Deep Learning for Minimum Mean-Square Error Approaches to Speech Enhancement

Aaron Nicolson*, Kuldip K. Paliwal

Signal Processing Laboratory, Griffith University, Brisbane, Queensland, Australia, 4160

Abstract

Recently, the focus of speech enhancement research has shifted from minimum mean-square error (MMSE) approaches, like the MMSE short-time spectral amplitude (MMSE-STSA) estimator, to state-of-the-art masking- and mapping-based deep learning approaches. We aim to bridge the gap between these two differing speech enhancement approaches. Deep learning methods for MMSE approaches are investigated in this work, with the objective of producing enhanced speech that achieves higher quality and intelligibility scores than current masking- and mapping-based deep learning approaches. Since the speech enhancement performance of an MMSE approach improves with the accuracy of the used *a priori* signal-to-noise ratio (SNR) estimator, a residual long short-term memory (ResLSTM) network is utilised here to accurately estimate the *a priori* SNR. MMSE approaches utilising the ResLSTM *a priori* SNR estimator are compared to recent masking- and mapping-based deep learning approaches, using subjective and objective measures of speech quality and intelligibility. The tested conditions include clean speech mixed with real-world non-stationary and coloured noise sources, at multiple SNR levels. MMSE approaches utilising the proposed *a priori* SNR estimator were able to achieve higher enhanced speech quality and intelligibility scores than recent masking- and mapping-based deep learning approaches. The results presented in this work show that the performance of an MMSE approach to speech enhancement significantly increases when utilising deep learning.

Availability: The proposed *a priori* SNR estimator is available at `https://github.com/anicolson/DeepXi`

Keywords: Speech enhancement; *A priori* SNR estimation; Minimum mean-square error (MMSE) approach; Residual long short-term memory (ResLSTM) network; Deep Xi; Speech separation.

1. Introduction

The minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator is the benchmark against which other speech enhancement methods are evaluated against [1]. Other prominent MMSE approaches to speech enhancement include the minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator [2] and the Wiener filter (WF) approach [3]. While once at the forefront of speech enhancement research, less attention has been paid to the aforementioned MMSE approaches as of late. The research focus of the speech enhancement community has turned to deep learning methods.

Deep learning methods have recently been employed for speech enhancement, and have demonstrated state-of-the-art performance [4]. Neural networks have been used as non-linear maps from noisy speech spectra to clean speech spectra. A denoising autoencoder (DAE) was pre-trained for this task using noisy and clean speech pairs [5]. A non-causal neural network clean speech spectrum estimator was proposed that produced enhanced speech with high objective quality scores [6], which later incorporated multi-objective learning and ideal binary mask (IBM)-based post-processing [7]. Neural networks have also been utilised to estimate time-frequency masks. A long short-term memory (LSTM) network was used recently to estimate the ideal ratio mask (IRM) [8].

We aim to bridge the gap between MMSE and deep learning approaches to speech enhancement, with the objective of producing enhanced speech that achieves higher quality and intelligibility scores than that of recent masking- and mapping-based deep learning approaches. Here, the performance improvement that deep learning methods can provide to the aforementioned MMSE approaches is investigated. Each MMSE approach requires the *a priori* signal-to-noise ratio (SNR) estimate of a noisy speech spectral component. The *a priori* SNR is formally described in Subsection 2.2. Since the performance of an MMSE approach to speech enhancement improves with the accuracy of the used *a priori* SNR estimator, deep learning methods are used here to accurately estimate the *a priori* SNR.

*Corresponding author

Email addresses: `aaron.nicolson@griffithuni.edu.au` (Aaron Nicolson), `k.paliwal@griffith.edu.au` (Kuldip K. Paliwal)

A priori SNR estimation is a difficult task, especially when considering the multitude of different noise sources. The decision-directed (DD) approach [1] to *a priori* SNR estimation was introduced with the MMSE-STSA estimator, and uses a weighted average of the *a priori* SNR estimate from the previous and current frames. The DD approach suffers from a frame delay problem [9], which is addressed by the two-step noise reduction (TSNR) technique [10]. Harmonic regeneration noise reduction (HRNR) [11] further improves upon the TSNR technique by computing an *a priori* SNR estimate from enhanced speech with artificially restored harmonics. Other *a priori* SNR estimates are computed using a maximum-likelihood approach. Selective cepstro-temporal smoothing (SCTS) [12] performs adaptive temporal smoothing on the cepstral representation of the maximum-likelihood estimate of the clean speech power spectrum, in order to estimate the *a priori* SNR.

It has been demonstrated that residual long short-term memory (ResLSTM) networks are proficient acoustic models [13]. Motivated by this, a causal ResLSTM network, and a non-causal residual bidirectional LSTM (ResBLSTM) network [14] are used here for *a priori* SNR estimation. Unlike previous *a priori* SNR estimators, the proposed estimators do not require a noise estimator. Recently, a recurrent neural network (RNN) was used to aid the DD approach in *a priori* SNR estimation [15]. The proposed estimators differ by directly estimating the *a priori* SNR. It was found that mapping the *a priori* SNR target values to the interval $[0, 1]$ improved the rate of convergence of the used stochastic gradient descent algorithm. We propose to use the cumulative distribution function (CDF) of the *a priori* SNR in dB as the map. By using the CDF, large sections of the distribution are not excluded.

In this work, MMSE approaches utilising deep learning are evaluated using subjective and objective measures of speech quality and intelligibility. The tested conditions include clean speech mixed with real-world non-stationary and coloured noise sources, at multiple SNR levels. The MMSE approaches utilising deep learning are compared to masking- and mapping-based deep learning approaches to speech enhancement. Frame-wise spectral distortion (SD) levels are used to evaluate the accuracy of the proposed *a priori* SNR estimators.

The paper is organised as follows: background knowledge is presented in Section 2, including the analysis, modification, and synthesis (AMS) procedure, *a priori* SNR, and MMSE approaches; the mapped *a priori* SNR training target is described in Section 3; the ResLSTM and ResBLSTM *a priori* SNR estimators are described in Section 4; the experiment setup is described in Section 5, including the objective and subjective testing procedures; the results and discussion are presented in Section 6; conclusions are drawn in Section 7.

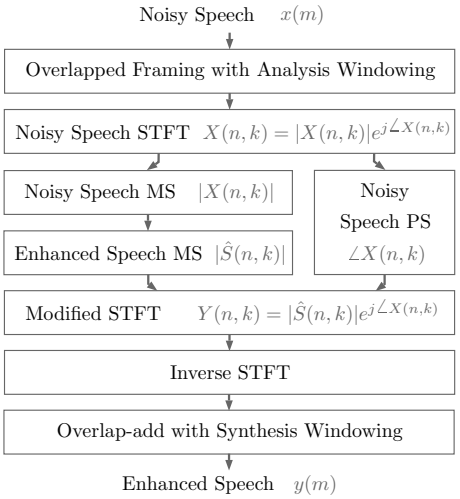

Figure 1: Block diagram of the short-time Fourier AMS speech enhancement framework.

2. Background

2.1. AMS Speech Enhancement Framework

The short-time Fourier analysis, modification, and synthesis (AMS) framework was used to produce enhanced speech. The AMS framework [16, 17] consists of three stages: (1) the analysis stage, where noisy speech undergoes short-time Fourier transform (STFT) analysis; (2) the modification stage, where the noisy speech STFT is compensated for noise distortion to produce the modified STFT; and (3) the synthesis stage, where the inverse STFT operation is followed by overlap-add synthesis to construct the enhanced speech. A block diagram of the AMS framework is shown in Figure 1.

An uncorrelated additive noise model is assumed:

$$x(m) = s(m) + d(m), \tag{1}$$

where $x(m)$, $s(m)$, and $d(m)$ denote the noisy speech, clean speech, and noise, respectively, and m denotes the discrete-time index. Noisy speech is analysed frame-wise using the running STFT [18]:

$$X(n,k) = \sum_{m=0}^{N_l-1} x(m + nN_s)w(m)e^{-j2\pi mk/N_l}, \tag{2}$$

where n denotes the frame index, k denotes the discrete-frequency index, N_l denotes the frame length in discrete-time samples, N_s denotes the frame shift in discrete-time samples, and $w(m)$ is the analysis window function.

