On training targets for deep learning approaches to clean speech magnitude spectrum estimation

Aaron Nicolson1,a) and Kuldip K. Paliwal2,b)
1Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston, Queensland, 4006, Australia
2Signal Processing Laboratory, Griffith University, Brisbane, Queensland 4111, Australia

ABSTRACT:
Estimation of the clean speech short-time magnitude spectrum (MS) is key for speech enhancement and separation. Moreover, an automatic speech recognition (ASR) system that employs a front-end relies on clean speech MS estimation to remain robust. Training targets for deep learning approaches to clean speech MS estimation fall into three categories: computational auditory scene analysis (CASA), MS, and minimum mean square error (MMSE) estimator training targets. The choice of the training target can have a significant impact on speech enhancement/separation and robust ASR performance. Motivated by this, the training target that produces enhanced/separated speech at the highest quality and intelligibility and that which is best for an ASR front-end is found. Three different deep neural network (DNN) types and two datasets, which include real-world nonstationary and coloured noise sources at multiple signal-to-noise ratio (SNR) levels, were used for evaluation. Ten objective measures were employed, including the word error rate of the Deep Speech ASR system. It is found that training targets that estimate the a priori SNR for MMSE estimators produce the highest objective quality scores. Moreover, it is established that the gain of MMSE estimators and the ideal amplitude mask produce the highest objective intelligibility scores and are most suitable for an ASR front-end. © 2021 Acoustical Society of America. https://doi.org/10.1121/10.0004823
(Received 12 September 2020; revised 8 April 2021; accepted 10 April 2021; published online 18 May 2021)
[Editor: B. Yegnanarayana]

I. INTRODUCTION
Speech corrupted by background noise (or noisy speech) can reduce the efficiency of communication between a speaker and listener. The linguistic content of noisy speech can be misinterpreted by the listener or background noise can cause fatigue. Factory welding, music, and nontarget speakers are all examples of background noise sources (Loizou, 2013). The performance of speech processing systems, such as automatic speech recognition (ASR), automatic speaker verification (ASV), and automatic speaker identification (ASI) systems, can also be impacted by noisy speech (Le Prell and Clavier, 2017; Nicolson and Paliwal, 2020b). A system that can remove background noise or segregate the target speech (or clean speech) is, thus, indispensable for a speech processing system. Speech corrupted by background noise and reverberation from surface reflections (or noisy-reverberant speech) and systems with multiple microphones pose a more difficult task for such systems (Tawara et al., 2019; Zhao et al., 2017). For simplicity, this study focuses on non-reverberant noisy speech recorded using a single-microphone (or single-channel noisy speech).

The objective of speech enhancement is to improve the perceived quality and intelligibility of noisy speech. Speech enhancement algorithms accomplish this task by suppressing background noise without distorting the speech (Loizou, 2013). They are used to suppress background noise during phone calls, conference calls, and in hearing aid devices. A popular approach is to use an estimator of the clean speech that is derived from statistical models and an optimisation criteria. The Wiener filter (WF) optimally estimates, in the mean squared error (MSE) sense, the discrete Fourier transform (DFT) coefficients of the clean speech—when the clean speech and noise DFT coefficients are assumed to be independent Gaussian random variables (Lim and Oppenheim, 1979). The minimum mean square error short-time spectral amplitude (MMSE-STSA) and minimum mean square error log-spectral amplitude (MMSE-LSA) estimators differ from the WF by optimally estimating the clean speech magnitude and log-magnitude spectra, respectively (Ephraim and Malah, 1984, 1985). Whereas the WF is a linear estimator that depends solely on the a priori signal-to-noise ratio (SNR), the MMSE-STSA and MMSE-LSA estimators are nonlinear estimators that depend on both the a priori and a posteriori SNRs. Other prominent estimators assume super-Gaussian clean speech and noise priors (Martin, 2005) or employ a perceptually motivated optimisation criteria (Loizou, 2005). We refer to estimators that use the MSE as the optimisation criteria collectively as minimum mean square error (MMSE) estimators henceforth.

An alternative to speech enhancement is speech separation (or segregation)—a special case of source separation in which the target speech is segregated from the noisy speech.

---
a)Electronic mail: aaron.nicolson@griffithuni.edu.au, ORCID: 0000-0002-7163-1809.
b)ORCID: 0000-0002-3553-3662.
Source separation is performed using computational auditory scene analysis (CASA)—the computational task of segregating mixtures of sound sources (Wang and Brown, 2006). One approach to speech separation is to classify each time-frequency (TF) component of noisy speech as either speech or noise dominant (Wang, 2005). This is realised by using the ideal binary mask (IBM) as the objective of CASA, which reduces speech separation to a binary classification problem. The IBM can be used to attenuate the noise-dominant TF components. In practice, however, a noisy speech TF component can contain a ratio of speech and noise. Instead of a hard label, a soft label can be used to segregate the target speech in a noisy speech TF component (Srinivasan et al., 2006). This is realised by using the ideal ratio mask (IRM) as the objective of CASA, which has the same form as the gain of the square-root Wiener filter (SRWF; Lim and Oppenheim, 1979).

Although subjective listening tests conducted under stringent conditions are the gold standard for the evaluation of enhanced/separated speech quality and intelligibility, they are costly and time consuming (Hu and Loizou, 2008). Thus, objective measures of quality and intelligibility are often used as an alternative. For objective quality, perceptual evaluation of speech quality (PESQ) is commonly used in the literature (Rix et al., 2001). However, PESQ was developed for the purpose of evaluating the distortions introduced by speech codecs and communication channels—not for the distortions introduced by speech enhancement/ separation. Perceptual objective listening quality analysis (POLQA) is a later generation of PESQ that considers enhanced/separated speech and wideband audio (Beerends et al., 2013). However, POLQA is not open source, causing its use to be less frequent than other objective quality measures. In Hu and Loizou (2008), several composite objective quality measures were developed specifically for the evaluation of the signal distortion (CSIG), background noise intrusiveness (CBAK), and overall signal quality (COVL) of enhanced/separated speech (where the “signal” is the clean speech). Their correlation with subjective scores was higher than that of previous objective quality measures, including PESQ. The signal-to-distortion ratio (SDR; Vincent et al., 2006) and the scale-invariant signal-to-distortion ratio (SI-SDR; Roux et al., 2019) are two commonly used objective quality measures that indicate the amount of distortion between the enhanced/separated and clean speech. For objective intelligibility, short-time objective intelligibility (STOI) is commonly used and demonstrates a high correlation with subjective scores (Taal et al., 2011). Extended short-time objective intelligibility (ESTOI) builds upon STOI by not assuming mutual independence between frequency bands and incorporating the spectral correlation between 400ms length spectrograms of the enhanced/separated and clean speech (Jensen and Taal, 2016). The word error rate (WER) of an ASR system can also be used to objectively evaluate the intelligibility of enhanced/separated speech (Thomas-Stonell et al., 1998).

Currently, deep learning approaches are at the forefront of speech enhancement and separation. Deep neural networks (DNNs) provide a nonlinear map from a given noisy speech representation to a target representation. DNNs were first employed for IBM estimation, where a feedforward neural network (FNN) provided learned features to linear support vector machines (SVMs; Wang and Wang, 2013). It was found that the use of a DNN enabled the system to generalise to unobserved speakers and noise sources. FNNs were later employed for IRM estimation, where it was found that FNN-IRM estimator is able to outperform a FNN-IBM estimator when used as a front-end for ASR (Narayanan and Wang, 2013). An FNN was later used to estimate the clean speech magnitude spectrum (MS), which was found to outperform the combination of the MMSE-LSA estimator and the improved minima controlled recursive averaging (IMCRA) noise estimation approach (Cohen, 2003; Xu et al., 2015). In a study by Wang et al. (2014), the IBM, IRM, MS, and ideal amplitude mask (IAM) were compared as training targets, where the IAM is the ratio of the clean speech MS to the noisy speech MS. It was found that the IRM and IAM as the training targets produced higher objective quality and intelligibility scores than the IBM and MS. However, the study was limited by its use of only PESQ and STOI as objective measures. In Nicolson and Paliwal (2019a), a deep learning approach to a priori SNR estimation improved the performance of MMSE estimators. It was found that the a priori SNR training target produces higher objective quality scores than the IRM, with the IRM producing higher objective intelligibility scores. However, this study was also limited as objective scores were computed using only a wideband extension of the PESQ (Morioka et al., 2005) and STOI.

The aforementioned training targets use only the magnitude of each TF component. However, there are a set of training targets that incorporate the phase of each TF component. The phase sensitive mask (PSM) is an extension of the IAM that includes the phase difference between each clean and noisy speech TF component (Erdogan et al., 2015). Results in Williamson et al. (2016) indicate that the PSM is able to outperform the IRM. The complex ideal ratio mask (cIRM) is a complex TF mask that uses both the real and imaginary components of the target speech and noise DFT coefficients (Williamson et al., 2016). Results indicate that the cIRM is able to outperform the PSM and IRM in terms of objective quality and intelligibility. In Pascual et al. (2017), clean speech time-domain samples were used as the training target. However, results in Williamson et al. (2016) indicate that the PSM, cIRM, and IRM are able to attain higher objective quality and intelligibility scores than clean speech time-domain samples.

Current methods used to increase the robustness of an ASR system include (1) using a deep learning approach to speech enhancement/separation as a front-end for noisy speech preprocessing, and (2) multi-condition training (Zhang et al., 2018). It has been found that using multi-condition training or a deep learning-based speech
enhancement/separation front-end significantly improves the robustness of an ASR system with the combination of both methods providing the best performance (Narayanan and Wang, 2013). In this study, we concentrate solely on deep learning-based speech enhancement/separation front-ends. Features used as input to current ASR acoustic models are derived from the MS (Wang et al., 2020; Kriman et al., 2020; Moritz et al., 2020). These include mel-scale filterbank, gammatone filterbank, and cepstral-domain features (Schluter et al., 2007). When using a deep learning-based speech enhancement/separation front-end, such features are computed from the enhanced/separated speech MS (Nicolson and Paliwal, 2019b). This means that estimating the short-time phase spectrum—to keep this study concise. We also propose to jointly estimate the a priori and a posteriori SNRs—to increase the performance of the MMSE-STSA and MMSE-LSA estimators. Additionally, we propose to use the gain of a MMSE estimator as a training target as motivated by the use of the IRM as the training target (which has the same form as the gain of the SRWF). We, therefore, investigate three classes of training targets for clean speech MS estimation: (1) CASA training targets, (2) MS training targets, and (3) MMSE estimator training targets (including a priori SNR, joint a priori and a posteriori SNR, and gain training targets). The DEMAND Voice Bank and Deep Xi datasets are included in the experiment setup, which consists of real-world nonstationary and coloured noise sources at multiple SNR levels. We also investigate which function is best for compressing the dynamic range of the training target values. We also assess multiple loss functions for the training targets. A temporal convolutional network (TCN), a recurrent neural network (RNN), and a multi-head attention network were used to evaluate each training target on different DNN architectures. Multiple objective quality and intelligibility measures are included in the experiment setup, including CSIG, CBAK, COVL, PESQ, segmental signal-to-noise ratio (SegSNR; Mermelstein, 1979), STOI, ESTOI, SDR, SI-SDR, and WER.

In this paper, we first describe the analysis, modification, and synthesis (AMS) framework (Sec. II). MMSE estimators are then described in Sec. III. In Sec. IV, the training targets are described. The experiment setup is described in Sec. V, and the results are discussed in Sec. VI. Conclusions are drawn in Sec. VII.

II. ANALYSIS, MODIFICATION, AND SYNTHESIS FRAMEWORK

The short-time Fourier AMS framework is used for speech enhancement/separation (Allen, 1977; Allen and Rabiner, 1977). The AMS framework consists of three

\[
X[l, k] = \sum_{n=0}^{N_s-1} x[n + lN_s]w[n]e^{-j2\pi nk/N_s},
\]

where \( l \) denotes the time-frame index, \( k \) denotes the discrete-frequency bin, \( N_s \) denotes the time-frame duration in discrete-time samples, \( N_s \) denotes the time-frame shift in discrete-time samples, and \( w[n] \) is a window function. As we aim to modify the noisy speech MS, the polar form of the noisy speech spectrum is used:

\[
X[l, k] = |X[l, k]|e^{j\angle X[l, k]},
\]

where \( |X[l, k]| \) and \( \angle X[l, k] \) denote the noisy speech magnitude and phase spectra, respectively. Similarly, the clean speech magnitude and phase spectra are denoted as \( |S[l, k]| \) and \( \angle S[l, k] \), respectively, and the noise magnitude and phase spectra are denoted as \( |D[l, k]| \) and \( \angle D[l, k] \), respectively. The modified spectrum is constructed by combining an estimate of the clean speech MS, \( |S[l, k]| \), with the noisy speech phase spectrum,

\[
Y[l, k] = |S[l, k]|e^{j\angle X[l, k]}.
\]

The synthesis stage involves applying the ISTFT to the modified spectrum. First, the inverse DFT is applied to the modified spectrum,

\[
y_f[l, n] = \frac{1}{N_d} \sum_{k=0}^{N_s-1} Y[l, k]e^{j2\pi nk/N_s},
\]

where \( y_f[l, n] \) is the framed enhanced/separated speech. The overlap-add method is subsequently applied to produce the final enhanced/separated speech (Crochiere, 1980),

\[
y[n] = \frac{\sum_{l=-\infty}^{l_n} y_f[l, n - lN_s]}{\sum_{l=-\infty}^{l_n} w[n - lN_s]}.
\]

In this work, the Hamming window function with a time-frame duration of 32 ms (\( N_d = 512 \)) and a time-frame shift of 16 ms (\( N_s = 256 \)) is used.

https://doi.org/10.1121/10.0004823
III. MMSE ESTIMATORS

The gain of a MMSE estimator is applied to the noisy speech MS to obtain a clean speech MS estimate for Eq. (4),
\[ \hat{S}[l, k] = G[l, k] \cdot |X[l, k]|. \tag{7} \]

The gain for the WF, SRWF, and constrained Wiener filter (CWF; Loizou, 2013), as well as for the MMSE-STSA and MMSE-LSA estimators are defined in Table I, where \( \nu[l, k] = \frac{\nu[l, k]}{(\xi[l, k] + 1)} \gamma[l, k] \), \( \xi[l, k] \) is the a priori SNR, and \( \gamma[l, k] \) is the a posteriori SNR. The a priori SNR is defined as
\[ \zeta[l, k] = \frac{\lambda_d[l, k]}{\hat{z}_d[l, k]}, \tag{8} \]
where \( \hat{z}_d[l, k] = \text{E}[|S[l, k]|^2] \) is the variance of the clean speech spectral component, and \( \lambda_d[l, k] = \text{E}[|D[l, k]|^2] \) is the variance of the noise spectral component. The a posteriori SNR is defined as
\[ \gamma[l, k] = \frac{|X[l, k]|^2}{\hat{z}_d[l, k]}, \tag{9} \]

IV. TRAINING TARGETS

For speech enhancement/separation, a DNN learns to map the noisy speech MS to a training target, which can be used to estimate the clean speech MS as defined in Eq. (4). Training targets with magnitude or power values tend to be difficult for a DNN to learn as noted in Wang et al. (2014). Hence, a function must be used to compress their dynamic range— which we refer to as a compression function henceforth (referred to as a mapping function in Nicolson and Paliwal, 2019a). Here, the training target is denoted as \( t[l, k] \). The training target can be computed from the clean speech and noise of the noisy speech in Eq. (1) as they are observed during training. The training targets described in this section belong to one of three categories: CASA, MS, and MMSE estimator training targets. Some of the compression functions described in this section require statistics of the distribution of their input (e.g., mean and standard deviation). These statistics are estimated from a sample of the training set as described in Sec. V D.

A. CASA training targets

CASA training targets, including the IBM, IRM, and IAM, are detailed here. The IBM is computed by applying a threshold to the instantaneous a priori SNR,
\[ t[l, k] = \text{IBM}[l, k] = \begin{cases} 1 & \text{if} \frac{|S[l, k]|^2}{|D[l, k]|^2} > 1, \\ 0 & \text{otherwise}, \end{cases} \tag{10} \]
where the threshold used here is equivalent to 0 dB (Wang, 2005). The IRM is computed as in Wang et al. (2014),
\[ t[l, k] = \text{IRM}[l, k] = \sqrt{\frac{|S[l, k]|^2}{|S[l, k]|^2 + |D[l, k]|^2}}, \tag{11} \]
The IAM is computed as in Erdogan et al. (2015),
\[ t[l, k] = \text{IAM}[l, k] = \frac{|S[l, k]|}{|X[l, k]|}, \tag{12} \]
where the values for the IAM[\( l, k \)] are clipped between zero and one. The masked magnitude spectrum approximation (mMSA) loss function can be used in combination with the IAM (Lu J. et al., 2017; Weninger et al., 2014)
\[ L_{\text{mMSA}} = \frac{1}{LK} \sum_{l=1}^{L} \sum_{k=0}^{K-1} \left| X[l, k] \right|^2 (\text{IAM}[l, k] - \text{IAM}[l, k])^2, \tag{13} \]
which is equivalent to the masked signal approximation (mSA) loss function from Zheng and Zhang [2019, Eq. (8)] without the phase terms.

The IBM, IRM, and IAM are applied elementwise to each TF component of the noisy speech MS, for example,
\[ |\hat{S}[l, k]| = |X[l, k]| \cdot \text{IAM}[l, k]. \tag{14} \]

B. MS training targets

Here, training targets derived from the clean speech MS are described. The magnitude or power of the clean speech spectrum has been found difficult for a DNN to learn (Xu et al., 2015; Wang et al., 2014). We confirmed this in preliminary testing, where the loss during training would not converge with the magnitude or power values of the clean speech spectrum as the training target. This is likely due to the large dynamic range of the magnitude and power values.

| Table I. Gain for each MMSE estimator. |
|----------------------------------------|
| MMSE estimator | Gain function, \( G[l, k] \) |
|----------------|--------------------------------|
| WF             | \( \frac{\nu[l, k]}{|\nu[l, k]| + 1} \) |
| SRWF           | \( \frac{\nu[l, k]}{|\nu[l, k]| + 1} \) |
| CWF            | \( \frac{\sqrt{\nu(n, k)} \cdot \text{exp} \left( \frac{-\nu(n, k)}{2} \right)}{|\nu(n, k)|} \cdot \text{exp} \left( \frac{-\nu(n, k)}{2} \right) \) |
| MMSE-STSA      | \( \frac{\sqrt{\nu(n, k)} \cdot \text{exp} \left( \frac{-\nu(n, k)}{2} \right)}{|\nu(n, k)|} \cdot \text{exp} \left( \frac{-\nu(n, k)}{2} \right) \) |
| MMSE-LSA       | \( \frac{\sqrt{\nu(n, k)} \cdot \text{exp} \left( \frac{-\nu(n, k)}{2} \right)}{|\nu(n, k)|} \cdot \text{exp} \left( \frac{-\nu(n, k)}{2} \right) \) |

3276 J. Acoust. Soc. Am. 149 (5), May 2021  
Aaron Nicolson and Kuldeep K. Paliwal
Hence, the clean speech MS must be compressed, which leads to the decibel value of the clean speech MS as a training target,
\[ r[l, k] = |S_{\text{dB}}[l, k]| = 20 \log_{10}(|S[l, k]|). \tag{15} \]

Standardising $|S_{\text{dB}}[l, k]|$ has also been found beneficial (Xu et al., 2015):
\[ r[l, k] = z(|S_{\text{dB}}[l, k]|) = \frac{|S_{\text{dB}}[l, k]| - \mu_k}{\sigma_k}, \tag{16} \]
where $\mu_k$ and $\sigma_k^2$ are the mean and variance of the $k$th discrete-frequency bin, respectively. Min-max normalisation has also been found to facilitate training (Wang et al., 2014),
\[ r[l, k] = |S[l, k]|' = \frac{|S[l, k]| - \min(|S[l, k]|)}{\max(|S[l, k]|) - \min(|S[l, k]|)}, \tag{17} \]
where $\min(\cdot)$ and $\max(\cdot)$ find the minimum and maximum values over all time-frames for the $k$th discrete-frequency bin, respectively. Applying min-max normalisation to the decibel values also facilitates training (Wang et al., 2014), where $|S_{\text{dB}}[l, k]|'$ is formed by replacing $|S[l, k]|$ in Eq. (17) with $|S_{\text{dB}}[l, k]|$. Power compression can also be applied to the clean speech MS as in Ephrat et al. (2018),
\[ r[l, k] = |S[l, k]|^\alpha, \tag{18} \]
where $\alpha = 0.3$ from Ephrat et al. (2018) is used in this work.

We also investigate the cumulative distribution function (CDF) of $|S_{\text{dB}}[l, k]|$ as a compression function as motivated by Nicolson and Paliwal (2019a). The distribution of $|S_{\text{dB}}[l, 64]|$ is shown in Fig. 1(a), which indicates that $|S_{\text{dB}}[l, k]|$ follows a normal distribution. Hence, we assume that $|S_{\text{dB}}[l, k]|$ follows a normal distribution, $|S_{\text{dB}}[l, k]| \sim \mathcal{N}(\mu_k, \sigma_k^2)$, where $\mu_k$ and $\sigma_k^2$ are the mean and variance of the $k$th discrete-frequency bin, respectively. The normal CDF of $|S_{\text{dB}}[l, k]|$ is given by
\[ r[l, k] = F(|S_{\text{dB}}[l, k]|) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{|S_{\text{dB}}[l, k]| - \mu_k}{\sigma_k \sqrt{2}} \right) \right], \tag{19} \]
where $\text{erf}(\cdot)$ is the error function. We also investigated using the CDF of $|S[l, k]|$ as a compression function, where $|S[l, k]|$ is distributed exponentially. However, in preliminary testing, we found that the loss during training would not converge.

C. MMSE estimator training targets

As in Nicolson and Paliwal (2019a), the instantaneous a priori SNR can be used as a training target. The instantaneous a priori SNR is computed by using the instantaneous values $|S[l, k]|^2$ and $|D[l, k]|^2$ in place of $\lambda_s[l, k]$ and $\lambda_d[l, k]$, respectively, in Eq. (8). However, the power values of the instantaneous a priori SNR are difficult to train (Nicolson and Paliwal, 2019a). This was confirmed in preliminary testing, where the loss during training would not converge with the power values of the instantaneous a priori SNR as the training target. This is likely due to the large dynamic range of the power values. Thus, a compression function must be used to facilitate training. As with the clean speech MS, the decibel value of the instantaneous a priori SNR can be used as a training target,
\[ r[l, k] = \xi_{\text{a priori}}[l, k] = 10 \log_{10}(\xi[l, k]). \tag{20} \]

An additional training target is formed by standardising $\xi_{\text{a priori}}[l, k]$, where $z(\xi_{\text{a priori}}[l, k])$ is formed by replacing $|S_{\text{dB}}[l, k]|$ in Eq. (16) with $\xi_{\text{a priori}}[l, k]$. We also apply min-max normalisation to $\xi[l, k]$ and $\xi_{\text{a priori}}[l, k]$, where $\xi[l, k]'$ and $\xi_{\text{a priori}}[l, k]'$ are formed by replacing $|S[l, k]|$ in Eq. (17) with $\xi[l, k]$ and $\xi_{\text{a priori}}[l, k]$, respectively.

In Nicolson and Paliwal (2019a), the CDF of $\xi_{\text{a priori}}[l, k]$ was used as the compression function as part of the Deep Xi framework. It was assumed that $\xi_{\text{a priori}}[l, k]$ is distributed normally: $\xi_{\text{a priori}}[l, k] \sim \mathcal{N}(\mu_k, \sigma_k^2)$, where $\mu_k$ and $\sigma_k^2$ are the mean and variance of the $k$th discrete-frequency bin, respectively.
The distribution of \( \zeta_{ab}[l, 64] \) is shown in Fig. 1(c). The normal CDF of \( \zeta_{ab}[l, k] \) is given by

\[
F(\zeta_{ab}[l, k]) = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{\zeta_{ab}[l, k] - \mu_k}{\sigma_k \sqrt{2}} \right) \right].
\] (21)

For the Deep Xi framework, the maximum likelihood (ML) \textit{a posteriori} SNR estimate is used with the MMSE-STSA and MMSE-LSA estimators: \( \hat{\gamma}[l, k] = \zeta[l, k] + 1 \). Whereas the MMSE-STSA and MMSE-LSA estimators largely rely on the \textit{a priori} SNR, a small improvement may be realised by more accurately estimating the \textit{a posteriori} SNR (Cappe, 1994). Motivated by this, we propose to jointly estimate both the \textit{a priori} and \textit{a posteriori} SNRs. The decibel value of the instantaneous \textit{a priori} and \textit{a posteriori} SNRs can be used as a training target,

\[
\hat{r}[l, k] = [\hat{\xi}_{ab}[l, k]; \gamma_{ab}[l, k]],
\] (22)

where \([::]\) is the concatenation operation and \( \gamma_{ab}[l,k] \) is computed as

\[
\gamma_{ab}[l,k] = 10 \log_{10}(\hat{\gamma}[l,k]).
\] (23)

We also apply the min-max normalisation to \([\hat{\zeta}[l,k]; \gamma[l,k]]\) and \([\hat{\xi}_{ab}[l,k]; \gamma_{ab}[l,k]]\), giving \([\hat{\zeta}[l,k]; \gamma[l,k]]\) and \([\hat{\xi}_{ab}[l,k]; \gamma_{ab}[l,k]]\), respectively. \( \gamma[l,k] \) and \( \gamma_{ab}[l,k] \) are formed by replacing \( |S[l,k]| \) in Eq. (17) with \( |\gamma[l,k]| \) and \( |\gamma_{ab}[l,k]| \), respectively.

We also propose to use the CDF of \( \hat{\gamma}_{ab}[l,k] \) and \( \gamma_{ab}[l,k] \) as a compression function to form a training target,

\[
\hat{r}[l, k] = [F(\hat{\gamma}_{ab}[l, k]); F(\gamma_{ab}[l, k])].
\] (24)

The distribution of \( \gamma_{ab}[l,64] \) is shown in Fig. 1(e), which indicates that \( \gamma_{ab}[l,k] \) follows a super-Gaussian distribution. Hence, we assume that \( \gamma_{ab}[l,k] \) follows a Laplace distribution, \( \gamma_{ab}[l,k] \sim \text{Laplace}(\zeta_k, b_k) \), where \( \zeta_k \) and \( b_k \) are the location and scale parameters of the \( k \)th discrete-frequency bin, respectively. The Laplace CDF of \( \gamma_{ab}[l,k] \) is given by

\[
F(\gamma_{ab}[l,k]) = \frac{1}{2} + \frac{1}{2} \text{sgn}(\gamma_{ab}[l,k] - \zeta_k) \times \left(1 - \exp \left(\frac{\gamma_{ab}[l,k] - \zeta_k}{b_k}\right) \right),
\] (25)

where \( \text{sgn} \) is the sign function.

We also propose to use the gain of an MMSE estimator as the training target, where the instantaneous \textit{a priori} and \textit{a posteriori} SNRs are used to compute the gain. The SRWF gain has already been applied as a training target in the literature in the form of the IRM. We, thus, extend this approach to include the gain of the WF (\( G_{WF}[l,k] \)) and CWF (\( G_{CWF}[l,k] \)), as well as the MMSE-STSA (\( G_{MMSE-STSA}[l,k] \)) and MMSE-LSA (\( G_{MMSE-LSA}[l,k] \)) estimators. The values for \( G_{MMSE-STSA}[l,k] \) and \( G_{MMSE-LSA}[l,k] \) are clipped between 0 and 1.

V. EXPERIMENT SETUP

A. DNN: ResNet-TCN

A modified version of the residual network temporal convolutional network (ResNet-TCN) from Zhang et al. (2020) is used to evaluate each training target. The set of hyperparameters for the ResNet-TCN used in this work are derived from Zhang et al. (2020). It is shown from input to output in Fig. 2. Its input at time-frame \( l \) is \( |X_l| = \{|X[l,0]|, |X[l,1]|, \ldots, |X[l,N_d/2]|\} \), which is the 257-point single-sided noisy speech MS that includes the DC and Nyquist discrete-frequency bins (where \( N_d = 512 \) is the number of time-domain samples for each time-frame and the number of DFT coefficients considered. Additionally, a time-frame shift of \( N_i = 256 \) is used). Its output at time-frame \( l \) is \( \hat{l}_l = \{\hat{l}_l[0], \hat{l}_l[1], \ldots, \hat{l}_l[N_d/2]\} \). Explicitly, the input and output sizes of ResNet-TCN for each time-frame is 257. The input is first transformed by \( FC \), a fully connected layer of size \( d_{model} = 256 \). Instead of applying layer normalisation (Ba et al., 2016), followed by the rectifier linear function to \( FC \), as in Zhang et al. (2020), we apply the rectifier linear activation function followed by layer normalisation without the scale and shift operations. This reduces overfitting as demonstrated in Xu et al. (2019). The \( FC \) layer is followed by \( B = 40 \) bottleneck residual blocks, where \( b = 1, 2, \ldots, B \) is the block index. Residual blocks are facilitated by adding the block’s input to its output, preventing the vanishing and exploding gradient problems (He et al., 2016).

Each block contains three one-dimensional causal dilated convolutional units. Here, we modify the pre-activation of the convolutional units in Zhang et al. (2020) by using the rectifier linear activation function, followed by layer normalisation without the scale and shift operations.
(again, following Xu et al., 2019). The kernel size, output size, and dilation rate for each convolutional unit are denoted in Fig. 2 as (kernel size, output size, dilation rate). The first and third convolutional units have a kernel size of one, whereas the second has a kernel size of \( k = 3 \). The first and second convolutional units have a dilation rate of one, whereas the second convolutional unit employs a dilation rate of \( d \), providing a receptive field over previous time steps. The dilation rate for the second convolutional unit, \( d \), is cycled as the block index, \( b \), increases: \( d = 2^{(b-1) \mod \left(\log_2(D)\right)+1} \), where mod is the modulo operation, and \( D = 16 \) is the maximum dilation rate. This dilation rate scheme is similar to the one used for the Conv-TasNet (Luo et al., 2015). 

For training, the Adam algorithm (Kingma and Ba, 2014) with default hyper-parameters is used for gradient descent optimisation. Gradients are clipped between \([-1, 1]\). A mini-batch size of eight variable-length noisy speech signals is used for each training iteration. Each example for a mini-batch is padded to the number of time-frames of the longest duration noisy speech signal from the mini-batch. The loss is masked for the padded components. The noisy speech signals for each mini-batch are computed on the fly as follows: each clean speech recording selected for the mini-batch is mixed with a random section of a randomly selected noise recording at a randomly selected SNR level from the set specified for the training set. The selection order for the clean speech recordings is randomised for each epoch. ResNet-TCN is implemented in TensorFlow 2.3.0 (Abadi et al., 2015) and available online. The training targets are also evaluated using a RNN and a multi-head attention DNN architecture as described in Appendix A 2.

**B. Output layer activation function and loss function**

In Table II, we specify the output layer activation function and loss function used for each training target. For \( \{t[l,k] \in \mathbb{R} : 0 \leq t[l,k] \leq 1\} \), the sigmoid activation function is applied to the output layer, and the MSE and binary cross-entropy (BCE) are both investigated as the loss function (except for IAM, where mMSA is also evaluated). For all other cases, no activation function is applied to the output layer and MSE is investigated as the loss function.

**C. Datasets**

Separate models for each training target are trained and evaluated on the following datasets.

**1. DEMAND Voice Bank dataset**

The DEMAND Voice Bank dataset (Valentini-Botinhao et al., 2016) has been used frequently so as to late to evaluate deep learning approaches to speech enhancement/separation (Nikzad et al., 2020). The training set includes clean speech recordings from 28 speakers of the Voice Bank corpus (Veaux et al., 2013; 11 572 recordings). It also includes two synthetic noise sources (speech-shaped noise and babble) as described in Valentini-Botinhao et al., 2016) as well as eight real-world noise recordings from the DEMAND dataset (Thienn et al., 2013). The clean speech and noise recordings are downsampled from 48 to 16 kHz. The noisy speech signals for the training set are formed by mixing each clean speech recording with a random section of a randomly selected noise recording at a random selected SNR level from the set \{0, 5, 10, 15\} (dB). This creates a set of 11 572 noisy speech signals for training. The test set includes 824 clean speech recordings of two speakers from the Voice Bank corpus—393 from p232 and 431 from p257 (Veaux et al., 2013). Both speakers are separate from those selected for the training set. A total of 20 different conditions are used to create the noisy speech, including 5 noise sources from the DEMAND dataset (separate from those included in the training set), and 4 SNR levels: \{2.5, 7.5, 12.5, 17.5\} (dB). The clean speech and noise recordings are downsampled from 48 to 16 kHz prior to mixing. Approximately 20 different sentences for each speaker are used per condition with the total number of noisy speech signals in the test set amounting to 824. For the training set, the minimum, average, and maximum duration of the recordings are 1.1, 2.5, and 9.8 sec, respectively. For the test set, the minimum, average, and maximum duration of the recordings are 1.2, 2.9, and 15.1 sec, respectively.

**2. Deep Xi dataset**

The Deep Xi dataset (Nicolson, 2020) is larger than the DEMAND Voice Bank dataset with a wider range of

| Target, \( t \) | Range | Output activation | Loss function |
|----------------|-------|-------------------|---------------|
| IBM            | [0, 1] | Sigmoid           | MSE or BCE    |
| IRM            | [0, 1] | Sigmoid           | MSE or BCE    |
| IAM*           | [0, 1] | Sigmoid           | MSE, BCE, or mMSA |
| \( S[0] \)     | [0, 1] | Sigmoid           | MSE or BCE    |
| \( S[\alpha] \) | \( -\infty, \infty \) | Linear         | MSE           |
| \( z[\alpha] \) | \( -\infty, \infty \) | Linear         | MSE           |
| \( S[\beta] \) | [0, 1] | Sigmoid           | MSE or BCE    |
| \( S[\gamma] \) | \( 0, \infty \) | Linear         | MSE           |
| \( F(S[\delta]) \) | [0, 1] | Sigmoid           | MSE or BCE    |
| \( F(S[\epsilon]) \) | [0, 1] | Sigmoid           | MSE or BCE    |
| \( z[\epsilon] \) | \( -\infty, \infty \) | Linear         | MSE           |
| \( z[\epsilon][\gamma] \) | \( 0, \infty \) | Sigmoid         | MSE or BCE    |
| \( z[\epsilon][\gamma][\delta] \) | \( -\infty, \infty \) | Linear         | MSE           |
| \( F(S[\zeta]) \) | [0, 1] | Sigmoid           | MSE or BCE    |
| \( G_{WF} \) | [0, 1] | Sigmoid           | MSE or BCE    |
| \( G_{CWF} \) | [0, 1] | Sigmoid           | MSE or BCE    |
| \( G_{MSE}^{\text{MMSE}} \) | [0, 1] | Sigmoid           | MSE or BCE    |
| \( G_{MSE}^{\text{LSA}} \) | [0, 1] | Sigmoid           | MSE or BCE    |
conditions for training. The test set also allows for the evaluation of individual noise sources and SNR levels. For the training and validation sets, 69 708 clean speech recordings from the CSTR VCTK corpus (Veaux et al., 2017; 41 169 recordings, speakers p232 and p257 are excluded) and the train-clean-100 set of the Librispeech corpus (Panayotov et al., 2015; 28 539 recordings) are used. The minimum, average, and maximum duration of the recordings are 1.4, 12.3, and 17.2 sec, respectively, for the train-clean-100 set and 1.2, 3.6, and 15.1 sec, respectively, for the CSTR VCTK corpus. For the training and validation sets, noise recordings from the QUT-NOISE dataset (Dean et al., 2010), Nonspeech dataset (Hu and Wang, 2010), RSG-10 dataset (voice babble, F16, and factory welding) are excluded as they are used for the test set; Steeneken and Geurtsen (1988), Urban Sound dataset (street music recording no. 26 270 is excluded as it is used for the test set; Salamon et al., 2014), Environmental Background Noise dataset (Saki et al., 2016), noise set from the MUSAN corpus (Snyder et al., 2015), multiple FreeSound packs,4 and coloured noise recordings (with an x value ranging from −2 to 2 in increments of 0.25). Noise recordings that are over 30 s in length are split into segments of 30 s or less. This gives a total of 17 458 noise recordings, each with a length less than or equal to 30 s. All clean speech and noise recordings are single channel with a sampling frequency of 16 kHz (recordings with a higher sampling frequency are downsampled to 16 kHz). The SNR levels from the set \{q ∈ \mathbb{Z} | 10 ≤ q ≤ 20\} (dB), where q is the SNR level, are used for the training set. For the validation set, 1 000 clean speech and noise recordings are randomly selected (without replacement) and removed from the aforementioned clean speech and noise sets. Each clean speech recording is paired with one of the noise recordings. The clean speech recording is then mixed with a random section of the noise recordings at a randomly selected SNR level from the set \{q ∈ \mathbb{Z} | 10 ≤ q ≤ 20\} (dB). This forms 1 000 noisy speech signals for the validation set. For the test set, recordings of four real-world noise sources, including two nonstationary and two coloured, are included in the test set. The two real-world nonstationary noise sources include voice babble from the RSG-10 noise dataset (Steeneken and Geurtsen, 1988) and street music (recording no. 26 270) from the Urban Sound dataset (Salamon et al., 2014). The two real-world coloured noise sources include F16 and factory welding from the RSG-10 noise dataset (Steeneken and Geurtsen, 1988). Ten clean speech recordings are randomly selected (without replacement) from the test-clean set of the Librispeech corpus for each of the four noise recordings. The clean speech recordings from the test-clean set have a duration of up to 34 s. All clean speech and noise recordings are single channel with a sampling frequency of 16 kHz (recordings with a higher sampling frequency are downsampled to 16 kHz). To create the noisy speech, a random section of the noise recording is selected and mixed with the clean speech at the following SNR levels: \{-5, 0, 5, 10, 15\} (dB). This creates a test set of 200 noisy speech signals.

D. Sample

The statistics required for some of the compression functions described in Sec. IV (μk, σk, min(·), max(·), ζk, and b2) are estimated from a sample of 1 000 randomly selected training examples (statistics are found separately for each training set). For the Laplace CDF in Eq. (25), ζk is assumed to be zero and b2 is estimated using values greater than ζk—as a larger proportion of \(γ_{nk}[l, k]\) values are in the right tail [see Figs. 1(e) and 1(f)].

E. ASR system

Project DeepSpeech is an ASR system trained solely on clean speech (i.e., no multi-condition training is used). It is an open source implementation of the Deep Speech ASR system (Hannun et al., 2014). It uses 26 mel-frequency cepstral coefficients (MFCCs) as its input. It is used to evaluate the objective intelligibility of the training targets. It is also used to evaluate the front-end performance of each training target for robust ASR.

F. Objective quality and intelligibility measures

The objective quality and intelligibility measures used in this study to assess each training target are given in Table III. Each objective measure requires the time-domain samples of the enhanced and clean speech. The WER is calculated by WER = 100\( | D(H, R) | / N \), where \( H \) is the hypothesis transcript, \( R \) is the reference transcript, and \( N \) is the number of words in \( R \). \( D(H, R) \) is the Levenshtein distance between \( H \) and \( R \). The WER is given as a fraction from zero to one or as a percentage.

VI. RESULTS AND DISCUSSION

In this section, we evaluate the speech enhancement/separation and robust ASR performance of each training target, as well as each compression and loss function. The training targets are evaluated on the DEMAND Voice Bank and Deep Xi dataset. Epochs 125 and 150 for the DEMAND Voice Bank and Deep Xi datasets were used for evaluation, respectively. The employed DNNs required more epochs for

| Measure        | Assesses | Range       |
|----------------|----------|-------------|
| CSIG (Hu and Loizou, 2008) | Quality | [1, 5]     |
| CBAK (Hu and Loizou, 2008)  | Quality | [1, 5]     |
| COVL (Hu and Loizou, 2008)  | Quality | [1, 5]     |
| PESQ (Rix et al., 2001)    | Quality | [−0.5, 4.5]|
| STOI (Taal et al., 2011)   | Intelligibility | [0, 100]%  |
| ESTOI (Jensen and Taal, 2016) | Intelligibility | [0, 100]%  |
| SDR (Vincent et al., 2006) | Quality | (−∞, ∞)   |
| SI-SDR (Roux et al., 2019) | Quality | (−∞, ∞)   |
| SegSNR (Mermelstein, 1979) | Quality | (−∞, ∞)   |
| WER | Intelligibility | [0, 100]%  |
the validation loss to converge on the Deep Xi dataset. The time-frame index, \( l \), and discrete-frequency bin, \( k \), are omitted from the notation hereafter for convenience. Results for \( |S|^j + \text{MSE} \) and \( [\xi^j; \gamma^j] + \text{MSE} \) were not included as the training loss would not converge after repeated training runs. For difference testing, the two-sample \( t \)-test (\( \alpha = 0.05 \)) is used henceforth. The speech enhancement/separation performances of the \( a \) priori SNR training targets are evaluated using the MMSE-LSA estimator and SRWF for conciseness (joint \( a \) priori and \( a \) posteriori SNR training targets are evaluated using only the MMSE-LSA estimator). The MMSE-LSA estimator is chosen as it scores highly for each tested objective measure as demonstrated in Appendix A1. The SRWF is chosen as its gain has the same form as the IRM. This will reveal the advantages and disadvantages of estimating the gain over estimating the \( a \) priori and \( a \) posteriori SNRs.

### A. CASA vs MS vs MMSE training targets

First, we compare the objective quality and intelligibility scores of the three training target categories using the mean objective scores in Table IV. MMSE training targets produced the best objective scores for each measure except STOI. MMSE training targets, namely, the joint \( a \) priori and \( a \) posteriori SNR training targets, attained the best scores for all objective measures of quality (except CSIG, which was attained by \( G_{\text{CWF}} \) and \( G_{\text{MMSE-STSA}} \)), demonstrating their capacity to produce enhanced/separated speech at a high quality. A significant difference exists between the CBAK scores of \( [F(\zeta_{db})]; F(\gamma_{db})] + \text{BCE} \) and all CASA and MS training targets, as well as \( F(\zeta_{db}) + \text{BCE} \) and all CASA and MS training targets, demonstrating that \( a \) priori SNR and joint \( a \) priori and \( a \) posteriori SNR training targets are the best at suppressing background noise. MMSE training targets, in particular the gain training target, achieved the best scores for ESTOI and WER, demonstrating that they also produce enhanced/separated speech with a high intelligibility. A CASA training target produced the highest STOI score (IAM + MSE), demonstrating its proficiency at producing highly intelligible enhanced/separated speech.

Comparing CASA and MS training targets, IAM + MSE produced better STOI, ESTOI, SDR, SI-SDR, SegSNR, and WER scores than any of the MS training targets. In fact, a significant difference between the SDR, SI-SDR, SegSNR, STOI, and ESTOI scores for IAM + MSE and all MS training targets exists. However, MS training targets, namely, \( |S|^{0.3} \), produced higher CSIG, CBAK, COVL, and PESQ scores than any of the CASA training targets. Moreover, a significant difference between the COVL and PESQ scores of \( |S|^{0.3} \) and all the CASA training targets exists. This indicates that CASA training targets produce more intelligible enhanced/separated speech than MS training targets. However, it is uncertain if CASA or MS training targets produce enhanced/separated speech at a higher quality as CASA training targets attain higher SDR, SI-SDR, and SegSNR scores, whereas MS training targets attain higher CSIG, CBAK, COVL, and PESQ scores. Considering that CSIG, CBAK, COVL, and PESQ are more correlated with subjective quality scores than SegSNR (Hu and Loizou, 2008), it stands to reason that the MS training targets produce higher quality enhanced/separated speech.

Comparing MS and MMSE training targets, \( [F(\zeta_{db}); F(\gamma_{db})] + \text{BCE} \) obtained better scores than any of the MS training targets for each of the objective measures. The advantage that MMSE training targets have over MS training targets is emphasised when considering that a significant difference between the CBAK, SDR, SI-SDR, and ESTOI scores for \( [F(\xi_{db}); F(\gamma_{db})] + \text{BCE} \) and all MS training targets exists. This indicates that MMSE training targets produce enhanced/separated speech at a higher quality and intelligibility than MS training targets.

Comparing CASA and MMSE training targets, \( G_{\text{CWF}} + \text{MSE} \) produced better CSIG, CBAK, COVL, PESQ, ESTOI, SDR, SI-SDR, SegSNR, and WER scores than the IRM and IAM. Moreover, when comparing the CB\( \text{AK} \) scores for \( G_{\text{CWF}} + \text{MSE} \) to those for the IRM and IAM, a statistically significant difference exists, demonstrating the main advantage that MMSE training targets hold over CASA training targets—background noise suppression. In summary, the results in Table IV indicate that \( a \) priori SNR and joint \( a \) priori and \( a \) posteriori SNR training targets are best for background noise suppression and produce the highest quality enhanced/separated speech, whereas gain training targets, as well as the IAM, cause the least speech distortion and produce the most intelligible enhanced/separated speech. Moreover, CASA training targets produce more intelligible enhanced/separated speech than MS training targets, and MS training targets produce enhanced/separated speech at a higher quality than CASA training targets.

### B. CASA training targets

In this subsection, we evaluate the mean objective scores of the CASA training targets presented in Table IV. Amongst the CASA training targets, the IBM produced the highest SDR, SI-SDR, and SegSNR, demonstrating that its all-or-nothing strategy succeeds at decreasing the amount of distortion between the clean and enhanced/separated speech. However, IRM + MSE and IAM + MSE produce better objective scores for all other measures. Moreover, a significant difference exists between the CSIG, CBAK, COVL, PESQ, STOI, ESTOI, and WER scores for the IBM and those for IRM + MSE and IAM + MSE, indicating that the IBM produces enhanced/separated speech at a relatively lower quality and intelligibility. Comparing the IRM and IAM, the IAM attains better CSIG, COVL, PESQ, STOI, ESTOI, and WER scores than the IRM, indicating that the IAM produces more intelligible enhanced/separated speech with less speech distortion than the IRM. However, the IRM attains higher CBAK, SDR, SI-SDR, and SegSNR scores than the IAM, suggesting that the IRM is superior at background noise suppression than the IAM.
TABLE IV. Mean objective scores for ResNet-TCN on the test set of the DEMAND Voice Bank dataset described in Sec. VC1. The highest score for each measure—except WER—appears in boldface. The lowest WER is in boldface. The values for the STOI, ESTOI, and WER are given as percentages.

| Category | Target | Loss | MMSE estimator | CSIG | CBAK | COVL | PESQ | STOI | ESTOI | SDR | SI-SDR | SegSNR | WER |
|----------|--------|------|----------------|------|------|------|------|------|-------|-----|-------|--------|-----|
| —        | Noisy speech | —    | —              | 5.30 | 2.47 | 2.73 | 1.99 | 91.53 | 78.31 | 8.68 | 8.39 | 1.71   | 30.03 |
| CASA     | IBM    | MSE  | —              | 1.99 | 2.96 | 1.96 | 2.11 | 92.97 | 81.64 | 19.27 | 17.56 | 9.11   | 36.60 |
| IBM      | BCE    | —    | 2.12           | 2.96 | 2.03 | 2.13 | 92.73 | 81.63 | 18.90 | 17.37 | 9.04   | 36.87 |
| IRM      | MSE    | —    | 4.17           | 3.29 | 3.46 | 2.73 | 93.74 | 84.05 | 17.72 | 16.92 | 8.04   | 25.52 |
| IRM      | BCE    | —    | 4.16           | 3.25 | 3.45 | 2.72 | 93.87 | 84.15 | 17.05 | 16.42 | 7.55   | 25.57 |
| IAM      | MSE    | —    | 4.20           | 3.28 | 3.49 | 2.74 | **93.94** | 84.22 | 17.38 | 16.70 | 7.66   | 25.48 |
| IAM      | BCE    | —    | 4.22           | 3.29 | 3.50 | 2.75 | 93.76 | 84.11 | 17.25 | 16.56 | 7.67   | 25.86 |
| IAM      | mMSA   | —    | 2.23           | 2.72 | 1.97 | 1.77 | 93.52 | 76.36 | 18.04 | 16.23 | 7.73   | 32.66 |

**MMSE**

| Category | Target | Loss | MMSE-LSA | CSIG | CBAK | COVL | PESQ | STOI | ESTOI | SDR | SI-SDR | SegSNR | WER |
|----------|--------|------|----------|------|------|------|------|------|-------|-----|-------|--------|-----|
| —        | Noisy speech | —    | MMSE-LSA | 2.34 | 1.60 | 1.77 | 1.57 | 75.28 | 49.88 | 8.20 | -4.21 | -1.81   | 58.85 |
| CASA     | IBM    | MSE  | MMSE-LSA | 4.11 | 3.06 | 3.40 | 2.68 | 91.16 | 79.72 | 12.50 | 10.89 | 4.80   | 28.31 |
| IAM      | MSE    | —    | MMSE-LSA | 4.07 | 3.05 | 3.38 | 2.70 | 90.99 | 79.22 | 11.91 | 9.93 | 4.69   | 29.59 |
| IAM      | BCE    | —    | MMSE-LSA | 4.02 | 2.94 | 3.27 | 2.51 | 91.15 | 79.47 | 11.91 | 10.16 | 4.30   | 28.88 |
| IAM      | mMSA   | —    | MMSE-LSA | 4.09 | 3.08 | 3.38 | 2.67 | 91.88 | 80.51 | 13.62 | 11.99 | 5.22   | 28.04 |
| IAM      | BCE    | —    | MMSE-LSA | 4.25 | 3.29 | 3.57 | 2.87 | 92.96 | 82.74 | 16.37 | 15.21 | 6.79   | 27.25 |

**IAM**

| Category | Target | Loss | MMSE-LSA | CSIG | CBAK | COVL | PESQ | STOI | ESTOI | SDR | SI-SDR | SegSNR | WER |
|----------|--------|------|----------|------|------|------|------|------|-------|-----|-------|--------|-----|
| —        | Noisy speech | —    | MMSE-LSA | 3.71 | 2.69 | 2.91 | 2.13 | 89.46 | 76.82 | 9.58 | 6.33  | 3.38   | 32.75 |
| IAM      | MSE    | —    | MMSE-LSA | 3.91 | 2.92 | 3.18 | 2.45 | 90.19 | 78.12 | 11.26 | 8.79  | 4.45   | 30.77 |

**WF MSE**

| Category | Target | Loss | MMSE-LSA | CSIG | CBAK | COVL | PESQ | STOI | ESTOI | SDR | SI-SDR | SegSNR | WER |
|----------|--------|------|----------|------|------|------|------|------|-------|-----|-------|--------|-----|
| —        | Noisy speech | —    | MMSE-LSA | 2.34 | 1.60 | 1.77 | 1.57 | 75.28 | 49.88 | 8.20 | -4.21 | -1.81   | 58.85 |
| CASA     | IBM    | MSE  | MMSE-LSA | 4.11 | 3.06 | 3.40 | 2.68 | 91.16 | 79.72 | 12.50 | 10.89 | 4.80   | 28.31 |
| IAM      | MSE    | —    | MMSE-LSA | 4.07 | 3.05 | 3.38 | 2.70 | 90.99 | 79.22 | 11.91 | 9.93 | 4.69   | 29.59 |
| IAM      | BCE    | —    | MMSE-LSA | 4.02 | 2.94 | 3.27 | 2.51 | 91.15 | 79.47 | 11.91 | 10.16 | 4.30   | 28.88 |
| IAM      | mMSA   | —    | MMSE-LSA | 4.09 | 3.08 | 3.38 | 2.67 | 91.88 | 80.51 | 13.62 | 11.99 | 5.22   | 28.04 |
| IAM      | BCE    | —    | MMSE-LSA | 4.25 | 3.29 | 3.57 | 2.87 | 92.96 | 82.74 | 16.37 | 15.21 | 6.79   | 27.25 |

**WF BCE**

| Category | Target | Loss | MMSE-LSA | CSIG | CBAK | COVL | PESQ | STOI | ESTOI | SDR | SI-SDR | SegSNR | WER |
|----------|--------|------|----------|------|------|------|------|------|-------|-----|-------|--------|-----|
| —        | Noisy speech | —    | MMSE-LSA | 3.71 | 2.69 | 2.91 | 2.13 | 89.46 | 76.82 | 9.58 | 6.33  | 3.38   | 32.75 |
| IAM      | MSE    | —    | MMSE-LSA | 3.91 | 2.92 | 3.18 | 2.45 | 90.19 | 78.12 | 11.26 | 8.79  | 4.45   | 30.77 |

For the IBM, a significant difference exists only between the CSIG scores for the MSE and BCE loss functions, suggesting that either loss function is suitable for the IBM. For the IRM, there is a significant difference between the SDR, SI-SDR, and SegSNR for the MSE and BCE loss functions. For the IAM, no significant difference exists between the objective scores for the MSE and BCE loss functions, suggesting that either loss function is suitable for the IAM. For the IAM, the mMSA loss function produced the highest SDR and SegSNR. However, the mMSA loss function was outperformed on the remaining objective measures with a significant difference existing between the CSIG, CBAK, COVL, PESQ, ESTOI, and SDR scores for the mMSA loss function and MSE and BCE loss functions. In **Luo et al. (2017)**, the mMSA loss function attained higher SDR scores than the MSE loss function. However, no other
objective measures [apart from signal-to-interference ratio (SIR) and signal-to-artifacts ratio (SAR)] were reported in Luo et al. (2017). This highlights the importance of using a range of objective measures for evaluation. For the mSA loss function in Zheng and Zhang [2019, Eq. (8)], which is the mMSA loss function with additional phase terms, the authors trained using the MSE loss function for the first portion of training and then the mSA loss function for the second portion of training. This may increase performance when using the mMSA loss function.

C. MS training targets

Here, we evaluate the mean objective scores of the MS training targets found in Table IV. For the MS training targets, multiple compression functions were investigated. Min-max normalisation and power compression of the clean speech MS was found to be best, with $|S|^1$ producing the best objective scores for STOI, SI-SDR, and WER and $|S|^{0.3}$ producing the best objective scores for CSIG, CBak, COVL, PESQ, STOI, SDR, and SegSNR. This suggests that applying a compression function to the exponentially distributed magnitude or power values is better than applying it to the normally distributed decibel values of the clean speech MS. We speculate that this is because decibel values above 0 dB are overcompressed. A portion of the values below 0 dB are a result of regions of silence, whereas values above 0 dB correspond to regions of speech. An example of the overcompression of values above 0 dB is shown in Fig. 1(b); the CDF compresses all values below 0 dB into the interval $[0,0.967]$, whereas values above 0 dB are compressed into the interval $(0.967,1]$. The upper bounds of the interval $[0,0.967]$ also indicate the proportion of values that are below 0 dB for the sample described in Sec. VD (i.e., 96.7% of the decibel values for the clean speech MS are below 0 dB). This likely causes a model to poorly estimate the formants of speech. For $|S|^{0.3}$, a significant difference was found between the objective scores of each loss function (for all measures except WER) with BCE performing best. This was also the case for $F(|S|)$. This indicates that the BCE loss function is best for MS training targets with compression applied to their decibel values. As the MSE loss function assumes that the variable to be estimated is distributed normally, less importance is given to the observed values that are further away from the mean. This is detrimental to the values of the clean speech MS above 0 dB, as after compression they would be located far away from the mean of $|S|^1$ and $F(|S|)$.

D. MMSE training targets

The mean objective scores attained by the MMSE training targets found in Table IV are evaluated in this subsection. Amongst all three MMSE training target subclasses (a priori SNR, joint a priori and a posteriori SNRs, and gain training targets), the results indicate that the joint a priori and a posteriori SNR training targets produce enhanced-separated speech at the highest quality (highest CBak, COVL, PESQ, SDR, SI-SDR, and SegSNR scores), and the gain training targets produce the most intelligible enhanced/separated speech (best CSIG, STOI, ESTOI, and WER scores). The a priori SNR training targets performed similarly to the joint a priori and a posteriori SNR training targets with a significant difference existing only between the SDR and SI-SDR values of $F(\zeta_{ab})$ + MMSE-LSA and $F(\zeta_{ab}+\gamma_{ab})$ + MMSE-LSA. Moreover, it is unclear if an improvement in the performance of the MMSE-LSA estimator is achieved when replacing the ML estimate of the a posteriori SNR with that estimated using a joint a priori and a posteriori SNR training target. This highlights how little of an impact the a posteriori SNR has on the MMSE-LSA estimator as described in Cappe (1994).

The best performing a priori SNR training target $F(\zeta_{ab})$, producing the best CSIG, CBak, COVL, PESQ, STOI, ESTOI, SDR, SI-SDR, and SegSNR scores $[\zeta(\bar{\zeta}_{ab})]$ produced the lowest WER and equaled the highest STOI score. This indicates that the CDF as the compression function is an apt choice for the decibel values of the instantaneous a priori SNR. This is the opposite finding to that found for the decibel values of the clean speech MS. Like the decibel values of the clean speech MS, the decibel values of the instantaneous a priori SNR are distributed normally. However, poor estimation of values from the tails of the distribution do not impact the performance as significantly as they correspond to very small and large power values of the instantaneous a priori SNR. The best performing joint a priori and a posteriori SNR training target was $[F(\zeta_{ab})+F(\gamma_{ab})]$, producing the best CSIG, COVL, PESQ, STOI, and WER scores, and $[\zeta_{ab}+\gamma_{ab}]$, producing the highest CBak, ESTOI, SDR, SI-SDR, and SegSNR scores. Amongst the gain training targets, $G_{WF}$ and $G_{MMSE-LSA}$ produced the most intelligible enhanced/separated speech (best STOI, ESTOI, and WER scores), $G_{WF}$, $G_{MMSE-LSA}$ were best at background noise suppression (best CBak, SDR, SI-SDR, and SegSNR scores), and $G_{WF}$ and $G_{MMSE-LSA}$ produced the highest quality enhanced/separated speech (best CSIG, COVL, and PESQ scores).

For $\zeta$, it can be seen that the BCE loss function significantly outperforms the MSE loss function. This is the only a priori SNR training target in which a compression function is applied to the power values and not to the decibel values. The power values of the instantaneous a priori SNR are distributed exponentially, even after min-max normalisation. This makes the MSE loss function suboptimal as it assumes a normal distribution. This also gives reason as to why the loss for $[\zeta;\gamma]$ + MSE would not converge during training. The MSE loss function was best for $\zeta_{ab}$, where a significant difference between the CSIG, COVL, PESQ, SDR, SI-SDR, SegSNR, and WER scores of the MSE and BCE loss functions exists.

For $F(\zeta_{ab})$, it is unclear which loss function is best as a significant difference existed only between the SDR of the MSE and BCE loss functions. For the joint a priori and a posteriori SNR training targets, it is unclear if the MSE or BCE loss function is best as a significant difference exists.
only between the SDR and SI-SDR scores of the MSE and BCE loss functions for \( [\hat{\gamma}'; \hat{\gamma}'] \) (there was no significant difference between the objective scores of the MSE and BCE loss functions for \( [F(\hat{\gamma}'); F(\gamma')] \)). This finding was also the case for the gain training targets as a significant difference exists only between the COVL and PESQ scores of the MSE and BCE loss functions for \( G_{\text{MMSE-LSA}} \) and between the SDR of the MSE and BCE loss functions for \( G_{\text{MMSE-LSE}} \).

### E. A priori SNR and joint a priori and a posteriori SNR training targets vs gain training targets

When comparing the mean objective scores (presented in Table IV) obtained by the best performing a priori SNR and joint a priori and a posteriori SNR training targets to those of the gain training targets, it is clear that the former produce enhanced/separated speech at a higher quality while the latter produce enhanced/separated speech at a higher intelligibility. For example, \( F(\hat{\gamma}'); \hat{\gamma}'] \) + SRWF + BCE produced higher CSIG, CBAM, COVL, PESQ, ESTOI, STOI, SDR, SI-SDR, and SegSNR scores than IRM + BCE (i.e., \( G_{\text{SRWF}} + \text{BCE} \)), whereas IRM + BCE produced better STOI and WER scores than \( F(\hat{\gamma}'); \hat{\gamma}'] \) + SRWF + BCE, where a statistically significant difference exists between the CBAM, COVL, PESQ, STOI, SDR, SI-SDR, and SegSNR scores. Moreover, \( [F(\hat{\gamma}'); F(\gamma')] \) + MMSE-LSA + BCE produced higher CSIG, CBAM, COVL, PESQ, SDR, SI-SDR, and SegSNR scores than \( G_{\text{MMSE-LSA}} \), whereas \( G_{\text{MMSE-LSA}} \) produced better STOI, ESTOI, and WER scores than \( [F(\hat{\gamma}'); F(\gamma')] \) + MMSE-LSA + BC, where a statistically significant difference exists between the CBAM, PESQ, and SegSNR scores.

To help understand this phenomenon, we compare the distribution of the IRM [Fig. 1(g)] to that of \( F(\hat{\gamma}'); \hat{\gamma}'] \) [Fig. 1(h)]. It can be seen that \( F(\hat{\gamma}') \) is distributed uniformly between zero and one, meaning that a model would equally observe all values during training. However, the IRM has a truncated bimodal distribution with peaks at zero and one, causing a model to more frequently observe values from these peaks. The values from the peak located at one correspond to the formants of the clean speech. This indicates that gain training targets bias a model to more accurately estimate the formants of the clean speech, leading to less speech distortion and more intelligible enhanced/separated speech (Bhat et al., 2017).

### F. Objective scores on alternative DNN architectures

The mean objective scores in Table IV are produced using the ResNet-TCN. To ensure that the previous findings are upheld on different DNN architectures, objective scores for each training target are found using an attention-based network and an RNN, namely, the multi-head attention network (MHANet) and bidirectional long short-term memory (BiLSTM)-Chimera++. The ResNet-TCN and MHANet are causal models, whereas BiLSTM-Chimera++ is a non-causal model. More details about each model, including hyperparameters, are given in Appendix A 2. The objective scores for MHANet and BiLSTM-Chimera++ on the DEMAND Voice Bank dataset are given in Tables V and VI, respectively.

For MHANet, the joint a priori and a posteriori SNR training targets produced the highest CBAM, COVL, PESQ, SI-SDR, and SegSNR scores. Moreover, the joint a priori and a posteriori SNR training targets produced the highest CSIG, CBAM, COVL, PESQ, ESTOI, SDR, SI-SDR, and SegSNR scores for BiLSTM-Chimera++. This supports the finding that joint a priori and a posteriori SNR training targets produce the highest quality enhanced/separated speech amongst the training targets. For MHANet, the gain training targets produced the best ESTOI and WER scores. Moreover, the gain training targets produced the best STOI and WER scores for BiLSTM-Chimera++. This supports the finding that gain training targets produce the most intelligible enhanced/separated speech amongst the training targets—with the IAM being the next best (IAM produced the highest STOI score for MHANet).

One notable inconsistency for the IAM is the difference between the objective scores for the mMSA loss function and MSE and BCE loss functions—the difference was significantly smaller for MHANet and BiLSTM-Chimera++ than for ResNet-TCN (for CSIG, CBAM, COVL, PESQ, and ESTOI). Another notable difference is that \( |S|^0.3 + \text{BiLSTM-Chimera++} \) equaled the highest CSIG and PESQ scores. Clean speech MS training targets, as well as IAM + mMSA, appear more difficult to learn as shown by their objective scores for the ResNet-TCN. Moreover, MHANet and BiLSTM-Chimera++ are more efficient at learning each of the training targets than ResNet-TCN as shown by the overall improvement in the objective scores from Table IV to Tables V and VI. Considering this, the learning efficiency of MHANet and BiLSTM-Chimera++ likely benefited IAM + mMSA and \( |S|^0.3 \) most of all as it would be harder to increase the objective scores of the training targets that demonstrated a high speech enhancement/separation performance on the ResNet-TCN.

### G. Deep Xi dataset objective scores

In this subsection, we evaluate the mean objective scores of the training targets on a second dataset, namely, the Deep Xi dataset (Table VII). Findings made on the Deep Xi dataset will be used either to support or refute the findings made in previous subsections on the DEMAND Voice Bank dataset. Due to their lack of performance on the DEMAND Voice Bank dataset, we exclude \( \hat{\xi}' \) and \( [\hat{\xi}'; \hat{\gamma}'] \) from this subsection. A priori SNR and joint a priori and a posteriori SNR training targets attained the highest CBAM, COVL, and PESQ scores, reinforcing that they produce enhanced/separated speech at the highest quality and are best at background noise suppression. Gain training targets, namely, \( G_{\text{MMSE-STA}} \), attained the highest CSIG, STOI, ESTOI, SDR, and SI-SDR scores, and IAM + MSE produced the best WER, supporting the claim that the IAM and gain training targets cause the least speech distortion and produce the most intelligible enhanced/separated speech.
For the IAM, the difference between the objective scores for the mMSA loss function and MSE and BCE loss functions was larger for the Deep Xi dataset than for the DEMAND Voice Bank dataset. It was expected that the larger training was larger for the Deep Xi dataset than for the DEMAND Voice Bank dataset. It was expected that the larger training targets on the Deep Xi dataset would enable the ResNet-TCN to better learn IAM+ mMSA, however, this was not the case. As for the DEMAND Voice Bank dataset, the CDF was the best compression function for the a priori SNR and joint a priori and a posteriori SNR training targets on the Deep Xi dataset.

**H. Objective scores on individual conditions**

Here, we evaluate the mean objective scores of the training targets on the individual conditions of the Deep Xi dataset. The conditions include multiple noise sources and
SNR levels. As the mean objective score for each condition is computed from a low number of samples (ten), we avoid difference testing in this subsection. Only the best performing CASA, MS, and MMSE training targets are considered. To keep the following evaluation succinct, we present only five of the aforementioned objective measures, namely, CSIG, CBAK, COVL, ESTOI, and WER. Moreover, only the MMSE-LSA estimator is used with the a\textit{priori} SNR training targets. Finally, only the MSE loss function is evaluated (except with $|S|'$ as it was only trained with the BCE loss function).

The mean CSIG scores are given in Table VIII. The highest CSIG scores were attained by $|S|'$ and $|S|'$ at lower SNR levels and MMSE training targets at higher SNR levels. This indicates that $|S|'$ and $|S|'$ and MMSE training targets do not heavily distort the speech at lower and higher

| Category | Target | Loss  | MMSE estimator | CSIG  | CBAK  | COVL  | PESQ  | STOI  | ESTOI | SDR  | SI-SDR | SegSNR | WER   |
|----------|--------|-------|----------------|-------|-------|-------|-------|-------|-------|-------|--------|--------|-------|
| MS       | —      | BCE   | MMS-LSA        | 3.50  | 2.47  | 2.73  | 1.99  | 91.53 | 78.31 | 8.68  | 8.39   | 1.71   | 30.03 |
| CASA     | IBM    | MSE   | —              | 1.89  | 3.02  | 1.94  | 2.23  | 94.22 | 82.70 | 19.00 | 17.52  | 9.10   | 25.54 |
|          | IBM    | BCE   | —              | 1.79  | 3.03  | 1.87  | 2.23  | 94.05 | 82.60 | 19.68 | 17.87  | 9.36   | 31.39 |
|          | IRM    | MSE   | —              | 4.25  | 3.40  | 3.58  | 2.89  | 94.21 | 85.17 | 18.67 | 17.70  | 8.85   | 23.84 |
|          | IRM    | BCE   | —              | 4.28  | 3.43  | 3.61  | 2.92  | 94.07 | 85.10 | 18.96 | 17.87  | 8.99   | 23.75 |
|          | IAM    | MSE   | —              | 4.21  | 3.30  | 3.52  | 2.81  | 93.90 | 84.66 | 17.44 | 16.79  | 7.86   | 24.05 |
|          | IAM    | BCE   | —              | 4.22  | 3.32  | 3.52  | 2.80  | 93.84 | 84.46 | 17.83 | 17.09  | 8.15   | 23.45 |
|          | IAM    | mMSA  | —              | 4.02  | 3.31  | 3.37  | 2.71  | 93.53 | 83.36 | 18.94 | 17.93  | 8.94   | 26.28 |
| MMSE     | $|S|$ | MSE   | MMSE-LSA       | 4.27  | 3.34  | 3.62  | 2.94  | 93.45 | 83.57 | 16.39 | 15.24  | 7.00   | 25.91 |
|          | $|S|$ | BCE   | —              | 3.90  | 2.93  | 3.19  | 2.49  | 89.72 | 76.42 | 10.76 | 8.96   | 4.58   | 32.46 |
|          | $|S|'$ | MSE   | —              | 3.96  | 3.01  | 3.27  | 2.58  | 89.95 | 77.32 | 11.57 | 9.85   | 4.90   | 30.17 |
|          | $|S|'$ | BCE   | —              | 4.01  | 3.08  | 3.33  | 2.66  | 91.04 | 79.09 | 12.73 | 10.89  | 5.45   | 29.48 |
|          | $|S|'$ | BCE   | —              | 4.06  | 3.12  | 3.39  | 2.72  | 91.02 | 79.42 | 13.17 | 11.34  | 5.61   | 28.57 |
|          | $|S|'$ | BCE   | —              | 4.34  | 3.40  | 3.70  | 3.05  | 93.24 | 83.82 | 16.45 | 14.97  | 7.22   | 25.30 |
|          | $|S|'$ | BCE   | —              | 3.66  | 2.71  | 2.89  | 2.13  | 88.41 | 74.78 | 8.73  | 5.70   | 3.82   | 33.33 |
|          | $|S|'$ | BCE   | —              | 3.83  | 2.87  | 3.10  | 2.38  | 89.21 | 76.05 | 10.65 | 8.24   | 4.40   | 30.95 |

TABLE VI. Mean objective scores for BiLSTM-Chimera++ on the test set of the DEMAND Voice Bank dataset described in Sec. V C.1. The highest score for each measure—except WER—appears in boldface. The lowest WER is in boldface. The values for the STOI, ESTOI, and WER are given as percentages.
SNR levels, respectively. From the objective scores in Table VIII, it is unclear if the noise source has a significant impact on the best performing training target. This will need investigating in future work—a larger sample size for each condition, as well as the evaluation of each training target over multiple training runs, is recommended.

The mean CBAK scores are given in Table IX. The highest CBAK scores were attained by $F(\xi_{\text{db}})$ and $[F(\xi_{\text{db}}):F(\gamma_{\text{db}})]$ for most conditions. This, again, demonstrates their proficiency at background noise suppression. Once more, $|S|^{0.3}$ performed well at low SNR levels—attaining the highest CBAK scores for voice babble, street music, and $F16$ at lower SNR levels. The mean COVL scores are given in Table X, where the highest COVL scores were attained by $|S|^{0.3}$ at lower SNR levels and MMSE training targets at higher SNR levels. The mean CSIG, CBAK, and COVL scores from Tables VIII–X indicate that MS training targets produce the highest quality enhanced/separated speech at lower SNR levels with the same being true for MMSE training targets at high SNR levels.

The mean STOI scores are given in Table XI; the highest STOI scores were attained by $G_{\text{MMSE--MMSE}}$ for most conditions, by IAM for some conditions, and by $|S|^{0.3}$ at lower SNR levels. Thus, MS training targets also produce highly intelligible enhanced/separated speech at lower SNR levels. At lower SNR levels, it appears easier to learn the mapping from the noisy speech MS to the clean speech MS rather than to a CASA or MMSE training target. One fact that could lead to answering this phenomenon is that unlike CASA and MMSE training targets, the distribution of the

| Category | Target | Loss | MMSE estimator | CSIG | CBAK | COVL | PESQ | STOI | ESTOI | SDR | SI-SDR | SegSNR | WER |
|----------|--------|------|----------------|------|------|------|------|------|-------|-----|--------|--------|-----|
| —        | Noisy speech | —    | —              | 2.26 | 1.80 | 1.67 | 1.24 | 77.75 | 56.12 | 5.00 | 0.25 | 58.88 |
| CASA     | IBM    | MSE  | —              | 1.46 | 2.19 | 1.34 | 1.34 | 80.58 | 64.17 | 10.27 | 9.57 | 56.87 |
| CASA     | IBM    | BCE  | —              | 1.42 | 2.19 | 1.32 | 1.34 | 81.04 | 64.14 | 10.53 | 9.91 | 55.51 |
| CASA     | IRM    | MSE  | —              | 3.08 | 2.43 | 2.36 | 1.73 | 85.68 | 70.19 | 10.47 | 10.09 | 4.16 |
| CASA     | IRM    | BCE  | —              | 3.11 | 2.43 | 2.37 | 1.74 | 85.52 | 70.03 | 10.48 | 10.03 | 4.14 |
| CASA     | IAM    | MSE  | —              | 3.10 | 2.46 | 2.39 | 1.76 | 85.70 | 70.56 | 10.72 | 10.18 | 4.34 |
| CASA     | IAM    | BCE  | —              | 3.11 | 2.44 | 2.38 | 1.74 | 85.50 | 69.97 | 10.48 | 9.97 | 4.16 |
| CASA     | IAM    | mMSA | —              | 1.29 | 1.95 | 1.17 | 1.15 | 68.27 | 48.37 | 7.16 | 5.44 | 3.74 |
| MS       | $|S'|$  | BCE  | —              | 3.14 | 2.43 | 2.41 | 1.79 | 85.78 | 70.60 | 9.84 | 9.13 | 3.66 |
| MS       | $|S'|$  | MSE  | —              | 3.07 | 2.36 | 2.35 | 1.72 | 82.38 | 66.56 | 7.18 | 6.00 | 2.95 |
| MS       | $z(|S'|)$ | MSE | —              | 2.83 | 2.18 | 2.14 | 1.57 | 79.19 | 61.02 | 5.42 | 3.94 | 1.95 |
| MS       | $|S'|$  | BCE  | —              | 2.99 | 2.30 | 2.26 | 1.64 | 81.44 | 65.14 | 6.67 | 5.76 | 2.76 |
| MS       | $|S'|$  | MSE  | —              | 2.99 | 2.32 | 2.27 | 1.65 | 81.63 | 65.14 | 7.10 | 5.99 | 3.02 |
| MS       | $|S'|$  | BCE  | —              | 2.85 | 2.15 | 2.11 | 1.47 | 80.58 | 63.53 | 5.09 | 3.29 | 1.72 |
| MS       | $|S'|$  | BCE  | —              | 2.95 | 2.24 | 2.23 | 1.60 | 81.18 | 64.45 | 6.01 | 4.56 | 2.12 |

J. Acoust. Soc. Am. 149 (5), May 2021
Aaron Nicolson and Kuldip K. Palwal 3287
TABLE VIII. Mean CSIG scores for each condition of the test set from the Deep Xi dataset described in Sec. V C 2. The MMSE-LSA estimator is used by the MMSE training targets. The highest score for each condition appears in boldface.

| Noise          | Target       | SNR  | 5   | 10  | 15  |
|----------------|--------------|------|-----|-----|-----|
| Voice babble   | Noisy speech | 1.91 | 2.24| 2.64| 3.08| 3.58|
|                | IRM + MSE    | 2.27 | 2.82| 3.35| 3.85| 4.30|
|                | IAM + MSE    | 2.25 | 2.83| 3.37| 3.89| **4.36**|
|                | $\|S_1\|^2 +$ BCE | 2.34 | 2.85| 3.35| 3.82| 4.23|
|                | $\|S_0\|^3 +$ MSE | **2.37** | **2.93** | **3.42** | **3.88** | 4.28|
|                | $F(\xi_{\text{MMSE-LSA}}) +$ MSE | 2.25 | 2.82| 3.36| 3.88| 4.35|
|                | $\{F(\xi_{\text{MMSE-LSA}}):F(\gamma_{\text{MMSE-LSA}})\} +$ MSE | 2.22 | 2.78| 3.34| 3.85| 4.34|
|                | $G_{\text{MMSE-LSA}} +$ MSE | 2.28 | 2.85| 3.40| **3.91** | **4.36**|
| Street music   | Noisy speech | 1.33 | 1.61| 1.97| 2.41| 2.93|
|                | IRM + MSE    | 1.85 | 2.33| 2.82| 3.29| 3.75|
|                | IAM + MSE    | 1.85 | 2.35| 2.85| 3.32| 3.76|
|                | $\|S_1\|^2 +$ BCE | **1.96** | 2.41| 2.87| 3.30| 3.70|
|                | $\|S_0\|^3 +$ MSE | 1.95 | 2.40| 2.85| 3.30| 3.70|
|                | $F(\xi_{\text{MMSE-LSA}}) +$ MSE | 1.91 | 2.44| 2.95| **3.46** | **3.92**|
|                | $\{F(\xi_{\text{MMSE-LSA}}):F(\gamma_{\text{MMSE-LSA}})\} +$ MSE | 1.87 | 2.39| 2.92| 3.43| **3.94**|
|                | $G_{\text{MMSE-LSA}} +$ MSE | 1.94 | **2.46** | **2.97** | 3.45| 3.89|
| F16            | Noisy speech | 1.24 | 1.58| 1.99| 2.46| 2.98|
|                | IRM + MSE    | 2.30 | 2.74| 3.16| 3.60| 4.04|
|                | IAM + MSE    | 2.34 | 2.78| 3.19| 3.61| 4.04|
|                | $\|S_1\|^2 +$ BCE | **2.51** | **2.91** | **3.30** | 3.69| 4.05|
|                | $\|S_0\|^3 +$ MSE | 2.50 | 2.89| 3.27| 3.66| 4.01|
|                | $F(\xi_{\text{MMSE-LSA}}) +$ MSE | 2.40 | 2.83| 3.24| **3.72** | **4.14**|
|                | $\{F(\xi_{\text{MMSE-LSA}}):F(\gamma_{\text{MMSE-LSA}})\} +$ MSE | 2.31 | 2.78| 3.21| 3.67| **4.09**|
|                | $G_{\text{MMSE-LSA}} +$ MSE | 2.36 | 2.81| 3.23| 3.66| 4.08|
| Factory        | Noisy speech | 1.48 | 1.77| 2.18| 2.65| 3.15|
|                | IRM + MSE    | 2.15 | 2.61| 3.04| 3.50| 3.91|
|                | IAM + MSE    | 2.13 | 2.61| 3.07| 3.53| 3.91|
|                | $\|S_1\|^2 +$ BCE | **2.25** | **2.68** | **3.10** | **3.53** | **3.88**|
|                | $\|S_0\|^3 +$ MSE | **2.25** | **2.68** | **3.08** | **3.47** | **3.85**|
|                | $F(\xi_{\text{MMSE-LSA}}) +$ MSE | 2.20 | 2.70| 3.16| 3.60| 3.94|
|                | $\{F(\xi_{\text{MMSE-LSA}}):F(\gamma_{\text{MMSE-LSA}})\} +$ MSE | 2.10 | 2.63| 3.10| 3.54| 3.95|
|                | $G_{\text{MMSE-LSA}} +$ MSE | 2.13 | **2.68** | **3.17** | **3.63** | **4.02**|

MS training target does not change with the SNR level. The scores for the WER are given in Table XII. It can be seen that there is no consistency between which training target performs best from condition to condition. Additionally, the STOI scores in Table XI are not indicative of the WER for each condition.

I. Robust ASR

In this subsection, we determine which training target, in the context of clean speech MS estimation, is most suited for an ASR front-end. Hence, we examine the WER in Tables IV, VII, V, and VI.

- In Table IV, $G_{\text{WF}}$ achieved the best WER, followed by $G_{\text{CWF}}$.
- In Table VII, the IAM achieved the best WER, followed by $G_{\text{MMSE-LSA}}$.
- In Table V, $G_{\text{CWF}}$ achieved the best WER, followed by the IAM;
- in Table VI, $G_{\text{MMSE-LSA}}$ achieved the best WER, followed by $G_{\text{MMSE-LSA}}$.

The preceding points indicate that gain training targets $G_{\text{WF}}$, IRM (i.e., $G_{\text{SRWF}}$), $G_{\text{CWF}}$, $G_{\text{MMSE-LSA}}$, and $G_{\text{MMSE-LSA}}$ as well as the IAM, are most suited for a front-end. The naive choice when selecting a training target is $S_{\text{MM}}$ as the clean speech MS is the quantity to be estimated and it employs the simplest compression function amongst the MS training targets. By replacing $S_{\text{MM}}$ with a more performative training target, such as $G_{\text{CWF}}$, a 4.39% reduction in the WER can be realised (Table V)—resulting in a significant improvement. It should also be noted that the training targets that attained the lowest WER also attained high objective intelligibility scores, reinforcing the notion that the WER is an objective measure of speech intelligibility (Thomas-Stonell et al., 1998). This implies that the same trait that results in high
TABLE X. Mean COVL scores for each condition of the test set from the Deep Xi dataset described in Sec. V.C.2. The MMSE-LSA estimator is used by the MMSE training targets. The highest score for each condition appears in boldface.

| Noise Target | SNR | -5 | 0  | 5  | 10 | 15 |
|--------------|-----|----|----|----|----|----|
| Voice babble | Noisy speech | 1.38 | 1.58 | 1.86 | 2.23 | 2.71 |
| | IRM + MSE | 1.61 | 2.03 | 2.50 | 3.00 | 3.49 |
| | IAM + MSE | 1.61 | 2.04 | 2.53 | 3.06 | 3.57 |
| | $|S|^0 + BCE$ | 1.66 | 2.06 | 2.52 | 3.00 | 3.44 |
| | $|S|^0 + MSE$ | **1.69** | **2.15** | **2.61** | **3.10** | **3.53** |
| | $F(\zeta_{ib}) + MSE$ | 1.61 | 2.05 | 2.55 | 3.08 | 3.58 |
| | $F(\zeta_{ib}) + MSE$ | **1.58** | **2.02** | **2.54** | **3.06** | **3.60** |
| | $G_{MMSE - STSA} + MSE$ | 1.63 | 2.07 | 2.57 | 3.09 | 3.59 |
| Street music | Noisy speech | 1.07 | 1.21 | 1.44 | 1.76 | 2.21 |
| | IRM + MSE | 1.36 | 1.71 | 2.12 | 2.57 | 3.04 |
| | IAM + MSE | 1.37 | 1.74 | 2.16 | 2.60 | 3.06 |
| | $|S|^0 + BCE$ | 1.44 | 1.78 | 2.17 | 2.59 | 3.01 |
| | $|S|^0 + MSE$ | **1.45** | **1.80** | **2.21** | **2.64** | **3.05** |
| | $F(\zeta_{ib}) + MSE$ | 1.41 | 1.82 | 2.26 | 2.77 | 3.26 |
| | $F(\zeta_{ib}) + MSE$ | **1.40** | **1.81** | **2.26** | **2.76** | **3.29** |
| | $G_{MMSE - STSA} + MSE$ | **1.43** | **1.83** | **2.27** | **2.75** | **3.21** |
| F16 | Noisy speech | 1.05 | 1.20 | 1.46 | 1.80 | 2.24 |
| | IRM + MSE | 1.66 | 2.03 | 2.42 | 2.86 | 3.31 |
| | IAM + MSE | 1.70 | 2.07 | 2.46 | 2.88 | 3.32 |
| | $|S|^0 + BCE$ | **1.83** | **2.19** | **2.57** | **2.98** | **3.36** |
| | $|S|^0 + MSE$ | **1.83** | **2.19** | **2.58** | **2.98** | **3.34** |
| | $F(\zeta_{ib}) + MSE$ | **1.74** | **2.12** | **2.55** | **3.06** | **3.49** |
| | $F(\zeta_{ib}) + MSE$ | **1.70** | **2.11** | **2.54** | **3.02** | **3.46** |
| | $G_{MMSE - STSA} + MSE$ | **1.72** | **2.11** | **2.51** | **2.95** | **3.39** |
| Factory | Noisy speech | 1.17 | 1.31 | 1.55 | 1.89 | 2.31 |
| | IRM + MSE | 1.53 | 1.87 | 2.25 | 2.69 | 3.12 |
| | IAM + MSE | 1.53 | 1.88 | 2.28 | 2.73 | 3.12 |
| | $|S|^0 + BCE$ | 1.61 | 1.94 | 2.31 | 2.74 | 3.09 |
| | $|S|^0 + MSE$ | **1.62** | **1.96** | **2.32** | **2.71** | **3.11** |
| | $F(\zeta_{ib}) + MSE$ | **1.59** | **1.99** | **2.40** | **2.85** | **3.19** |
| | $F(\zeta_{ib}) + MSE$ | **1.53** | **1.93** | **2.36** | **2.79** | **3.21** |
| | $G_{MMSE - STSA} + MSE$ | **1.54** | **1.94** | **2.37** | **2.84** | **3.24** |

J. Acoust. Soc. Am. 149 (5), May 2021 Aaron Nicolson and Kuldeep K. Paliwal 3289

objective intelligibility scores also results in a low WER—causing a low amount of speech distortion.

J. Choice of objective measures

When investigating a proposed speech enhancement/separation method, it is important to consider multiple objective measures. For example, attaining the highest STOI and ESTOI scores is not an indication that a method will attain the lowest WER and be best for robust ASR. Moreover, if this study was conducted using a single objective measure, for example, SDR, one would assume that the mMSA loss function is better than the MSE and BCE loss functions for the IAM—which would be an incorrect assumption in retrospect.

Another observation is that although it is clear that SDR, SI-SDR, and SegSNR measure the amount of distortion between the clean and enhanced/separated speech, it can be unclear how this translates to speech enhancement/separation performance. Do they measure overall quality? Speech distortion (which is different than the amount of distortion between the clean and enhanced/separated speech) Background noise intrusiveness? Or even intelligibility? For example, SegSNR is considered an objective measure of quality; however, it has been shown to only moderately correlate with subjective scores of overall quality, speech distortion, or even background noise intrusiveness (Hu and Loizou, 2008). Moreover, a method that attains a high CBAK score can be considered good at background noise suppression, but it is unclear what a method is good at when it attains a high SegSNR. Therefore, we stress that SDR, SI-SDR, and SegSNR should be complimented with objective measures that correlate highly with subjective scores. And when feasible, conclusions drawn using objective measures should be validated using subjective testing (this study...
TABLE XII. Mean WER for each condition of the test set from the Deep XI dataset described in Sec. V C 2. The MMSE-LSA estimator is used by the MMSE training targets. The lowest WER for each condition appears in boldface.

| Noise            | Target       | SNR |
|------------------|--------------|-----|
|                  |              | −5  | 0  | 5  | 10 | 15 |
| Voice babble     | Noisy speech | 100 | 89 | 79 | 21 | 12 | 25 |
|                  | IRM + MSE    | 91  | 60 | 43 | 32 | 12 | 15 |
|                  | IAM + MSE    | 87  | 44 | 62 | 28 | 12 | 10 |
|                  | $|S|^2$ + BCE  | 92  | 98 | 64 | 32 | 21 | 11 |
|                  | $|S|^0.5$ + MSE| 91  | 22 | 56 | 29 | 14 | 36 |
|                  | $F(\xi_m)$ + MSE| 90  | 89 | 54 | 68 | 30 | 08 |
|                  | $F(\xi_m)$ + MSE| 87  | 56 | 98 | 28 | 32 | 13 |
|                  | $G_{MMSE}$ + MSE| 92  | 80 | 73 | 28 | 12 | 11 |
| Street music     | Noisy speech | 98  | 05 | 88 | 60 | 56 | 26 |
|                  | IRM + MSE    | 83  | 57 | 52 | 24 | 14 | 14 |
|                  | IAM + MSE    | 83  | 24 | 51 | 32 | 20 | 98 |
|                  | $|S|^2$ + BCE  | 81  | 86 | 56 | 25 | 67 | 17 |
|                  | $|S|^0.5$ + MSE| 84  | 54 | 64 | 38 | 27 | 48 |
|                  | $F(\xi_m)$ + MSE| 80  | 28 | 51 | 50 | 29 | 06 |
|                  | $F(\xi_m)$ + MSE| 83  | 75 | 59 | 38 | 29 | 12 |
|                  | $G_{MMSE}$ + MSE| 79  | 31 | 55 | 83 | 26 | 33 |
| F16              | Noisy speech | 98  | 83 | 99 | 60 | 63 | 17 |
|                  | IRM + MSE    | 89  | 38 | 56 | 66 | 23 | 94 |
|                  | IAM + MSE    | 86  | 71 | 49 | 82 | 25 | 82 |
|                  | $|S|^2$ + BCE  | 84  | 49 | 60 | 11 | 30 | 34 |
|                  | $|S|^0.5$ + MSE| 88  | 95 | 57 | 60 | 36 | 72 |
|                  | $F(\xi_m)$ + MSE| 95  | 83 | 58 | 92 | 27 | 62 |
|                  | $F(\xi_m)$ + MSE| 90  | 66 | 50 | 05 | 24 | 82 |
|                  | $G_{MMSE}$ + MSE| 91  | 17 | 47 | 66 | 26 | 87 |
| Factory          | Noisy speech | 97  | 66 | 97 | 14 | 66 | 70 |
|                  | IRM + MSE    | 90  | 78 | 65 | 78 | 47 | 02 |
|                  | IAM + MSE    | 90  | 02 | 67 | 29 | 50 | 89 |
|                  | $|S|^2$ + BCE  | 93  | 78 | 63 | 14 | 53 | 06 |
|                  | $|S|^0.5$ + MSE| 86  | 30 | 69 | 24 | 56 | 98 |
|                  | $F(\xi_m)$ + MSE| 92  | 35 | 77 | 96 | 62 | 36 |
|                  | $F(\xi_m)$ + MSE| 91  | 76 | 67 | 23 | 45 | 57 |
|                  | $G_{MMSE}$ + MSE| 95  | 05 | 74 | 30 | 50 | 37 |

would be considered unsuitable due to the vast number of different methods).

VII. CONCLUSION

In this study, we compare CASA, MS, and MMSE training targets for clean speech MS estimation with the aim of determining which produces enhanced/separated speech at the highest quality and intelligibility and which is most suitable for an ASR front-end. We find that a priori SNR and joint a priori and a posteriori SNR training targets produce the highest quality enhanced/separated speech. We also find that gain training targets and the IAM not only produce the most intelligible enhanced/separated speech but are also the most suited for an ASR front-end.

One finding was that if the clean speech MS is to be used as the training target, a compression function applied to its magnitude or power values is recommended rather than its decibel values. Unless careful consideration is taken of the distribution of its decibel values, a compression function could overcompress values above 0 dB, which correspond to speech formants and a very small proportion of the overall distribution. MS training targets were also found to perform the best at lower SNR levels (−5 and 0 dB). Although it is unclear why this phenomenon occurs, we anticipate that the following may lead to an answer: Unlike the other training targets, the distribution of the clean speech MS does not change with the SNR. Another finding was that the CDF was the best compression function for a priori SNR and joint a priori and a posteriori SNR training targets. The uniform distribution that the CDF produces gives equal importance to all values during training as opposed to the truncated bimodal distribution of the gain training targets, which gives more importance to noise and speech-dominant values during training. This causes a priori SNR and joint a priori and a posteriori SNR training targets to produce the best objective quality scores and gain training targets to produce the best objective intelligibility scores.

Recommendations for future work include a comprehensive study on training targets that include phase information. This includes training targets that estimate the clean speech phase spectrum, complex spectrum, discrete-time samples, as well as training targets in the modulation domain. The phase spectrum can be estimated by using the derivatives of the phase in the time and frequency directions (Masuyama et al., 2020). The most prominent training target for clean speech complex spectrum estimation is the cIRM (Williamson et al., 2016). Aside from directly estimating the clean speech discrete-time samples, training targets for the Kalman filter and augmented Kalman filter are emerging in the literature (Roy et al., 2020a; Roy et al., 2020b). Training targets are also being explored in the modulation domain (Yan et al., 2020). Attention should also be paid to the variability of the objective scores produced by a training target over multiple training runs. Additionally, a dataset that contains more samples per condition would allow for a statistical analysis of the performance of the training targets over individual conditions. The appropriate output layer activation function for use with the CDF compression function should also be investigated—as the CDF produces a uniform distribution between zero and one [see Fig. 1(h)]. This leads to using a piecewise linear function (with the same form as the uniform CDF) as the activation function of the output layer. This may lead to an improvement in the performance over the currently used sigmoid function. And, finally, a compression function for the gain training targets and the IAM should be investigated. This is to determine how much of an impact clipping has on the speech enhancement performance of these training targets.

APPENDIX A

1. MMSE estimator comparison

In this section, we provide objective scores for multiple MMSE estimators. Often, the gain function of the MMSE
TABLE XIII. Mean objective scores of multiple MMSE estimators on the test set of the DEMAND Voice Bank dataset described in Sec. VC1. MHANet with \( F(\xi_{db}); F(\gamma_{db}) \) as the training target (using BCE as the loss function) is used to estimate the \textit{a priori} and \textit{a posteriori} SNRs. The highest score for each measure—except WER—appears in boldface. The lowest WER is in boldface. The values for the STOI, ESTOI, and WER are given as percentages.

| MMSE estimator | CSIG  | CBAK  | COVL  | PESQ  | STOI  | ESTOI | SDR   | SI-SDR | SegSNR | WER   |
|----------------|-------|-------|-------|-------|-------|-------|-------|--------|--------|-------|
| Noisy speech  | 3.50  | 2.47  | 2.73  | 1.99  | 91.53 | 78.31 | 8.68  | 8.39   | 1.71   | 30.03 |
| WF            | 3.86  | 3.45  | 3.35  | 2.83  | 93.51 | 84.42 | 20.22 | 18.51  | 9.89   | 26.59 |
| SRWF          | 4.31  | 3.41  | 3.62  | 2.89  | 94.05 | 84.96 | 18.06 | 17.27  | 8.47   | 25.03 |
| CWF           | 4.38  | 3.49  | 3.72  | 3.01  | 94.02 | 84.94 | 18.98 | 17.77  | 8.84   | 24.31 |
| MMSE-STSA     | 4.34  | 3.47  | 3.67  | 2.95  | 93.99 | 84.88 | 18.62 | 17.62  | 8.87   | 24.66 |
| MMSE-LSA      | 4.35  | 3.52  | 3.70  | 3.00  | 93.98 | 84.92 | 19.29 | 18.07  | 9.30   | 24.67 |

A single head self-attention has the added advantage of reducing the number of parameters required for each compression function, with \( \text{W}_{p} \in \mathbb{R}^{2048 \times 256} \). The addition of the learned position encoding provides an improvement in the objective scores obtained by the MHANet [the objective scores between MHANet-1.0c (MHANet with no position encoding) and MHANet-1.1c (MHANet with the learned position encoding) are available online].

The BiLSTM network of Chimera+++ from Wang et al. (2018) is used in this study to evaluate the performance of each training target on an RNN. The twin heads of Chimera+++ are replaced by a single head in this work. To indicate this difference, we denote the network as BiLSTM-Chimera+++.

The objective scores for the MHANet and BiLSTM-Chimera+++ are given in Tables V and VI, respectively.

1Note that the parameters required for each compression function are omitted for convenience. For example, \( z([S_{m}[l,k]]; \mu_{1}, \sigma_{1}) \) is simplified to \( z([S_{m}[l,k]]) \).

2Deep Xi is available at https://github.com/anicolson/DeepXi (Last viewed 4/1/2021).

3For a comparison of the ResNet-TCN used here to that used in Zhang et al. (2020), please see https://github.com/anicolson/DeepXi (Last viewed 4/1/2021).

4Freesound packs that are used include 147, 199, 247, 379, 622, 643, 1133, 1563, 1840, 2432, 4366, 4439, 4780, 8420, 14826, 15046, 15097, 15598, 16204, 17266, 17403, 17430, 17468, 17579, 19093, 20237, 20241, 21558, 22953, and 24590.

5Project DeepSpeech is available at https://github.com/mozilla/DeepSpeech (model 0.7.4 is used) (Last viewed 8/25/2020).

2. Alternate DNN architectures

The MHANet from Nicolson and Paliwal (2020a) is used in this study to evaluate the performance of each training target on an attention-based network. The MHANet uses multi-head self-attention to more efficiently model the long-term dependencies of noisy speech than TCNs and RNNs. The MHANet employs masked self-attention, ensuring that inference is causal. The MHANet used here is identical to the one used in Nicolson and Paliwal (2020a, Table Appendix A 2, configuration D), except that a learned position encoding of size 256 is added after the first layer [see “first layer” in Nicolson and Paliwal, 2020a, Fig. 2 (left)] as in Devlin et al. (2019). The time-frame index indicates the position. The position encoding is learned using weight matrix \( \text{W}_{p} \) with a maximum length of 2048 time-frames.
(POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part I—Temporal alignment,” J. Acoust. Eng. Soc. 61(6), 366–384, available at https://www.aes.org/ltd/browse.cfm?elid=16829.

Bhat, G. S., Shankar, N., Reddy, C. K. A., and Panahi, I. M. S. (2017). “Formant frequency-based speech enhancement technique to improve intelligibility for hearing aid users with smartphone as an assistive device,” in 2017 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT), pp. 32–35.

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.

Cohen, I. (2003). “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” IEEE Trans. Speech Audio Process. 11(5), 466–475.

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

Dean, D. B., Sridharan, S., Vogt, R. J., and Mason, M. W. (2010). “The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms,” in INTERSPEECH-2010, pp. 3110–3113.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (Association for Computational Linguistics, Minneapolis, MN), pp. 4171–4186, available at https://www.aclweb.org/anthology/N19-1423 (Last viewed 4/25/2021).

Ephraim, Y., and 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.

Ephraim, Y., and 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.

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., and Rubinstein, M. (2018). “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” ACM Trans. Graph. 37(4), 112.

Erdogan, H., Hershey, J. R., Watanebe, S., and Le Roux, J. (2015). “Phase-sensitive and recognition-booster speech separation using deep recurrent neural networks,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 708–712.

Hannun, A., Case, C., Casper, J., Catanaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., and Ng, A. (2014). “Deep speech: Scaling up end-to-end speech recognition,” arXiv:1412.5567 [cs.CL].

He, K., Zhang, X., Ren, S., and Sun, J. (2016). “Deep residual learning for image recognition,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Hu, G., and Wang, D. (2010). “A tandem algorithm for pitch estimation and voiced speech segregation,” IEEE Trans. Audio, Speech, Lang. Process. 18(8), 2067–2079.

Hu, Y., and Loizou, P. C. (2008). “Evaluation of objective quality measures for speech enhancement,” IEEE Trans. Audio, Speech, Lang. Process. 16(1), 229–238.

Jensen, J., and Taal, C. H. (2016). “An algorithm for predicting the intelligibility of speech masked by modulated noise masks,” IEEE/ACM Trans. Audio, Speech, Lang. Process. 24(11), 2009–2022.

Kingma, D. P., and Ba, J. (2014). “Adam: A method for stochastic optimization,” arXiv:1412.6980 [cs.LG].

Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., and Zhang, Y. (2020). “Quartznet: Deep automatic speech recognition with 1D time-channel separable convolutions,” in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6124–6128.

Le Prell, C. G., and Claver, O. H. (2017). “Effects of noise on speech recognition: Challenges for communication by service members,” Hear. Res. 349, 76–89.

Lim, J. S., and Oppenheim, A. V. (1979). “Enhancement and bandwidth compression of noisy speech,” Proc. IEEE 67(12), 1586–1604.

Loizou, P. C. (2005). “Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum,” IEEE Trans. Speech Audio Process. 13(5), 857–869.
Salamon, J., Jacoby, C., and Bello, J. P. (2014). “A dataset and taxonomy for urban sound research,” in Proceedings of the 22nd ACM International Conference on Multimedia (Association for Computing Machinery, New York), pp. 1041–1044.

Schnitzer, R., Beznjak, I., Wagner, H., and Ney, H. (2007). “Gammatone features and feature combination for large vocabulary speech recognition,” in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 4, pp. IV-649–IV-652.

Snyder, D., Chen, G., and Povey, D. (2015). “MUSAN: A music, speech, and noise corpus,” arXiv:1510.08484 [cs.SD].

Srinivasan, S., Roman, N., and Wang, D. (2006). “Binary and ratio time-frequency masks for robust speech recognition,” Speech Commun. 48(11), 1486–1501.

Steeneken, H. J., and Geurtsen, F. W. (1998). “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., and Jensen, J. (2011). “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Trans. Audio, Speech, Lang. Process. 19(7), 2125–2136.

Thiemann, J., Ito, N., and Vincent, E. (2013). “The diverse environments multi-channel acoustic noise database (DEMAND): A database of multi-channel environmental noise recordings,” Proc. Migs. Acoust. 19(1), 03081.

Thomas-Stonell, N., