DeepMMSE: A Deep Learning Approach to MMSE-based Noise Power Spectral Density Estimation

Qiquan Zhang, Aaron Nicolson, Mingjiang Wang, Kuldip K. Paliwal, and Chenxu Wang

过去常使用MMSE的方法作为功率图的评估指标
然而，由于现有的先验信噪比(SNR)估计方法，它们缺乏跟踪高非平稳噪声源的能力。
这是由一个潜在的假设造成的，即噪声信号的变化速度比语音信号慢。

***Abstract*—An accurate noise power spectral density (PSD) tracker is an indispensable component of a single-channel speech enhancement system. Bayesian-motivated minimum mean-square error (MMSE)-based noise PSD estimators have been the most prominent in recent time. However, they lack the ability to track highly non-stationary noise sources due to current methods of *a priori* signal-to-noise (SNR) estimation. This is caused by the underlying assumption that the noise signal changes at a slower rate than the speech signal. As a result, MMSE-based noise PSD trackers exhibit a large tracking delay and produce noise PSD estimates that require bias compensation. Motivated by this, we propose an MMSE-based noise PSD tracker that employs a temporal convolutional network (TCN) *a priori* SNR estimator. The proposed noise PSD tracker, called DeepMMSE makes no assumptions about the characteristics of the noise or the speech, exhibits no tracking delay, and produces an accurate estimate that requires no bias correction. Our extensive experimental investigation shows that the proposed DeepMMSE method outperforms state-of-the-art noise PSD trackers and demonstrates the ability to track abrupt changes in the noise level. Furthermore, when employed in a speech enhancement framework, the proposed DeepMMSE method is able to outperform state-of-the-art noise PSD trackers, as well as multiple deep learning approaches to speech enhancement.**

***Availability*: DeepMMSE is available at: https://github.com/anicolson/DeepXi.**

***Index Terms*—Noise PSD tracking, minimum mean-square error (MMSE), DeepMMSE, speech enhancement, noise estimation, Deep Xi.**

I. Introduction

OVER the past decade, there has been a growing focus on speech processing applications, such as digital hearing aids, automatic speech recognition, speaker verification, and mobile communications. In real-world environments, speech may be degraded by non-stationary or coloured background disturbances, with examples including train, car, factory, and street noise. Background disturbances can severely affect the performance of a speech processing system. However, their robustness can be increased by first enhancing the noisy speech. A speech enhancement algorithm aims to suppress the noise and improve the overall perceived quality and intelligibility of the noisy speech. These algorithms can be broadly divided into two categories: single-channel [1]–[4] and multi-channel [5]–[8]. In this paper, we focus on single-channel speech enhancement algorithms.

Most single-channel speech enhancement algorithms require an estimate of the noise power spectral density (PSD). Over- or under-estimation of the noise PSD can lead to speech distortion or a large amount of residual noise in the enhanced speech. An early method for noise PSD estimation exploits a voice activity detector (VAD) to update the noise PSD estimate during speech absence [9]–[12]. Such methods are effective under fairly stationary noise conditions, but often fail to track non-stationary noise sources during speech activity. Moreover, it remains a particularly difficult task to achieve an accurate VAD in low signal-to-noise ratio (SNR) conditions.

A more advanced noise tracking method is minimum statistics (MS) [13], [14]. The MS method was motivated by the observation that the power of the noisy speech in each frequency bin often decays to the power level of the noise, even during speech presence. Thus, the MS method tracks the minimum value of the smoothed noisy speech power for each frequency bin within a finite time window. The found minimum is multiplied by a bias correction factor to give the final noise PSD estimate [14]. The length of the time window is a crucial parameter for the performance of the MS method. If the length is too small, the MS method may track the noisy speech PSD instead, leading to over-estimation. However, if the length is too large, a significant tracking delay may occur, especially for a fast increase in the noise level.

Another group of noise PSD tracking methods are time-recursive averaging algorithms, including the minima controlled recursive averaging (MCRA) method [15], the improved MCRA method (IMCRA) [16], and MCRA-2 [17]. These methods update the noise PSD estimate by recursively averaging the previously estimated noise PSD and the current noisy periodogram using a smoothing parameter, where the smoothing parameter is adjusted by the speech presence probability (SPP). The main distinction between the three algorithms is the method of SPP estimation. For MCRA and MCRA-2, the SPP estimate is found by taking the ratio between the smoothed noisy speech spectrum and its local minimum, and then thresholding against a certain value. MCRA and

Manuscript received date; revised date. This work was supported by the Basic Research Discipline Layout Project of Shenzhen under Grant JCYJ20170412151226061 and Grant JCYJ20170808110410773. (Joint corresponding authors: Aaron Nicolson, Mingjiang Wang.)

Qiquan Zhang and Mingjiang Wang are with the School of Electronic & Information Engineering, Harbin Institute of Technology, Shenzhen 518055, China (email: zhangqiquan_hit@163.com; mjwang@hit.edu.cn).

Aaron Nicolson and Kuldip K. Paliwal are with the Signal Processing Laboratory, Griffith University, Brisbane, Queensland 4111, Australia (email: aaron.nicolson@griffithuni.edu.au; k.paliwal@griffith.edu.au).

Chenxu Wang is with the School of Information Science & Engineering, Harbin Institute of Technology, Weihai 264209, China (email: wangchenxu@hit.edu.cn).

MRCA-2 use the MS method [14] and the continuous spectral minimum tracking technique [18] to search for the minimum, respectively. Moreover, fixed and frequency-dependent thresholds are employed in MCRA and MCRA-2, respectively. In IMCRA, the SPP estimate is computed from both an *a priori* speech absence probability (SAP) estimate and an *a priori* SNR estimate. The calculation of the *a priori* SAP involves two iterations of smoothing and minimum tracking. As these methods rely on minimum tracking, they also have difficulties tracking rapidly increasing noise levels.

Another class of noise PSD trackers is formulated from the perspective of Bayesian statistics. Several minimum mean-square error (MMSE)-based Bayesian noise PSD estimators exist in the literature [19]–[23], and are described in Section III. These noise trackers are based on an MMSE estimate of the noise periodogram, which is computed using the *a priori* and *a posteriori* SNR. Their noise tracking performance is predominantly affected by the accuracy of the *a priori* SNR estimate. Each MMSE-based noise PSD tracker uses a different method of *a priori* SNR estimation, with many based on temporal cepstral smoothing (TCS) [21], the decision-directed (DD) approach, or the maximum-likelihood (ML) method [24]. However, these methods assume that the noise changes at a slower rate than the speech signal. As a consequence, a delay is introduced during rapid changes of the instantaneous *a priori* SNR. Therefore, current MMSE-based noise PSD trackers are capable of tracking moderately non-stationary noise sources, but are unable to track highly non-stationary noise sources. In summary, the performance of current MMSE-based noise PSD estimators is limited by the method of *a priori* SNR estimation.

Recently, a deep learning approach to *a priori* SNR estimation, called Deep Xi [25], was used to significantly increase the performance of MMSE approaches to speech enhancement [24], [26]. Unlike its predecessors, it does not make any assumptions about the characteristics of the speech or noise, does not exhibit any tracking delay, and does not rely on bias compensation. The Deep Xi framework utilises a residual long short-term memory (ResLSTM) recurrent neural network (RNN) to estimate the *a priori* SNR directly from the noisy speech magnitude spectrum of a given time-frame (Deep Xi-ResLSTM). Deep Xi-ResLSTM was shown to be more accurate than previous *a priori* SNR estimators [24], [27]–[29]. As shown in [25], the speech enhancement performance of the *a priori* SNR as the training target was similar to that of the ideal ratio mask (IRM) as the training target. Furthermore, both the *a priori* SNR and the IRM as the training target outperformed the clean speech magnitude spectrum as the training target.

Here, we propose an MMSE-based noise PSD estimator that employs the Deep Xi framework for *a priori* SNR estimation. This is motivated by the following observations: 1) the bottleneck of current MMSE-based noise PSD estimators is the method of *a priori* SNR estimation, 2) Deep Xi-ResLSTM is significantly more accurate than previous *a priori* SNR estimatoin methods, and 3) the Deep Xi framework exhibits no tracking delay, does not make assumptions about the characteristics of the noise or the speech, and does not require bias compensation. We also improve the Deep Xi framework by replacing the ResLSTM network with a temporal convolutional network (TCN). It outperforms the ResLSTM network, while using 5x fewer parameters and avoiding the complexity of training a recurrent architecture. To the best of our knowledge, this is the first time a deep learning MMSE-based noise PSD estimator has been proposed.

The remainder of this paper will be structured as follows. Section II gives a detailed explanation of the used signal model and notation. A brief overview and analysis of the state-of-the-art MMSE-based noise PSD trackers is given in Section III. In Section V, the proposed MMSE-based noise PSD tacker is described. The experimental setup is described in Section VI. The results and discussion are presented in Section VII. In Section VIII, we provide the conclusions and future considerations.

II. Signal Model And Notation

In the time-domain, the noisy speech signal is given by

$$y[n] = s[n] + d[n], \tag{1}$$

where $s[n]$ and $d[n]$ denote the clean speech and uncorrelated additive noise, respectively, and n denotes the discrete-time index. The noisy speech is then analysed frame-wise using the short-time Fourier transform (STFT):

$$Y_l[k] = S_l[k] + D_l[k], \tag{2}$$

where $Y_l[k]$, $S_l[k]$, and $D_l[k]$ denote the complex-valued STFT coefficients of the noisy speech, the clean speech and the noise, respectively, for time-frame index l and discrete-frequency index k. We apply the standard assumption that $S_l[k]$ and $D_l[k]$ are statistically independent across time and frequency, and follow conditional zero-mean Gaussian distributions with spectral variances $E\{|S_l(k)|^2\} = \lambda_s[l,k]$, and $E\{|D_l(k)|^2\} = \lambda_d[l,k]$, where $E\{\cdot\}$ represents the statistical expectation operator. For convenience, l and k are omitted from the notation unless otherwise indicated. The polar form is used to express Y, S, and D: $Y = Re^{j\phi}$, $S = Ae^{j\varphi}$, and $D = Ne^{j\theta}$, where R, A, and N are the noisy speech, clean speech, and noise magnitude spectrums, respectively, and ϕ, φ, and θ are the noisy speech, clean speech, and noise phase spectrums, respectively. The *a priori* SNR and the *a posteriori* SNR are defined as

$$\xi = \frac{\lambda_s}{\lambda_d}, \tag{3}$$

and

$$\gamma = \frac{R^2}{\lambda_d}, \tag{4}$$

respectively. Here, an estimated quantity of a variable is denoted by a hat symbol in order to differentiate from the true value, e.g., $\widehat{N^2}$ is an estimate of the instantaneous noise spectral power N^2.

III. Overview of MMSE-based noise PSD estimators

The MMSE-based noise PSD estimators reported in [19]–[23] are briefly summarised in this section. They are all based on an MMSE estimate of the noise periodogram, which is defined as the conditional expectation: $E\{N^2|Y,\lambda_s,\lambda_d\}$. Using Bayes' rule, this can be written as

$$\begin{aligned}\widehat{N^2} &= E\left\{N^2|Y,\lambda_s,\lambda_d\right\},\\ &= \frac{\int_0^{+\infty}\int_0^{2\pi} n^2 f\left(Y|n,\theta,\lambda_s\right) f\left(n,\theta|\lambda_d\right) d\theta dn}{\int_0^{+\infty}\int_0^{2\pi} f\left(Y|n,\theta,\lambda_s\right) f\left(n,\theta|\lambda_d\right) d\theta dn}.\end{aligned} \tag{5}$$

Under the assumption that the STFT coefficients of clean speech and noise follow complex-Gaussian distributions, $f\left(Y|n,\theta,\lambda_s\right)$ and $f\left(n,\theta|\lambda_d\right)$ can be written as

$$f\left(Y|n,\theta,\lambda_s\right) = \frac{1}{\pi\lambda_s}\exp\left\{-\frac{\left|Y-ne^{j\theta}\right|^2}{\lambda_s}\right\}, \tag{6}$$

and

$$f\left(n,\theta|\lambda_d\right) = \frac{n}{\pi\lambda_d}\exp\left\{-\frac{n^2}{\lambda_d}\right\}. \tag{7}$$

Inserting Equations (6) and (7) into Equation (5), and using the integral relationship of the zeroth-order modified Bessel function of the first kind, $I_0(\cdot)$ [30, Eq. 8.406.3, 8.411.1], we obtain

$$\begin{aligned}\widehat{N^2} &= E\left\{N^2|Y,\lambda_s,\lambda_d\right\},\\ &= \frac{\int_0^{+\infty} n^3\exp\left(-n^2/\lambda\right) I_0\left(2nY/\lambda_d\right) dn}{\int_0^{+\infty} n\exp\left(-n^2/\lambda\right) I_0\left(2nY/\lambda_d\right) dn},\end{aligned} \tag{8}$$

where λ satisfies the relation $\lambda = \lambda_s\lambda_d/\left(\lambda_s+\lambda_d\right)$. Using [30, Eq. 6.631.1, 8.406.3, 9.212.1], the solution of Equation (8) is obtained as

$$\widehat{N^2} = \left[\frac{1}{(1+\xi)^2} + \frac{\xi}{(1+\xi)\gamma}\right] R^2. \tag{9}$$

A temporal recursive smoothing operation is then applied to $\widehat{N^2}$ in order to obtain the final noise PSD estimate:

$$\widehat{\lambda_d}[l,k] = \alpha_d\widehat{\lambda_d}[l-1,k] + (1-\alpha_d)\widehat{N^2}[l,k]. \tag{10}$$

From Equation (9), it is obvious that the MMSE-based Bayesian noise PSD estimator depends on the *a priori* and *a posteriori* SNR. As shown in [19]–[23], the noise PSD tracking performance is predominately affected by the accuracy of the *a priori* SNR estimate, $\hat{\xi}$.

In [19], the decision-directed (DD) approach [1] was employed to estimate the *a priori* SNR and a heuristic bias correction method was proposed to compensate the bias introduced by the DD estimator. In [20], a limited maximum likelihood (ML) estimator [1] was used to estimate the *a priori* SNR. A more sophisticated bias compensation term was also derived to correct the bias introduced by the ML estimate. Temporal cepstrum smoothing (TCS) was used in [21] to obtain a more accurate estimate of the *a priori* SNR, where a bias correction factor was also derived. In [22], the limited ML estimator was interpreted as a hard-decision, VAD-like estimator [20]. This was subsequently refined by using a soft-decision speech presence probability (SPP), which required no bias compensation term. The SPP was calculated using a fixed optimal *a priori* SNR. In [23], an MMSE speech spectral power estimator that incorporates speech presence uncertainty (SPU) and a bias correction factor was proposed to improve the DD approach, producing an improved tracking capability (ImMMSE).

IV. Deep Xi framework

Deep Xi is a deep learning approach to *a priori* SNR estimation [25]. The Deep Xi framework consists of two stages. For the first stage, a deep neural network (DNN) estimates a mapped version of the *a priori* SNR, $\hat{\bar{\boldsymbol{\xi}}}_l = \{\hat{\bar{\xi}}_l[0],\hat{\bar{\xi}}_l[1],\cdots,\hat{\bar{\xi}}_l[K-1]\}$, from the noisy speech magnitude spectrum, $\mathbf{R}_l = \{R[l,0],R[l,1],\cdots,R[l,K-1]\}$, where K is the number of discrete-frequency bins for each time-frame. For the second stage, the *a priori* SNR estimate, $\hat{\boldsymbol{\xi}}_l$, is computed from the mapped *a priori* SNR estimate, $\hat{\bar{\boldsymbol{\xi}}}_l$. The mapped *a priori* SNR, $\bar{\boldsymbol{\xi}}_l$, and the computation of the *a priori* SNR estimate during the second stage is described in the next subsection.

A. Mapped a priori SNR training target

The training target for a DNN within the Deep Xi framework is the mapped *a priori* SNR, as described in [25]. The mapped *a priori* SNR is a mapped version of the instantaneous *a priori* SNR. For the instantaneous case, the clean speech and noise of the noisy speech in Equation (1) are known completely. This means that $\lambda_s[l,k]$ and $\lambda_d[l,k]$ in Equation (3) can be replaced with the squared magnitude of the clean speech and noise spectral components, respectively.

In [25], the instantaneous *a priori* SNR (in dB), $\xi_{\mathrm{dB}}[l,k] = 10\log_{10}(\xi_l[k])$, 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 (CDF) of $\xi_{\mathrm{dB}}[l,k]$ was used as the map. It can be seen from [25, Fig. 2 (top)] that the distribution of ξ_{dB} for a given frequency component follows a normal distribution. It was thus assumed that $\xi_{\mathrm{dB}}[l,k]$ is distributed normally with mean μ_k and variance σ_k^2: $\xi_{\mathrm{dB}}[l,k] \sim \mathcal{N}(\mu_k,\sigma_k^2)$. The map is given by

$$\bar{\xi}_l[k] = \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{\xi_{\mathrm{dB}}[l,k]-\mu_k}{\sigma_k\sqrt{2}}\right)\right], \tag{11}$$

where $\bar{\xi}_l(k)$ is the mapped *a priori* SNR. Following [25], the statistics of $\xi_{\mathrm{dB}}[l,k]$ for each noisy speech spectral component are found over a sample of the training set.* During inference, the *a priori* SNR estimate is found from the mapped *a priori* SNR estimate as follows:

$$\hat{\xi}_l[k] = 10^{\left(\left(\sigma_k\sqrt{2}\mathrm{erf}^{-1}(2\hat{\bar{\xi}}_l[k]-1)+\mu_k\right)/10\right)}. \tag{12}$$

*The sample mean and variance of $\xi_{\mathrm{dB}}[l,k]$ for each noisy speech spectral component are found over 1 250 noisy speech signals created from the clean speech and noise training sets (Section VI-B). 250 randomly selected (without replacement) clean speech recordings are mixed with random sections of randomly selected (without replacement) noise recordings. Each of these are mixed at five different SNR levels: -5 to 15 dB, in 5 dB increments.

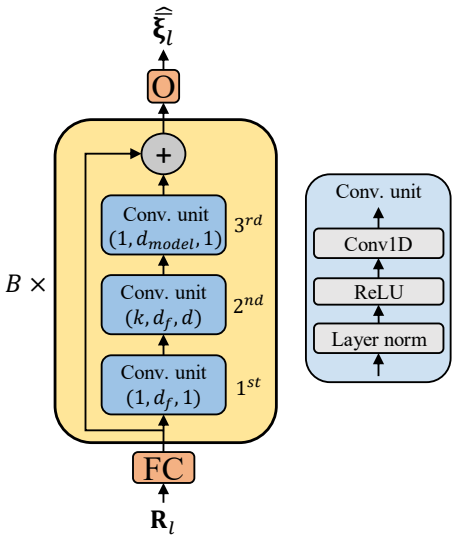


Fig. 1. Deep Xi-TCN. It consists of a fully-connected first layer, **FC**, followed by B residual blocks, and then a fully-connected output layer, **O** that employs sigmoidal units. The kernel size, output size, and dilation rate for each convolutional unit is denoted as (**kernel size**, **output size**, **dilation rate**).

V. Proposed DeepMMSE method

When estimating the *a priori* SNR for Equation (9), the methods described in Subsection III assume that the noise changes at a slower rate than the speech. As a consequence, they are only able to track moderately non-stationary noise sources, and lack the ability to track highly non-stationary noise sources. Motivated by this, we propose an MMSE-based noise PSD estimator that employs the Deep Xi framework for *a priori* SNR estimation. The Deep Xi framework does not exploit any underlying assumptions about the speech or noise, and produces an *a priori* SNR estimate with negligible bias. The proposed noise PSD tracker includes the following four stages:

1) The *a priori* SNR estimate, $\hat{\xi}$, of the noisy speech magnitude spectrum, R, is first found using Deep Xi-TCN (available from https://github.com/anicolson/DeepXi). Deep Xi-TCN is described in Section V-A.
2) As in [25], the instantaneous *a posteriori* SNR can be computed from the instantaneous *a priori* SNR: $\gamma = \xi + 1$. This is used to estimate the (instantaneous) *a posteriori* SNR, $\hat{\gamma}$, from the (instantaneous) *a priori* SNR estimate given by Deep Xi.[†]
3) Using $\hat{\xi}$ and $\hat{\gamma}$, the noise periodogram estimate, $\widehat{N^2}$, is found using Equation (9).
4) The final noise PSD estimate, $\widehat{\lambda_d}$, is found by applying a first-order temporal recursive smoothing operation to $\widehat{N^2}$, with a smoothing factor of α_d, as in Equation (10).

Due to the combination of Deep Xi-TCN and the MMSE noise periodogram estimator, we refer to the proposed method as DeepMMSE henceforth.

[†]This converts the estimator from Equation (9) from an affine estimator to a linear estimator equivalent to the Wiener filter. As a consequence, the estimator depends solely on the *a priori* SNR.

A. Deep Xi-TCN

In [25], a ResLSTM network was used within the Deep Xi framework to estimate the *a priori* SNR. Here, we replace the ResLSTM network with a TCN [31]. The advantages that the TCN offers over the ResLSTM network include a reduction in training time and a reduction in the number of required parameters. Deep Xi-TCN is shown in Figure 1, and is described from input to output as follows. The input to the TCN is the noisy speech magnitude spectrum for the l^{th} frame, $\mathbf{R}_l$. The input is first transformed by **FC**, a fully-connected layer of size d_{model} that includes layer normalisation followed by the rectified linear unit (ReLU) activation function. The **FC** layer is followed by B bottleneck residual blocks, where $b = 1, 2, \ldots, B$ is the block index.

As in [32], each block contains three one-dimensional causal dilated convolutional units. Each convolutional unit is pre-activated by layer normalisation [33] followed by the ReLU activation function [34]. The kernel size, output size, and dilation rate for each convolutional unit is denoted in Figure 1 as (**kernel size**, **output size**, **dilation rate**). The first and third convolutional units in each block have a kernel size of 1, whilst the second convolutional unit has a kernel size of k. The first and second convolutional units have an output size of d_f, whilst the third convolutional unit has an output size of d_{model}.

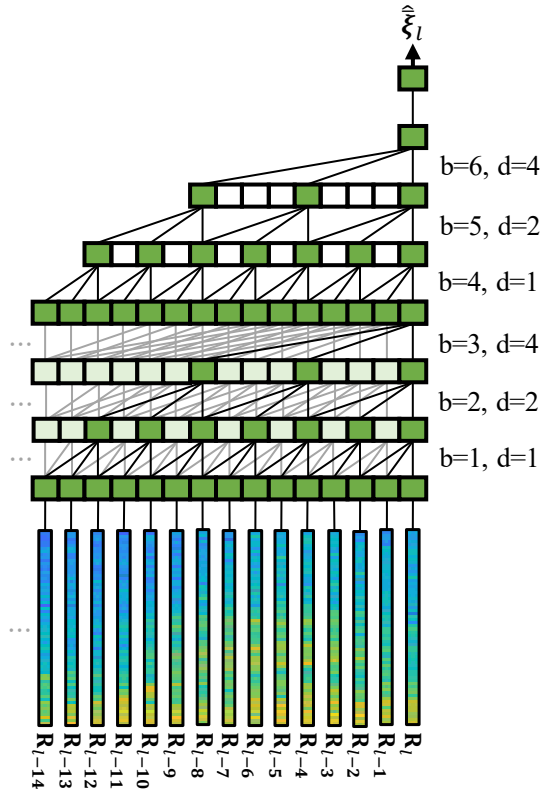


Fig. 2. Example of the contextual field of Deep Xi-TCN, with $D = 4$, $B = 6$, and $k = 3$.

The first and third convolutional units have a dilation rate of 1, while the second convolutional unit employs a dilation rate of d, providing a contextual field over previous time steps.

As in [35], the dilation rate d is cycled as the block index b increases: $d = 2^{(b-1 \bmod (\log_2(D)+1)}$, 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 2, 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 block is followed by the output layer, **O**, which is a fully-connected layer with sigmoidal units. The **O** layer estimates the mapped *a priori* SNR for each spectral component of the l^{th} time-frame, $\hat{\boldsymbol{\xi}}_l$. The mapped *a priori* SNR is described in Subsection IV-A. The following hyperparameters are chosen for the network, as a compromise between training time and performance: $d_{model} = 256$, $d_f = 64$, and $B = 40$. As in [32], k is set to 3, and D is set to 16. The TCN expends fewer parameters than the ResLSTM network, at approximately 2 million parameters as opposed to 10 million parameters. The training time for the TCN is also significantly less than that of the ResLSTM network, at approximately 40 minutes per epoch as opposed to 10 hours per epoch. Details about the training strategy for Deep Xi-TCN are given in Subsection VI-C. The *a priori* SNR estimation accuracy of Deep Xi-TCN is compared to that of Deep Xi-ResLSTM in Table I.

VI. Experimental Settings

A. Signal processing

A square-root-Hann window function is used for spectral analysis and synthesis [36]–[38], with a frame-length of 32 ms (512 time-domain samples) and a frame-shift of 16 ms (256 time-domain samples). The noise PSD is estimated from the 257-point single-sided noisy speech PSD, which included both the DC frequency component and the Nyquist frequency component.

B. Training Set

Here, we describe the clean speech and noise recordings used to train Deep Xi-TCN. The clean speech recordings from the following speech corpora are included in the training set: the *train-clean-100* set from the Librispeech corpus [39] (28 539 utterances), the CSTR VCTK corpus [40] (42 015 utterances), and the si^* and sx^* training sets from the TIMIT corpus [41] (3 696 utterances). This gives a total of 74 250 clean speech recordings. 5% of the clean speech recordings are randomly selected and used as the validation set. Thus, 70 537 clean speech recordings are used in the training set, and 3 713 clean speech recordings are used in the validation set. The noise recordings from the following noise datasets are included in the training set: the QUT-NOISE dataset [42], the Nonspeech dataset [43], the Environmental Background Noise dataset [44], [45], the noise set from the MUSAN corpus [46], multiple FreeSound packs,[‡] and coloured noise recordings (with an α value ranging from -2 to 2 in increments of 0.25). This gives a total of 2 382 noise recordings. All clean speech and noise recordings are single-channel, with a sampling frequency of 16 kHz (recordings with a sampling frequency higher than 16 kHz are down-sampled to 16 kHz). A description of how the noisy speech is created from the clean speech and noise recordings is given in Subsection VI-C.

[‡]Freesound packs that are 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.

C. Training strategy

The following strategy is employed to train the TCN:

- Cross-entropy as the loss function.
- The *Adam* algorithm [47] 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.
- A total of 175 epochs is used to train the TCN, where the number of training examples in an epoch is equal to the number of clean speech files in the training set (70 537).
- A mini-batch size of 10 noisy speech signals.
- The noisy speech 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).

D. Test set

Recordings of eight different noise sources are included in the test set. The first is a computer-generated noise source, specifically modulated white Gaussian noise. It is created by modulating Gaussian noise as follows:

$$f(n) = 1 + \sin\left(2\pi n \frac{f_{mod}}{f_s}\right), \tag{13}$$

where f_{mod} and f_s denote the modulation and sampling frequency, respectively, and n is the time-domain sample index. In this study, we set the modulation frequency to $f_{mod} = 0.5$ Hz. Five of the eight recordings are of real-world non-stationary noise sources, including *passing train*, *passing car*, and *traffic* from mutliple FreeSound packs, *street music* (recording no. 26 270) from the Urban Sound dataset [48], and *voice babble* from the RSG-10 noise dataset [49]. Two of the eight recordings are of real-world coloured noise sources, including *factory* and *F16* from the RSG-10 noise dataset [49]. 30 clean speech recordings are randomly selected[§] (without replacement) from the TSP speech corpus [50] for each noise recording.[¶] To create the noisy speech, a random section of the noise recording is mixed with the clean speech at the following SNR levels: -5 to 15 dB, in 5 dB increments. 150 noisy speech files are available in the test set for each noise source. The noisy speech signals are single channel, with a sampling frequency of 16 kHz.

[§]For the noise recordings used in Table I, only 10 clean speech recordings are randomly selected.

[¶]Only adult speakers are included from the TSP speech corpus.

TABLE I
A priori SNR ESTIMATE SD LEVELS ATTAINED BY EACH OF THE *a priori* SNR ESTIMATORS.

Noises	Methods	Input SNR (dB) -5	0	5	10	15
Voice babble	DD [24]	18.5	17.7	17.2	17.0	17.2
	TSNR [27]	18.4	17.5	17.0	16.9	17.1
	HRNR [28]	19.5	18.9	18.5	18.4	18.6
	SCTS [29]	17.5	16.8	16.5	16.5	16.9
	Deep Xi-ResLSTM	14.5	13.9	13.3	12.8	12.4
	Deep Xi-TCN	**13.6**	**13.0**	**12.4**	**11.9**	**11.6**
Street music	DD [24]	19.9	18.6	17.6	17.0	16.8
	TSNR [27]	19.7	18.4	17.4	16.8	16.6
	HRNR [28]	19.8	18.7	17.9	17.5	17.5
	SCTS [29]	18.6	17.4	16.6	16.2	16.2
	Deep Xi-ResLSTM	13.5	13.1	12.7	12.3	12.0
	Deep Xi-TCN	**12.7**	**12.3**	**11.9**	**11.5**	**11.2**
F16	DD [24]	22.1	20.5	19.2	18.2	17.5
	TSNR [27]	21.8	20.2	18.9	17.9	17.2
	HRNR [28]	20.7	19.4	18.4	17.7	17.3
	SCTS [29]	20.8	19.2	18.0	17.1	16.6
	Deep Xi-ResLSTM	13.3	12.7	12.3	12.0	11.7
	Deep Xi-TCN	**12.0**	**11.6**	**11.2**	**10.9**	**10.6**
Factory	DD [24]	24.0	22.2	20.7	19.4	18.5
	TSNR [27]	23.7	22.0	20.4	19.2	18.3
	HRNR [28]	23.0	21.4	20.1	19.1	18.4
	SCTS [29]	22.4	20.7	19.3	18.2	17.4
	Deep Xi-ResLSTM	13.8	13.2	12.7	12.4	12.1
	Deep Xi-TCN	**13.1**	**12.5**	**12.1**	**11.8**	**11.5**

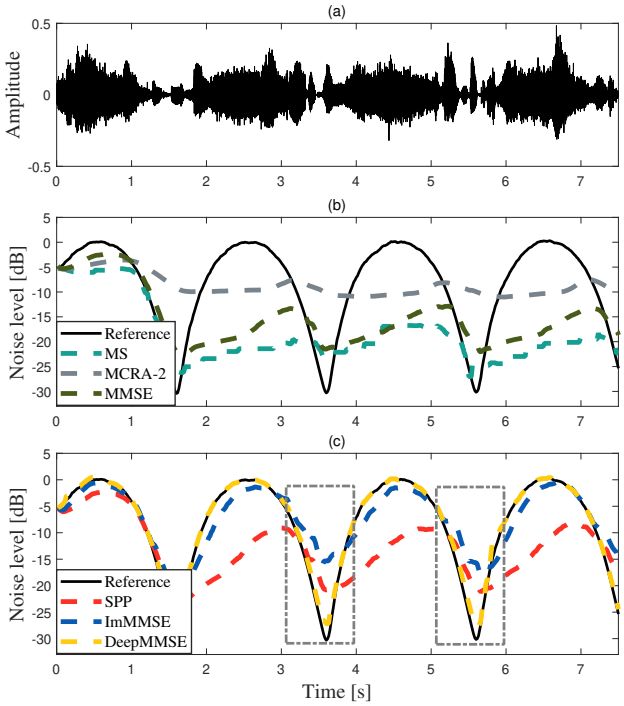


Fig. 3. (a) Speech corrupted by modulated Gaussian noise at an SNR level of 0 dB. (b)-(c) Noise PSD tracking performance of the noise PSD trackers, including the proposed DeepMMSE method. The noise PSDs are averaged over all frequency bins.

E. Evaluation measures

***A priori* SNR estimation accuracy**: The frame-wise spectral distortion (SD) [51] is used to evaluate the accuracy of an *a priori* SNR estimator, with lower levels indicating a better accuracy. The SD is defined as the root-mean-square difference between the *a priori* SNR estimate in dB, $\hat{\xi}_{\mathrm{dB}}[l,k]$, and the instantaneous case in dB, $\xi_{\mathrm{dB}}[l,k]$, for the l^{th} time-frame:[‖]

$$D_l^2 = \frac{1}{K/2+1}\sum_{k=0}^{K-1}\left[\xi_{\mathrm{dB}}[l,k] - \hat{\xi}_{\mathrm{dB}}[l,k]\right]^2, \qquad (14)$$

where K indicates the number of frequency-bins. Average SD levels are found over all frames for the test condition.

Noise PSD estimation accuracy: The spectral estimation accuracy of the noise PSD trackers is evaluated directly using the symmetric mean log-spectral distortion (LogErr) between the reference noise PSD, $\lambda_d[l,k]$, and the estimated noise PSD, $\hat{\lambda}_d[l,k]$, as follows [52]:

$$\mathrm{LogErr} = \frac{1}{LK}\sum_{l=0}^{L-1}\sum_{k=0}^{K-1}\left|10\log_{10}\left[\frac{\lambda_d[l,k]}{\hat{\lambda}_d[l,k]}\right]\right|, \qquad (15)$$

where L denotes the number of time-frames and K indicates the number of frequency-bins. As in previous works, the reference noise PSD, $\lambda_d[l,k]$, is obtained from the noise periodogram using a first-order recursive operation with a constant smoothing factor $\alpha_s = 0.8$ [20]–[23], [53]. This smoothing factor is also used by each of the MMSE-based noise PSD trackers (MMSE, SPP, ImMMSE, and DeepMMSE) to compute the final noise PSD estimate, $\widehat{\lambda_d}[l,k]$.

Speech enhancement performance: The noise PSD estimators are incorporated into the following speech enhancement framework:

1) First, the noise PSD estimate, $\widehat{\lambda_d}$, is computed from the noisy speech magnitude spectrum, R, using the noise PSD tracker. A constant smoothing factor of $\alpha_d = 0.8$ is used for all noise PSD trackers, except for the proposed DeepMMSE method which used $\alpha_d = 0$, as no smoothing is required for its noise PSD estimate.
2) Next, the *a posteriori* SNR estimate, $\hat{\gamma} = Y^2/\widehat{\lambda_d}$, is used to find the *a priori* SNR estimate. As in [14], [17], [20], [22], [23], the DD approach is used to estimate the *a priori* SNR for the MS, MCRA-2, MMSE, SPP, and ImMMSE methods.[**] The smallest allowable value is set to $\xi_{min} = -15$ dB for the DD approach, as in [26]. For the proposed DeepMMSE method, the ML estimate of the *a priori* SNR is employed: $\hat{\xi} = \max\left(\hat{\gamma} - 1, 0\right)$. Utilising the ML estimate with DeepMMSE produces better speech enhancement results than using the DD approach. Moreover, using the *a priori* SNR estimate from Deep Xi directly with an MMSE clean speech spectrum estimator is identical to the work in [25]. It should be

[‖] $\xi_{dB}[l,k]$ and $\hat{\xi}_{dB}[l,k]$ values that are less than -60 dB, or greater than 40 dB are clipped to -60 dB and 40 dB, respectively. Note that these were the clipping values used in [25], with the authors noting that they were reported incorrectly in that work.

[**] A smoothing factor of 0.98 is used for the MS, MCRA-2, MMSE, SPP methods, as in [14], [17], [20], [22]. A smoothing factor of 0.94 is used for the ImMMSE method, as in [23].

noted that the speech enhancement performance of the previous noise PSD estimators (MS, MCRA-2, MMSE, SPP, and ImMMSE) is worse when the ML estimate of the *a priori* SNR is used over the DD approach.

3) $\hat{\gamma}$ and $\hat{\xi}$ are then used by the MMSE clean speech spectrum estimator with generalised Gamma priors from [2] to enhance the noisy speech magnitude spectrum. The MMSE clean speech spectrum estimator assumes that the speech DFT coefficients follow a generalized-Gamma distribution with parameters $\gamma = 1$ and $\nu = 0.6$.

The objective quality and intelligibility of the resultant enhanced speech is then evaluated. The perceptual evaluation of speech quality (PESQ) is the metric used to evaluate the objective speech quality [54]. Disturbance processing and cognitive modelling are primarily used to determine the PESQ score. The PESQ score ranges from -0.5 to 4.5, with a higher PESQ score implying better speech quality. The short-time objective intelligibility (STOI) measure [55], [56] is used to evaluate the objective speech intelligibility. It is based on the correlation coefficient between the temporal envelops of the clean speech and enhanced speech in short-time regions. The STOI score ranges from 0 to 1 (or 0 to 100%), with a higher STOI score indicating better speech intelligibility.

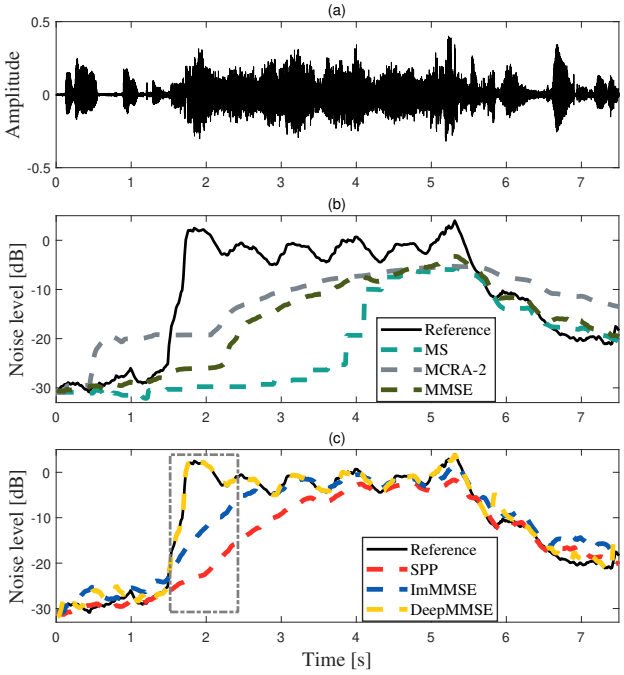


Fig. 4. (a) Speech corrupted by passing train noise at an SNR level of 0 dB. (b)-(c) Noise PSD tracking performance of the noise PSD trackers, including the proposed DeepMMSE method. The noise PSDs are averaged over all frequency bins.

VII. Results and Discussions

A. A priori SNR estimation accuracy

Before evaluating the tracking performance of the proposed DeepMMSE method, we determine if Deep Xi-TCN is a more accurate *a priori* SNR estimator than Deep Xi-ResLSTM. DeepXi-TCN is also compared to previous *a priori* SNR estimators, like the DD approach [24], the two-step noise reduction (TSNR) technique [27], harmonic regeneration noise reduction (HRNR) [28], and selective cepstro-temporal smoothing (SCTS) [29]. As seen in Table I, Deep Xi-TCN attains the lowest SD level for each tested condition. Deep Xi-TCN is able to outperform Deep Xi-ResLSTM in terms of *a priori* SNR estimation accuracy, while consuming fewer parameters and requiring less time to train.

TABLE II
Noise PSD estimation accuracy in terms of LogErr for various noise types and at different SNR levels. The lowest LogErr for each tested condition is indicated in boldface.

Noises	Methods	Input SNR (dB) -5	0	5	10	15
Mod. white	MS [14]	19.58	17.85	16.16	14.69	13.13
	MCRA-2 [17]	8.11	8.03	7.92	8.15	8.53
	MMSE [20]	10.88	10.34	9.67	9.12	8.56
	SPP [22]	9.22	8.65	7.98	7.45	7.16
	ImMMSE [23]	3.23	3.27	3.60	4.21	5.03
	DeepMMSE	**0.45**	**0.62**	**0.84**	**1.15**	**1.50**
Voice babble	MS [14]	6.95	6.84	6.76	6.77	6.64
	MCRA-2 [17]	4.47	4.55	5.10	6.44	8.71
	MMSE [20]	4.14	4.23	4.32	4.54	4.59
	SPP [22]	3.74	3.65	3.72	3.94	4.10
	ImMMSE [23]	2.94	3.30	3.68	4.30	5.04
	DeepMMSE	**1.27**	**1.50**	**2.10**	**2.46**	**2.87**
Factory	MS [14]	5.29	5.26	5.20	5.11	5.02
	MCRA-2 [17]	3.58	3.47	3.41	3.63	4.44
	MMSE [20]	3.02	3.07	3.17	3.32	3.54
	SPP [22]	2.71	2.70	2.72	2.83	3.07
	ImMMSE [23]	2.13	2.22	2.34	2.76	3.40
	DeepMMSE	**0.49**	**0.68**	**0.96**	**1.34**	**1.76**
Passing Car	MS [14]	21.11	17.85	15.25	14.86	12.55
	MCRA-2 [17]	9.84	8.71	8.84	9.95	11.99
	MMSE [20]	16.36	14.61	12.78	11.07	9.65
	SPP [22]	6.45	5.54	5.14	5.13	5.60
	ImMMSE [23]	3.48	3.53	3.94	4.68	6.01
	DeepMMSE	**1.14**	**1.64**	**2.41**	**3.52**	**4.97**
Passing Train	MS [14]	12.58	12.10	11.68	10.10	10.25
	MCRA-2 [17]	7.33	6.66	6.21	6.31	7.24
	MMSE [20]	8.40	8.53	8.72	8.66	8.47
	SPP [22]	4.64	4.70	4.85	4.97	5.01
	ImMMSE [23]	1.72	2.01	2.46	3.20	4.20
	DeepMMSE	**0.90**	**1.28**	**1.85**	**2.62**	**3.63**
Traffic	MS [14]	10.56	10.39	10.11	9.64	9.07
	MCRA-2 [17]	8.01	7.17	6.45	6.23	6.84
	MMSE [20]	5.57	5.70	5.83	5.97	5.98
	SPP [22]	3.46	3.53	3.68	3.85	4.07
	ImMMSE [23]	1.39	1.67	2.14	2.74	3.59
	DeepMMSE	**0.70**	**0.99**	**1.43**	**2.15**	**3.20**

B. Noise PSD tracking evaluation

An important feature of any noise PSD estimator is the ability to track non-stationary noise sources. In this subsection, the noise PSD tracking performance of the proposed DeepMMSE method is evaluated against five state-of-the-art noise PSD trackers, including the MS [14], MCRA-2 [17], MMSE [20], SPP [22], and ImMMSE [23] methods. Three clean speech recordings from the test set are concatenated and corrupted with non-stationary modulated Gaussian noise at an SNR level of 0 dB, as shown Figure 3 (a). Figures 3 (b) and (c) show the

TABLE III
PESQ SCORES OF THE ENHANCED SPEECH PRODUCED BY EACH OF THE NOISE PSD TRACKERS, AS WELL AS LSTM-IRM [57] AND XU2017 [58]. THE HIGHEST PESQ SCORE FOR EACH TESTED CONDITION IS INDICATED IN BOLDFACE.

		Input SNR (dB)				
Noises	Methods	-5	0	5	10	15
Mod. white	Unprocessed	1.28	1.56	1.92	2.30	2.69
	MS [14]	1.31	1.63	2.03	2.40	2.79
	MCRA-2 [17]	1.22	1.63	2.03	2.34	2.63
	MMSE [20]	1.27	1.66	2.08	2.47	2.88
	SPP [22]	1.27	1.70	2.14	2.51	2.92
	ImMMSE [23]	1.50	1.97	2.37	2.69	2.99
	LSTM-IRM [57]	1.76	2.24	2.64	2.95	3.19
	Xu2017 [58]	1.53	2.11	2.58	2.90	3.16
	DeepMMSE	**2.28**	**2.70**	**2.96**	**3.24**	**3.47**
Voice babble	Unprocessed	1.00	1.43	1.83	2.21	2.58
	MS [14]	1.23	1.63	2.08	2.47	2.88
	MCRA-2 [17]	1.22	1.62	2.06	2.43	2.77
	MMSE [20]	1.23	1.66	2.12	2.54	2.94
	SPP [22]	1.20	1.64	2.11	2.54	2.95
	ImMMSE [23]	1.24	1.69	2.13	2.53	2.90
	LSTM-IRM [57]	1.25	1.80	2.29	2.67	2.96
	Xu2017 [58]	1.45	2.05	2.51	2.95	3.29
	DeepMMSE	**1.54**	**2.11**	**2.61**	**3.06**	**3.38**
Factory	Unprocessed	1.06	1.29	1.68	2.09	2.49
	MS [14]	1.29	1.69	2.10	2.49	2.85
	MCRA-2 [17]	1.28	1.72	2.15	2.52	2.79
	MMSE [20]	1.38	1.81	2.24	2.59	2.91
	SPP [22]	1.38	1.84	2.26	2.62	2.94
	ImMMSE [23]	1.42	1.87	2.31	2.67	2.96
	LSTM-IRM [57]	1.25	1.84	2.25	2.59	2.88
	Xu2017 [58]	1.39	1.99	2.50	2.85	3.15
	DeepMMSE	**1.78**	**2.25**	**2.63**	**2.99**	**3.31**
Passing Train	Unprocessed	1.56	1.76	1.99	2.30	2.63
	MS [14]	1.64	1.83	2.09	2.39	2.72
	MCRA-2 [17]	1.55	1.80	2.03	2.33	2.62
	MMSE [20]	1.66	1.87	2.15	2.46	2.78
	SPP [22]	1.79	1.98	2.29	2.64	2.95
	ImMMSE [23]	1.80	2.07	2.42	2.75	3.05
	LSTM-IRM [57]	1.95	2.23	2.58	2.92	3.19
	Xu2017 [58]	2.01	2.30	2.68	2.98	3.17
	DeepMMSE	**2.28**	**2.57**	**2.90**	**3.14**	**3.32**
Passing Car	Unprocessed	1.47	1.68	1.90	2.21	2.56
	MS [14]	1.53	1.75	2.02	2.33	2.67
	MCRA-2 [17]	1.44	1.73	2.00	2.32	2.61
	MMSE [20]	1.56	1.79	2.09	2.42	2.75
	SPP [22]	1.68	1.92	2.24	2.61	2.94
	ImMMSE [23]	1.68	1.99	2.35	2.68	3.00
	LSTM-IRM [57]	1.99	2.33	2.65	2.94	3.23
	Xu2017 [58]	1.84	2.26	2.67	3.03	3.34
	DeepMMSE	**2.23**	**2.55**	**2.85**	**3.15**	**3.45**
Traffic	Unprocessed	1.36	1.57	1.73	2.04	2.40
	MS [14]	1.32	1.52	1.79	2.13	2.51
	MCRA-2 [17]	1.22	1.48	1.80	2.16	2.49
	MMSE [20]	1.35	1.55	1.87	2.23	2.60
	SPP [22]	1.36	1.72	2.04	2.48	2.84
	ImMMSE [23]	1.44	1.74	2.16	2.56	2.92
	LSTM-IRM [57]	1.64	1.99	2.32	2.62	2.90
	Xu2017 [58]	1.63	2.05	2.39	2.68	2.95
	DeepMMSE	**2.10**	**2.50**	**2.80**	**3.06**	**3.28**

TABLE IV
STOI SCORES (IN %) OF THE ENHANCED SPEECH PRODUCED BY EACH OF THE NOISE PSD TRACKERS, AS WELL AS LSTM-IRM [57] AND XU2017 [58]. THE HIGHEST STOI SCORE FOR EACH TESTED CONDITION IS INDICATED IN BOLDFACE.

		Input SNR (dB)				
Noises	Methods	-5	0	5	10	15
Mod. white	Unprocessed	70.0	79.26	85.97	91.73	95.55
	MS [14]	69.88	79.12	85.87	91.71	95.56
	MCRA-2 [17]	66.86	77.20	84.28	90.20	93.93
	MMSE [20]	68.17	78.48	85.52	91.39	95.41
	SPP [22]	66.91	78.08	85.56	91.54	95.40
	ImMMSE [23]	68.33	78.81	86.53	92.14	95.44
	LSTM-IRM [57]	77.50	85.97	91.01	94.56	96.32
	Xu2017 [58]	70.12	81.31	88.48	91.36	95.47
	DeepMMSE	**80.58**	**87.91**	**91.97**	**95.13**	**97.0**
Voice babble	Unprocessed	60.16	72.42	83.01	90.74	95.46
	MS [14]	58.15	72.11	83.73	91.66	95.99
	MCRA-2 [17]	57.04	71.13	82.97	90.87	95.17
	MMSE [20]	56.07	70.11	82.63	91.30	96.16
	SPP [22]	55.59	69.68	82.07	90.83	95.98
	ImMMSE [23]	56.33	69.98	82.08	90.47	95.37
	LSTM-IRM [57]	64.2	78.55	88.0	93.52	96.53
	Xu2017 [58]	62.50	74.77	83.82	90.14	94.74
	DeepMMSE	**66.0**	**80.0**	**89.54**	**94.88**	**97.54**
Factory	Unprocessed	57.85	69.92	80.88	89.18	94.49
	MS [14]	56.63	70.57	82.27	90.51	95.34
	MCRA-2 [17]	55.78	69.83	81.79	90.00	94.66
	MMSE [20]	54.90	68.53	82.00	90.70	95.47
	SPP [22]	54.94	68.58	81.91	90.82	95.72
	ImMMSE [23]	56.43	69.69	81.79	90.54	95.38
	LSTM-IRM [57]	62.4	76.82	86.64	92.49	95.92
	Xu2017 [58]	61.0	74.20	83.89	90.50	94.74
	DeepMMSE	**68.29**	**81.47**	**89.11**	**93.91**	**96.81**
Passing Train	Unprocessed	70.38	77.89	84.65	90.24	94.42
	MS [14]	70.68	78.19	84.88	90.32	94.44
	MCRA-2 [17]	69.49	77.14	83.75	89.04	93.00
	MMSE [20]	69.93	77.98	84.86	90.45	94.53
	SPP [22]	69.53	77.60	84.93	90.53	94.55
	ImMMSE [23]	69.68	77.79	84.92	90.52	94.52
	LSTM-IRM [57]	74.27	80.77	87.39	92.56	95.80
	Xu2017 [58]	71.21	78.66	85.61	90.59	94.0
	DeepMMSE	**77.30**	**83.67**	**88.74**	**93.40**	**95.90**
Passing Car	Unprocessed	69.36	77.68	84.76	90.39	94.52
	MS [14]	69.82	78.12	85.07	90.51	94.54
	MCRA-2 [17]	68.68	76.99	83.83	89.14	93.04
	MMSE [20]	68.83	77.82	84.94	90.62	94.65
	SPP [22]	68.59	77.59	85.11	90.77	94.71
	ImMMSE [23]	69.05	77.69	85.18	90.67	94.57
	LSTM-IRM [57]	78.73	84.70	89.88	94.20	96.92
	Xu2017 [58]	71.41	82.04	88.89	93.21	96.05
	DeepMMSE	**77.30**	**85.27**	**90.51**	**94.60**	**97.10**
Traffic	Unprocessed	67.42	75.39	83.12	89.31	93.92
	MS [14]	66.71	75.69	83.50	89.52	94.00
	MCRA-2 [17]	66.54	75.35	82.74	88.40	92.66
	MMSE [20]	65.76	75.31	83.29	89.42	94.00
	SPP [22]	65.48	74.92	83.62	89.78	94.24
	ImMMSE [23]	65.99	74.95	83.43	89.53	94.15
	LSTM-IRM [57]	72.50	81.67	88.43	92.89	**95.73**
	Xu2017 [58]	72.55	81.38	87.07	90.89	93.81
	DeepMMSE	**76.74**	**84.73**	**89.91**	**93.52**	95.71

averaged noise PSD estimates produced by the MS, MCRA-2, MMSE, SPP, ImMMSE, and DeepMMSE methods. The reference noise PSD is also included in Figures 3 (b) and (c). It is obvious that the MS, MCRA-2, and MMSE methods poorly track the modulated Gaussian noise PSD. The SPP method yields an improvement in tracking performance over the MS, MCRA-2, and MMSE methods, but still produces a noise PSD estimate with large bias. The ImMMSE method outperforms the SPP method, demonstrating the ability to track regions of moderately-changing noise levels. The proposed DeepMMSE method produces a noise PSD estimate with negligible bias, and is able to track the abrupt changes in the noise level, e.g., in the time-spans from 3–4, and 5–6 seconds, as highlighted in Figure 3 (c).