In polar form, the STFT of the noisy speech is expressed as

$$X(n,k) = |X(n,k)|e^{j\angle X(n,k)}, \tag{3}$$

where $|X(n,k)|$ denotes the short-time noisy speech magnitude spectrum and $\angle X(n,k)$ denotes the short-time

noisy speech phase spectrum. The noisy speech magnitude spectrum is enhanced, while the noisy speech phase spectrum remains unchanged. The enhanced speech magnitude spectrum is an estimate of the clean speech magnitude spectrum, and is denoted by $|\hat{S}(n,k)|$. The modified STFT is constructed by combining the enhanced speech magnitude spectrum with the noisy speech phase spectrum:

$$Y(n,k) = |\hat{S}(n,k)|e^{j\angle X(n,k)}. \tag{4}$$

The enhanced speech is constructed by applying the inverse STFT operation to the modified STFT, followed by least-squares overlap-add synthesis [19, 20]:

$$y(m) = \frac{\sum_{n=-\infty}^{\infty} w(m-nN_s)y_f(n,m-nN_s)}{\sum_{n=-\infty}^{\infty} w^2(m-nN_s)}, \tag{5}$$

where $y_f(n,m-nN_s)$ is the framed enhanced speech, after the inverse STFT operation has been applied to the modified STFT.

2.2. A Priori SNR

MMSE approaches to speech enhancement require *a priori* knowledge of the clean speech and noise. The true, unobserved SNR, or the *a priori* SNR of a noisy speech spectral component, is used for gain computation and is defined by

$$\xi(n,k) = \frac{\lambda_s(n,k)}{\lambda_d(n,k)}, \tag{6}$$

where $\lambda_s(n,k) = \mathrm{E}\{|S(n,k)|^2\}$ is the variance of the clean speech spectral component, and $\lambda_d(n,k) = \mathrm{E}\{|D(n,k)|^2\}$ is the variance of the noise spectral component. As the clean speech and noise are unobserved during enhancement, the *a priori* SNR must be estimated from the observed noisy speech. When training a supervised learning algorithm to estimate the *a priori* SNR, clean speech and noise are given. As a result, the variance of the clean speech and noise spectral components are replaced by the squared magnitude of the clean speech and noise spectral components, respectively. The *a priori* SNR computed from given clean speech and noise has been called the local *a priori* SNR previously [21].

2.3. MMSE Approaches to Speech Enhancement

The minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [1] is the optimal MMSE clean speech magnitude spectrum estimator. It uses both the *a priori* and *a posteriori* SNR of a given noisy speech spectral component to compute the gain function. The *a posteriori* SNR is given by

$$\gamma(n,k) = \frac{|X(n,k)|^2}{\lambda_d(n,k)}. \tag{7}$$

The MMSE-STSA estimator gain function is given by

$$G_{\text{MMSE-STSA}}(n,k) = \frac{\sqrt{\pi}}{2}\frac{\sqrt{\nu(n,k)}}{\gamma(n,k)}\exp\Big(\frac{-\nu(n,k)}{2}\Big)\Big((1+\nu(n,k))I_0\Big(\frac{\nu(n,k)}{2}\Big)+\nu(n,k)I_1\Big(\frac{\nu(n,k)}{2}\Big)\Big), \tag{8}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively, and $\nu(n,k)$ is given by

$$\nu(n,k) = \frac{\xi(n,k)}{\xi(n,k)+1}\gamma(n,k). \tag{9}$$

The minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator minimises the mean-squared error between the clean and enhanced speech log-magnitude spectra [2]. The MMSE-LSA gain function is given by

$$G_{\text{MMSE-LSA}}(n,k) = \frac{\xi(n,k)}{\xi(n,k)+1}\exp\left\{\frac{1}{2}\int_{\nu(n,k)}^{\infty}\frac{e^{-t}}{t}dt\right\}. \tag{10}$$

The integral in Equation 10 is known as the exponential integral.

The Wiener filter (WF) approach to estimating the clean speech magnitude spectrum [3] minimises the mean-squared error between the clean and enhanced speech complex discrete Fourier transform (DFT) coefficients. The WF gain function is given by

$$G_{\text{WF}}(n,k) = \frac{\xi(n,k)}{\xi(n,k)+1}. \tag{11}$$

The recently popularised ideal ratio mask (IRM) [8] is the square-root WF (SRWF) approach gain function [22] computed from given clean speech and noise[1]:

$$G_{\text{SRWF}}(n,k) = \sqrt{\frac{\xi(n,k)}{\xi(n,k)+1}}. \tag{12}$$

3. Mapped *A Priori* SNR Training Target

In preliminary experiments, it was found that mapping the *a priori* SNR (in dB) training target values for the k^{th} noisy speech spectral component, $\xi_{\text{dB}}(n,k)$, to the interval $[0,1]$ improved the rate of convergence of the used stochastic gradient descent algorithm. The cumulative distribution function (CDF) of $\xi_{\text{dB}}(n,k)$ was used as the map. It is assumed that $\xi_{\text{dB}}(n,k)$ is distributed normally with

[1] Comparing the speech enhancement performance of the IRM to the SRWF approach is trivial, as they have the same mathematical form.

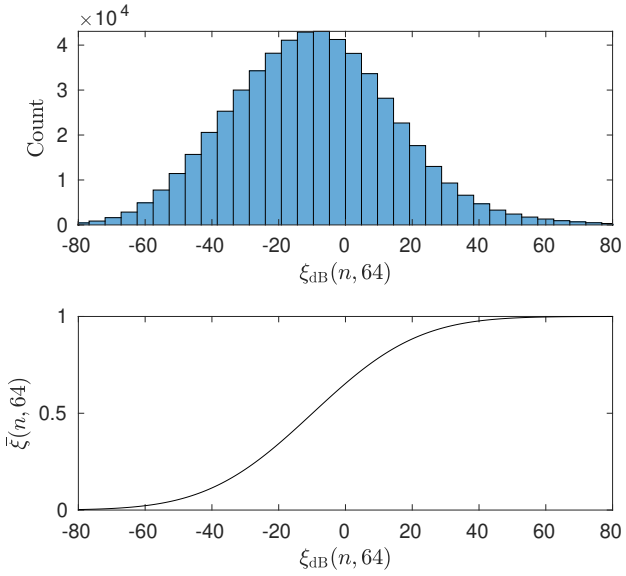

Figure 2: (Top) The distribution of $\xi_{\text{dB}}(n, 64)$, over a sample of the training set. (Bottom) The CDF of $\xi_{\text{dB}}(n, 64)$, assuming that $\xi_{\text{dB}}(n, 64)$ is distributed normally (the sample mean and variance were found over the sample of the training set).

mean μ_k and variance σ_k^2: $\xi_{\text{dB}}(n, k) \sim \mathcal{N}(\mu_k, \sigma_k^2)$. Thus, the map is given by

$$\bar{\xi}(n, k) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\xi_{\text{dB}}(n, k) - \mu_k}{\sigma_k\sqrt{2}}\right)\right], \qquad (13)$$

where $\bar{\xi}(n, k)$ is the mapped *a priori* SNR.

The statistics of $\xi_{\text{dB}}(n, k)$ for the k^{th} noisy speech spectral component were found over a sample of the training set[2]. As an example, the distribution of $\xi_{\text{dB}}(n, 64)$ found over the aforementioned sample is shown in Figure 2 (top). It can be seen that it follows a normal distribution. A poorly chosen logistic map will force large sections of the distribution to the endpoints of the target interval, $[0, 1]$. The CDF of $\xi_{\text{dB}}(n, 64)$ over the sample is shown in Figure 2 (bottom), and is used to map the distribution of $\xi_{\text{dB}}(n, 64)$ to the interval $[0, 1]$.

4. ResLSTM & ResBLSTM *A Priori* SNR Estimators

A residual long short-term memory (ResLSTM) network [13] is used to estimate the *a priori* SNR for the MMSE approaches, as shown in Figure 3. A ResLSTM consists of multiple residual blocks, with each block learning a residual function with reference to its input [23]. Residual connections allow for deep, powerful architectures [24]. The input to the ResLSTM is the magnitude spectrum of the n^{th} noisy speech frame, $|X(n, k)|$, for $k = 0, 1, ..., N_l/2$, where N_l is the frame length in discrete-time samples. The ResLSTM estimates the *a priori* SNR[3] for each of the noisy speech magnitude spectrum components.

The ResLSTM consists of 5 residual blocks, with each block containing a long short-term memory (LSTM) cell [25, 26], **F**, with a cell size of 512. LSTM cells are capable of learning both short and long-term temporal dependencies. Using LSTM cells within the residual blocks enables the ResLSTM to be a proficient sequence-based model. The residual connection is from the input of the residual block to after the LSTM cell activation [27]. **FC** is a fully-connected layer with 512 Rectified Linear Units (ReLU) [28]. Layer normalisation is used before the activation function of **FC** [29]. The output layer, **O**, is a fully-connected layer with sigmoidal units.

Shown in Figure 4 is the non-causal residual bidirectional long short-term memory (ResBLSTM) network *a priori* SNR estimator. The ResBLSTM network is identical to the ResLSTM network, except that the residual blocks include both a forward and backward LSTM cell (**F** and **B**, respectively) [14], each with a cell size of 512. While the concatenation of the forward and backward cell activations before the residual connection is standard for a ResBLSTM [30], the summation of the activations is used in this work[4]. This was to maintain a cell and residual connection size of 512, and to avoid the use of long short-term memory projection (LSTMP) cells [32]. The residual connection was applied from the input of the residual block to after the summation of the forward and backward cell activations.

Details about the training strategy for the ResLSTM and ResBLSTM *a priori* SNR estimators are given in Subsection 5.3. Training time, memory usage, and speech enhancement performance were considered when selecting the hyperparameters for the ResLSTM and ResBLSTM networks[5].

[2]The sample mean and variance of $\xi_{\text{dB}}(n, k)$ for the k^{th} noisy speech spectral component were found over 1 250 noisy speech signals created from the training clean speech and noise sets (Section 5). 250 clean speech signals from the training clean speech set were mixed with random sections of randomly selected (without replacement) noise signals from the training noise set. Each of these were mixed at five different SNR levels: -5 to 15 dB, in 5 dB increments.

[3]$\hat{\xi}(n, k)$ values are obtained by applying the inverse of Equation 13 $\left(\hat{\xi}_{\text{dB}}(n, k) = \sigma_k\sqrt{2}\text{erf}^{-1}(2\hat{\bar{\xi}}(n, k) - 1) + \mu_k\right)$, followed by $10^{(\hat{\xi}_{\text{dB}}(n,k)/10)}$ to the $\hat{\bar{\xi}}(n, k)$ values.

[4]Following the intuition that residual networks behave like ensembles of relatively shallow networks [31], the summation of the forward and backward activations can be viewed as an ensemble of the activations with no weighting.

[5]The time taken for the completion of one training epoch was approximately 9 and 18 hours for the ResLSTM and ResBLSTM networks, respectively (NVIDIA GTX 1080 Ti GPUs were used).

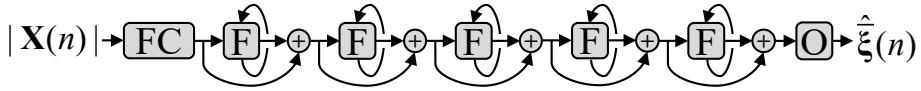

Figure 3: ResLSTM *a priori* SNR estimator. **FC** is a fully-connected layer with 512 ReLUs. The output layer, **O**, is a fully-connected layer with sigmoidal units. Each residual block contains an LSTM cell, **F**.

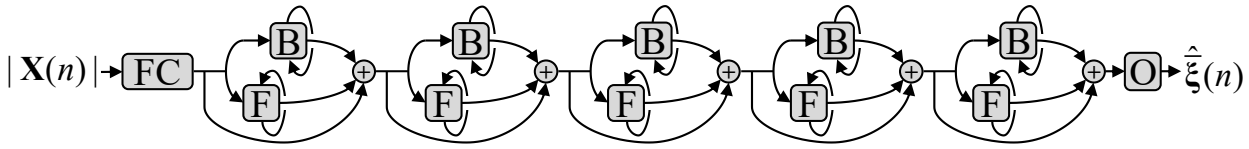

Figure 4: ResBLSTM *a priori* SNR estimator. **FC** is a fully-connected layer with 512 ReLUs. The output layer, **O**, is a fully-connected layer with sigmoidal units. Each residual block contains a forward and backward LSTM cell, **F** and **B**, respectively.

5. Experiment Setup

5.1. Signal Processing, Noise Estimation, and a posteriori SNR estimation

The Hamming window function was used as the analysis and synthesis window function [33, 34, 35], with a frame length of 32 ms ($N_l = 512$) and a frame shift of 16 ms ($N_s = 256$). The *a priori* SNR was estimated from the 257-point single-sided noisy speech magnitude spectrum, which included both the DC frequency component and the Nyquist frequency component. The MMSE-based noise estimator with speech presence probability (SPP) from [36] was used by the DD, TSNR, HRNR, and SCTS *a priori* SNR estimation methods. The *a posteriori* SNR was estimated using the observed noisy speech and the noise estimator when the DD approach, TSNR, HRNR, and SCTS *a priori* SNR estimation methods were used. When the ResLSTM and ResBLSTM *a priori* SNR estimators were used, the *a posteriori* SNR was estimated from the *a priori* SNR estimate using the following relationship: $\hat{\gamma}(n,k) = \hat{\xi}(n,k) + 1$.

5.2. Training Set

The *train-clean-100* set from the Librispeech corpus [37] (28 539 utterances), the CSTR VCTK Corpus [38] (42 015 utterances), and the si^* and sx^* training sets from the TIMIT corpus [39] (3 696 utterances) were included in the clean speech training set. The QUT-NOISE dataset [40], the Nonspeech dataset [41], the Environmental Background Noise dataset [42, 43], the noise set from the MUSAN corpus [44], multiple FreeSound packs[6], and coloured noise recordings (with an α value ranging from -2 to 2 in increments of 0.25) were included in the noise training set (2 382 recordings). All clean speech and noise signals were single-channel, with a sampling frequency of 16 kHz. The noise corruption procedure for the training set is described in Subsection 5.3.

5.3. Training Strategy

The following strategy was employed for neural network training:

- Cross-entropy as the loss function.
- The *Adam* algorithm [45] for gradient descent optimisation.
- 5% of the clean speech training set was used as a validation set.
- Each clean speech signal from a mini-batch was mixed with a random section of a randomly selected noise signal from the noise training set at a randomly selected SNR level (-10 to 20 dB, in 1 dB increments) to create the noisy speech signals for the mini-batch.
- A mini-batch size of 10 noisy speech signals.
- The selection order for the clean speech signals was randomised before each epoch.
- A total of 10 epochs were used to train the ResLSTM and ResBLSTM network.
- The LSTM network and training procedure recently proposed to estimate the IRM (LSTM-IRM estimator) [8] was replicated here. The LSTM-IRM estimator used here differed in a few ways from the original configuration: the IRM was estimated for the noisy speech magnitude spectrum (as described in Subsection 5.1), and the aforementioned training set was used to train the LSTM network for 10 epochs.

5.4. Test Set

Four recordings of four real-world noise sources, including two non-stationary and two coloured, were included in the test set. The two real-world non-stationary noise sources included *voice babble* from the RSG-10 noise dataset [46] and *street music*[7] from the Urban Sound

[6] Freesound packs 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.

[7] Street music signal number 26 270 was used from the Urban Sound dataset.

dataset [47]. The two real-world coloured noise sources included *F16* and *factory* (welding) from the RSG-10 noise dataset [46]. 10 clean speech signals were randomly selected (without replacement) from the TSP speech corpus[8] [48] for each of the four noise signal. To create the noisy speech, a random section of the noise signal was mixed with the clean speech at the following SNR levels: -5 to 15 dB, in 5 dB increments. This created a test set of 200 noisy speech files. The noisy speech signals were single channel, with a sampling frequency of 16 kHz.

