al.* proposed a deep learning-based KF for speech enhancement. A DNN containing three hidden layers is adopted for estimating the LPCs for each noisy speech frame. For training the DNN, only four noise recordings and four SNR levels were used, thus limiting its capability to generalise to unobserved conditions.

In addition, the noise covariance is estimated during speech pauses of the noisy speech, which does not account for conditions that have time-varying amplitudes.

Motivated by the performance of Deep Xi in combination with statistical model-based SEAs [26], a residual network (ResNet) [31] was incorporated within the Deep Xi framework to estimate parameters for the AKF [17]. Later on, the DeepMMSE framework [27] was used to estimate parameters for the KF [32]. In both methods [17], [32], the noise parameters for the AKF and KF are computed from the estimated noise PSD derived from Deep Xi and DeepMMSE, respectively. However, Deep Xi and DeepMMSE do not address clean speech LPC estimation directly from the noisy speech. Rather, a whitening filter is constructed with its coefficients computed from the estimated noise. The whitening filter is then applied to each noisy speech frame, yielding pre-whitened speech. The clean speech LPC parameters are then computed from the pre-whitened speech. It is shown that the estimated clean speech LPC parameters [32] produce a higher SD level than competing methods. This indicates that the whitening techniques in [17], [32] do not adequately address clean speech LPC estimation. It is also demonstrated that the biased clean speech LPC estimates in [17], [32] impact the quality and intelligibility of the enhanced speech in real-world noise conditions.

In [33], Yu *et al.* adopted a DNN and an LSTM network to estimate the clean speech and noise LPCs, respectively, as well as multi-band spectral subtraction (MB-SS) post-processing [4] for coloured-noise AKF-based speech enhancement (LSTM-CKFS). To estimate the prediction error variances for the AR processes of the AKF, the authors employed a maximum likelihood (ML) approach [34]. Due to training the LSTM network with a small amount of training data [30], LSTM-CKFS lacks the ability to generalise to unobserved conditions. The bias present in the LPC estimates of LSTM-CKFS produces a significant residual background noise in the resultant AKF enhanced speech [33]. The authors attempted to reduce the residual background noise through post processing [4]. For MB-SS, the noise spectrum is updated during speech pauses, which is not appropriate for noise conditions that have time-varying amplitudes.

In light of the shortcomings of existing deep learning-based KF and AKF methods (presented in Table 1) this paper introduces DeepLPC, a deep learning framework for accurately estimating the clean speech and noise LPC parameters. Specifically, the DeepLPC maps each frame of the noisy speech magnitude spectrum to the clean speech and noise LPC power spectra. The autocorrelation metrics (constructed from the estimated LPC power spectra using the inverse Fourier transform) are then solved by the Levinson-Durbin recursion, yielding the clean speech and noise LPC parameters. The proposed method aims to mitigate the weaknesses of previously proposed deep learning-based KFs and AKFs, by providing an improved estimate of the clean speech LPCs.

The structure of this paper is as follows: background knowledge is presented in Section II, including the signal model, the AKF SEA, and an overview of Deep Xi-KF and Deep Xi-AKF. In Section III, we describe the proposed SEA, which includes the proposed DeepLPC framework for LPC estimation. Following this, Section IV describes the experimental setup. The experimental results are then presented and discussed in Section V. Finally, Section VI gives some concluding remarks.

TABLE 1. Summary of existing deep learning-based LPC estimation methods for the KF as well as AKF.

Methods	Summary	Limitations
DNN-LPC [29]	A DNN [15] is used to estimate the speech LPC parameters.	Due to training the DNN with a small dataset, its generalisation capabilities may be reduced.
DeepXi-AKF [17]	AKF is constructed with the noise and speech LPC parameters derived from a Deep Xi-ResNet framework and whitening technique [16], respectively.	The whitening technique gives a biased estimate of the speech LPC parameters, which impacts the quality and intelligibility of enhanced speech.
DeepXi-KF [32]	KF is constructed with the noise variance and speech LPC parameters derived from the DeepMMSE framework [27] and whitening technique [16], respectively.	As in [17], the biased speech LPC parameters derived from the whitening technique impact the quality and intelligibility of enhanced speech.
LSTM-CKFS [33]	AKF is constructed with the speech and noise LPC parameters derived from an LSTM network and an ML-based approach [34].	LSTM-CKFS [33] exhibits a high amount of bias.

II. BACKGROUND

A. SIGNAL MODEL

The noisy speech $y(n)$, at discrete-time sample n, is assumed to be given by:

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

where $s(n)$ is the clean speech and $v(n)$ is uncorrelated additive coloured noise. A 32 ms rectangular window with 50% overlap is used to convert $y(n)$ into frames, denoted by $y(n, l)$:

$$y(n, l) = s(n, l) + v(n, l), \tag{2}$$

where $l\epsilon\{0, 1, \ldots, L-1\}$ is the frame index with L being the total number of frames in an utterance, and $n\epsilon\{0, 1, \ldots, N-1\}$ where N is the total number of samples within each frame.

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

$$Y(l, m) = S(l, m) + V(l, m), \tag{3}$$

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

B. AKF FOR SPEECH ENHANCEMENT

For simplicity, the frame index is omitted in the AKF recursive equations. Each frame of the clean speech and noise signal in (2) can be represented with p^{th} and q^{th} order AR models, as in [35, Chapter 8]:

$$s(n) = -\sum_{i=1}^{p} a_i s(n-i) + w(n), \tag{4}$$

$$v(n) = -\sum_{k=1}^{q} b_k v(n-k) + u(n), \tag{5}$$

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

Equations (2), (4) and (5) can be used to form the following augmented state-space model (ASSM) of the AKF, as in [13]:

$$\boldsymbol{x}(n) = \boldsymbol{\Phi}\boldsymbol{x}(n-1) + \boldsymbol{r}\boldsymbol{g}(n), \tag{6}$$

$$y(n) = \boldsymbol{c}^\top \boldsymbol{x}(n). \tag{7}$$

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:

$$\boldsymbol{\Phi}_s = \begin{bmatrix} -a_1 & -a_2 & \ldots & -a_{p-1} & -a_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \tag{8}$$

$$\boldsymbol{\Phi}_v = \begin{bmatrix} -b_1 & -b_2 & \ldots & -b_{q-1} & -b_q \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, \tag{9}$$

3) $\boldsymbol{r} = \begin{bmatrix} \boldsymbol{r}_s & 0 \\ 0 & \boldsymbol{r}_v \end{bmatrix}$, where $\boldsymbol{r}_s = \begin{bmatrix} 1\ 0 \ldots 0 \end{bmatrix}^\top$, $\boldsymbol{r}_v = \begin{bmatrix} 1\ 0 \ldots 0 \end{bmatrix}^\top$,
4)

$$\boldsymbol{g}(n) = \begin{bmatrix} w(n) \\ u(n) \end{bmatrix}, \tag{10}$$

5) $\boldsymbol{c}^\top = \begin{bmatrix} \boldsymbol{c}_s^\top & \boldsymbol{c}_v^\top \end{bmatrix}$, where $\boldsymbol{c}_s = \begin{bmatrix} 1\ 0 \ldots 0 \end{bmatrix}^\top$ and $\boldsymbol{c}_v = \begin{bmatrix} 1\ 0 \ldots 0 \end{bmatrix}^\top$ are $p \times 1$ and $q \times 1$ vectors,
6) $y(n)$ is the noisy measurement at sample n.

For each frame, the AKF recursively computes an unbiased linear MMSE estimate, $\hat{\boldsymbol{x}}(n|n)$ at sample n, given $y(n)$, by using the following Equations [16]:

$$\hat{\boldsymbol{x}}(n|n-1) = \boldsymbol{\Phi}\hat{\boldsymbol{x}}(n-1|n-1), \tag{11}$$

$$\boldsymbol{\Psi}(n|n-1) = \boldsymbol{\Phi}\boldsymbol{\Psi}(n-1|n-1)\boldsymbol{\Phi}^\top + \boldsymbol{Q}\boldsymbol{r}\boldsymbol{r}^\top, \tag{12}$$

$$\boldsymbol{K}(n) = \boldsymbol{\Psi}(n|n-1)\boldsymbol{c}(\boldsymbol{c}^\top \boldsymbol{\Psi}(n|n-1)\boldsymbol{c})^{-1}, \tag{13}$$

$$\hat{\boldsymbol{x}}(n|n) = \hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^\top \hat{\boldsymbol{x}}(n|n-1)], \tag{14}$$

$$\boldsymbol{\Psi}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^\top]\boldsymbol{\Psi}(n|n-1), \tag{15}$$

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, $\boldsymbol{h}^\top \hat{\boldsymbol{x}}(n|n)$ gives the output of the AKF, $\hat{s}(n|n)$, where $\boldsymbol{h} = \begin{bmatrix} 1\ 0\ 0 \ldots 0 \end{bmatrix}^\top$ is a $(p+q) \times 1$ column vector. As in [16], $\hat{s}(n|n)$ is given by:

$$\hat{s}(n|n) = [1 - K_0(n)]\hat{s}(n|n-1) + K_0(n)[y(n) - \hat{v}(n|n-1)], \tag{16}$$

where $K_0(n)$ is the 1^{st} component of $\boldsymbol{K}(n)$, given by [16]:

$$K_0(n) = \frac{\alpha^2(n) + \sigma_w^2}{\alpha^2(n) + \sigma_w^2 + \beta^2(n) + \sigma_u^2}, \tag{17}$$

where $\alpha^2(n) = \boldsymbol{c}_s^\top \boldsymbol{\Phi}_s \boldsymbol{\Psi}_s(n-1|n-1)\boldsymbol{\Phi}_s^\top \boldsymbol{c}_s$ and $\beta^2(n) = \boldsymbol{c}_v^\top \boldsymbol{\Phi}_v \boldsymbol{\Psi}_v(n-1|n-1)\boldsymbol{\Phi}_v^\top \boldsymbol{c}_v$ are the transmission of *a posteriori* error variances of the speech and the noise augmented dynamic model from the previous sample, $n-1$, respectively [16].