Another example is shown in Figure 4 (a), where the clean speech signal used for Figure 3 (a) is corrupted with passing train noise at an SNR level of 0 dB. It is observed again that the proposed DeepMMSE method exhibits an excellent tracking performance, even for rapidly changing noise levels. As shown in Figure 4 (b), the MS, MCRA-2, and MMSE methods again show very large tracking delays. The SPP and ImMMSE methods are able to track the noise PSD during slow changes of the noise level, but again are not able to handle abrupt increases in the noise level, e.g., in the time-span from 1.5–2 seconds, as highlighted in Figure 4 (c). It can also be seen that the noise PSD estimate of the proposed Deep MMSE method has the least bias.

C. Noise PSD estimation accuracy

Here, we evaluate the noise PSD estimation accuracy of the proposed DeepMMSE method. The LogErr results for each of the noise PSD estimators are given in Table II. It can be seen that both the MS and MCRA-2 methods exhibit poor noise PSD estimation accuracy for all tested conditions. The MMSE method demonstrates a reasonable noise PSD estimation accuracy, except for noise sources that include rapidly changing noise levels (modulated Gaussian, passing train, and passing car in Table II). The SPP method is not as severely affected by rapidly changing noise levels, and exhibits an improvement in noise PSD estimation accuracy over the MMSE method. The ImMMSE algorithm demonstrates a high noise PSD estimation accuracy, especially at lower SNR levels. However, the proposed DeepMMSE method achieves the highest noise PSD estimation accuracy for each tested condition.

D. Speech enhancement performance

Here, the speech enhancement performance of the proposed DeepMMSE method is evaluated within the framework described in Subsection VI-E. The objective quality (PESQ) and intelligibility (STOI) scores of the enhanced speech produced by each of the noise PSD trackers are used for evaluation. Table III presents the PESQ scores attained by each of the noise PSD estimators. It can be seen that the proposed DeepMMSE method achieves the highest objective quality scores for each tested condition. The STOI scores obtained by each of the noise PSD estimators are presented in Table IV. It can be seen that the proposed DeepMMSE method produces the most objectively intelligible enhanced speech for each tested condition.

The proposed DeepMMSE method is able to outperform the other noise PSD trackers at all tested SNR levels, and for both real-world non-stationary (e.g. voice babble) and coloured noise sources (e.g. factory), as well as for computer-generated non-stationary noise (e.g. modulated Gaussian). In addition, the proposed DeepMMSE method is also compared to two recent deep learning approaches to speech enhancement: LSTM-IRM, and Xu2017. LSTM-IRM is an LSTM network that estimates the ideal ratio mask (IRM) [57], and Xu2017 is a multilayer perceptron (MLP) clean speech spectrum estimator that incorporates multi-objective learning and ideal binary mask (IBM) post-processing (available from https://github.com/yongxuUSTC/DNN-for-speech-enhancement) [58]. It can be seen in Tables III and IV that the proposed DeepMMSE method is able to produce enhanced speech with higher quality and intelligibility scores than LSTM-IRM and Xu2017 for all tested conditions (except for traffic noise at 15 dB).

E. Evaluation of enhanced speech spectrograms

In this section, the enhanced speech spectrograms produced by each of the noise PSD trackers are evaluated. A clean speech recording from the test set (Figure 5 (a)) is mixed with traffic noise at an SNR level of 5 dB, producing the noisy speech (Figure 5 (b)). The noisy speech is then enhanced by each of the five state-of-the-art noise PSD trackers (Figures 5 (c)-(g)), LSTM-IRM (Figure 5 (h)), Xu2017 (Figures 5 (i)), and the proposed DeepMMSE method (Figure 5 (j)). The enhanced speech produced by the MS method (Figure 5 (c)) exhibits a significant amount of residual noise and includes some musical noise. The enhanced speech produced by MCRA-2 (Figure 5 (d)) exhibits less residual noise, but more musical noise than that of the MS method. The enhanced speech produced by MMSE (Figure 5 (e)) exhibits less musical noise and speech distortion, but more residual noise than MCRA-2. The enhanced speech produced by SPP (Figure 5 (f)) exhibits less residual noise, but slightly more musical noise than MMSE. The enhanced speech produced by ImMMSE (Figure 5 (g)) exhibits less musical noise than SPP. The enhanced speech produced by LSTM-IRM (Figure 5 (h)) exhibits significantly less musical noise and speech distortion than ImMMSE. The enhanced speech produced by Xu2017 (Figure 5 (i)) exhibits less residual noise, but more speech distortion than LSTM-IRM. The enhanced speech produced by DeepMMSE (Figure 5 (j)) exhibits less residual noise and speech distortion than Xu2017. Highlighted in Figure 5 (i) are the regions of speech that Xu2017 heavily distorted. It can be seen in Figure 5 (j) that DeepMMSE did not heavily distort these same regions. It can also be seen that DeepMMSE produced enhanced speech with less residual noise than LSTM-IRM. These observations are reflected in the objective quality and intelligibility results in Tables III and IV, respectively.

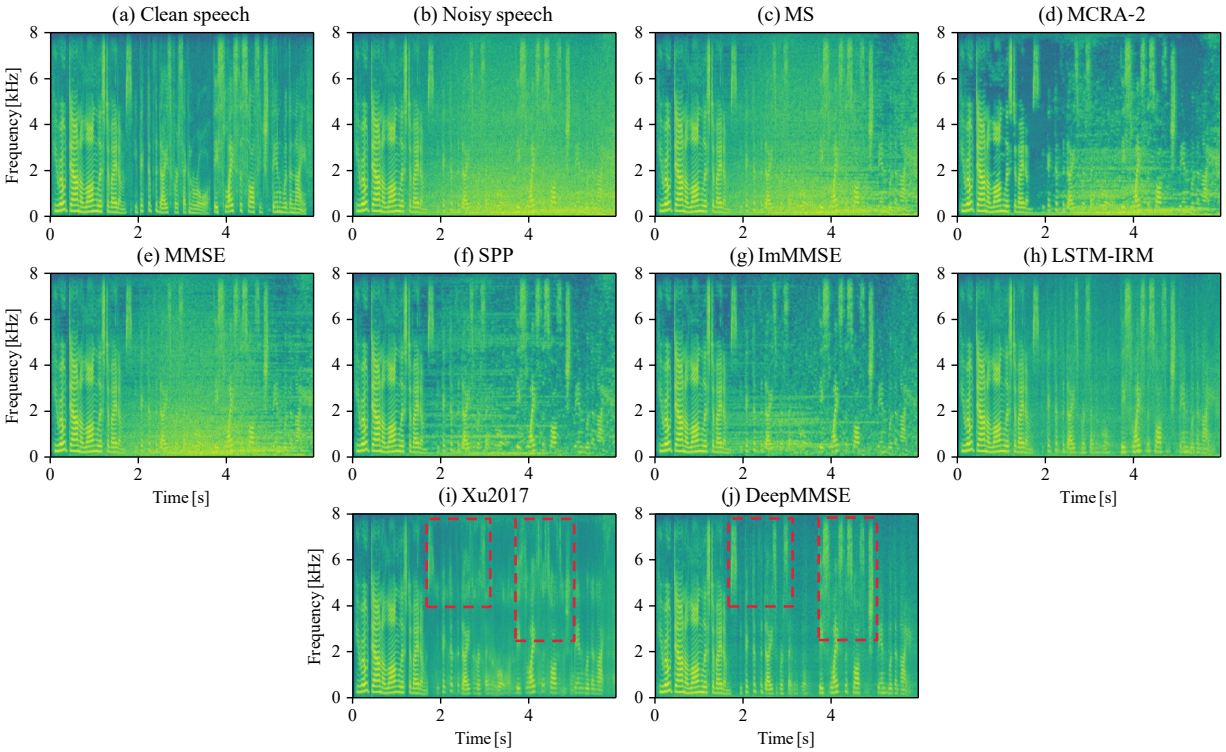


Fig. 5. The spectrograms of (a) the clean speech, (b) the noisy speech (clean speech mixed with traffic noise at an SNR level of 5 dB), and the enhanced speech produced by each of the noise PSD trackers: (c) MS, (d) MCRA-2, (e) MMSE, (f) SPP, (g) ImMMSE, (h) LSTM-IRM, (i) Xu2017, and (j) DeepMMSE (proposed).

VIII. Conclusion

A critical component of any speech enhancement system is the estimation of the noise PSD. The performance of an MMSE-based noise PSD tracking method heavily depends on the accuracy of the used *a priori* SNR estimator. Motivated by this, we investigate how an MMSE-based noise PSD tracker can benefit from a deep learning approach to *a priori* SNR estimation. The proposed noise PSD tracking method, called DeepMMSE, is evaluated using a variety of real-world non-stationary and coloured noise sources at multiple SNR levels. Our results show that the proposed DeepMMSE method is able to significantly outperform other noise PSD trackers in terms of spectral tracking accuracy. As the proposed method does not exploit any assumptions about the characteristics of the speech or noise, it is able to track sudden changes in the noise level. Furthermore, the proposed DeepMMSE method is able to yield higher speech enhancement objective quality and intelligibility scores than other noise PSD tracking methods, as well as two deep learning approaches to speech enhancement.

Recently, perceptually guided training has been exploited to increase speech enhancement performance [59]–[62]. This may be investigated in future work to obtain a further improvement in performance. In this work, the *a posteriori* SNR estimate is calculated from the *a priori* SNR estimate. Further improvements in performance may be obtained by using a deep learning approach to estimate the *a posteriori* SNR directly. Moreover, the focus of this paper is on modifying only the magnitude of the STFT components. Researchers have found that phase provides additional information which can help distinguish noise outliers from speech [63]. Incorporating this into the proposed estimator may improve its performance.

Acknowledgement

The authors thank the Circuit and Systems (signal processing) Group at Delft University of Technology for providing an implementation of the MMSE speech spectrum estimator with generalised Gamma priors.

References

[1] 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.

[2] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[3] R. C. Hendriks, T. Gerkmann, and J. Jensen, "DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synthesis Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, Jan. 2013.

[4] N. Dionelis and M. Brookes, "Phase-aware single-channel speech enhancement with modulation-domain Kalman filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 5, pp. 937–950, May 2018.

[5] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Berlin: Springer, 2008.

[6] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*. Berlin: Springer, 2013.

[7] Z. Wang, J. L. Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 1–5.

[8] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

[9] J.-H. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[10] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 697–710, Apr. 2013.

[11] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 7378–7382.

[12] G. Gelly and J. Gauvain, "Optimization of RNN-Based Speech Activity Detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 646–656, Mar. 2018.

[13] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proc. Euro. Signal Processing Conf.(EUSIPCO)*, 1994, pp. 1182–1185.

[14] ——, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[15] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[16] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sept. 2003.

[17] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *ELSEVIER Speech commun.*, vol. 48, no. 2, pp. 220–230, Feb. 2006.

[18] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proc. Eurospeech*, Sept. 1995, pp. 1513–1516.

[19] R. Yu, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Taipei, Taiwan*, Apr. 2009, pp. 4421–4424.

[20] 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., Dollas, TX*, Mar. 2010, pp. 4266–4269.

[21] T. Gerkmann and R. C. Hendriks, "Improved MMSE-based noise PSD tracking using temporal cepstrum smoothing," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 105–108.

[22] ——, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[23] Q. Zhang, M. Wang, Y. Lu, L. Zhang, and M. Idrees, "A novel fast nonstationary noise tracking approach based on MMSE spectral power estimator," *Digital Signal Processing*, vol. 88, pp. 41–52, May 2019.

[24] 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.

[25] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Communication*, vol. 111, pp. 44 – 55, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639318304308

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

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

[28] C. Plapous, C. Marro, and P. Scalart, "Speech enhancement using harmonic regeneration," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, March 2005, pp. 157–160.

[29] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 4897–4900.

[30] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products, 7th ed., D. Zwillinger and V. Moll*. New York: Academic Press, Feb. 2007.

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

[32] 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, 2016. [Online]. Available: http://arxiv.org/abs/1610.10099

[33] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[34] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. USA: Omnipress, 2010, pp. 807–814.

[35] Y. Luo and N. Mesgarani, "TasNet: surpassing ideal time-frequency masking for speech separation," *CoRR*, vol. abs/1809.07454, 2018. [Online]. Available: http://arxiv.org/abs/1809.07454

[36] J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, Sept 1993.

[37] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[38] K. Paliwal and K. Wojcicki, "Effect of analysis window duration on speech intelligibility," *IEEE Signal Processing Letters*, vol. 15, pp. 785–788, 2008.

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

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

[41] 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.

[42] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proceedings Interspeech 2010*, 2010, pp. 3110–3113.

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

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

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

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

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

[48] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[49] H. J. Steeneken and F. W. Geurtsen, "Description of the RSG-10 noise database," *Report IZF 1988-3, TNO Institute for Perception, Soesterberg, The Netherlands*, 1988.

[50] P. Kabal, "TSP speech database," *McGill University, Database Version*, 2002.

[51] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," in *Proceedings ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, April 1991, pp. 661–664 vol. 1.

[52] R. C. Hendriks, J. Jensen, and R. Heusdens, "DFT domain subspace based noise tracking for speech enhancement," in *Proc. Interspeech 2007*, 2007, pp. 830–833.

[53] M. Azarpour, G. Enzner, and R. Martin, "Binaural noise PSD estimation for binaural speech enhancement," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7068–7072.

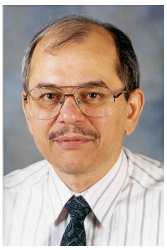
[54] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of

narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.
[55] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 4214–4217.
[56] ——, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.
[57] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.
[58] Y. Xu, J. Du, Z. Huang, L. Dai, and C. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," *CoRR*, vol. abs/1703.07172, 2017. [Online]. Available: http://arxiv.org/abs/1703.07172
[59] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 81–85.
[60] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1780–1792, Oct 2018.
[61] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5074–5078.
[62] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*, 2019, pp. 2031–2041.
[63] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

Qiquan Zhang was born in Yunnan Province, China in 1993. He received his B.S. degree in electronic science & technology from Harbin Institute of Technology Weihai Campus, Shandong Province, China, in 2015. He is currently working towards a Ph.D. degree in Electrical Science & Technology at the Shenzhen key laboratory of Internet of Things terminal technology, Harbin Institute of Technology, Shenzhen City. His research interests are digital speech and audio signal processing, speech enhancement algorithms, and microphone array signal processing.

Aaron Nicolson was born in Brisbane, Australia in 1994. He received his BEng degree (Class 1A Hons.) from Griffith University, Brisbane, Australia, in 2016. He is currently studying for a PhD degree at the Signal Processing Laboratory, Griffith University, Brisbane, Australia. His research includes machine learning, speech enhancement, speech recognition, and speaker recognition.

Mingjiang Wang was born in Heilongjiang Province, China in 1968. He received the B.S. and M.S. degrees in semiconductor physics & devices from Harbin Institute of Technology, Heilongjiang Province, China, in 1990 and 1993 respectively. In July 1998, he received the Ph.D. in electronic engineering from Fudan University, Shanghai City, China. From 1993 to 1995, he had worked as an Associate Professor at Southeast University, Jiangsu Province, China. He worked as a senior engineer at Huawei Technologies Co. Ltd from 1998 to 2000. Since September 2009, He is working as the professor at Harbin Institute of technology Shenzhen campus, in electronic and information engineering department. He is the director of Shenzhen key laboratory of internet of things terminal technology. His research interests involves the low-power loss chip design, speech signal processing, speech coding, and speech enhancement. His current research interests focus on audio/image deep learning algorithms for AI processing chip design.



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, India, in 1971, and the PhD 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 at a number of organizations including Tata Institute of Fundamental Research, Bombay, India Norwegian Institute of Technology, Trondheim, Norway, University of Keele, U.K., AT&T Bell Laboratories, Murray Hill, New Jersey, U.S.A., AT&T Shannon Laboratories, Florham Park, New Jersey, U.S.A., and Advanced Telecommunication Research Laboratories, Kyoto, Japan. Since July 1993, he has been a professor at Griffith University, Brisbane, Australia, in the School of Microelectronic Engineering. His current research interests include speech recognition, speech coding, speaker recognition, speech enhancement, face recognition, image coding, pattern recognition and artificial neural networks. Prof. Paliwal is a Fellow of Acoustical Society of India. He has served the IEEE Signal Processing Society's Neural Networks Technical Committee as a founding member from 1991 to 1995 and the Speech Processing Technical Committee from 1999 to 2003. He was an Associate Editor of the IEEE Transactions on Speech and Audio Processing during the periods 1994-1997 and 2003-2004. He is in the Editorial Board of the IEEE Signal Processing Magazine. He also served as an Associate Editor of the IEEE Signal Processing Letters from 1997 to 2000. 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 (published by Elsevier), and "Speech and Speaker Recognition: Advanced Topics" (published by Kluwer). He has received the IEEE Signal Processing Society's best (senior) paper award in 1995 for his paper on LPC quantization. He served as the Editor-in-Chief of the Speech Communication journal.

Chenxu Wang was born in Henan Province, China in 1977. He received his B.S. and M.S. degrees in Electronics Science & Technology from Harbin Institute of Technology, Heilongjiang Province, China, in 2001 and 2003 respectively. In July 2014, he received his Ph.D. in Microelectronics & Solid State Electronics from Harbin Institute of Technology, Heilongjiang Province, China. From 2004 until now, he had worked as a Lecturer, an Associate Professor successively in Harbin Institute of Technology, Weihai, Shandong Province, China. His research includes digital IC design and digital signal processing. His current research is focused on deep learning algorithm acceleration for AI implementation.