5.5. Spectral Distortion

The frame-wise spectral distortion (SD) [49] between the *a priori* SNR and its estimate for the n^{th} frame is defined as the root-mean-square difference between the *a priori* SNR in dB, $\xi_{\mathrm{dB}}(n,k)$, and the estimated *a priori* SNR in dB, $\hat{\xi}_{\mathrm{dB}}(n,k)$[9]:

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

Average SD levels were obtained over the test set.

5.6. Objective Evaluation

Objective measures were used to evaluate both the quality and intelligibility of the enhanced speech. Each objective measure evaluated the enhanced speech with respect to the corresponding clean speech. Average objective scores were obtained over the test set. The objective measures that were used included:

- The mean opinion score of the objective listening quality (MOS-LQO) [50] was used for objective quality evaluation, where the wideband perceptual evaluation of quality (Wideband PESQ) [51] was the objective model used to obtain the MOS-LQO.
- The short-time objective intelligibility (STOI) measure was used for objective intelligibility evaluation [52, 53].

5.7. Subjective Evaluation

Subjective testing was performed to evaluate the quality of the enhanced speech produced by the speech enhancement methods. The mean subjective preference (%) was used as the subjective quality measure. Mean subjective preference (%) scores were determined from a series of AB listening tests [54]. Each AB listening test involved a stimuli pair. Each stimulus was either clean, noisy, or enhanced speech. The enhanced speech stimuli were produced by the MMSE-LSA estimator utilising the DD approach, Xu2017 [6, 7], and the MMSE-LSA estimator utilising the ResBLSTM *a priori* SNR estimator. Therefore, each stimulus belonged to one of the following classes: clean speech, noisy speech, enhanced speech produced by the MMSE-LSA estimator utilising the DD approach, Xu2017 enhanced speech, or enhanced speech produced by the MMSE-LSA estimator utilising the ResBLSTM *a priori* SNR estimator.

After listening to a stimuli pair, the listeners' preference was determined by selecting 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. Pair-wise scoring was used, with a score of +1 awarded to the preferred class, and 0 to the other. If the listener had an equal preference for both stimuli, each class was awarded a score of +0.5. Participants could re-listen to the stimuli pair before selecting an option.

Table 1: *A priori* SNR estimation SD levels for each of the *a priori* SNR estimators. The lowest SD for each noise source and at each SNR level is shown in boldface. The tested conditions include clean speech mixed with real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources, at multiple SNR levels.

		SNR level (dB)				
Noise	$\hat{\xi}(n,k)$	-5	0	5	10	15
Voice babble	DD[1]	18.5	17.7	17.2	17.0	17.2
	TSNR[10]	18.4	17.5	17.0	16.9	17.1
	HRNR[11]	19.5	18.9	18.5	18.4	18.6
	SCTS[12]	17.5	16.8	16.5	16.5	16.9
	ResLSTM	14.5	13.9	13.3	12.8	12.4
	ResBLSTM	**12.7**	**12.1**	**11.6**	**11.2**	**10.9**
Street music	DD[1]	19.9	18.6	17.6	17.0	16.8
	TSNR[10]	19.7	18.4	17.4	16.8	16.6
	HRNR[11]	19.8	18.7	17.9	17.5	17.5
	SCTS[12]	18.6	17.4	16.6	16.2	16.2
	ResLSTM	13.5	13.1	12.7	12.3	12.0
	ResBLSTM	**11.8**	**11.4**	**11.1**	**10.7**	**10.5**
F16	DD[1]	22.1	20.5	19.2	18.2	17.5
	TSNR[10]	21.8	20.2	18.9	17.9	17.2
	HRNR[11]	20.7	19.4	18.4	17.7	17.3
	SCTS[12]	20.8	19.2	18.0	17.1	16.6
	ResLSTM	13.3	12.7	12.3	12.0	11.7
	ResBLSTM	**11.5**	**11.0**	**10.7**	**10.4**	**10.2**
Factory	DD[1]	24.0	22.2	20.7	19.4	18.5
	TSNR[10]	23.7	22.0	20.4	19.2	18.3
	HRNR[11]	23.0	21.4	20.1	19.1	18.4
	SCTS[12]	22.4	20.7	19.3	18.2	17.4
	ResLSTM	13.8	13.2	12.7	12.4	12.1
	ResBLSTM	**13.0**	**12.2**	**11.7**	**11.3**	**11.0**

Two utterances[10] from the test set were used as the

[8] Only adult male and female speakers were included from the TSP speech corpus.

[9] $\xi_{dB}(n,k)$ and $\hat{\xi}_{dB}(n,k)$ values that were less than -40 dB, or greater than 60 dB were clipped to -40 dB and 60 dB, respectively.

[10] Using the entirety of the test set was not feasible.

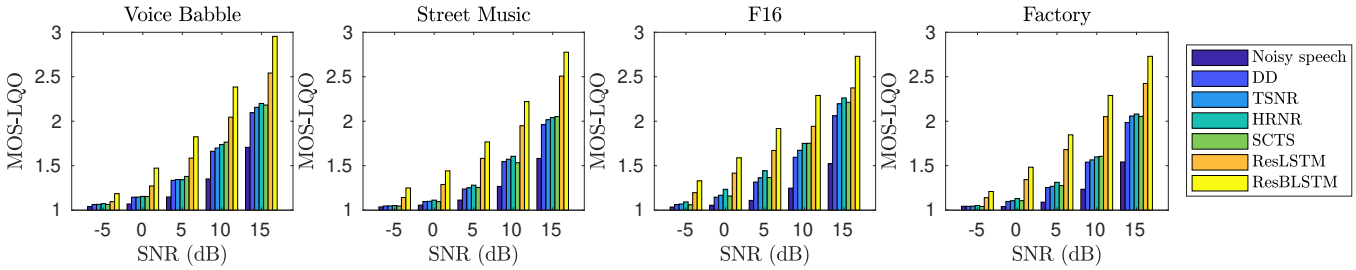

Figure 5: MMSE-STSA estimator objective quality (MOS-LQO) scores for each *a priori* SNR estimator. The tested conditions include clean speech mixed with real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources, at multiple SNR levels.

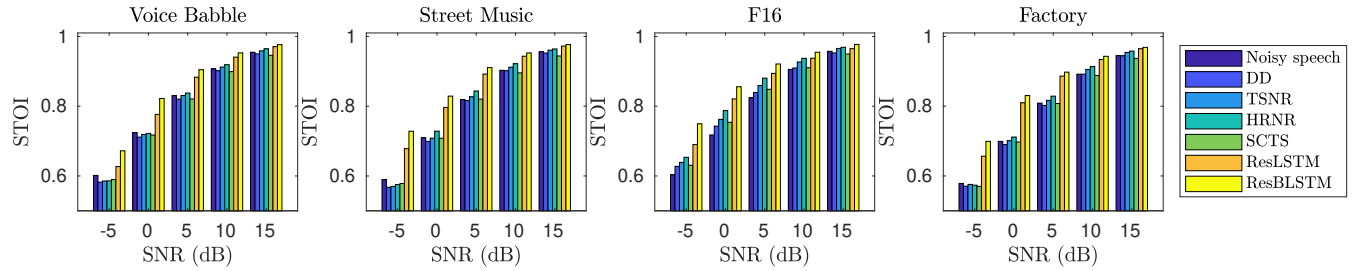

Figure 6: MMSE-STSA estimator objective intelligibility (STOI) scores for each *a priori* SNR estimator. The tested conditions include clean speech mixed with real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources, at multiple SNR levels.

clean speech stimuli: utterance 35_10, as uttered by male speaker MF, and utterance 01_03, as uttered by female speaker FA. *Voice babble* from the test set was mixed with the clean speech stimuli at an SNR level of 5 dB, producing the noisy speech stimuli. The enhanced speech stimuli for each of the speech enhancement methods was produced from the noisy speech stimuli. For each utterance, all possible stimuli pair combinations were presented to the listener (i.e. double-blind testing). Each participant listened to a total of 40 stimuli pair combinations. A total of five English-speaking listeners participated. Each listening test was conducted in a separate session, in a quiet room using closed circumaural headphones at a comfortable listening level.