Equation (16) reveals that $K_0(n)$ has a significant impact on $\hat{s}(n|n)$. In practice, the inaccurate estimates of ($\{a_i\}, \sigma_w^2$) and ($\{b_k\}, \sigma_u^2$) introduce bias into $K_0(n)$, which impacts $\hat{s}(n|n)$.

C. REVIEW OF Deep Xi-AKF AND Deep Xi-KF

This section briefly summarises Deep Xi-AKF and Deep Xi-KF [17], [32], including their limitations. Deep Xi-AKF and Deep Xi-KF both employ an MMSE-based noise PSD estimator, called DeepMMSE [27]. DeepMMSE utilises deep learning to estimate the noise PSD estimate, $\hat{\lambda}_v(l, m)$. Specifically, DeepMMSE leverages the Deep Xi framework to estimate the *a priori* and *a posteriori* SNR for the MMSE noise periodogram estimator [36], [37], where $|\hat{V}(l, m)|^2$ is the noise periodogram estimate. To obtain $\hat{\lambda}_v(l, m)$ from $|\hat{V}(l, m)|^2$, DeepMMSE employs first-order recursive smoothing:

$$\hat{\lambda}_v(l, m) = \eta\hat{\lambda}_d[l-1, k] + (1-\eta)|\hat{V}(l, m)|^2, \tag{18}$$

where η is the smoothing factor.

For Deep Xi-AKF [17], ($\{b_k\}, \sigma_u^2$) are computed from $\hat{\lambda}_v(l, m)$, where $\eta = 0.9$. However, ($\{a_i\}, \sigma_w^2$) are still unknown. As in [16], ($\{a_i\}, \sigma_w^2$) ($p = 10$) are computed frame-wise from pre-whitened speech $y_w(n, l)$, as presented in [17, Section III-A]. The AKF is then constructed with the estimated ($\{a_i\}, \sigma_w^2$) and ($\{b_k\}, \sigma_u^2$) for speech enhancement. In Deep Xi-KF [32], the noise variance, σ_v^2 is estimated from $\hat{\lambda}_v(l, m)$, where $\eta = 0$ was used in Equation (18) [32, Section 3.1]. As in [17], ($\{a_i\}, \sigma_w^2$) ($p = 10$) are computed frame-wise from pre-whitened speech, $y_w(n, l)$. The KF is then constructed with the estimated, σ_v^2 and ($\{a_i\}, \sigma_w^2$) for speech enhancement.

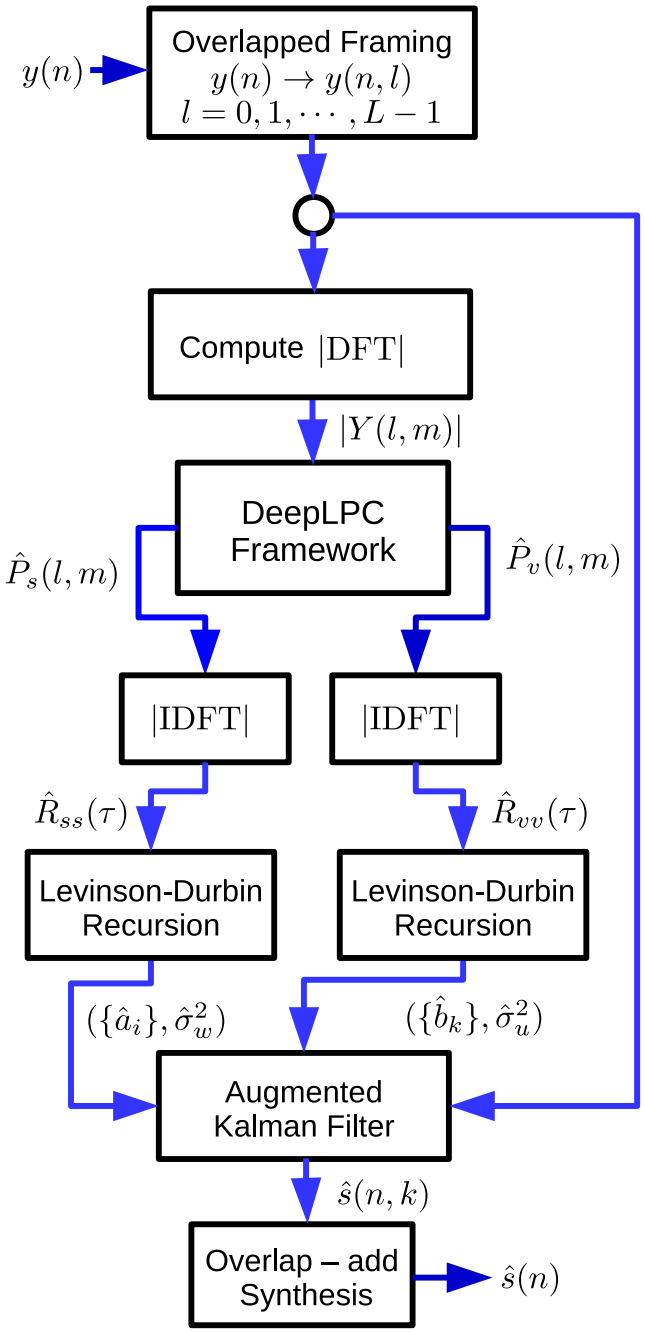

FIGURE 1. Block diagram of the proposed SEA.

It was observed in [16] that biased speech LPC estimates are produced when using the whitening filter [32]. This indicates that Deep Xi-AKF and Deep Xi-KF also produce biased speech LPC estimates. The biased estimates of $(\{a_i\}, \sigma_w^2)$ will thus impact the quality and intelligibility of the enhanced speech produced by the AKF and the KF. Moreover, directly estimating the noise LPC parameters rather than using DeepMMSE could result in less bias.

III. PROPOSED SPEECH ENHANCEMENT ALGORITHM

To address the shortcomings of Deep Xi-AKF and Deep Xi-KF highlighted in the previous section, we propose the

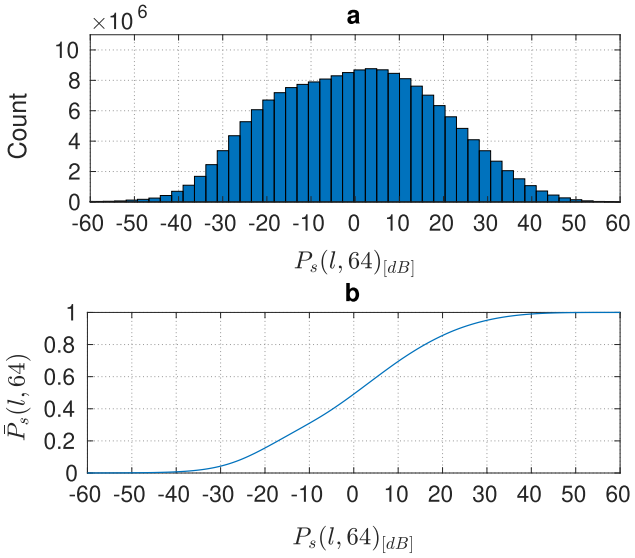

FIGURE 2. (a) The distribution of $P_s(l,64)_{[dB]}$ over a sample of the training set. (b) The CDF of $P_s(l,64)_{[dB]}$, assuming that $P_s(l,64)_{[dB]}$ is distributed normally. The sample of the training set is described in Section IV-B.

DeepLPC framework. DeepLPC jointly estimates the clean speech and noise LPC power spectra (LPC-PS), denoted as $P_s(l,m)$ and $P_v(l,m)$, respectively. The clean speech and noise LPC-PS estimates are used to compute the clean speech and noise LPC estimates.

Figure 1 shows the block diagram of the proposed SEA. As in Section II-A, $y(n)$ is first converted into frames, $y(n,l)$. Following this, the short-time noisy speech magnitude spectrum, $|\boldsymbol{Y}_l|$, is computed. Next, DeepLPC jointly estimates $P_s(l,m)$ and $P_v(l,m)$ from $|\boldsymbol{Y}_l|$. During training, $P_s(l,m)$ and $P_v(l,m)$ are computed as in [35, Chapter 9]:

$$P_s(l,m) = \frac{\sigma_w^2}{\left|1 + \sum_{i=1}^{p} a_i e^{-j2\pi im/M}\right|^2}, \tag{19}$$

$$P_v(l,m) = \frac{\sigma_u^2}{\left|1 + \sum_{k=1}^{q} b_k e^{-j2\pi km/M}\right|^2}, \tag{20}$$

where $(\{a_i\}, \sigma_w^2)$ $(p = 16)$ and $(\{b_k\}, \sigma_u^2)$ $(q = 16)$ are computed from the clean speech $(s(n,l))$ and noise $(v(n,l))$ using the autocorrelation method [35]. The chosen clean speech and noise LPC order ($p = 16$ and $q = 16$, respectively) is based on the findings that higher-order LPCs are required to accurately estimate the short-term correlation information of wideband (16 kHz) speech [38, Section 6.2.1-6.2.2 and Figure (6.2)].

To facilitate the convergence of the stochastic gradient descent algorithm, the dynamic range of $P_s(l,m)$ and $P_v(l,m)$ must be compressed. For this, we follow the same method used to compress the dynamic range of the instantaneous *a priori* SNR in [26]. We first convert $P_s(l,m)$ and $P_v(l,m)$ into the log-spectral domain, i.e., $P_s(l,m)_{[dB]} = 10\log_{10}(P_s(l,m))$ and $P_v(l,m)_{[dB]} = 10\log_{10}(P_v(l,m))$.

