Speech Communication 111 (2019) 44–55

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

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

ARTICLE INFO

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

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 intelligible enhanced speech at a high quality. 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 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. MMSE approaches utilising the proposed *a priori* SNR estimator are 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.

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.

* Corresponding author.
E-mail addresses: aaron.nicolson@griffithuni.edu.au (A. Nicolson), k.paliwal@griffith.edu.au (K.K. Paliwal).

https://doi.org/10.1016/j.specom.2019.06.002
Received 14 December 2018; Received in revised form 18 May 2019; Accepted 6 June 2019
Available online 13 June 2019
0167-6393/

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 cepstro-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 short-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:

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

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 (Vary and Martin, 2006):

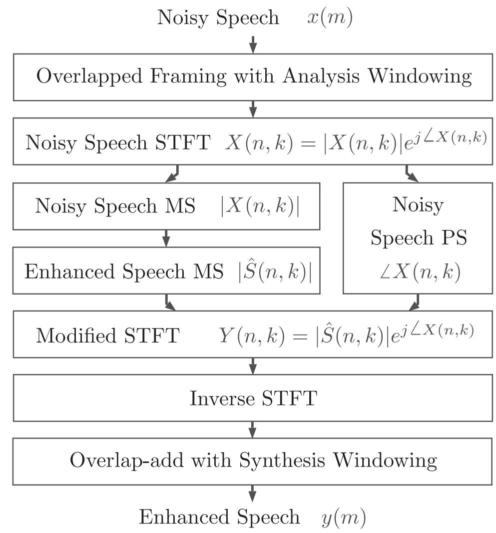

Fig. 1. Block diagram of the short-time Fourier AMS speech enhancement framework.

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

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 $|X(n,k)|$ 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 $|\hat{S}(n,k)|$. 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 (Griffin and Lim, 1984; Crochiere, 1980):

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

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*

An MMSE approach to speech enhancement utilises the *a priori* SNR to compute a gain function. The gain function is subsequantly 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

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

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 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-STSA) 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

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

The MMSE-STSA estimator gain function is given by

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

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 MSE between the clean and enhanced speech log-magnitude spectra (Ephraim and Malah, 1985). The MMSE-LSA gain function is given by

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

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

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

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:

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

3. Mapped *a priori* SNR training target

In preliminary experiments, it was found that mapping the oracle *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 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

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

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

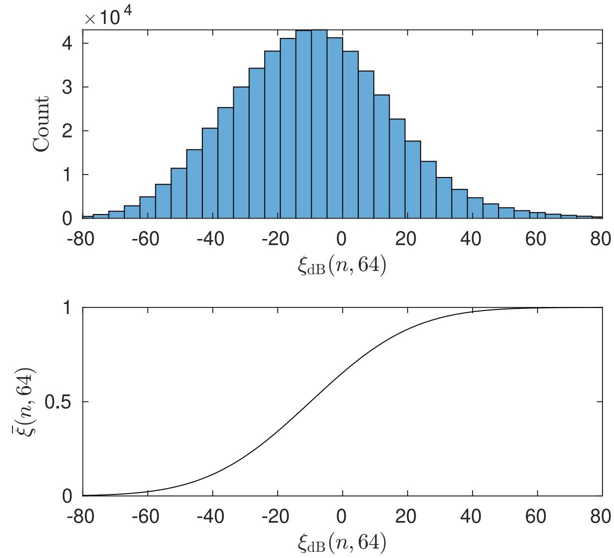

Fig. 2. (Top) The distribution of $\xi_{\text{dB}}(n, 64)$, over a sample of the training set. (Bottom) The CDF of $\xi_{\text{dB}}(n, 64)$, assuming that $\xi_{\text{dB}}(n, 64)$ is distributed normally (the sample mean and variance were found over the sample of the training set).

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.[1] As an example, the distribution of $\xi_{\text{dB}}(n, 64)$ 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 $\xi_{\text{dB}}(n, 64)$ over the sample is shown in Fig. 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 (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 n^{th} noisy speech frame, $|X(n,k)|$, for $k = 0, 1, \ldots, N_l/2$, where N_l is the frame length in discrete-time samples. The ResLSTM estimates the *a priori* SNR[2] 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

[1] 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 randomly selected (without replacement) 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.

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

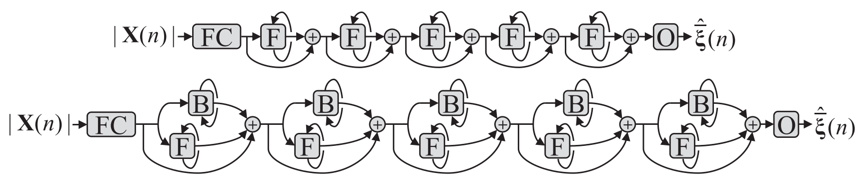

Fig. 3. ResLSTM (top) and ResBLSTM (bottom) *a priori* SNR estimators. **FC** is a fully-connected layer. The output layer, **O**, is a fully-connected layer with sigmoidal units. **F** and **B**, denote forward and backward LSTM cells, respectively.

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 *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.[3] 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 *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.[4]

5. 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 ($N_l = 512$) 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 Gerkmann and Hendriks (2012) was used by the DD, TSNR, HRNR, and SCTS *a priori* SNR estimation methods. The *a posteriori* SNR was estimated using both 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 (Panayotov et al., 2015) (28 539 utterances), the CSTR VCTK Corpus (Veaux et al., 2017) (42 015 utterances), and the *si** and *sx** 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,[5] 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 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*[6] 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[7] (Kabal, 2002) 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) (Paliwal and Atal, 1991) is defined as the root-mean-square difference between the *a priori* SNR

[3] 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.

[4] 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] 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.

[6] Street music recording number 26 270 was used from the Urban Sound dataset.

[7] Only adult speakers were included from the TSP speech corpus.

estimate in dB, $\hat{\xi}_{\rm dB}(n,k)$, and the oracle case in dB, $\xi_{\rm dB}(n,k)$, for the n^{th} frame[8]:

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

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) (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[9] 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.

[8] $\xi_{dB}(n,k)$ 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.

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

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

Noise	$\hat{\xi}(n,k)$	SNR level (dB)				
		−5	0	5	10	15
Voice babble	DD (Ephraim and Malah, 1984)	18.5	17.7	17.2	17.0	17.2
	TSNR (Plapous et al., 2004)	18.4	17.5	17.0	16.9	17.1
	HRNR (Plapous et al., 2005)	19.5	18.9	18.5	18.4	18.6
	SCTS (Breithaupt et al., 2008)	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 (Ephraim and Malah, 1984)	19.9	18.6	17.6	17.0	16.8
	TSNR (Plapous et al., 2004)	19.7	18.4	17.4	16.8	16.6
	HRNR (Plapous et al., 2005)	19.8	18.7	17.9	17.5	17.5
	SCTS (Breithaupt et al., 2008)	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 (Ephraim and Malah, 1984)	22.1	20.5	19.2	18.2	17.5
	TSNR (Plapous et al., 2004)	21.8	20.2	18.9	17.9	17.2
	HRNR (Plapous et al., 2005)	20.7	19.4	18.4	17.7	17.3
	SCTS (Breithaupt et al., 2008)	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 (Ephraim and Malah, 1984)	24.0	22.2	20.7	19.4	18.5
	TSNR (Plapous et al., 2004)	23.7	22.0	20.4	19.2	18.3
	HRNR (Plapous et al., 2005)	23.0	21.4	20.1	19.1	18.4
	SCTS (Breithaupt et al., 2008)	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**

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

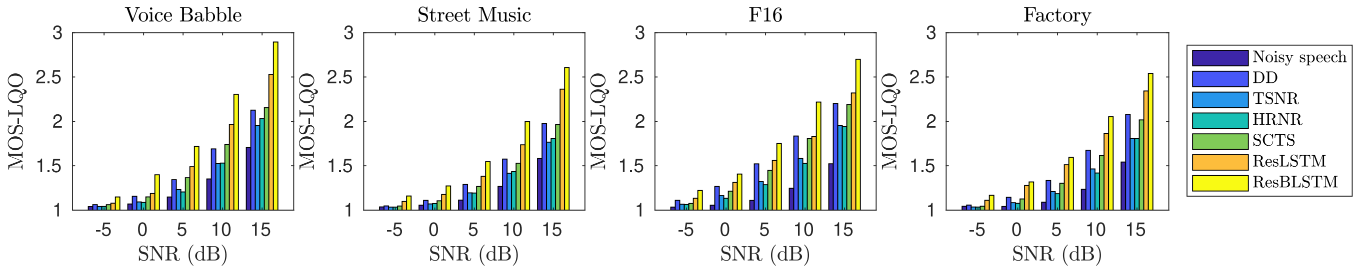

Fig. 4. MMSE-STSA estimator objective quality (MOS-LQO) scores for each *a priori* SNR estimator. The tested conditions include real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources at multiple SNR levels.

