Deep learning for minimum mean-square error approaches to speech enhancement

Aaron Nicolson*, Kuldip K. Paliwal
Signal Processing Laboratory, Griffith University, Brisbane, Queensland 4111, Australia

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 (Ephraim and Malah, 1984). Other prominent MMSE approaches to speech enhancement include the minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator (Ephraim and Malah, 1985) and the Wiener filter (WF) approach (Loizou, 2013). 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 (Zhang et al., 2018). Neural networks have been used as non-linear maps from noisy speech spectra to clean speech spectra. A denoising autoencoder (DAE) was pretrained for this task using noisy and clean speech pairs (Lu et al., 2013). A non-causal neural network clean speech spectrum estimator was proposed that produced enhanced speech with high objective quality scores (Xu et al., 2015), which later incorporated multi-objective learning and ideal binary mask (IBM)-based post-processing (Xu et al., 2017). 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) (Chen and Wang, 2017).

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 Section 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.

A priori SNR estimation is a difficult task, especially when considering the multitude of different noise sources. The decision-directed (DD) approach (Ephraim and Malah, 1984) 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 (Cappe, 1994), which is addressed by the two-step noise reduction (TSNR) technique (Plapous et al., 2004). Harmonic regeneration noise reduction (HRNR) (Plapous et al., 2005) 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 cepstrum-temporal smoothing (SCTS) (Breithaupt et al., 2008) 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-term memory (ResLSTM) networks are proficient acoustic models (Kim et al., 2017). Motivated by this, a causal ResLSTM network, and a non-causal residual bidirectional LSTM (ResBLSTM) network (Schuster and Paliwal, 1997) 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 (Xia and Stern, 2018). The proposed estimators differ by directly estimating the a priori SNR. This was accomplished by using the oracle case as the training target, where the oracle case is defined as the a priori SNR computed from the clean speech and noise. It was found that mapping the oracle 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 oracle 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 real-world non-stationary and coloured noise sources at multiple SNR levels. The MMSE approaches utilising deep learning are compared to recent 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 speech enhancement performance of the mapped a priori SNR, the IRM, and the clean speech magnitude spectrum as the training target is also evaluated.

The paper is organised as follows: background knowledge is presented in Section 2, including the analysis, modification, and synthesis (AMS) procedure, and MMSE approaches to speech enhancement; 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.

2. Background

2.1. AMS speech enhancement framework

The short-time Fourier analysis, modification, and synthesis (AMS) framework is used here to produce the enhanced speech. The AMS framework (Allen, 1977; Allen and Rabiner, 1977) 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 Fig. 1.

An uncorrelated additive noise model is assumed:

\[ x(m) = s(m) + d(m), \]

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 (Vary and Martin, 2006):

![Block diagram of the short-time Fourier AMS speech enhancement framework.](image)

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

\[
X(n, k) = \sum_{n=0}^{N-1} x(m + nN_s)e^{-j2\pi nk/N_s},
\]

where \( n \) denotes the frame index, \( k \) denotes the discrete-frequency index, \( N_s \) denotes the frame length in discrete-time samples, \( N_d \) 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

\[
X(n, k) = \left| X(n, k) \right|e^{j\angle X(n, k)},
\]

where \( |X(n, k)| \) and \( \angle X(n, k) \) denote the short-time magnitude and phase spectrum of the noisy speech, respectively. 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 \( \hat{\left| S(n, k) \right|} \). The modified STFT is constructed by combining the enhanced speech magnitude spectrum with the noisy speech phase spectrum:

\[
Y(n, k) = \left| \hat{S}(n, k) \right|e^{j\angle X(n, k)};
\]

The enhanced speech is constructed by applying the inverse STFT operation to the modified STFT, followed by least-squares overlap-add synthesis (Griffin and Lim, 1984; Crochiere, 1980):

\[
y(m) = \sum_{m=-\infty}^{\infty} x(m - nN_s)y_f(n, m - nN_s).
\]

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

An MMSE approach to speech enhancement utilises the a priori SNR to compute a gain function. The gain function is subsequently applied to the magnitude spectrum of the noisy speech, which produces the enhanced speech magnitude spectrum. The a priori SNR of a noisy speech spectral component is defined as

\[
\xi(n, k) = \frac{\lambda_s(n, k)}{\lambda_d(n, k)},
\]

where \( \lambda_s(n, k) = E[|S(n, k)|^2] \) is the variance of the clean speech spectral component, and \( \lambda_d(n, k) = E[|D(n, k)|^2] \) is the variance of the noise.
spectral component. As the clean speech and noise are unobserved during speech 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, the clean speech and noise are given (the oracle case). 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 oracle case has been called the local a priori SNR previously (Plapous et al., 2006).