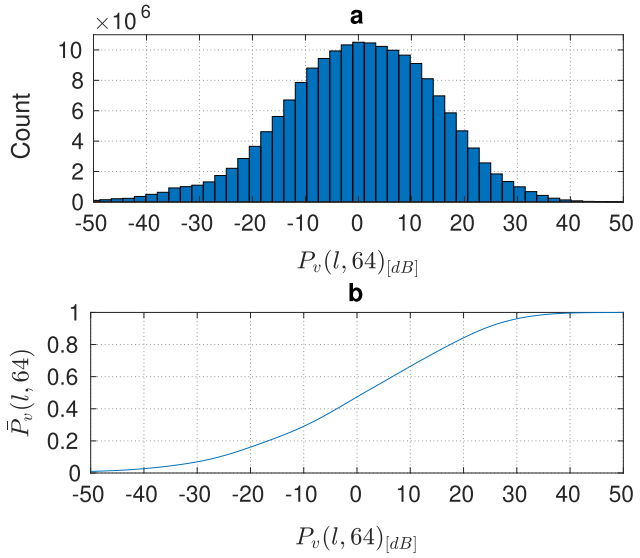

FIGURE 3. (a) The distribution of $P_v(l,64)_{[dB]}$ over a sample of the training set. (b) The CDF of $P_v(l,64)_{[dB]}$, assuming that $P_v(l,64)_{[dB]}$ is distributed normally. The sample of the training set is described in Section IV-B.

Next, we utilise the cumulative distribution function (CDF) of $P_s(l,m)_{[dB]}$ and $P_v(l,m)_{[dB]}$ to compress their dynamic range to the interval [0, 1]. As an example, we observe that $P_s(l,64)_{[dB]}$ and $P_v(l,64)_{[dB]}$ follow a Gaussian distribution, as shown in Figures 2 (a) and 3 (a), respectively. Therefore, we assume that $P_s(l,m)_{[dB]}$ and $P_v(l,m)_{[dB]}$ are distributed normally with mean, μ_s and μ_v, and variance σ_s^2 and σ_v^2, respectively ($P_s(l,m)_{[dB]} \sim \mathcal{N}(\mu_s, \sigma_s^2)$ and $P_v(l,m)_{[dB]} \sim \mathcal{N}(\mu_v, \sigma_v^2)$). The statistics of $P_s(l,m)_{[dB]}$ and $P_v(l,m)_{[dB]}$, i.e., (μ_s, σ_s^2) and (μ_v, σ_v^2) for each frequency bin m were found over a sample of the training set, as described in Section IV-B. The CDF of $P_s(l,64)_{[dB]}$ and $P_v(l,64)_{[dB]}$ over the sample is shown in Figure 2 (b) and Figure 3 (b), respectively. The CDFs of $P_s(l,m)_{[dB]}$ and $P_v(l,m)_{[dB]}$ are used to form the training targets for DeepLPC:

$$\bar{P}_s(l,m) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{P_s(l,m)_{[dB]} - \mu_s}{\sigma_s\sqrt{2}}\right)\right], \quad (21)$$

$$\bar{P}_v(l,m) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{P_v(l,m)_{[dB]} - \mu_v}{\sigma_v\sqrt{2}}\right)\right]. \quad (22)$$

$\bar{P}_s(l,m)$ and $\bar{P}_v(l,m)$ are concatenated to form the final training target for DeepLPC:

$$\boldsymbol{\zeta}_l = \{\bar{P}_s(l,0), \bar{P}_s(l,1), \ldots, \bar{P}_s(l,M-1), \bar{P}_v(l,0), \bar{P}_v(l,1), \ldots, \bar{P}_v(l,M-1), \quad (23)$$

where $\boldsymbol{\zeta}_l$ is of size $M \times 2$.

During inference, $\hat{\boldsymbol{\zeta}}_l$ is first split into $\hat{\bar{P}}_s(l,m)$ and $\hat{\bar{P}}_v(l,m)$. The clean speech and noise LPC-PS estimates are then computed from $\hat{\bar{P}}_s(l,m)$ and $\hat{\bar{P}}_v(l,m)$:

$$\hat{P}_s(l,m) = 10^{((\sigma_s\sqrt{2}\text{erf}^{-1}(2\hat{\bar{P}}_s(l,m)-1)+\mu_s)/10)}, \quad (24)$$

$$\hat{P}_v(l,m) = 10^{((\sigma_v\sqrt{2}\text{erf}^{-1}(2\hat{\bar{P}}_v(l,m)-1)+\mu_v)/10)}. \quad (25)$$

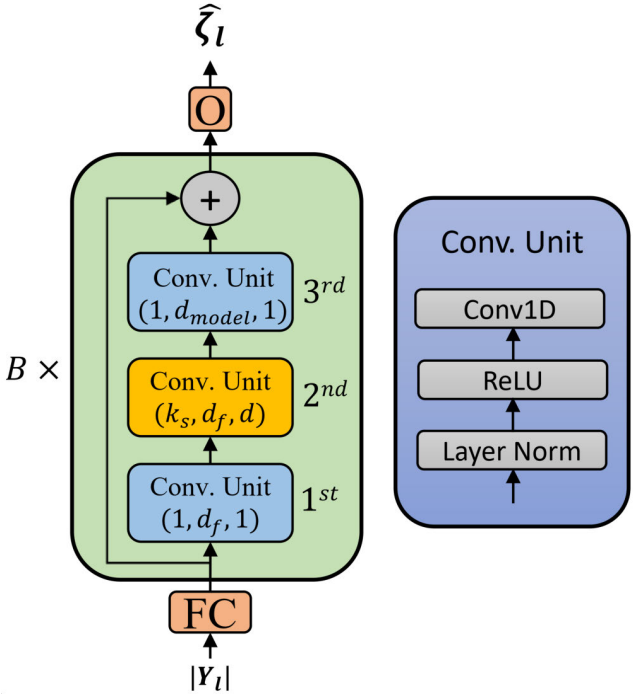

FIGURE 4. ResNet-TCN within the proposed DeepLPC framework. The ResNet-TCN consists of a fully-connected first layer, FC, followed by B residual blocks, and then a fully-connected output layer, O that employs sigmoidal units.

The |IDFT| of $\hat{P}_s(l,m)$ and $\hat{P}_v(l,m)$ yields an estimate of the autocorrelation matrices, $\hat{R}(\tau)$ and $\hat{H}(\tau)$, where τ is the autocorrelation lag. With $\hat{R}(\tau)$ and $\hat{H}(\tau)$, $\{a_i\}$ and $\{b_k\}$ can be represented using the Yule-Walker equation [35, Chapter 8]:

$$\begin{bmatrix} \hat{R}(0) & \hat{R}(1) & \ldots & \hat{R}(p-1) \\ \hat{R}(1) & \hat{R}(0) & \ldots & \hat{R}(p-2) \\ \hat{R}(2) & \hat{R}(1) & \ldots & \hat{R}(p-3) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{R}(p-1) & \hat{R}(p-2) & \ldots & \hat{R}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} -\hat{R}(1) \\ -\hat{R}(2) \\ -\hat{R}(3) \\ \vdots \\ -\hat{R}(p) \end{bmatrix}, \quad (26)$$

$$\begin{bmatrix} \hat{H}(0) & \hat{H}(1) & \ldots & \hat{H}(p-1) \\ \hat{H}(1) & \hat{H}(0) & \ldots & \hat{H}(p-2) \\ \hat{H}(2) & \hat{H}(1) & \ldots & \hat{H}(p-3) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{H}(q-1) & \hat{H}(q-2) & \ldots & \hat{H}(0) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_q \end{bmatrix} = \begin{bmatrix} -\hat{H}(1) \\ -\hat{H}(2) \\ -\hat{H}(3) \\ \vdots \\ -\hat{H}(q) \end{bmatrix}. \quad (27)$$

The matrices $\hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{H}}$ in Equations (26)-(27) are *Toepiltz* matrices, where all diagonal elements are identical. The special *Toepiltz* structure of the autocorrelation matrix allows the use of the Levinson-Durbin recursion to solve the Yule-Walker equation [35, Chapter 8]. Solving the Equations (26) and (27) using the Levinson-Durbin recursion [35, Chapter 8], yields $(\{\hat{a}_i\}, \hat{\sigma}_w^2)$ $(p = 16)$ and $(\{\hat{b}_k\}, \hat{\sigma}_u^2)$ $(q = 16)$ for the AKF. For more information about solving the Yule-Walker equation using Levinson-Durbin recursion, we refer the readers to [35, Section 8.2.2].

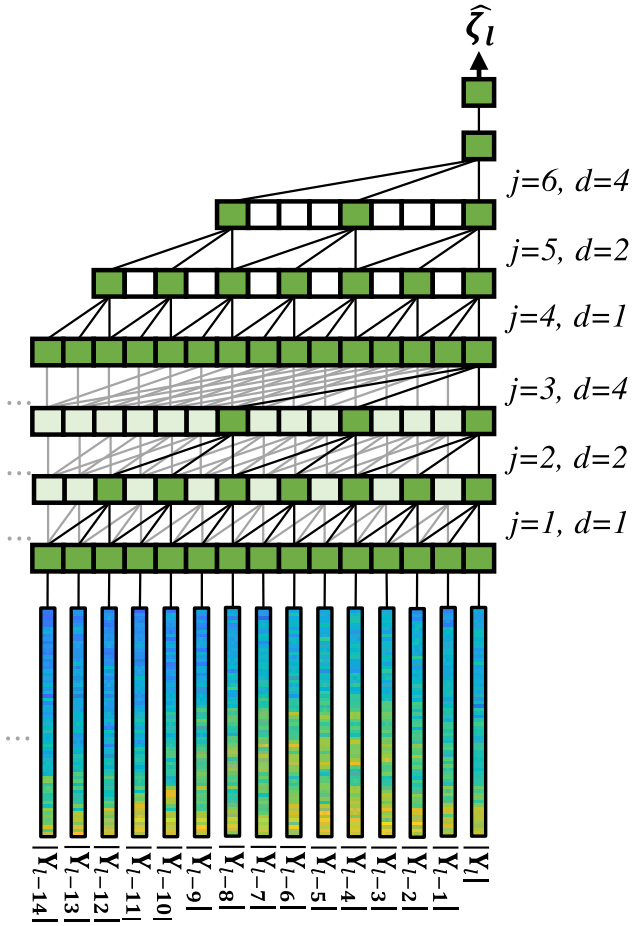

FIGURE 5. Example of the contextual field of DeepLPC-ResNet-TCN, where $D = 4$, $B = 6$, and $k_s = 3$ are used.

A. DeepLPC-ResNet-TCN

The ResNet-TCN from [27] is chosen to estimate $\boldsymbol{\zeta}_l$ from $|\boldsymbol{Y}_l|$ for DeepLPC, and is detailed in this section. The ResNet-TCN is shown in Figure 4. The input, $|\boldsymbol{Y}_l|$ is first passed through **FC**, a fully-connected layer of size d_{model}, followed by the rectified linear unit (ReLU) activation function [39] and layer normalization (LN) [40] (in [27], LN is followed by ReLU). For all LN operations of the ResNet-TCN, the center and scale parameters omitted—to prevent overfitting (this differs to the ResNet-TCN used in [27]). **FC** is followed by B bottleneck residual blocks, where $j\epsilon\{1, 2, \ldots, B\}$ is the block index. As in [41], each block contains three one-dimensional convolutional units. Each convolutional unit is pre-activated by LN [40] followed by the ReLU activation function [39]. The kernel size, output size, and dilation rate for each convolutional unit is denoted as (**kernel size**, **output size**, **dilation rate**), as shown in Figure 4.

The first and third convolutional units in each block have a kernel size of one, whilst the second convolutional unit has a kernel size of k_s. The output size of the first and second convolutional unit is d_f, while the third one is d_{model}. A dilation rate of one is set for the first and the third convolutional units, and d for the second convolutional unit. The second convolutional unit provides a contextual field over previous time steps. The dilation rate d is cycled as the block index j increases: $d = 2^{(j-1 \bmod (\log_2(D)+1)}$, where mod is the modulo operation, and D is the maximum dilation rate. An example of how the dilation rate is cycled is shown in Figure 5, with $D = 4$, and $B = 6$. It can be seen that the dilation rate is reset after block three. This also demonstrates the contextual field gained by the use of causal dilated convolutional units. The last residual block is followed by the output layer, **O**, which is a fully-connected layer with sigmoidal units. The output layer gives an estimate of $\hat{\boldsymbol{\zeta}}_l$. The hyperparameters used in [27] were used in this study: $d_{model} = 256$, $d_f = 64$, $B = 40$, $k_s = 3$, and $D = 16$. The training strategy as well as a complexity and convergence analysis of ResNet-TCN are detailed in Sections IV-D and IV-E.

IV. SETUP OF THE SPEECH ENHANCEMENT EXPERIMENT

A. TRAINING & VALIDATION SET

In this paper, two datasets are used for training DeepLPC: the DEMAND Voice Bank [51] and Deep Xi datasets,[1] which have been used previously in [17], [26], [27], [32]. The details of the dataset are summarized in Table. 2. All clean speech and noise recordings are single-channel with a sampling frequency of 16 kHz. For the Deep Xi dataset, the noisy speech for the validation set is created by corrupting each of the 3 713 clean speech recordings with a random section of a randomly selected noise recording (from the set of 813 noise recordings) at a randomly selected SNR level (−10 to +20 dB, in 1 dB increments). The number of validation examples is thus equal to the number of clean speech recordings (3 713). As in [51], we do not use a validation set for the DEMAND Voice Bank dataset during training.

B. TRAINING SET SAMPLE

For the Deep Xi dataset, 2 500 randomly selected clean speech recordings were mixed with 2 500 randomly selected noise recordings with SNR levels: −10 dB to +20 dB in 1 dB increments, giving 2 500 noisy speech signals. For each frequency bin, m, the sample mean and variances, (μ_s, σ_s^2) and (μ_v, σ_v^2) were computed from 2 500 concatenated clean speech recordings and scaled noises, respectively. This sample is used in Figures 2 and 3. The same is done to produce the sample for the DEMAND Voice Bank dataset, except that a sample size of 1 500 is used.

C. TEST SET

For the objective experiments, the NOIZEUS dataset was used to evaluate the performance of DeepLPC trained with the Deep Xi dataset. In addition, DeepLPC trained with the DEMAND Voice Bank dataset is evaluated using the DEMAND Voice Bank test set. The details of the NOIZEUS and DEMAND Voice Bank test sets are given in Table 3. All the clean speech and noise recordings in Table 3 are single-channel with a sampling frequency of 16 kHz. Note that the

[1] Specifically, an earlier version of the open-source Deep Xi dataset [53].

TABLE 2. Summary of the training and validation datasets used in this paper.

Dataset	No. of Speech Recordings	No. of Noise Recordings	No. of Train Data	No. of Validation data
Deep Xi dataset (training set)	74 250 clean speech recordings are taken from the Librispeech corpus [42] (28 539), CSTR VCTK corpus [43] (42 015), and TIMIT corpus [44] (3 696).	16 243 noise recordings are taken from the following datasets: QUT-NOISE [45], Nonspeech [46], Environmental Background Noise [47], [48], MUSAN [49], FreeSound packs[2], and coloured noise recordings.	70 537 clean speech recordings and 15 430 noise recordings are used to construct the training set as described in Section IV-D.	5% of speech and noise recordings, i.e., 3 713 clean speech recordings and 813 noise recordings are used to construct the validation set.
DEMAND Voice Bank dataset (training set)	The Voice Bank corpus comprises of 11 572 clean speech recordings [50].	10 noise recordings, including 2 synthetic noises (*speech shaped* and *babble*) [51] and 8 real-world noises from DEMAND dataset [52].	11 572 training examples are generated, as described in Section IV-D.	As in [51], no validation data is used in training.

[2] Freesound packs (https://freesound.org/) that were used: 147, 199, 247, 379, 622, 643, 1 133, 1 563, 1 840, 2 432, 4 366, 4 439, 15 046, 15 598, 21 558.

TABLE 3. Summary of the test datasets used in this paper.

Dataset	No. of Speech Recordings	No. of Noise Recordings	Generation of Test Dataset
NOIZEUS dataset	30 clean speech utterances belonging to six speakers (three male and three female) are taken from the NOIZEUS corpus [1, Chapter 12].	2 real-world non-stationary (*voice babble, street*) and 2 real-world coloured (*factory, f16* noise recordings are taken from [47], [48].	The noisy speech for the test set is formed by mixing the clean speech with *voice babble, street, factory*, and *f16* noise recordings at multiple SNR levels varying from -5dB to +15 dB, in 5 dB increments. This provides 30 examples per condition with 20 total conditions.
DEMAND Voice Bank (test set)	824 clean speech recordings of two speakers from Voice Bank Corpus–393 from *p232* and 431 from *p257* [50].	5 noise recordings selected from the DEMAND dataset [52].	The noisy speech for the test set is formed by mixing 824 clean speech recordings with 5 noise recordings at multiple SNR levels: $\{2.5, 7.5, 12.5, 17.5\}$. This provides 824 examples per condition with 20 total conditions.

speech and the noise recordings in both test sets are different from those used in the training and validation sets.

D. TRAINING STRATEGY

The following training strategy was employed to train DeepLPC-ResNet-TCN:

- Mean square error is used as the loss function.
- The *Adam* algorithm with default hyperparameters is adopted for the gradient descent optimisation [54].
- Gradients are clipped between $[-1, 1]$.
- The number of training examples in an epoch is equal to the number of clean speech recordings used in the training set, i.e., 70 537 for Deep Xi dataset and 11 572 for DEMAND Voice Bank dataset.
- A mini-batch size of 8 training examples is used.
- For the Deep Xi dataset, the noisy speech signals are generated on the fly as follows: each clean speech recording is corrupted with a randomly selected noise recording at a randomly selected SNR level (-10 to +20 dB, in 1 dB increments).
- For the DEMAND Voice Bank dataset, the noisy speech signals for the training set are formed by mixing each clean speech recording with a random section of a randomly selected noise recording at a random SNR level from the set $\{0, 5, 10, 15\}$ (dB). This creates a set of 11 572 noisy speech signals for training.
- For the Deep Xi dataset, we employ early stopping with a patience of 30 epochs. Using this strategy, training was terminated at epoch 168.
- No early stopping is set for the DEMAND Voice Bank dataset as it has no validation set (Table 2). Instead, a total of 125 epochs were used to train DeepLPC.

E. COMPLEXITY AND CONVERGE ANALYSIS OF DeepLPC-ResNet-TCN

The complexity of a DNN depends on the number of training parameters, where ResNet-TCN has 2.1 million parameters. This is markedly less than other models used for speech enhancement, such as the residual LSTM (ResLSTM) from [26], which employs 10 million parameters. This also allows for a significant speedup in training time, with ResNet-TCN taking 40 minutes per epoch when compared to 8 hours per epoch for the ResLSTM network on the Deep Xi dataset (using an NVIDIA GTX 1080 Ti GPU).

Next, we analyze the convergence of the mean squared error between the predicted and true values for the training and validation data sets of DeepLPC-ResNet-TCN, as shown in Figure 6. It can be seen that the mean squared error reduces for the training as well as the validation data after each epoch, until converging at around epoch 140. As the early stopping criterion with a patience of 30 is used, epoch 138 is chosen for testing.

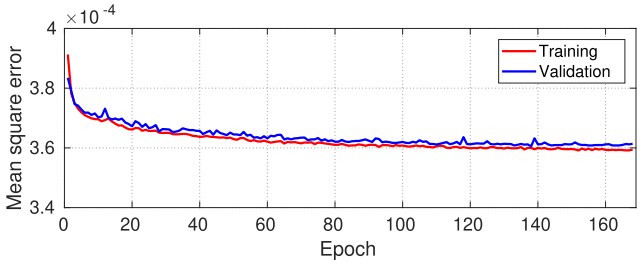

FIGURE 6. Mean square error between the predicted and true values for the training and validation data sets of DeepLPC-ResNet-TCN.