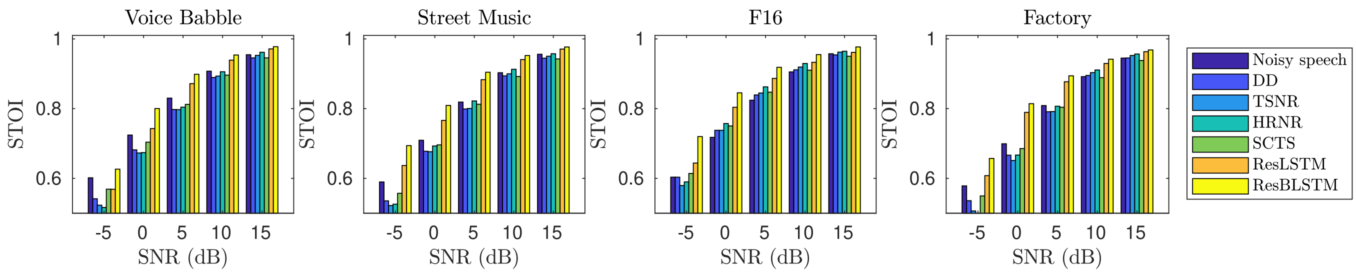

Fig. 5. MMSE-STSA estimator objective intelligibility (STOI) scores for each *a priori* SNR estimator. The tested conditions include real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources at multiple SNR levels.

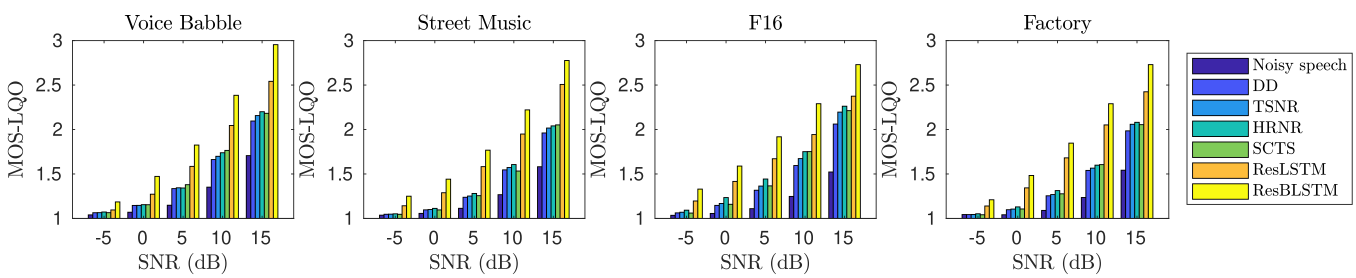

Fig. 6. MMSE-LSA estimator objective quality (MOS-LQO) scores for each *a priori* SNR estimator. The tested conditions include real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources at multiple SNR levels.

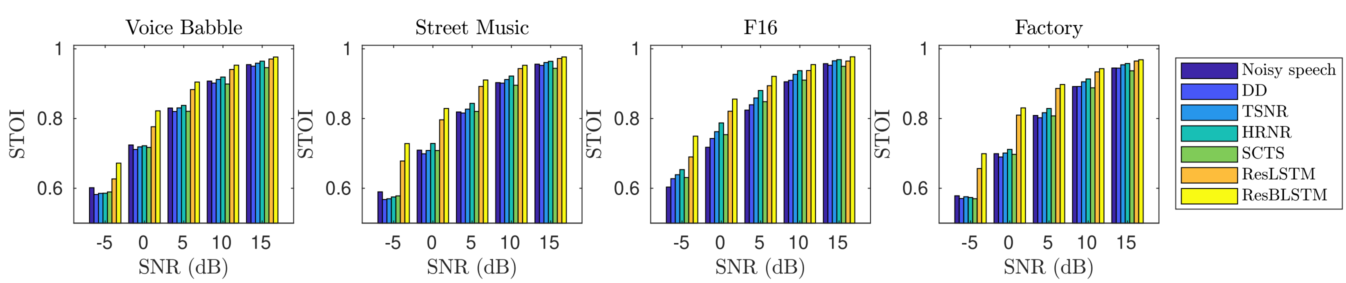

Fig. 7. MMSE-LSA estimator objective intelligibility (STOI) scores for each *a priori* SNR estimator. The tested conditions include real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources at multiple SNR levels.

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

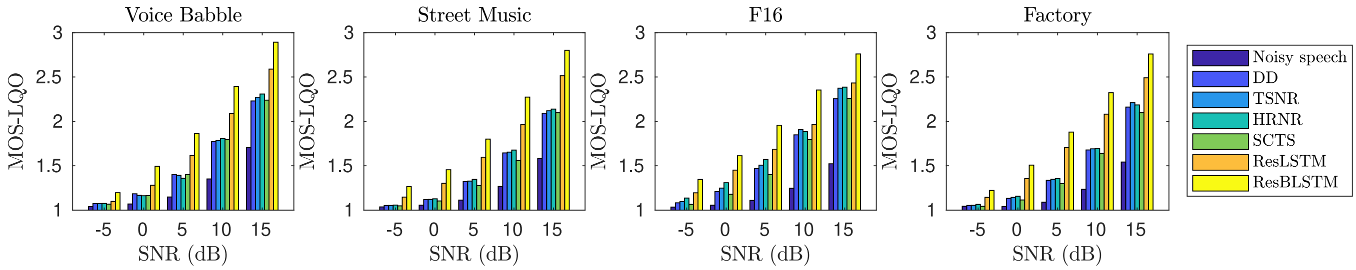

Fig. 8. WF approach objective quality (MOS-LQO) scores for each *a priori* SNR estimator. The tested conditions include real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources at multiple SNR levels.

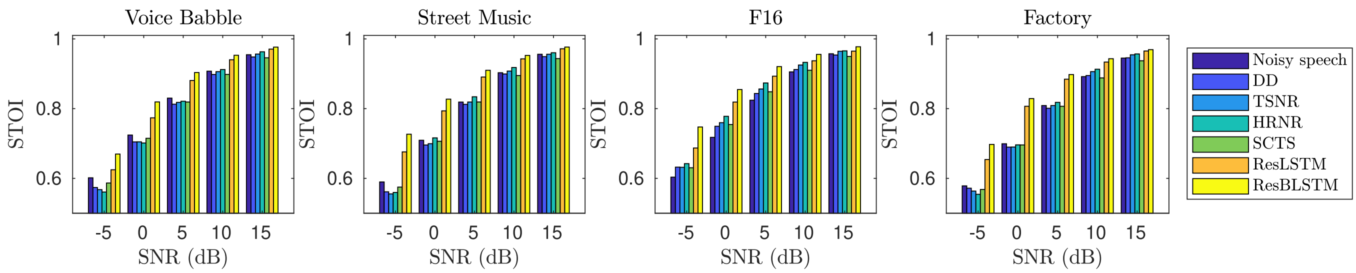

Fig. 9. WF approach objective intelligibility (STOI) scores for each *a priori* SNR estimator. The tested conditions include real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources at multiple SNR levels.

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 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 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 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 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-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,[10] 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 Section 6.4 as it achieved the highest average objective quality score, and the second highest average objective intelligibility score.

Table 2
The average improvement over the MMSE approach in the preceding row is shown for both objective quality (MOS-LQO) and intelligibility (STOI).

$\hat{\xi}$	Gain	MOS-LQO	STOI
ResLSTM	WF	–	–
	MMSE-STSA	+0.10	+1.76%
	MMSE-LSA	+0.02	−0.15%
ResBLSTM	WF	+0.07	+1.37%
	MMSE-STSA	+0.13	+1.19%
	MMSE-LSA	+0.02	−0.08%

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.[11] The SRWF approach, MMSE-STSA 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).

[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].

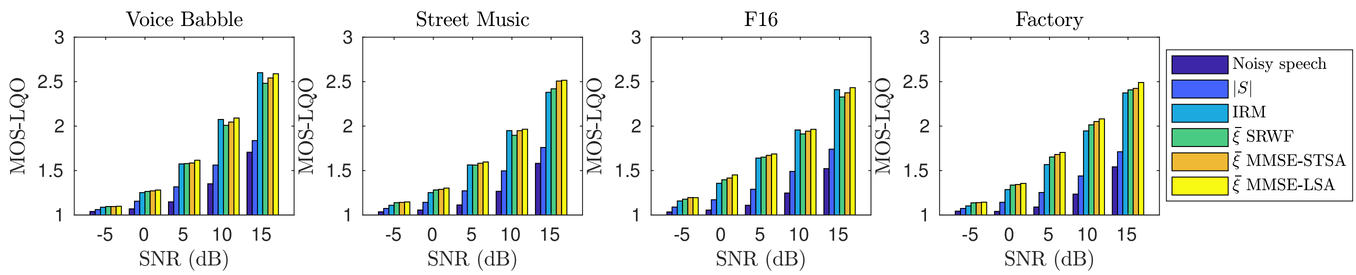

Fig. 10. Objective quality (MOS-LQO) scores for each training target. The tested conditions include real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources at multiple SNR levels.