2.3. MMSE approaches to speech enhancement

The minimum mean-square error short-time spectral amplitude (MMSE-LSA) estimator (Ephraim and Malah, 1984) optimally estimates (in the mean-square error (MSE) sense) the magnitude spectrum of the clean speech. 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

\[ g(n, k) = \frac{|X(n, k)|^2}{\tilde{d}(n, k)}. \]

The MMSE-LSA estimator gain function is given by

\[ G_{\text{MMSE-LSA}}(n, k) = \frac{\sqrt{2}}{2} \frac{\sqrt{\nu(n, k)}}{7(n, k)} \exp\left(-\frac{\nu(n, k)}{2}\right) \times \left(1 + \nu(n, k)I_0\left(\frac{\nu(n, k)}{2}\right) + \nu(n, k)I_1\left(\frac{\nu(n, k)}{2}\right)\right), \]

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

\[ \nu(n, k) = \frac{\xi(n, k)}{\xi(n, k) + 1} \gamma(n, k). \]

The minimum mean-square error log-spectral amplitude (MMSE-LSA) estimator minimises the MSE between the clean and enhanced speech log-magnitude spectrum (Ephraim and Malah, 1985). The MMSE-LSA gain function is given by

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

The integral in Eq. (10) is known as the exponential integral.

The Wiener filter (WF) approach to estimating the clean speech magnitude spectrum (Loizou, 2013) minimises the MSE between the clean and enhanced speech complex discrete Fourier transform (DFT) coefficients. The gain function for the WF approach is given by

\[ G_{\text{WF}}(n, k) = \frac{\xi(n, k)}{\xi(n, k) + 1}. \]

The recently popularised ideal ratio mask (IRM) (Chen and Wang, 2017) is the square-root WF (SRWF) approach gain function (Lim and Oppenheim, 1979) computed from given clean speech and noise:

\[ G_{\text{SRWF}}(n, k) = \frac{\xi(n, k)}{\xi(n, k) + 1}. \]

3. Mapped a priori SNR training target

In preliminary experiments, it was found that mapping the oracle a priori SNR (in dB) training target values for the kth noisy speech spectral component, \( \tilde{\xi}(n, k) \), to the interval [0, 1] improved the rate of convergence of the used stochastic gradient descent algorithm. The cumulative distribution function (CDF) of \( \tilde{\xi}(n, k) \) was used as the map. It is assumed that \( \tilde{\xi}(n, k) \) is distributed normally with mean \( \mu_k \) and variance \( \sigma_k^2 \), \( \tilde{\xi}(n, k) \sim \mathcal{N}(\mu_k, \sigma_k^2) \). Thus, the map is given by

\[ \tilde{\xi}(n, k) = \frac{1}{2} \left( 1 + \text{erf} \left( \frac{\tilde{\xi}(n, k) - \mu_k}{\sigma_k \sqrt{2}} \right) \right). \]

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

Fig. 2. (Top) The distribution of \( \tilde{\xi}(n, 64) \), over a sample of the training set. (Bottom) The CDF of \( \tilde{\xi}(n, 64) \), assuming that \( \tilde{\xi}(n, 64) \) is distributed normally (the sample mean and variance were found over the sample of the training set).

The statistics of \( \tilde{\xi}(n, k) \) for the kth noisy speech spectral component were found over a sample of the training set. As an example, the distribution of \( \tilde{\xi}(n, 64) \) found over the aforementioned sample is shown in Fig. 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 \( \tilde{\xi}(n, 64) \) over the sample is shown in Fig. 2 (bottom), and is used to map the distribution of \( \tilde{\xi}(n, 64) \) to the interval [0,1].

4. ResLSTM & ResBLSTM a priori SNR estimators

A residual long short-term memory (ResLSTM) network (Kim et al., 2017) is used to estimate the a priori SNR for the MMSE approaches, as shown in Fig. 3 (top). A ResLSTM consists of multiple residual blocks, with each block learning a residual function with reference to its input (He et al., 2015). Residual connections allow for deep, powerful architectures (He et al., 2016). The input to the ResLSTM is the magnitude spectrum of the nth noisy speech frame, \( |X(n, k)| \), for \( k = 0, 1, \ldots, N_k/2 \), where \( N_k \) is the frame length in discrete-time samples. The ResLSTM estimates the a priori SNR 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 ( Hochreiter and Schmidhuber, 1997; Gers et al., 1999), 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 (Wu et al., 2016). FC is a fully-connected layer with 512
Rectified Linear Units (ReLUs) (Nair and Hinton, 2010). Layer normalisation is used before the activation function of FC (Ba et al., 2016). The output layer, O, is a fully-connected layer with sigmoidal units. Shown in Fig. 3 (bottom) is the non-causal residual bidirectional long short-term memory (ResBLSTM) network \textit{a priori} SNR estimator. The ResBLSTM is identical to the ResLSTM, except that the residual blocks include both a forward and backward LSTM cell (F and B, respectively) (Schuster and Paliwal, 1997), 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 (Hanson et al., 2018), the summation of the activations is used in this work.\footnote{Following the intuition that residual networks behave like ensembles of relatively shallow networks (Veit et al., 2016), the summation of the forward and backward activations can be viewed as an ensemble of the activations with no weighting.} 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 (Sak et al., 2014). 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 \textit{a priori} SNR estimators are given in Section 5.3. Training time, memory usage, and speech enhancement performance were considered when selecting the hyperparameters for the ResLSTM and ResBLSTM networks.\footnote{The time taken for the completion of one training epoch for the ResLSTM and the ResBLSTM networks was approximately 9 and 18 hours, respectively (NVIDIA GTX 1080 Ti GPUs were used).}

5. \textbf{Experiment setup}

5.1. Signal processing, noise estimation, and a posteriori SNR estimation

The Hamming window function was used for analysis and synthesis (Picone, 1993; Huang et al., 2001; Paliwal and Wojcicki, 2008), with a frame length of 32 ms ($N_f = 512$) and a frame shift of 16 ms ($N_s = 256$). The \textit{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 Gerkmann and Hendriks (2012) was used by the DD, TSNR, HRNR, and SCTS \textit{a priori} SNR estimation methods. The \textit{a posteriori} SNR was estimated using both the observed noisy speech and the noise estimator when the DD approach, TSNR, HRNR, and SCTS \textit{a priori} SNR estimation methods were used. When the ResLSTM and ResBLSTM \textit{a priori} SNR estimators were used, the \textit{a posteriori} SNR was estimated from the \textit{a priori} SNR estimate using the following relationship: $\hat{\gamma}(n, k) = \hat{\xi}(n, k) + 1$.

5.2. Training set

The \textit{train-clean}100 set from the LibriSpeech corpus (Panayotov et al., 2015) (28,539 utterances), the CSTR VCTK Corpus (Veaux et al., 2017) (42,015 utterances), and the st$^1$ and sx$^1$ training sets from the TIMIT corpus (Garofolo et al., 1993) (3,696 utterances) were included in the clean speech training set. The QUT-NOISE dataset (Dean et al., 2010), the Nonspeech dataset (Hu, 2004), the Environmental Background Noise dataset (Saki et al., 2016; Saki and Kehtarnavaz, 2016), the noise set from the MUSAN corpus (Snyder et al., 2015), multiple FreeSound packs,\footnote{FreeSound packs that were used: 147, 199, 247, 379, 622, 643, 1133, 1563, 1,840, 2,432, 4,366, 4,439, 15,046, 15,598, 21,558.} and coloured noise recordings (with an \textit{a} 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 Section 5.3.

5.3. Training strategy

The following strategy was employed for neural network training:

- Cross-entropy as the loss function.
- The Adam algorithm (Kingma and Ba, 2014) for gradient descent optimisation.
- 5\% of the clean speech training set was used as a validation set.
- For each mini-batch, each clean speech signal 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.
- 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 networks.
- The LSTM-IRM estimator (Chen and Wang, 2017) was replicated here, and used the noisy speech magnitude spectrum (as described in Section 5.1) as its input. It was trained for 10 epochs using the aforementioned training set.

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 (Steeneken and Geurtsen, 1988) and street music\footnote{Street music recording number 26270 was used from the Urban Sound dataset.} from the Urban Sound dataset (Salamon et al., 2014). The two real-world coloured noise sources included F16 and factory (welding) from the RSG-10 noise dataset (Steeneken and Geurtsen, 1988). 10 clean speech signals were randomly selected (without replacement) from the TSP speech corpus\footnote{Only adult speakers were included from the TSP speech corpus.} (Kabal, 2002) for each of the four noise signals. 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) (Paliwal and Atal, 1991) is defined as the root-mean-square difference between the \textit{a priori} SNR...
estimate in dB, $\hat{\xi}_{dn}(n, k)$, and the oracle case in dB, $\xi_{dn}(n, k)$, for the $n^{th}$ frame:\footnote{\(\hat{\xi}_{dn}(n, k)\) and \(\xi_{dn}(n, k)\) values that were less than $-40$ dB, or greater than $60$ dB were clipped to $-40$ dB and $60$ dB, respectively.}\footnote{Using the entirety of the test set was not feasible.}

\[
D_n^2 = \frac{1}{N_f/2 + 1} \sum_{k=0}^{N_f/2} [\hat{\xi}_{dn}(n, k) - \xi_{dn}(n, k)]^2.
\]

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) (ITU-T Recommendation P.800.1, 2006) was used for objective quality evaluation, where the wideband perceptual evaluation of quality (Wideband PESQ) (ITU-T Recommendation P.862.2, 2007) was the objective model used to obtain the MOS-LQO.
- The short-time objective intelligibility (STOI) measure was used for objective intelligibility evaluation (Taal et al., 2010; 2011).

5.7. Subjective evaluation

Subjective testing was used 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 (So and Paliwal, 2011). 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 (Xu et al., 2015; 2017), 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.

Two utterances\footnote{Using the entirety of the test set was not feasible.} from the test set were used as the 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 are used to evaluate the accuracy of each a priori SNR estimator. For 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. Evaluating the results presented by Xia and Stern (2018), the RNN-assisted DD approach (a deep learning-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.

6.2. 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 are shown in Figs. 4 and 5, 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 an average MOS-LQO improvement

<table>
<thead>
<tr>
<th>Noise Source</th>
<th>SNR Level (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-5</td>
</tr>
<tr>
<td>DD (Ephraim and Malah, 1984)</td>
<td>18.5</td>
</tr>
<tr>
<td>TSNR (Plapous et al., 2004)</td>
<td>18.4</td>
</tr>
<tr>
<td>Voice babble</td>
<td>19.5</td>
</tr>
<tr>
<td>Street music</td>
<td>17.5</td>
</tr>
<tr>
<td>ResLSTM</td>
<td>14.5</td>
</tr>
<tr>
<td>ResBLSTM</td>
<td>12.7</td>
</tr>
<tr>
<td>DD (Ephraim and Malah, 1984)</td>
<td>19.9</td>
</tr>
<tr>
<td>TSNR (Plapous et al., 2004)</td>
<td>19.7</td>
</tr>
<tr>
<td>Voice babble</td>
<td>19.8</td>
</tr>
<tr>
<td>Street music</td>
<td>18.6</td>
</tr>
<tr>
<td>ResLSTM</td>
<td>13.5</td>
</tr>
<tr>
<td>ResBLSTM</td>
<td>11.8</td>
</tr>
<tr>
<td>DD (Ephraim and Malah, 1984)</td>
<td>22.1</td>
</tr>
<tr>
<td>TSNR (Plapous et al., 2004)</td>
<td>21.8</td>
</tr>
<tr>
<td>Voice babble</td>
<td>20.7</td>
</tr>
<tr>
<td>Street music</td>
<td>20.8</td>
</tr>
<tr>
<td>ResLSTM</td>
<td>13.3</td>
</tr>
<tr>
<td>ResBLSTM</td>
<td>11.5</td>
</tr>
<tr>
<td>DD (Ephraim and Malah, 1984)</td>
<td>24.0</td>
</tr>
<tr>
<td>TSNR (Plapous et al., 2004)</td>
<td>23.7</td>
</tr>
<tr>
<td>Voice babble</td>
<td>23.0</td>
</tr>
<tr>
<td>Street music</td>
<td>22.4</td>
</tr>
<tr>
<td>ResLSTM</td>
<td>13.8</td>
</tr>
<tr>
<td>ResBLSTM</td>
<td>13.0</td>
</tr>
</tbody>
</table>
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 ResBLSTM a priori SNR estimators achieved an average 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 are shown in Figs. 6 and 7, 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 an average 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 average 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 average MOS-LQO improvement of 0.13 and 0.32, 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 are shown in Figs. 8 and 9, 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 an average 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 WP 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 average STOI improvement of 5.5% and 8.5%, respectively, compared to when the DD approach was used.

6.2.4. Comparison of MMSE approaches

A comparison of each MMSE approach utilising the proposed a priori SNR estimators is shown in Table 2. It can be seen that both the MMSE-STS and MMSE-LSA estimators outperformed the WF approach. As described previously, when the MMSE-STS and MMSE-LSA estimators are optimal MMSE clean speech magnitude spectrum estimators, 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-STS and MMSE-LSA estimators. This gives reason as to why the MMSE-STS and MMSE-LSA estimators outperformed the WF approach. The MMSE-LSA estimator was selected for the speech enhancement comparison in Section 6.4 as it achieved the highest average objective quality score, and the second highest average objective intelligibility score.

6.3. Comparison of training targets

Here, the speech enhancement performance of the (mapped) a priori SNR, the IRM, and the clean speech magnitude spectrum as the training target is evaluated. The training strategy described in Section 5.3 was used to train an identical ResLSTM network for each training target. The SRWF approach, MMSE-STS estimator, and the MMSE-LSA estimator are used to evaluate the a priori SNR training target. The SRWF approach is used instead of the WF approach as it has the same form as the IRM. The objective quality and intelligibility scores achieved by each training target are shown in Figs. 10 and 11, respectively. The a priori SNR training target achieved the highest objective quality scores, for both the real-world non-stationary and coloured noise sources (except for voice babble at 15 dB). However, the IRM training target achieved the highest objective intelligibility scores, for both the real-world non-stationary and coloured noise sources (except for factory at 0 dB).

---

**Table 2**

The average improvement over the MMSE approach in the preceding row is shown for both objective quality (MOS-LQO) and intelligibility (STOI).

<table>
<thead>
<tr>
<th>( \xi )</th>
<th>Gain</th>
<th>MOS-LQO</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResLSTM</td>
<td>WF</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ResLSTM</td>
<td>MMSE-STS</td>
<td>+0.10</td>
<td>+1.67%</td>
</tr>
<tr>
<td>ResLSTM</td>
<td>MMSE-LSA</td>
<td>+0.02</td>
<td>–0.15%</td>
</tr>
<tr>
<td>ResBLSTM</td>
<td>WF</td>
<td>+0.07</td>
<td>+1.37%</td>
</tr>
<tr>
<td>ResBLSTM</td>
<td>MMSE-STS</td>
<td>+0.13</td>
<td>+1.19%</td>
</tr>
<tr>
<td>ResBLSTM</td>
<td>MMSE-LSA</td>
<td>+0.02</td>
<td>–0.08%</td>
</tr>
</tbody>
</table>

---

10 Specifically, the MMSE-LSA estimator is the optimal clean speech log-magnitude spectrum estimator.

11 The cross-entropy loss function was used when optimising for the mapped a priori SNR and the IRM. In contrast, the quadratic loss function was used when optimising for the clean speech MS, as its values are not bounded to the interval [0,1].
It can be seen in Table 3 and in Figs. 10 and 11 that the a priori SNR and the IRM both outperform the clean speech magnitude spectrum as the training target. These results are consistent with those reported in the literature. A study on training targets by Wang et al. (2014) found that the IRM as the training target produces significantly higher objective quality and intelligibility scores than the clean speech magnitude spectrum (as indicated by FFT-MAG in Wang et al., 2014) for both real-world non-stationary and coloured noise sources at multiple SNR levels (−5, 0, and 5 dB). It has also been shown by Zhao et al. (2016) that higher objective intelligibility scores are obtained when the IRM is used instead of the clean speech magnitude spectrum as the training target, for voice babble at multiple SNR levels (−5, 0, and 5 dB) (as shown by Fig. 2 in Wang et al., 2014).

As can be seen in Table 3, there is a trade-off between enhanced speech quality and intelligibility when selecting between the IRM and the a priori SNR as the training target. If it is desired to produce enhanced speech that is more intelligible, the IRM should be chosen as the training target. If it is desired for the enhanced speech to have a higher quality, the a priori SNR should be chosen as the training target. A further trade-off between enhanced speech quality and intelligibility can be made through the selection of the MMSE approach. Amongst the MMSE approaches, the SRWF approach produces the most intelligible enhanced speech, but with the worst quality. On the contrary, the MMSE-LSA estimator produces the least intelligible enhanced speech, but with the highest quality. The MMSE-STSA estimator offers a compromise between the SRWF approach and the MMSE-LSA estimator.

6.4. Comparison of speech enhancement methods

Here, an MMSE approach utilising deep learning is compared to both a masking- and a mapping-based deep learning approach to speech enhancement. The MMSE-LSA estimator, utilising the ResLSTM and ResBLSTM a priori SNR estimators, is compared to the LSTM-IRM estimator from Chen and Wang (2017), and the non-causal neural network clean speech spectrum estimator\(^\text{12}\) that uses multi-objective learning and IBM-based post-processing from Xu et al. (2015, 2017), referred to as Xu2017 in this subsection. The MMSE-LSA estimator utilising the DD approach is also compared, to represent earlier speech enhancement methods.

6.4.1. Objective scores

The objective quality and intelligibility scores achieved by each of the speech enhancement methods for each tested condition are shown in Figs. 12 and 13, 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.

Table 4 details the average improvement that the proposed a priori SNR estimators hold over the other speech enhancement methods. 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

---

\(^{12}\) Five past and five future frames are used as part of its input feature vector.
quality and intelligibility scores amongst all the speech enhancement methods.

The advantages and disadvantages of each deep learning approach to speech enhancement can be seen in Table 4, as well as Figs. 12 and 13. 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.

When considering the training target results from Section 6.3, it can be deduced that most of the performance improvement gained by the MMSE-LSA estimator utilising the ResLSTM a priori SNR estimator over the LSTM-IRM estimator is due to the differing model and training strategy, and not the training target. However, the opposite is likely true for Xu2017. The results from Section 6.3 indicate that most of the performance improvement gained by the MMSE-LSA estimator utilising the ResBLSTM a priori SNR estimator over Xu2017 is due to the training target, and not the model, training strategy, or post-processing.

The LSTM-IRM estimator from Chen and Wang (2017) uses the quadratic loss function instead of the cross entropy loss function employed by the proposed a priori SNR estimators.

The average improvement over the speech enhancement method in the preceding row is shown for both objective quality (MOS-LQO) and intelligibility (STOI).

### Table 4

<table>
<thead>
<tr>
<th>Method</th>
<th>Casual MOS-LQO</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMSE-LSA; DD $\xi$ (Ephraim and Malah, 1984)</td>
<td>Yes</td>
<td>–</td>
</tr>
<tr>
<td>LSTM-IRM est. (Chen and Wang, 2017)</td>
<td>Yes</td>
<td>+0.01</td>
</tr>
<tr>
<td>MMSE-LSA; ResLSTM $\xi$</td>
<td>Yes</td>
<td>+0.22</td>
</tr>
<tr>
<td>Xu2017 (Xu et al., 2015; 2017)</td>
<td>No</td>
<td>+0.01</td>
</tr>
<tr>
<td>MMSE-LSA; ResBLSTM $\xi$</td>
<td>No</td>
<td>+0.21</td>
</tr>
</tbody>
</table>

Fig. 12. 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 real-world non-stationary (voice babble and street music) and coloured (F16 and factory) noise sources at multiple SNR levels.

Fig. 13. 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 real-world non-stationary (voice babble and street music) and coloured (F16 and factory) noise sources at multiple SNR levels.

Fig. 14. 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 Section 5.7. Voice babble at an SNR level of 5 dB was the condition used for the subjective tests.

6.4.2. Subjective quality scores

Subjective quality scores were obtained for 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 Section 5.7. Voice babble at an SNR level of 5 dB was the condition used for the subjective tests. The mean subjective preference (%) for each of the speech enhancement methods is shown in Fig. 14. It can be seen that the enhanced speech produced by the MMSE-LSA estimator utilising the ResBLSTM...
6.4.3. Enhanced speech spectrograms

Shown in Fig. 15 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 Fig. 15 (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 (Fig. 15 (e)). Xu2017 incorrectly suppressed some formant information (Fig. 15 (d)). The MMSE-LSA estimator utilising the DD approach demonstrated poor noise suppression (Fig. 15 (e)).

6.5. Areas requiring further investigation

One factor that affects the performance of the MMSE-STSA and MMSE-LSA estimators is the a posteriori SNR estimation accuracy. 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 loss function. A recent trend has been to include the STOI measure in the loss function (Fu et al., 2018; Zhao et al., 2018). The speech enhancement performance of the proposed a priori
SNR estimators may be 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).

Declaration of Competing Interest
The authors declare that they have no conflict of interest.

References
Hu, G., 2004. 100 nonspeech environmental sounds. The Ohio State University, Department of Computer Science and Engineering.