F. SD LEVEL EVALUATION

The frame-wise spectral distortion (SD) (dB) [29] is used to evaluate the accuracy of the LPC estimates produced by DeepLPC. Specifically, the estimated clean speech LPCs are evaluated. SD for l^{th} frame, D_l (in dB) is defined as the root-mean-square-difference between the LPC power spectrum estimate in dB, $\hat{P}_s(l,m)_{[dB]}$, and the oracle case in dB, $P_s(l,m)_{[dB]}$ as:

$$D_l = \sqrt{\frac{1}{M}\sum_{m=0}^{M-1}\left[P_s(l,m)_{[dB]} - \hat{P}_s(l,m)_{[dB]}\right]^2}. \quad (28)$$

G. OBJECTIVE QUALITY AND INTELLIGIBILITY MEASURES

Objective measures are used to evaluate the quality and intelligibility of the enhanced speech with respect to the corresponding clean speech. The objective quality and intelligibility measures used in this paper are given in Table 4.

H. SPECTROGRAM EVALUATION

We also compare the enhanced speech spectrograms of the proposed SEA to that of recent AKF SEAs to visually analyse the level of *residual* noise as well as *distortion*. For this purpose, we generate a set of stimuli by corrupting utterances *sp05* and *sp27* from the NOIZEUS corpus [1, Chapter 12]. The reference transcript for utterance *sp05* is: "*Wipe the grease off his dirty face*", and is corrupted with *voice babble* at 5 dB. The reference transcript for utterance *sp27* is: "*Bring your best compass to the third class*", and is corrupted with *factory* at 5dB. Utterance *sp05* and *sp27* were uttered by a male and female, respectively.

I. SUBJECTIVE EVALUATION

Subjective evaluation was carried out through a series of blind AB listening tests [5, Section 3.3.4]. To perform the tests, the stimuli set described in Section IV-H was used. The enhanced speech produced by eight SEAs as well as the corresponding clean speech and noisy speech signals were played as stimuli pairs to the listeners. Specifically, the test is performed on a total of 180 stimuli pairs (90 for each utterance) played in a random order to each listener, excluding the comparisons between the same method.

After listening to a stimuli pair, the listeners' preference was determined by the selection of one of three options. The first and second options indicated a preference for one of the two stimuli, while the third option indicated an equal preference for both stimuli. For pairwise scoring, 100% of the score is given to the preferred method, whilst 0% is given to the other. 50% of the score is given to each method for equal preference. The participants could re-listen to the stimuli pair if required. Ten English speaking listeners participate in the blind AB listening tests.[3] The average of the scores given by the listeners, termed as mean subjective preference score (%), is used to subjectively compare the SEAs.

TABLE 4. Objective measures, what each assesses, and the range of their scores. For each measure, higher is better.

Measure	Assesses	Range
CSIG [55]	Quality	$[1, 5]$
CBAK [55]	Quality	$[1, 5]$
COVL [55]	Quality	$[1, 5]$
PESQ [56]	Quality	$[-0.5, 4.5]$
STOI [57]	Intelligibility	$[0, 100]\%$
SI-SDR [58]	Quality	$[-\infty, \infty]$
SegSNR [59]	Quality	$[-\infty, \infty]$

J. SPECIFICATIONS OF THE COMPETITIVE SEAs

The performance of the proposed SEA is compared to the following SEAs (the following notation is used for convenience: (p, q) is the order of $\{a_i\}$ and $\{b_k\}$, (σ_w^2, σ_u^2) are the prediction error variances of the speech and noise AR models, w_f is the analysis frame duration (ms), and s_f is the analysis frame shift (ms)).

1) **Noisy**: speech corrupted with additive noise.
2) **AKF-Oracle:** AKF, where $(\{a_i\}, \sigma_w^2)$ and $(\{b_k\}, \sigma_u^2)$ are computed from the clean speech and the noise signal, where $p = 16$, $q = 16$, $w_f = 32$ ms, $s_f = 16$ ms, and a rectangular window is used for framing.
3) **DNN-LPC-KF:** KF-based SEA, where $(\{\hat{a}_i\}, \hat{\sigma}_w^2)$ are estimated using a DNN (C-DNN3) [29] and $\hat{\sigma}_v^2$ is computed from the first noisy speech frame, $y(0, l)$ (with the assumption that it is silent), $p = 12$, $w_f = 20$ ms, $s_f = 0$ ms. A rectangular window is used for framing.
4) **LSTM-CKFS** [33]: AKF constructed using $(\{\hat{a}_i\}, \hat{\sigma}_w^2)$ and $(\{\hat{b}_k\}, \hat{\sigma}_u^2)$ computed using an LSTM and ML-based approaches, followed by post subtraction using the multi-band SS method [4], where $p = 12$, $q = 12$, $w_f = 20$ ms, $s_f = 0$ ms, and a rectangular window is used for framing.
5) **EEUE-FCNN** [28]: End-to-end utterance enhancement using a fully convolutional neural network.
6) **IAM-IFD** [24]: Phase-aware DNN for speech enhancement, where $w_f = 20$ ms, $s_f = 5$ ms, and the Hamming window is used for analysis and synthesis.
7) **Deep Xi-KF** [32]: KF-based SEA, where $\hat{\sigma}_v^2$ is estimated using the DeepMMSE framework [27] and $(\{\hat{a}_i\}, \hat{\sigma}_w^2)$ are computed from pre-whitened speech corresponding to each noisy speech frame, where $p = 10$,

[3]The AB listening tests were conducted with approval from the Griffith University's Human Research Ethics Committee: database protocol number 2018/671.

TABLE 5. Average SD (dB) level comparison for each of the LPC estimation methods on NOIZEUS dataset as described in Table 3. The boldface represent the lowest SD level.

		SNR level (dB)				
Noise	Methods	-5	0	5	10	15
Voice babble	Noisy	22.05	18.29	14.86	13.80	11.87
	DNN-LPC [29]	16.72	15.98	13.24	12.76	10.79
	LSTM-CKFS [33]	15.91	14.51	12.11	11.89	9.23
	Deep Xi-KF [32]	14.95	13.88	11.81	10.31	9.11
	DeepLPC	**11.89**	**10.49**	**8.73**	**7.33**	**6.51**
Street	Noisy	20.21	16.39	14.43	13.88	12.45
	DNN-LPC [29]	13.41	12.25	11.68	11.18	10.87
	LSTM-CKFS [33]	12.57	11.05	10.78	10.35	9.86
	Deep Xi-KF [32]	11.66	10.51	9.74	9.21	8.95
	DeepLPC	**9.21**	**8.74**	**7.59**	**6.91**	**5.89**
Factory	Noisy	29.46	25.21	21.16	18.36	16.83
	DNN-LPC [29]	18.74	17.15	16.47	15.79	14.67
	LSTM-CKFS [33]	16.39	15.91	14.61	13.60	13.12
	Deep Xi-KF [32]	15.10	14.98	13.87	12.72	12.33
	DeepLPC	**12.29**	**10.89**	**9.48**	**8.21**	**7.89**
F16	Noisy	28.81	24.56	20.54	17.78	15.32
	DNN-LPC [29]	18.93	17.78	16.55	15.23	13.22
	LSTM-CKFS [33]	16.78	15.36	14.65	13.13	12.78
	Deep Xi-KF [32]	14.21	13.01	12.59	11.96	10.81
	DeepLPC	**12.13**	**10.46**	**9.49**	**8.63**	**7.83**

$w_f = 32$ ms, $s_f = 16$ ms, and a rectangular window is used for framing.

8) **Deep Xi-ResNet-TCN-MMSE-LSA**: A ResNet-TCN is incorporated within the Deep Xi framework to estimate the *a priori* SNR. The estimated *a priori* SNR is then employed by the MMSE-LSA estimator [9], where $w_f = 32$ ms, $s_f = 16$ ms, and a square-root-Hann window is used for analysis and synthesis.
9) **Proposed:** AKF constructed from $(\{\hat{a}_i\}, \hat{\sigma}_w^2)$ and $(\{\hat{b}_k\}, \hat{\sigma}_u^2)$ computed using DeepLPC framework, where $p = 16$, $q = 16$, $w_f = 32$ ms, $s_f = 16$ ms, and a rectangular window is used for framing.

V. RESULTS AND DISCUSSIONS

A. SD LEVEL COMPARISON

The average SD levels (found over all frames for each test condition on NOIZEUS dataset) attained by DeepLPC are given in Table 5.

It can be seen that for both real-world non-stationary (*voice babble* and *street*) and coloured (*factory* and *f16*) noise conditions, DeepLPC is able to produce lower SD levels than the competing deep learning-based LPC estimation methods. Amongst the competing methods, Deep Xi-KF [17] produced the lowest SD levels. The SD levels for noisy speech indicate the upper bounds of the SD level.

The average SD level for each method found on the DEMAND Voice Bank test set is shown in Table 6. It can be seen that DeepLPC produced a lower SD level than the competing methods. Deep Xi-KF [32] produced the next lowest SD level, followed by LSTM-CKFS [33], and DNN-LPC [29]. In light of the comparative study, the lower SD levels attained by DeepLPC demonstrate that it produces clean speech LPC estimates with less bias than previous methods.

TABLE 6. Average SD (dB) level comparison for each of the LPC estimation methods on DEMAND Voice Bank test set as described in Table 3. The boldface represent the lowest SD level.

Methods	Average SD Level (dB)
Noisy	20.67
DNN-LPC [29]	15.13
LSTM-CKFS [33]	13.26
Deep Xi-KF [32]	11.35
DeepLPC	**8.13**

B. OBJECTIVE QUALITY EVALUATION

Figure 7 shows the average PESQ score for each SEA over a range of conditions. It can be seen that AKF-Oracle exhibits the highest PESQ score for all of the tested conditions. This is due to $(\{a_i\}, \sigma_w^2)$ and $(\{b_k\}, \sigma_u^2)$ being computed from the clean speech and the noise signal, which is unobserved in practice. Thus, AKF-Oracle provides an indication of the upper-bound for the AKF in terms of PESQ score. Conversely, the average PESQ score for Noisy indicates the lower bound of the PESQ score for each of the tested conditions. The proposed SEA consistently produces a higher PESQ score than the competing SEAs across the tested conditions. Deep Xi-ResNet-TCN-MMSE-LSA produced the next best PESQ scores for each of the tested conditions. In light of this comparative study, it is evident that the proposed SEA produces higher quality enhanced speech than that of the competing SEAs across the tested conditions. This is due to the lower bias exhibited by the LPC estimates of DeepLPC, as demonstrated in the previous section.