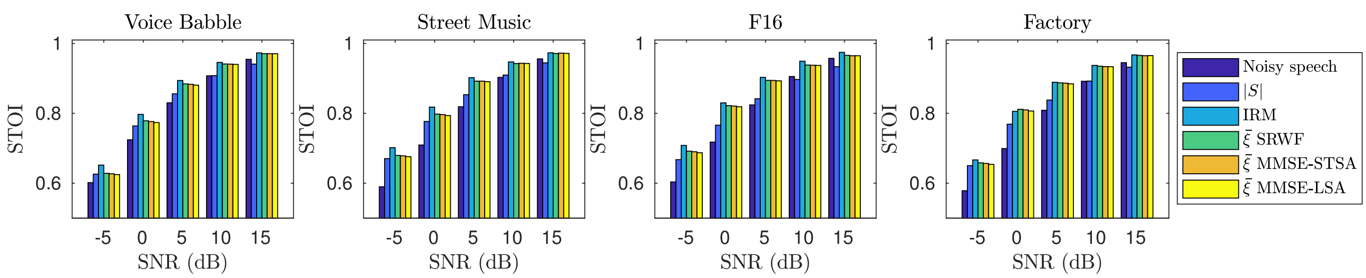

Fig. 11. Objective intelligibility (STOI) scores for each training target. The tested conditions include real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources at multiple SNR levels.

Table 3
The average improvement over the training target in the preceding row is shown for both objective quality (MOS-LQO) and intelligibility (STOI).

Target	MOS-LQO	STOI
$\|S\|$	–	–
IRM	+0.33	+3.49%
$\bar{\xi}$ SRWF	+0.01	−0.89%
$\bar{\xi}$ MMSE-STSA	+0.03	−0.09%
$\bar{\xi}$ MMSE-LSA	+0.02	−0.15%

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[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.

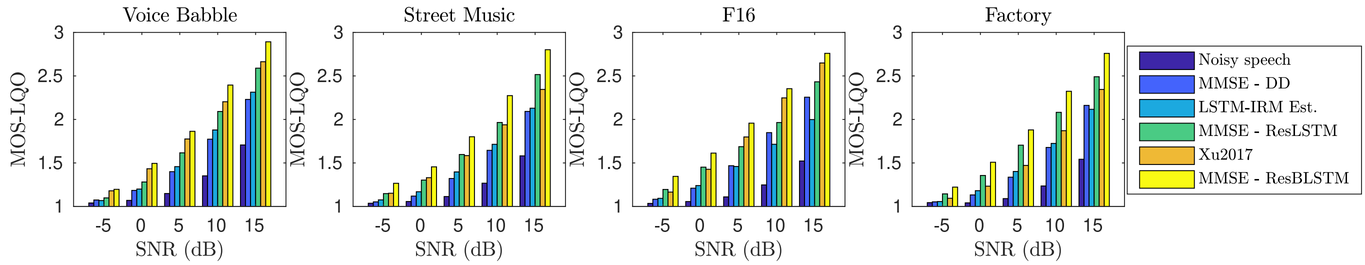

Fig. 12. Objective quality (MOS-LQO) scores for the MMSE-LSA estimator utilising the DD approach, the LSTM-IRM estimator, Xu2017, and the MMSE-LSA estimator utilising both the ResLSTM and ResBLSTM *a priori* SNR estimators. The tested conditions include real-world non-stationary (*voice babble* and *street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