6. Results and Discussion

6.1. A Priori SNR Estimation Accuracy

The *a priori* SNR estimation SD levels for each of the *a priori* SNR estimators is shown in Table 1. The SD levels were used to evaluate the accuracy of each *a priori* SNR estimator. For the real-world non-stationary noise sources, the ResLSTM *a priori* SNR estimator produced lower SD levels than the previous *a priori* SNR estimation methods (DD, TSNR, HRNR, and SCTS), with an average SD reduction of 4.7 dB when compared to the DD approach. The ResBLSTM *a priori* SNR estimator achieved an average SD reduction of 6.4 dB when compared to the DD approach, showing improved accuracy when causality is not a requirement. The proposed *a priori* SNR estimators also produced the lowest SD levels for the real-world coloured noise sources. The ResLSTM and ResBLSTM *a priori* SNR estimators achieved an average SD reduction of 7.6 and 8.9 dB, respectively, when compared to the DD approach.

The proposed *a priori* SNR estimators significantly outperform the previous *a priori* SNR estimation methods at all tested conditions. Evaluating the results in [15], the RNN-assisted DD approach (a deep learing-based *a priori* SNR estimator) could only outperform the DD approach at higher SNR levels (5 dB and greater for signal-to-distortion ratio (SDR)). Here, the ResLSTM and ResBLSTM *a priori* SNR estimators significantly outperform the DD approach for all conditions (in terms of SD level).

6.2. Objective Scores for the MMSE Approaches Utilising Deep Learning

6.2.1. MMSE-STSA Estimator Utilising Deep Learning

The objective quality and intelligibility scores for the MMSE-STSA estimator utilising each of the *a priori* SNR estimators is shown in Figure 5 and Figure 6, respectively. The MMSE-STSA estimator achieved the highest objective quality scores when deep learning was used, for both the real-world non-stationary and coloured noise sources. The MMSE-STSA estimator utilising the ResLSTM and ResBLSTM *a priori* SNR estimators achieved a MOS-LQO improvement of 0.30 and 0.52, respectively, compared to when the DD approach was used. The highest objective intelligibility scores were achieved by the MMSE-STSA estimator when deep learning was used, for both the real-world non-stationary and coloured noise sources. The MMSE-STSA estimator utilising the ResLSTM and

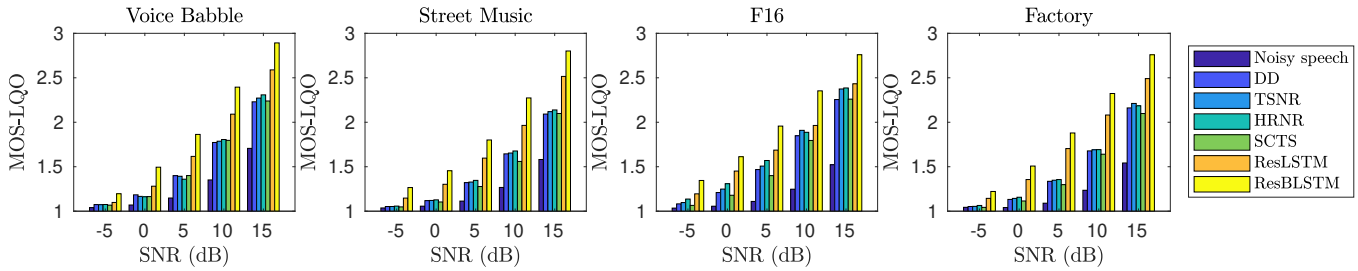

Figure 7: MMSE-LSA estimator objective quality (MOS-LQO) scores for each *a priori* SNR estimator. The tested conditions include clean speech mixed with real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources, at multiple SNR levels.

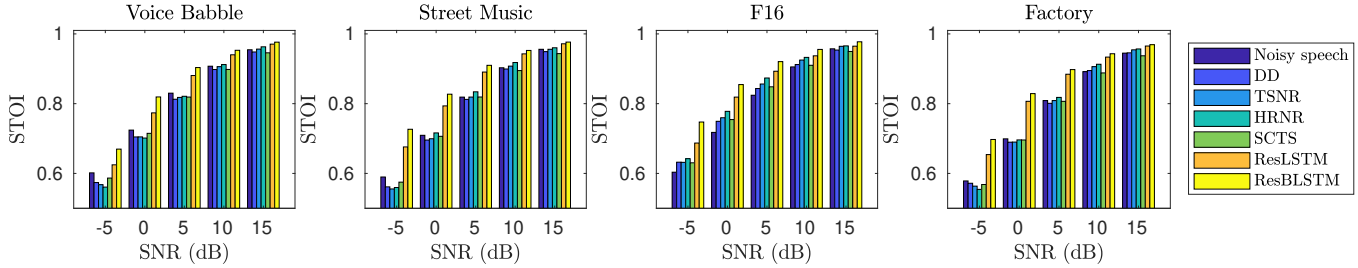

Figure 8: MMSE-LSA estimator objective intelligibility (STOI) scores for each *a priori* SNR estimator. The tested conditions include clean speech mixed with real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources, at multiple SNR levels.

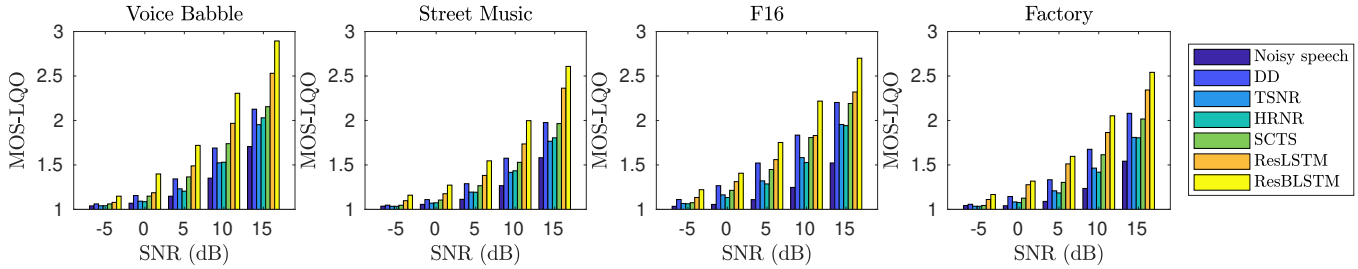

Figure 9: WF approach objective quality (MOS-LQO) scores for each *a priori* SNR estimator. The tested conditions include clean speech mixed with real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources, at multiple SNR levels.

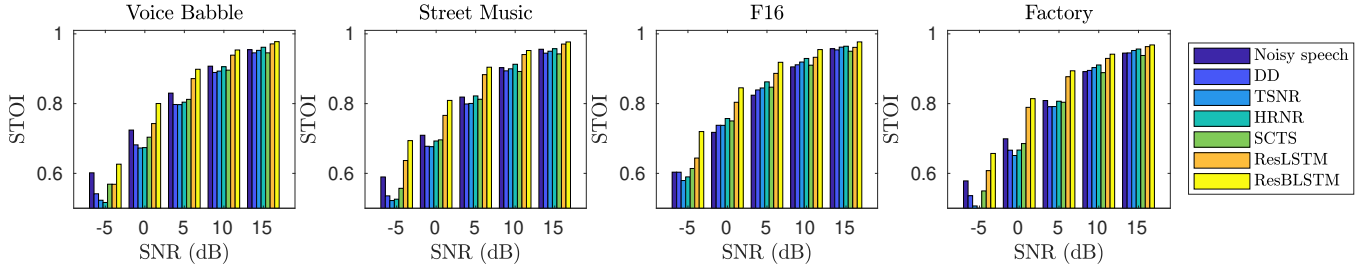

Figure 10: WF approach objective intelligibility (STOI) scores for each *a priori* SNR estimator. The tested conditions include clean speech mixed with real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources, at multiple SNR levels.

ResBLSTM *a priori* SNR estimators achieved an STOI improvement of 5.8% and 8.2%, respectively, compared to when the DD approach was used. The MMSE-STSA estimator utilising either of the proposed *a priori* SNR estimators achieved higher objective intelligibility scores than noisy speech, a feat that it struggled to achieve consistently with the other *a priori* SNR estimation methods. It can be seen that there is a correlation between *a priori* SNR estimation accuracy (given by the SD levels) and speech enhancement performance (given by the objective quality and intelligibility scores).

6.2.2. MMSE-LSA Estimator Utilising Deep Learning

The objective quality and intelligibility scores for the MMSE-LSA estimator utilising each of the *a priori* SNR estimators is shown in Figure 7 and Figure 8, respectively. The MMSE-LSA estimator achieved the highest objective quality scores when deep learning was used, for both the real-world non-stationary and coloured noise sources. The MMSE-LSA estimator utilising the ResLSTM and ResBLSTM *a priori* SNR estimators achieved a MOS-LQO improvement of 0.23 and 0.45, respectively, compared to when the DD approach was used. The objective intelligibility scores show that deep learning enabled the MMSE-LSA estimator to produce the most intelligible enhanced speech, for both the real-world non-stationary and coloured noise sources. The MMSE-LSA estimator utilising the ResLSTM and ResBLSTM *a priori* SNR estimators achieved an STOI improvement of 5.8% and 8.3%, respectively, compared to when the DD approach was used.

6.2.3. WF Approach Utilising Deep Learning

The objective quality and intelligibility scores for the WF approach utilising each of the *a priori* SNR estimators is shown in Figure 9 and Figure 10, respectively. The WF approach achieved the highest objective quality scores when deep learning was used, for both the real-world non-stationary and coloured noise sources. The WF approach utilising the ResLSTM and ResBLSTM *a priori* SNR estimators achieved a MOS-LQO improvement of 0.13 and 0.32, respectively, compared to when the DD approach was used. The objective intelligibility scores show that deep learning enabled the WF approach to produce the most intelligible enhanced speech, for both the real-world non-stationary and coloured noise sources. The WF approach utilising the ResLSTM and ResBLSTM *a priori* SNR estimators achieved an STOI improvement of 5.5% and 8.5%, respectively, compared to when the DD approach was used.

6.2.4. Comparison of MMSE approaches

The average objective quality and intelligibility scores over all of the tested conditions for each of the MMSE approaches utilising the ResLSTM and ResBLSTM *a priori* SNR estimators is shown in Table 2. It can be seen that both the MMSE-STSA and MMSE-LSA estimators outperformed the WF approach. As described previously, the MMSE-STSA and MMSE-LSA estimators are optimal MMSE clean speech magnitude spectrum estimators[11], whereas the WF approach is the optimal MMSE clean speech complex DFT coefficient estimator. The target in this work is the clean speech magnitude spectrum, which favours the MMSE-STSA and MMSE-LSA estimators. This gives reason as to why the MMSE-STSA and MMSE-LSA estimators outperformed the WF approach. The MMSE-LSA estimator was selected for the speech enhancement comparison in Subsection 6.3 as it achieved the highest average objective quality and intelligibility scores.

Table 2: Average objective quality (MOS-LQO) and intelligibility (STOI) scores for the MMSE approaches over all of the tested conditions when utilising the ResLSTM and ResBLSTM *a priori* SNR estimators. The MMSE approach that achieves the highest average objective score for each of the proposed *a priori* SNR estimators is shown in boldface.

$\hat{\xi}$	Objective measure	**MMSE approach**		
		MMSE-STSA	MMSE-LSA	WF
ResLSTM	MOS-LQO	1.71	**1.74**	1.61
	STOI	**0.85**	**0.85**	0.83
ResBLSTM	MOS-LQO	1.93	**1.96**	1.80
	STOI	**0.88**	**0.88**	0.86

6.3. Comparison of Speech Enhancement Methods

Here, an MMSE approach utilising deep learning is compared to both a masking- and a mapping-based deep learning method for speech enhancement. The MMSE-LSA estimator, utilising the ResLSTM and ResBLSTM *a priori* SNR estimators, is compared to an LSTM-IRM estimator [8], and a non-causal neural network clean speech spectrum estimator[12] that uses multi-objective learning and IBM-based post-processing [6, 7], referred to as Xu2017 in this subsection. The MMSE-LSA estimator utilising the DD approach is also compared, to represent earlier speech enhancement methods.

6.3.1. Objective Scores

The average objective quality and intelligibility scores over all of the tested conditions for each of the speech enhancement methods is shown in Table 3. The MMSE-LSA estimator utilising the causal ResLSTM *a priori* SNR estimator achieved the highest average objective quality and intelligibility scores amongst the causal speech enhancement methods. It also achieved a higher average intelligibility score than Xu2017 (a non-causal system). The MMSE-LSA estimator utilising the non-causal ResBLSTM *a priori* SNR estimator achieved the highest average objective quality and intelligibility scores amongst all of the speech enhancement methods.

[11] Specifically, the MMSE-LSA estimator is the optimal clean speech *log*-magnitude spectrum estimator.

[12] Five past and five future frames are used as part of its input feature vector.

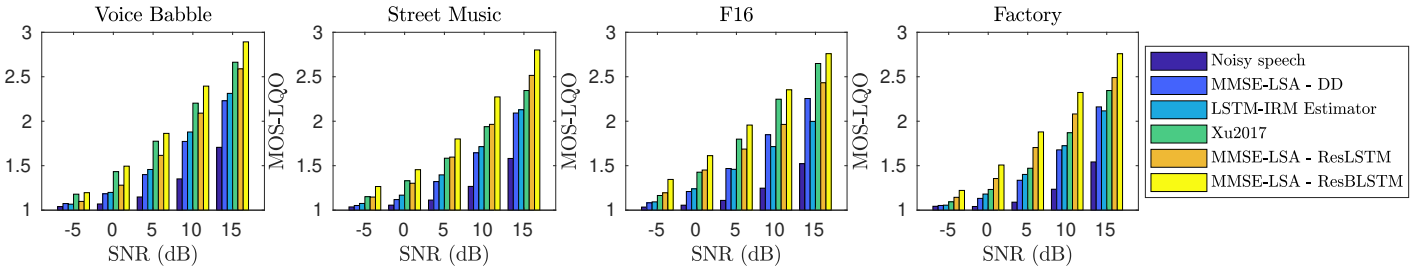

Figure 11: Objective quality (MOS-LQO) scores for the MMSE-LSA estimator utilising the DD approach, the LSTM-IRM estimator, Xu2017, and the MMSE-LSA estimator utilising both the ResLSTM and ResBLSTM *a priori* SNR estimators. The tested conditions include clean speech mixed with real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources, at multiple SNR levels.

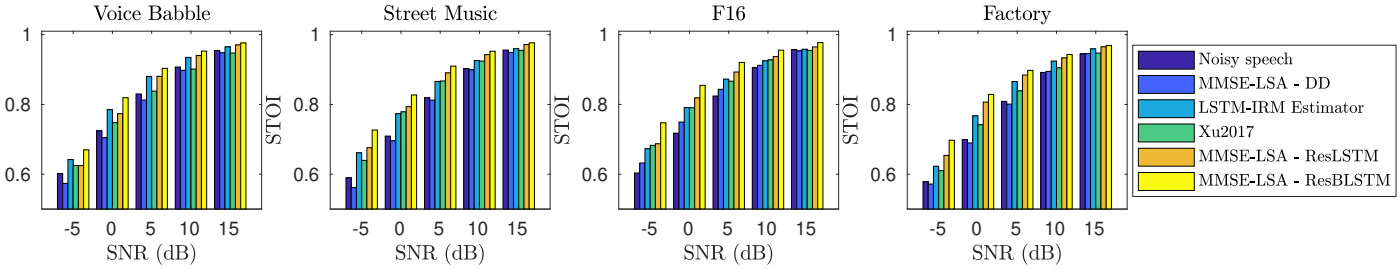

Figure 12: Objective intelligibility (STOI) scores for the MMSE-LSA estimator utilising the DD approach, the LSTM-IRM estimator, Xu2017, and the MMSE-LSA estimator utilising both the ResLSTM and ResBLSTM *a priori* SNR estimators. The tested conditions include clean speech mixed with real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources, at multiple SNR levels.

The objective quality and intelligibility scores achieved by each of the speech enhancement methods for each tested condition is shown in Figure 11 and 12, respectively. The MMSE-LSA estimator utilising the non-causal ResBLSTM *a priori* SNR estimator produced enhanced speech with higher objective quality and intelligibility scores than the LSTM-IRM estimator and Xu2017 for both real-world non-stationary and coloured noise sources. The MMSE-LSA estimator utilising the causal ResLSTM *a priori* SNR estimator achieved higher objective intelligibility scores than Xu2017 for all conditions, and the LSTM-IRM estimator for all noise sources other than *voice babble*. It also achieved higher objective quality scores than the LSTM-IRM estimator for all conditions, and Xu2017 for *street music* at high SNR levels, for *F16* at low SNR levels, and for *factory* at all SNR levels. It is important to stress that Xu2017 is a non-causal system, whilst the ResLSTM *a priori* SNR estimator is a causal system.

The advantages and disadvantages of each deep learning approach to speech enhancement can be seen in Table 3, as well as Figure 11 and 12. The advantage of Xu2017 is that it can produce enhanced speech with high objective quality scores. However, it produces enhanced speech with low objective intelligibility scores. The reverse is true for the LSTM-IRM estimator. It produces enhanced speech with low objective quality scores, but high objective intelligibility scores. On the other hand, the MMSE-LSA estimator utilising the proposed *a priori* SNR estimators is able to produce enhanced speech with both high objective quality and intelligibility scores.

Table 3: Average objective quality (MOS-LQO) and intelligibility (STOI) scores over all of the tested conditions for each of the speech enhancement methods. The highest average objective scores for both causal and non-causal speech enhancement methods are shown in boldface.

Method	Casual	MOS-LQO	STOI
Noisy speech	-	1.21	0.80
MMSE-LSA - DD	Yes	1.51	0.79
LSTM-IRM estimator	Yes	1.52	0.84
MMSE-LSA - ResLSTM	Yes	**1.74**	**0.85**
Xu2017	No	1.75	0.82
MMSE-LSA - ResBLSTM	No	**1.96**	**0.88**

6.3.2. Subjective Quality Scores

Subjective quality scores were obtained for the enhanced speech produced by the MMSE-LSA estimator utilising the DD approach, Xu2017, and the MMSE-LSA estimator utilising the ResBLSTM *a priori* SNR estimator. Details about the subjective testing procedure and the subjective test set are given in Subsection 5.7. *Voice babble* from the test set was used to produce the noisy speech at an SNR level of 5 dB. The mean subjective preference (%) for each of the speech enhancement methods is shown in Figure 13. It can be seen that the enhanced speech produced by the MMSE-LSA estimator utilising the ResBLSTM *a priori* SNR estimator (marked as MMSE-LSA (DL), where DL stands for deep learning) was preferred

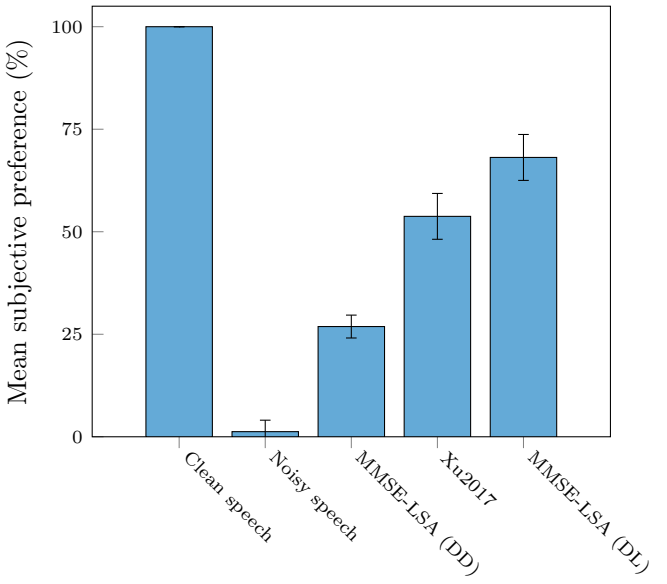

Figure 13: Mean subjective preference (%) scores for the MMSE-LSA estimator utilising the DD approach (MMSE-LSA (DD)), Xu2017, and the MMSE-LSA estimator utilising the ResBLSTM *a priori* SNR estimator (MMSE-LSA (DL), where DL stands for deep learning). The subjective testing procedure is described in Subsection 5.7. *Voice babble* from the test set was used to produce the noisy speech at an SNR level of 5 dB.

by listeners over Xu2017 enhanced speech.

6.3.3. Enhanced Speech Spectrograms

Shown in Figure 14 is the resultant enhanced speech magnitude spectrograms produced by the MMSE-LSA estimator utilising the DD approach, Xu2017, and the MMSE-LSA estimator utilising the ResBLSTM *a priori* SNR estimator. The clean and noisy speech magnitude spectrograms are shown in Figure 14 (a) and (b), respectively. The MMSE-LSA estimator utilising the ResBLSTM *a priori* SNR estimator was able to suppress most of the noise with little formant distortion (Figure 14 (e)). Xu2017 incorrectly suppressed some formant information (Figure 14 (d)). The MMSE-LSA estimator utilising the DD approach demonstrated poor noise suppression (Figure 14 (e)).

6.3.4. Areas Requiring Further Investigation

One factor that affects the performance of both the MMSE-STSA and MMSE-LSA estimators is the estimation of the *a posteriori* SNR. In this work, the *a posteriori* SNR estimate is computed using the *a priori* SNR estimate. Further performance gains may be achieved if deep learning methods are used to estimate the *a posteriori* SNR directly. Another area for investigation is the optimisation target and loss function. A recent trend has been to include the STOI objective intelligibility measure in the loss function [55, 56]. The speech enhancement performance of the proposed *a priori* SNR estimators may be

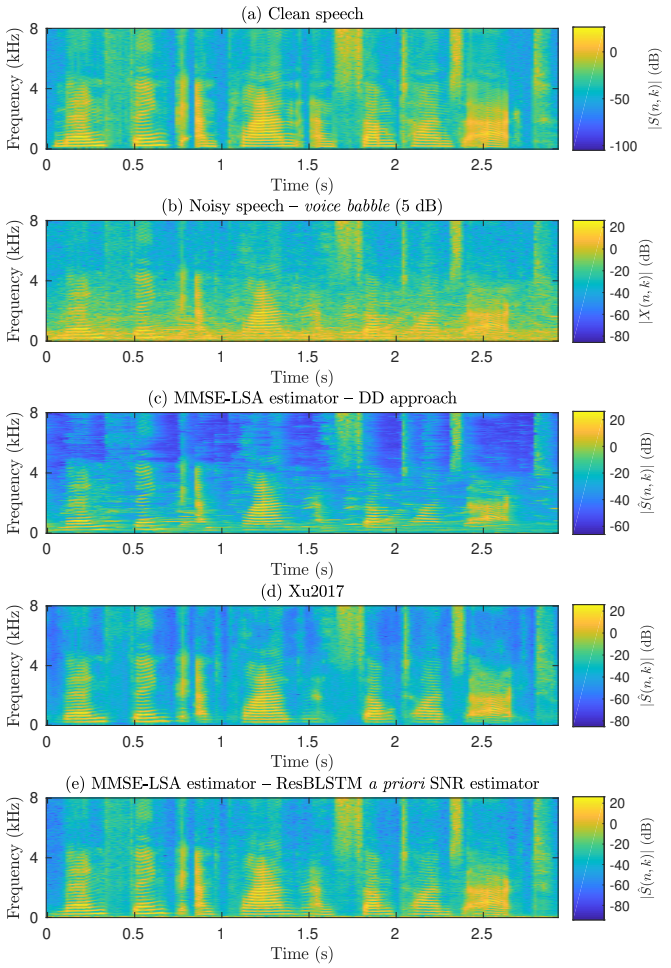

Figure 14: (a) Clean speech magnitude spectrogram of female *FF* uttering sentence 32_10, "Men think and plan and sometimes act" from the test set. (b) A recording of *voice babble* mixed with (a) at an SNR level of 5 dB. (c) Enhanced speech magnitude spectrogram produced by the MMSE-LSA estimator utilising the DD approach. (d) Enhanced speech magnitude spectrogram produced by Xu2017. (e) Enhanced speech magnitude spectrogram produced by the MMSE-LSA estimator utilising the ResBLSTM *a priori* SNR estimator.

improved if a perceptually motivated measure is integrated into the loss function.

7. Conclusion

Deep learning methods for MMSE approaches to speech enhancement are investigated in this work. A causal ResLSTM and a non-causal ResBLSTM are used here to accurately estimate the *a priori* SNR for the MMSE approaches. It was found that MMSE approaches utilising deep learning are able to produce enhanced speech that achieves higher quality and intelligibility scores than recent masking- and mapping-based deep learning approaches, for both real-world non-stationary and coloured noise sources. MMSE approaches utilising deep learning are currently being investigated for robust automatic speech recognition (ASR).

References

[1] Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing 32 (1984) 1109–1121.

[2] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, IEEE Transactions on Acoustics, Speech, and Signal Processing 33 (1985) 443–445.

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

[4] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, B. Schuller, Deep learning for environmentally robust speech recognition: An overview of recent developments, ACM Trans. Intell. Syst. Technol. 9 (2018) 49:1–49:28.

[5] X. Lu, Y. Tsao, S. Matsuda, C. Hori, Speech enhancement based on deep denoising autoencoder, in: Proceedings Interspeech 2013, 2013, pp. 436–440.

[6] Y. Xu, J. Du, L. Dai, C. Lee, A regression approach to speech enhancement based on deep neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 23 (2015) 7–19.

[7] Y. Xu, J. Du, Z. Huang, L. Dai, C. Lee, Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement, CoRR abs/1703.07172 (2017).

[8] J. Chen, D. Wang, Long short-term memory for speaker generalization in supervised speech separation, The Journal of the Acoustical Society of America 141 (2017) 4705–4714.

[9] O. Cappe, Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor, IEEE Transactions on Speech and Audio Processing 2 (1994) 345–349.

[10] C. Plapous, C. Marro, L. Mauuary, P. Scalart, A two-step noise reduction technique, in: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 1, 2004, pp. I–289. doi:10.1109/ICASSP.2004.1325979.

[11] C. Plapous, C. Marro, P. Scalart, Speech enhancement using harmonic regeneration, in: Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., volume 1, 2005, pp. I/157–I/160 Vol. 1. doi:10.1109/ICASSP.2005.1415074.

[12] C. Breithaupt, T. Gerkmann, 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, 2008, pp. 4897–4900. doi:10.1109/ICASSP.2008.4518755.

[13] J. Kim, M. El-Khamy, J. Lee, Residual LSTM: design of a deep recurrent architecture for distant speech recognition, CoRR abs/1701.03360 (2017).

[14] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (1997) 2673–2681.

[15] Y. Xia, R. Stern, A priori snr estimation based on a recurrent neural network for robust speech enhancement, in: Proc. Interspeech 2018, 2018, pp. 3274–3278. URL: http://dx.doi.org/10.21437/Interspeech.2018-2423. doi:10.21437/Interspeech.2018-2423.

[16] J. Allen, Short term spectral analysis, synthesis, and modification by discrete fourier transform, IEEE Transactions on Acoustics, Speech, and Signal Processing 25 (1977) 235–238.

[17] J. B. Allen, L. R. Rabiner, A unified approach to short-time fourier analysis and synthesis, Proceedings of the IEEE 65 (1977) 1558–1564.

[18] P. Vary, R. Martin, Digital Speech Transmission: Enhancement, Coding And Error Concealment, John Wiley & Sons, Inc., USA, 2006.

[19] D. Griffin, J. Lim, Signal estimation from modified short-time fourier transform, IEEE Transactions on Acoustics, Speech, and Signal Processing 32 (1984) 236–243.

[20] R. Crochiere, A weighted overlap-add method of short-time fourier analysis/synthesis, IEEE Transactions on Acoustics, Speech, and Signal Processing 28 (1980) 99–102.

[21] C. Plapous, C. Marro, P. Scalart, Improved signal-to-noise ratio estimation for speech enhancement, IEEE Transactions on Audio, Speech, and Language Processing 14 (2006) 2098–2108.

[22] J. S. Lim, A. V. Oppenheim, Enhancement and bandwidth compression of noisy speech, Proceedings of the IEEE 67 (1979) 1586–1604.

[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015).

[24] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, CoRR abs/1603.05027 (2016).

[25] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (1997) 1735–1780.

[26] F. A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: continual prediction with lstm, in: 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470), volume 2, 1999, pp. 850–855 vol.2. doi:10.1049/cp:19991218.

[27] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean, Google's neural machine translation system: Bridging the gap between human and machine translation, CoRR abs/1609.08144 (2016).

[28] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, Omnipress, USA, 2010, pp. 807–814. URL: http://dl.acm.org/citation.cfm?id=3104322.3104425.

[29] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450 (2016).

[30] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, Y. Zhou, Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks, Bioinformatics 34 (2018) 4039–4045.

[31] A. Veit, M. J. Wilber, S. J. Belongie, Residual networks are exponential ensembles of relatively shallow networks, CoRR abs/1605.06431 (2016).

[32] H. Sak, A. Senior, F. Beaufays, Long short-term memory recurrent neural network architectures for large scale acoustic modeling, in: Proceedings Interspeech 2014, 2014, pp. 338–342.

[33] J. W. Picone, Signal modeling techniques in speech recognition, Proceedings of the IEEE 81 (1993) 1215–1247.

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

[35] K. Paliwal, K. Wojcicki, Effect of analysis window duration on speech intelligibility, IEEE Signal Processing Letters 15 (2008) 785–788.

[36] T. Gerkmann, R. C. Hendriks, Unbiased mmse-based noise power estimation with low complexity and low tracking delay, IEEE Transactions on Audio, Speech, and Language Processing 20 (2012) 1383–1393.

[37] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An asr corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206–5210. doi:10.1109/ICASSP.2015.7178964.

[38] 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).

[39] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1, NASA STI/Recon Technical Report N 93 (1993).

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

[41] G. Hu, 100 nonspeech environmental sounds, The Ohio State University, Department of Computer Science and Engineering (2004).

[42] F. Saki, A. Sehgal, I. Panahi, 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), 2016, pp. 2204–2208. doi:10.1109/ICASSP.2016.7472068.

[43] F. Saki, 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), 2016, pp. 736–739. doi:10.1109/EMBC.2016.7590807.

[44] D. Snyder, G. Chen, D. Povey, MUSAN: A music, speech, and noise corpus, CoRR abs/1510.08484 (2015).

[45] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, CoRR abs/1412.6980 (2014).

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

[47] J. Salamon, C. Jacoby, J. P. Bello, A dataset and taxonomy for urban sound research, in: Proceedings of the 22Nd ACM International Conference on Multimedia, MM '14, ACM, New York, NY, USA, 2014, pp. 1041–1044. URL: http://doi.acm.org/10.1145/2647868.2655045. doi:10.1145/2647868.2655045.

[48] P. Kabal, TSP speech database, McGill University, Database Version (2002).

[49] K. K. Paliwal, 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, 1991, pp. 661–664 vol. 1. doi:10.1109/ICASSP.1991.150426.

[50] ITU-T Recommendation P.800.1, Mean opinion score (MOS) terminology, 2006.

[51] ITU-T Recommendation P.862.2, Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs, 2007.

[52] C. H. Taal, R. C. Hendriks, R. Heusdens, 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, 2010, pp. 4214–4217. doi:10.1109/ICASSP.2010.5495701.

[53] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of timefrequency weighted noisy speech, IEEE Transactions on Audio, Speech, and Language Processing 19 (2011) 2125–2136.

[54] S. So, K. K. Paliwal, Modulation-domain kalman filtering for single-channel speech enhancement, Speech Communication 53 (2011) 818 – 829.

[55] S. Fu, T. Wang, Y. Tsao, X. Lu, H. Kawai, End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2018) 1570–1584.

[56] Y. Zhao, B. Xu, R. Giri, T. Zhang, Perceptually guided speech enhancement using deep neural networks, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5074–5078. doi:10.1109/ICASSP.2018.8462593.