C. OBJECTIVE INTELLIGIBILITY EVALUATION

Figure 8 shows the average STOI score for each SEA. As in Section V-B, the AKF-Oracle method achieves the highest STOI score for each tested condition. Amongst the SEAs, the proposed method attained the highest average STOI score for each tested condition. When analyzing the performance of the competing SEAs, Deep Xi-ResNet-TCN-MMSE-LSA attained the next best average STOI scores for the tested conditions. Again, the proposed method outperformed the competing SEAs, producing more intelligible enhanced speech for the tested conditions. The lower bias exhibited by the LPC estimates of DeepLPC results in not only higher quality enhanced speech, but also more intelligible enhanced speech than the competing SEAs.

D. OBJECTIVE EVALUATION FOR MULTIPLE OBJECTIVE MEASURES

In this section, we perform an analysis of the performance improvement of the proposed method over the competing methods for the objective measures described in Table 4, including CSIG, CBAK, COVL, PESQ, STOI, SegSNR, and SI-SDR. The mean objective evaluation results for NOIZEUS dataset and DEMAND Voice Bank test set are shown

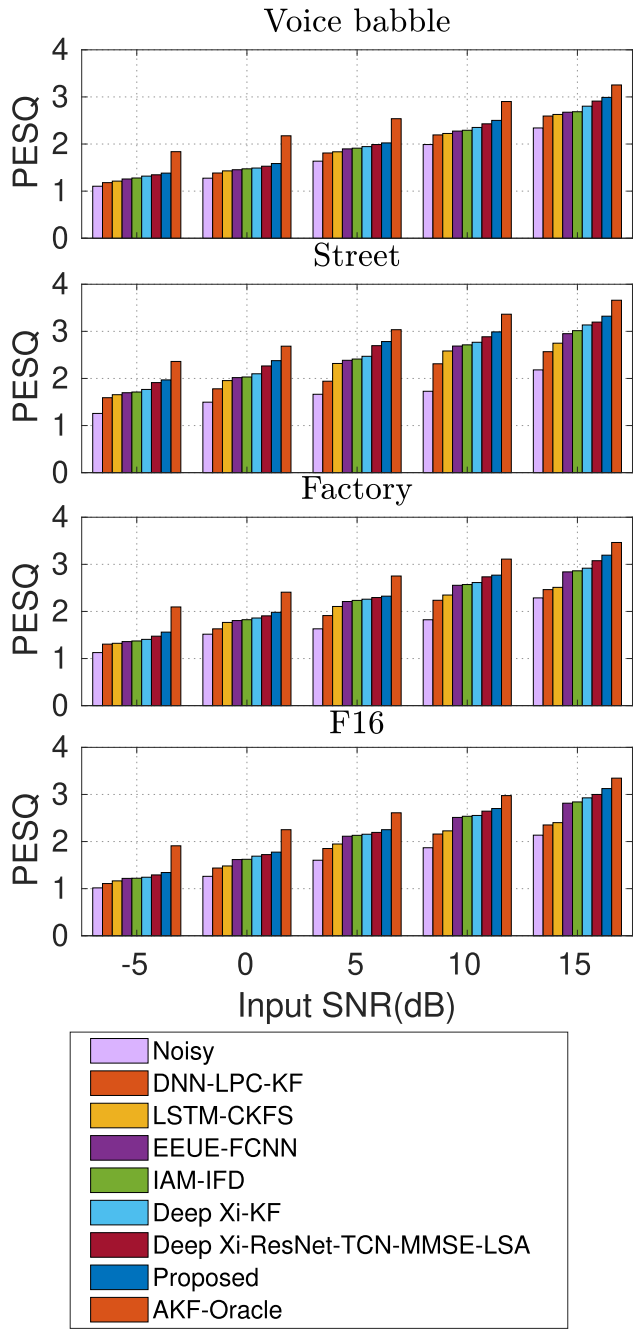

FIGURE 7. Average PESQ score for each SEA found over all frames for each condition in NOIZEUS dataset (Table 3).

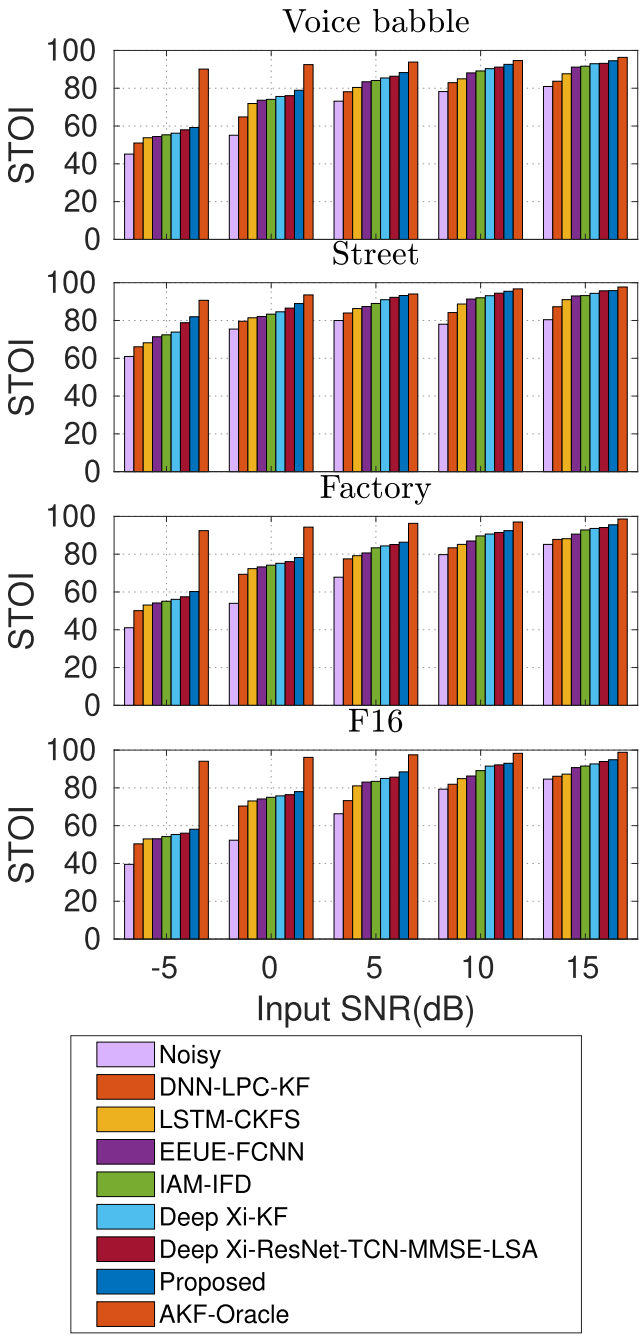

FIGURE 8. Average STOI score for each SEA found over all frames for each condition in NOIZEUS dataset (Table 3).

in Tables 7 and 8, respectively. It can be seen that AKF-Oracle produces the highest scores for all measures, which can be thought of as the upper boundary of performance. Noisy produced the lowest scores for all measures, indicating the lower boundary of performance. When comparing the performance of the proposed method, it shows a consistent CSIG, CBAK, COVL, PESQ, STOI, SegSNR, and SI-SDR score improvement over the competing methods. This again demonstrates that DeepLPC produces enhanced speech at a higher quality and intelligibility than any of the competing methods.

E. SPECTROGRAM ANALYSIS

This section analyzes the enhanced speech spectrograms produced by each of the SEAs for the test conditions specified in Section IV-H. Specifically, Figure 9 (i) shows the spectrogram of clean speech (male utterance *sp05*). The clean speech is corrupted by *voice babble* noise at an SNR level of 5 dB to create the noisy speech shown in Figure 9 (ii). This is a particularly tough condition for speech enhancement since the background noise exhibits characteristics similar to the speech produced by the target speaker. The enhanced speech produced by DNN-LPC-KF is shown in

TABLE 7. Mean objective scores on the NOIZEUS dataset in terms of CSIG, CBAK, COVL, PESQ, STOI, SegSNR, and SI-SDR. Apart from AKF-Oracle, the highest score amongst the methods for each measure is given in boldface.

Methods	CSIG	CBAK	COVL	PESQ	STOI	SegSNR	SI-SDR
Noisy speech	2.41	2.27	2.12	1.64	67.87	0.89	6.39
DNN-LPC-KF	2.59	2.49	2.34	1.89	74.60	6.33	10.72
LSTM-CKFS	2.63	2.55	2.42	1.99	77.58	6.54	11.15
EEUE-FCNN	2.76	2.66	2.56	2.05	79.45	6.93	11.59
IAM-IFD	2.95	2.72	2.63	2.11	80.64	7.01	11.88
Deep Xi-KF	3.11	2.83	2.72	2.16	81.89	7.14	12.15
Deep Xi-ResNet-TCN-MMSE-LSA	3.38	3.02	2.81	2.22	82.05	7.67	13.39
Proposed	**3.49**	**3.17**	**2.95**	**2.35**	**84.71**	**8.78**	**14.44**
AKF-Oracle	4.21	4.07	3.97	2.74	95.18	10.87	16.43

TABLE 8. Mean objective scores on the DEMAND Voice Bank test set in terms of CSIG, CBAK, COVL, PESQ, STOI, SegSNR, and SI-SDR. Apart from AKF-Oracle, the highest score amongst the competing methods for each measure is given in boldface.

Methods	CSIG	CBAK	COVL	PESQ	STOI	SegSNR	SI-SDR
Noisy speech	3.50	2.47	2.73	1.99	91.53	1.71	8.39
DNN-LPC-KF	3.61	2.79	2.91	2.57	91.79	7.33	16.68
LSTM-CKFS	3.63	2.85	2.93	2.61	91.87	7.44	16.75
EEUE-FCNN	3.67	2.87	3.02	2.64	92.03	7.64	17.19
IAM-IFD	3.74	2.96	3.11	2.69	92.13	7.67	17.32
Deep Xi-KF	3.90	3.09	3.23	2.74	92.34	8.21	17.55
Deep Xi-ResNet-TCN-MMSE-LSA	4.19	3.40	3.52	2.83	93.31	9.03	17.72
Proposed	**4.27**	**3.48**	**3.61**	**2.97**	**94.44**	**9.19**	**17.98**
AKF-Oracle	4.54	4.12	4.19	3.21	96.13	11.43	20.22

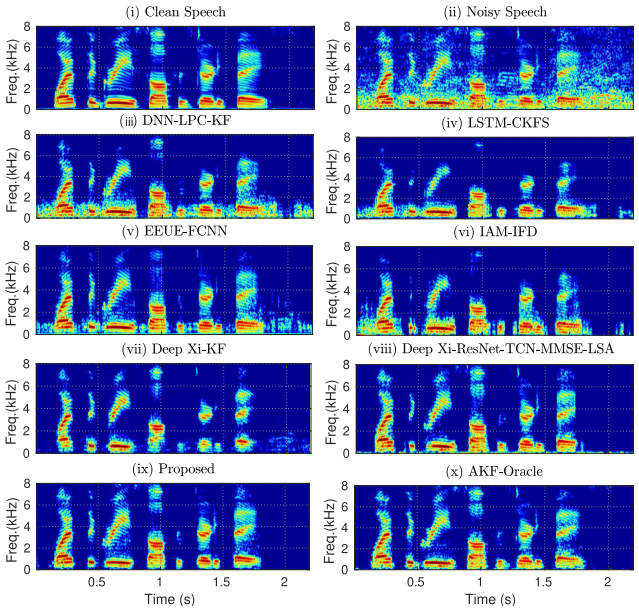

FIGURE 9. Spectrograms of: (i) clean speech (utterance sp05), (ii) noisy speech ((a) corrupted with 5 dB *voice babble* noise), and (iii)-(x) enhanced speech produced by each SEA.

Figure 9 (iii). It can be seen that DNN-LPC-KF significantly reduces the amount of background noise in the noisy speech. The enhanced speech produced by LSTM-CKFS [33] (Figure 9 (iv)) contains less residual background noise than that of DNN-LPC-KF (Figure 9 (iii)); however, suffers from significant speech distortion. Figure 9 (v) shows the enhanced speech produced by EEUE-FCNN [28]. This method produced less distorted speech than LSTM-CKFS [33] (Figure 9 (iv)); however, residual background noise still remains. The enhanced speech produced by IAM-IFD [24] (Figure 9 (vi)) shows less speech distortion and residual background noise than EEUE-FCNN (Figure 9 (v)). Less residual background noise is present in the enhanced speech produced by Deep Xi-KF [32] (Figure 9 (vii)) than IAM-IFD [24] (Figure 9 (vi)), however, the speech is more distorted. Deep Xi-ResNet-TCN-MMSE-LSA produced less distorted speech (Figure 9 (viii)) than that of Deep Xi-KF [32] (Figure 9 (vii)). The enhanced speech produced by the proposed method is shown in Figure 9 (ix). It can be seen that it produces less residual background noise and speech distortion than Deep Xi-ResNet-TCN-MMSE-LSA (Figure 9 (viii)). Finally, the enhanced speech produced by the AKF-Oracle method is shown in Figure 9 (x). The enhanced speech of AKF-Oracle is most similar to the clean speech in Figure 9 (i). This is due to AKF-Oracle using the clean speech and noise (unobserved in practice) for LPC parameter estimation.

Figure 10 shows the enhanced speech spectrograms of each SEA for a real-world coloured noise condition. The spectrogram of the clean speech (female utterance *sp27*) is shown in Figure 10 (i). The clean speech is corrupted by factory noise at an SNR level of 5 dB to generate the noisy speech shown in Figure 10 (ii). The enhanced speech produced by DNN-LPC-KF (Figure 10 (iii)) contains significant residual background noise. The enhanced speech produced by LSTM-CKFS [33] (Figure 10 (iv)) shows less residual background noise than DNN-LPC-KF (Figure 10 (iii)), however, it still suffers from significant speech distortion. The enhanced speech of EEUE-FCNN has less speech distortion and residual background noise (Figure 10 (v)) than LSTM-CKFS [33] (Figure 10 (iv)). IAM-IFD [24] produced enhanced speech with less residual background noise (Figure 10 (vi)) than EEUE-FCNN (Figure 10 (v)), although there remains noticeable speech distortion. The enhanced speech produced by Deep Xi-KF [32] (Figure 10 (vii)) contains less residual background noise, as well as speech distortion than IAM-IFD [24] (Figure 10 (vi)). Less residual background noise as well as speech distortion remains in the enhanced speech produced by Deep

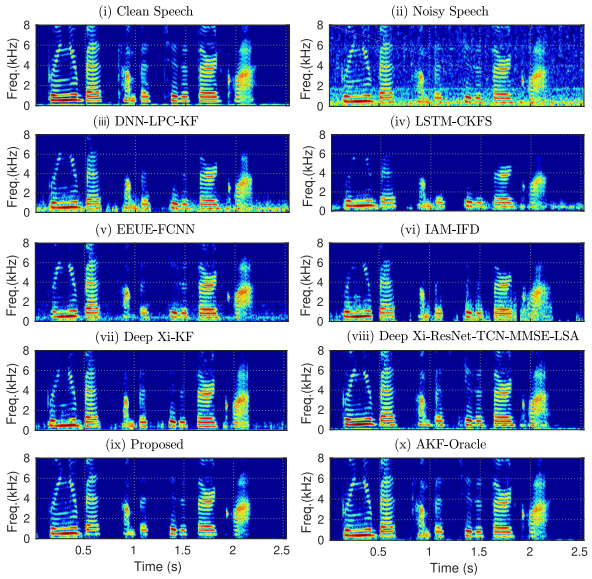

FIGURE 10. Spectrograms of: (i) clean speech (utterance sp27), (ii) noisy speech ((a) corrupted with 5 dB *factory* noise), and (iii)-(x) enhanced speech produced by each SEA.

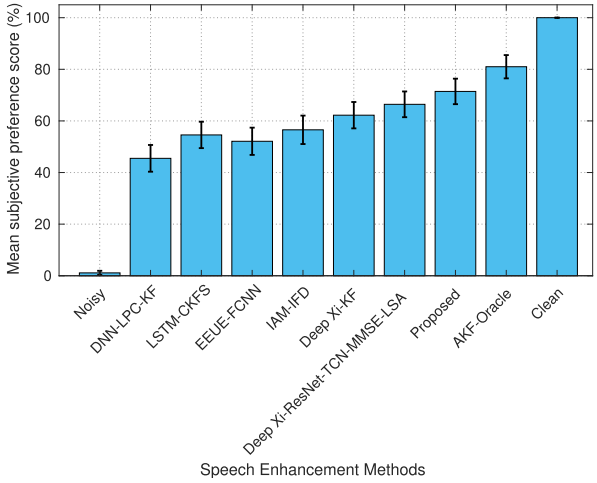

FIGURE 11. The mean preference score (%) comparison between the proposed and benchmark SEAs for male utterance sp05 corrupted with 5 dB non-stationary *voice babble* noise. The error bars indicate the standard deviation of the scores.

Xi-ResNet-TCN-MMSE-LSA (Figure 10 (viii)) than that of Deep Xi-KF [32] (Figure 10 (vii)). The enhanced speech produced by the proposed method (Figure 10 (ix)) contains the least amount of residual background noise, as well as speech distortion amongst the SEAs and is most similar to the enhanced speech produced by AKF-Oracle (Figure 10 (x)).

F. SUBJECTIVE EVALUATION

The mean subjective preference scores (%) for each SEA are shown in Figures 11 and 12. The non-stationary (*voice babble*) noise experiment in Figure 11 reveals that the proposed method is widely preferred (71%) by the listeners to that of the competing methods, apart from the clean speech (100%) and AKF-Oracle (81%). Deep Xi-ResNet-TCN-MMSE-LSA is found to be the next most preferred method (66%), followed by Deep Xi-KF (62%), IAM-IFD (56%), LSTM-CKFS

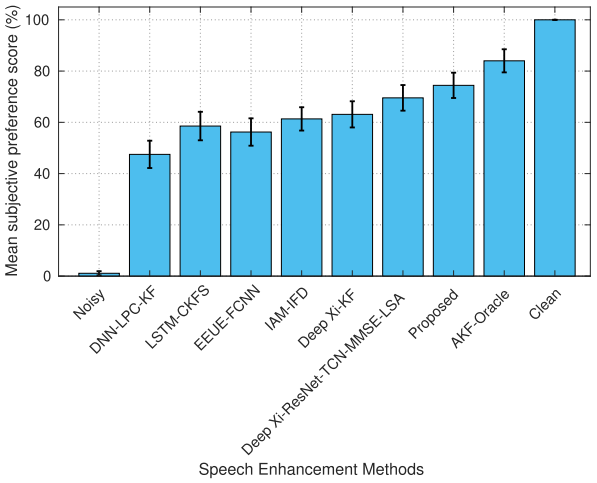

FIGURE 12. The mean preference score (%) comparison between the proposed and benchmark SEAs for female utterance sp27 corrupted with 5 dB coloured *factory* noise. The error bars indicate the standard deviation of the scores.

(57%), and then EEUE-FCNN (52%). LSTM-CKFS [33] was preferred by the listeners more than EEUE-FCNN [28], even though EEUE-FCNN attained higher objective scores (Section V-B and V-C). This may be due to the fact that LSTM-CKFS [33] demonstrates superior noise suppression in regions of speech than EEUE-FCNN [28], as indicated in [16]. DNN-LPC-KF was given the lowest preference score (46%) amongst the SEAs.

The blind AB listening test results for the coloured (*factory*) noise condition is shown in Figure 12. It can be seen that the proposed method achieves a better preference score (74%) than the competing methods, except for clean speech (100%) and AKF-Oracle (84%). As in the previous experiment, Deep Xi-ResNet-TCN-MMSE-LSA was the next most preferred method (70%), followed by Deep Xi-KF [32] (63%), IAM-IFD [24] (61%), LSTM-CKFS [33] (59%), EEUE-FCNN [28] (56%), and then DNN-LPC-KF (48%). In light of the blind AB listening tests, it is evident to say that the enhanced speech produced by the proposed method exhibits the best perceived quality amongst all of the competing SEAs, for both male and female utterances corrupted by real-world non-stationary and coloured noise sources.

VI. CONCLUSION

We propose the DeepLPC framework to jointly estimate clean speech and noise LPC parameters for the AKF. Specifically, DeepLPC maps each frame of the noisy speech magnitude spectrum to the LPC power spectrum of the clean speech and noise signal. Applying the inverse Fourier transform to the estimated LPC power spectra yields the corresponding autocorrelation matrices. Then, the application of the Levinson-Durbin recursion to the autocorrelation matrices yields the LPC estimates and the prediction error variances of the speech and noise signal. The Deep Xi and DEMAND Voice Bank datasets were used to train DeepLPC separately to ensure generalizing capability to unseen conditions. NOIZEUS dataset is used to evaluate the performance

of DeepLPC trained with Deep Xi dataset. In addition, DeepLPC trained with DEMAND Voice Bank dataset is evaluated using DEMAND Voice Bank test set. It was shown that the estimated clean speech LPCs produced by the proposed DeepLPC framework exhibit a lower SD level than recent deep learning methods for both of the test sets. Moreover, the AKF constructed with the clean speech and noise LPC parameters derived from the DeepLPC leads to the capability of speech enhancement in real-world noise conditions. Extensive objective and subjective testing on NOIZEUS and DEMAND Voice Bank test sets demonstrate that the proposed method outperforms the competing deep learning-based methods in various noise conditions for a wide range of SNR levels.

Though the proposed DeepLPC framework achieves the lowest SD level for LPC estimation, there is still room for improvement. For example, in the proposed DeepLPC framework, we have incorporated ResNet-TCN. However, the multi-head self-attention network (MHANet) has been shown to outperform ResNet-TCN for speech enhancement [60]. Motivated by this, the DeepLPC framework will employ the MHANet in future work to facilitate a further improvement in LPC estimation for AKF-based speech enhancement.

REFERENCES

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

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 4, Apr. 1979, pp. 208–211.

[4] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 4, May 2002, pp. 4160–4164.

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

[6] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, vol. 2, May 1996, pp. 629–632.

[7] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, "A two-step noise reduction technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, May 2004, pp. 289–292.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[10] K. Paliwal, B. Schwerin, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Commun.*, vol. 54, no. 2, pp. 282–305, Feb. 2012.

[11] B. M. Mahmmod, A. R. Ramli, T. Baker, F. Al-Obeidat, S. H. Abdulhussain, and W. A. Jassim, "Speech enhancement algorithm based on super-Gaussian modeling and orthogonal polynomials," *IEEE Access*, vol. 7, pp. 103485–103504, 2019.

[12] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 12, Apr. 1987, pp. 177–180.

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

[14] G. J. Brown and D. Wang, *Separation of Speech by Computational Auditory Scene Analysis*. Berlin, Germany: Springer, 2005.

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[16] 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 Commun.*, vol. 105, pp. 62–76, Dec. 2018.

[17] S. K. Roy, A. Nicolson, and K. K. Paliwal, "Deep learning with augmented Kalman filter for single-channel speech enhancement," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Oct. 2020, pp. 1–5.

[18] S. K. Roy, W.-P. Zhu, and B. Champagne, "Single channel speech enhancement using subband iterative Kalman filter," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2016, pp. 762–765.

[19] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

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

[21] N. Saleem, M. I. Khattak, M. Al-Hasan, and A. B. Qazi, "On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks," *IEEE Access*, vol. 8, pp. 160581–160595, 2020.

[22] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 708–712.

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

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

[25] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

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

[27] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1404–1415, 2020.

[28] S.-W. Fu, T.-W. 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 Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[29] C. Pickersgill, S. So, and B. Schwerin, "Investigation of DNN prediction of power spectral envelopes for speech coding & ASR," in *Proc. 17th Speech Sci. Technol. Conf. (SST)*, Sydney, NSW, Australia, Dec. 2018, pp. 1–5.

[30] H. Yu, Z. Ouyang, W.-P. Zhu, B. Champagne, and Y. Ji, "A deep neural network based Kalman filter for time domain speech enhancement," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–5.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[32] S. K. Roy, A. Nicolson, and K. K. Paliwal, "A deep learning-based Kalman filter for speech enhancement," in *Proc. Interspeech*, Oct. 2020, pp. 2692–2696.

[33] H. Yu, W.-P. Zhu, and B. Champagne, "Speech enhancement using a DNN-augmented colored-noise Kalman filter," *Speech Commun.*, vol. 125, pp. 142–151, Dec. 2020.

[34] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech Language Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[35] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*. Hoboken, NJ, USA: Wiley, 2006.

[36] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4266–4269.

[37] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[38] S. So, "Efficient block quantisation for image and speech coding," Ph.D. dissertation, Griffith Univ., Brisbane, QLD, Australia, 2005. [Online]. Available: https://research-repository.griffith.edu.au/handle/10072/366625
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," *CoRR*, vol. abs/1502.01852, Feb. 2015.
[40] L. J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, vol. abs/1607.06450, Jul. 2016.
[41] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. van den Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," *CoRR*, vol. abs/1610.10099, Mar. 2016.
[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
[43] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," Centre Speech Technol. Res., Univ. Edinburgh, Edinburgh, U.K., Tech. Rep., 2017.
[44] 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, Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. 4930, Feb. 1993, vol. 93.
[45] 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 *Proc. Interspeech*, 2010, pp. 3110–3113.
[46] G. Hu, "100 nonspeech environmental sounds," Dept. Comput. Sci. Eng., Ohio State Univ., Columbus, OH, USA, Tech. Rep., 2004.
[47] 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 *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2204–2208.
[48] F. Saki and N. Kehtarnavaz, "Automatic switching between noise classification and speech enhancement for hearing aid devices," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 736–739.
[49] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, Oct. 2015.
[50] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA Held Jointly With Conf. Asian Spoken Lang. Res. Eval. (O-COCOSDA/CASLRE)*, Nov. 2013, pp. 1–4.
[51] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. 9th ISCA Speech Synth. Workshop*, Sep. 2016, pp. 146–152.
[52] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust.*, vol. 19, no. 1, 2013, Art. no. 035081.
[53] A. Nicolson, "Deep Xi dataset," *IEEE Dataport*, 2020, doi: 10.21227/3adt-pb04.
[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.
[55] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[56] 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," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2001, pp. 749–752.
[57] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[58] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR—Half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 626–630.
[59] P. Mermelstein, "Evaluation of a segmental SNR measure as an indicator of the quality of ADPCM coded speech," *J. Acoust. Soc. Amer.*, vol. 66, no. 6, pp. 1664–1667, Dec. 1979.
[60] A. Nicolson and K. K. Paliwal, "Masked multi-head self-attention for causal speech enhancement," *Speech Commun.*, vol. 125, pp. 80–96, Dec. 2020.

SUJAN KUMAR ROY (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science and engineering from the University of Rajshahi, Bangladesh, in 2008 and 2010, respectively, and the Master of Applied Science (M.A.Sc.) degree in electrical and computer engineering from Concordia University, Canada, in May 2016. He is currently pursuing the Ph.D. degree with the School of Engineering, Griffith University, Brisbane, Australia. His research interests include speech processing, machine learning, and data science.

AARON NICOLSON was born in Brisbane, Australia, in 1994. He received the B.Eng. (Hons.) and Ph.D. degrees from Griffith University, Brisbane, in 2016 and 2020, respectively. He is currently a Postdoctoral Research Fellow with the Australian eHealth Research Centre, CSIRO. His research interests include speech, natural language, image, and multimodal processing using deep learning.

KULDIP K. PALIWAL was born in Aligarh, India, in 1952. He received the B.S. degree from Agra University, Agra, India, in 1969, the M.S. degree from Aligarh Muslim University, Aligarh, in 1971, and the Ph.D. degree from Bombay University, Bombay, India, in 1978. He has been carrying out research in the area of speech processing, since 1972. He has worked with a number of organizations, including Tata Institute of Fundamental Research, Bombay; the India Norwegian Institute of Technology, Trondheim, Norway; the University of Keele, U.K.; AT&T Bell Laboratories, Murray Hill, NJ, USA; AT&T Shannon Laboratories, Florham Park, NJ, USA; and Advanced Telecommunication Research Laboratories, Kyoto, Japan. Since July 1993, he has been a Professor with the School of Microelectronic Engineering, Griffith University, Brisbane, Australia. His current research interests include speech recognition, speech coding, speaker recognition, speech enhancement, face recognition, image coding, pattern recognition, and artificial neural networks. He has published more than 300 articles in these research areas. He is currently a fellow of the Acoustical Society of India. He has served as a Founding Member for the IEEE Signal Processing Society's Neural Networks Technical Committee, from 1991 to 1995, and the Speech Processing Technical Committee, from 1999 to 2003. He received the IEEE Signal Processing Society's Best (Senior) Paper Award for his paper on LPC quantization, in 1995. He was the General Co-Chair of the Tenth IEEE Workshop on Neural Networks for Signal Processing (NNSP2000). He has co-edited two books, *Speech Coding and Synthesis* (Elsevier) and *Speech and Speaker Recognition: Advanced Topics* (Kluwer). He was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, from 1994 to 1997 and from 2003 to 2004. He is on the Editorial Board of the *IEEE Signal Processing Magazine*. He also served as an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS, from 1997 to 2000. He served as the Editor-in-Chief of the *Speech Communication Journal* (Elsevier), from 2005 to 2011.