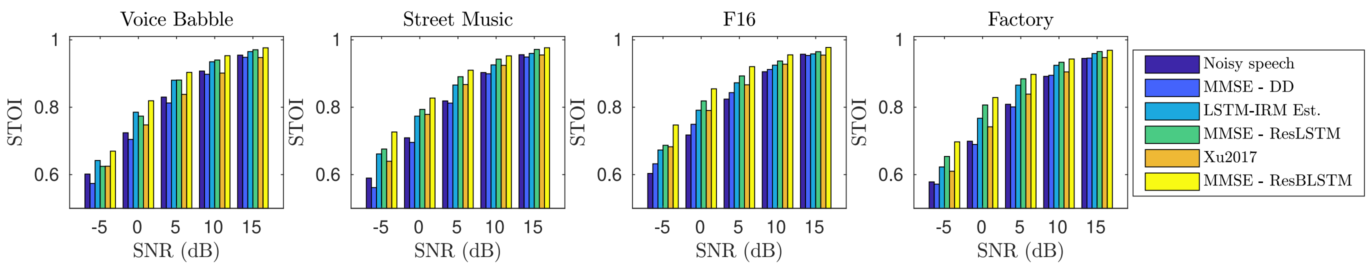

Fig. 13. Objective intelligibility (STOI) scores for the MMSE-LSA estimator utilising the DD approach, the LSTM-IRM estimator, Xu2017, and the MMSE-LSA estimator utilising both the ResLSTM and ResBLSTM *a priori* SNR estimators. The tested conditions include real-world non-stationary (*voice babble* and *street music*) and coloured (*F16* and *factory*) noise sources at multiple SNR levels.

Table 4
The average improvement over the speech enhancement method in the preceding row is shown for both objective quality (MOS-LQO) and intelligibility (STOI).

Method	Casual	MOS-LQO	STOI
MMSE-LSA; DD $\hat{\xi}$ (Ephraim and Malah, 1984)	Yes	–	–
LSTM-IRM est. (Chen and Wang, 2017)	Yes	+0.01	+4.52%
MMSE-LSA; ResLSTM $\hat{\xi}$	Yes	+0.22	+1.28%
Xu2017 (Xu et al., 2015; 2017)	No	+0.01	−2.59%
MMSE-LSA; ResBLSTM $\hat{\xi}$	No	+0.21	+5.08%

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,[13] 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.

[13] 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.

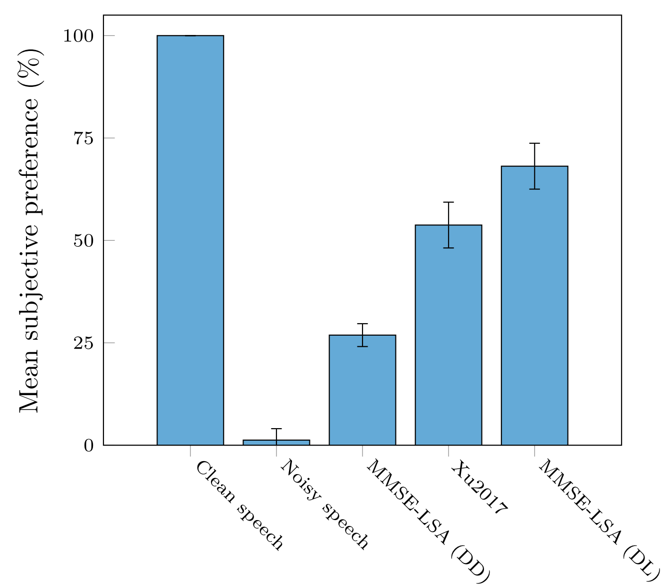

Fig. 14. Mean subjective preference (%) scores for the MMSE-LSA estimator utilising the DD approach (MMSE-LSA (DD)), Xu2017, and the MMSE-LSA estimator utilising the ResBLSTM *a priori* SNR estimator (MMSE-LSA (DL), where DL stands for deep learning). The subjective testing procedure is described in 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

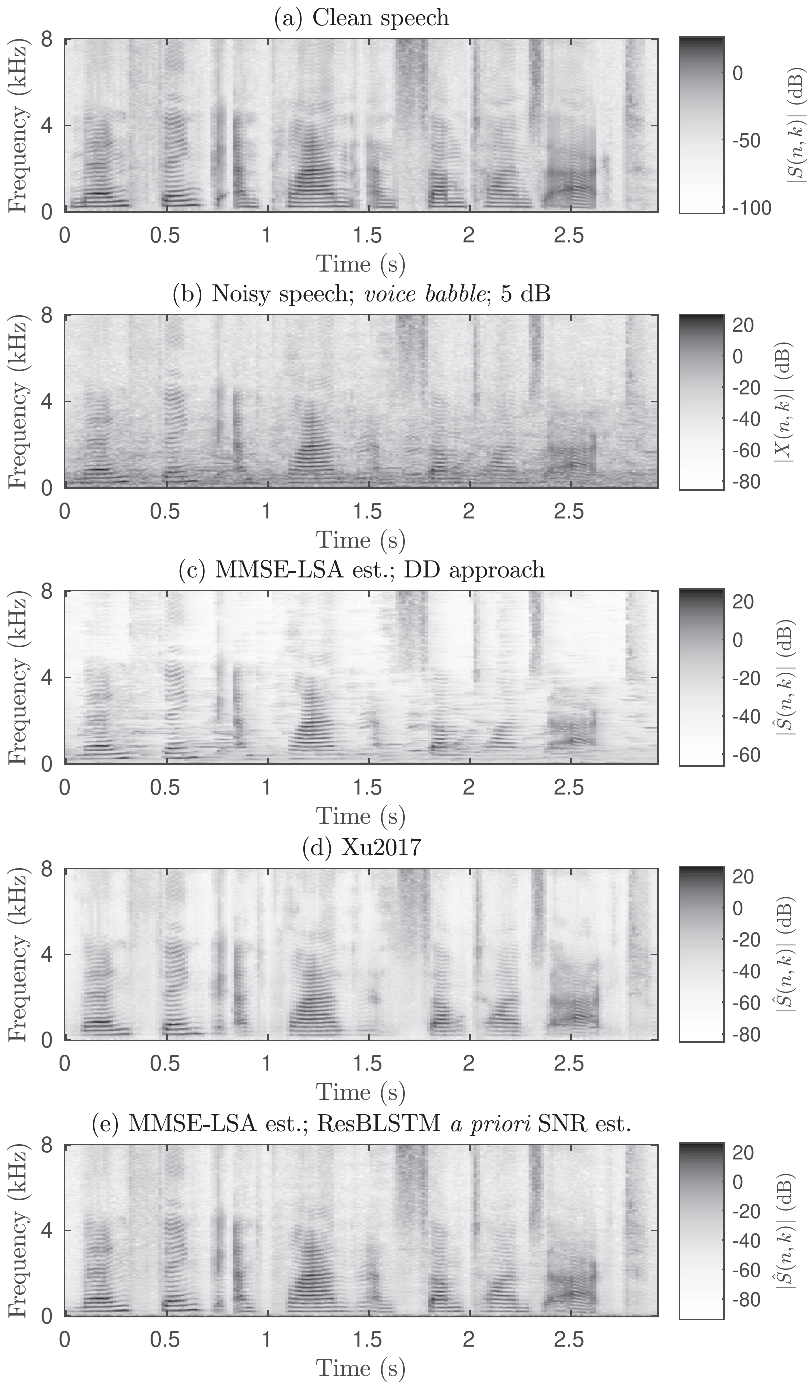

Fig. 15. (a) Clean speech magnitude spectrogram of female *FF* uttering sentence 32_10, "Men think and plan and sometimes act" from the test set. (b) A recording of *voice babble* mixed with (a) at an SNR level of 5 dB. (c) Enhanced speech magnitude spectrogram produced by the MMSE-LSA estimator utilising 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.

a priori SNR estimator (marked as MMSE-LSA (DL), where DL stands for deep learning) was preferred by listeners over Xu2017 enhanced speech.

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

Allen, J., 1977. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. IEEE Trans. Acoust. Speech Signal Proc. 25 (3), 235–238. doi:10.1109/TASSP.1977.1162950.

Allen, J.B., Rabiner, L.R., 1977. A unified approach to short-time Fourier analysis and synthesis. Proc. IEEE 65 (11), 1558–1564. doi:10.1109/PROC.1977.10770.

Ba, J. L., Kiros, J. R., Hinton, G. E., 2016. Layer normalization. 1607.06450.

Breithaupt, C., Gerkmann, T., Martin, R., 2008. A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing. In: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4897–4900. doi:10.1109/ICASSP.2008.4518755.

Cappe, O., 1994. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. 2 (2), 345–349. doi:10.1109/89.279283.

Chen, J., Wang, D., 2017. Long short-term memory for speaker generalization in supervised speech separation. J. Acoust. Soc. Am. 141 (6), 4705–4714. doi:10.1121/1.4986931.

Crochiere, R., 1980. A weighted overlap-add method of short-time Fourier analysis/synthesis. IEEE Trans. Acoust. Speech Signal Process. 28 (1), 99–102. doi:10.1109/TASSP.1980.1163353.

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

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32 (6), 1109–1121. doi:10.1109/TASSP.1984.1164453.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33 (2), 443–445. doi:10.1109/TASSP.1985.1164550.

Fu, S., Wang, T., Tsao, Y., Lu, X., Kawai, H., 2018. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. IEEE/ACM Tran. Audio SpeechLang. Process. 26 (9), 1570–1584. doi:10.1109/TASLP.2018.2821903.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., 1993. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1. NASA STI/Recon Technical Report N. 93.

Gerkmann, T., Hendriks, R.C., 2012. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Lang.Process. 20 (4), 1383–1393. doi:10.1109/TASL.2011.2180896.

Gers, F.A., Schmidhuber, J., Cummins, F., 1999. Learning to forget: continual prediction with LSTM. In: Ninth International Conference on Artificial Neural Networks (ICANN), 2, pp. 850–855 vol.2. doi:10.1049/cp:19991218.

Griffin, D., Lim, J., 1984. Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust. Speech Signal Process. 32 (2), 236–243. doi:10.1109/TASSP.1984.1164317.

Hanson, J., Paliwal, K., Litfin, T., Yang, Y., Zhou, Y., 2018. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics 34 (23), 4039–4045. doi:10.1093/bioinformatics/bty481.

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

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

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735.

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

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

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

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

Kabal, P., 2002. TSP Speech Database. McGill University, Database Version.

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

Kingma, D.P., Ba, J., 2014. Adam: a Method for Stochastic Optimization. CoRR abs/1412.6980.

Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604. doi:10.1109/PROC.1979.11540.

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

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

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. Omnipress, USA, pp. 807–814.

Paliwal, K., Wojcicki, K., 2008. Effect of analysis window duration on speech intelligibility. IEEE Signal Process. Lett. 15, 785–788. doi:10.1109/LSP.2008.2005755.

Paliwal, K.K., Atal, B.S., 1991. Efficient vector quantization of LPC parameters at 24 bits/frame. In: Proceedings ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing, pp. 661–664 vol. 1. doi:10.1109/ICASSP.1991.150426.

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

Picone, J.W., 1993. Signal modeling techniques in speech recognition. Proc. IEEE 81 (9), 1215–1247. doi:10.1109/5.237532.

Plapous, C., Marro, C., Mauuary, L., Scalart, P., 2004. A two-step noise reduction technique. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1, pp. 289–292. doi:10.1109/ICASSP.2004.1325979.

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

Plapous, C., Marro, C., Scalart, P., 2006. Improved signal-to-noise ratio estimation for speech enhancement. IEEE Trans. Audio Speech Lang.Process. 14 (6), 2098–2108. doi:10.1109/TASL.2006.872621.

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

Saki, F., Kehtarnavaz, N., 2016. 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), pp. 736–739. doi:10.1109/EMBC.2016.7590807.

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

Salamon, J., Jacoby, C., Bello, J.P., 2014. A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM International Conference on Multimedia. ACM, New York, NY, USA, pp. 1041–1044. doi:10.1145/2647868.2655045.

Schuster, M., Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 45 (11), 2673–2681. doi:10.1109/78.650093.

Snyder, D., Chen, G., Povey, D., 2015. MUSAN: a music, speech, and noise corpus. CoRR abs/1510.08484.

So, S., Paliwal, K.K., 2011. Modulation-domain Kalman filtering for single-channel speech enhancement. Speech Commun. 53 (6), 818–829. doi:10.1016/j.specom.2011.02.001.

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

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4214–4217. doi:10.1109/ICASSP.2010.5495701.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang.Process. 19 (7), 2125–2136. doi:10.1109/TASL.2011.2114881.

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

Veaux, C., Yamagishi, J., MacDonald, K., et al., 2017. CSTR VCTK Corpus: English Multi--Speaker Corpus for CSTR Voice Cloning Toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR).

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

Wang, Y., Narayanan, A., Wang, D., 2014. On training targets for supervised speech separation. IEEE/ACM Trans. Audio SpeechLang. Process. 22 (12), 1849–1858. doi:10.1109/TASLP.2014.2352935.

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

Xia, Y., Stern, R., 2018. A priori SNR estimation based on a recurrent neural network for robust speech enhancement. In: Proc. Interspeech 2018, pp. 3274–3278. doi:10.21437/Interspeech.2018-2423.

Xu, Y., Du, J., Dai, L., Lee, C., 2015. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio SpeechLang. Process. 23 (1), 7–19. doi:10.1109/TASLP.2014.2364452.

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

Zhang, Z., Geiger, J., Pohjalainen, J., Mousa, A.E.-D., Jin, W., Schuller, B., 2018. Deep learning for environmentally robust speech recognition: an overview of recent developments. ACM Trans. Intell. Syst. Technol. 9 (5), 1–28. doi:10.1145/3178115.

Zhao, Y., Wang, D., Merks, I., Zhang, T., 2016. DNN-based enhancement of noisy and reverberant speech. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6525–6529. doi:10.1109/ICASSP.2016.7472934.

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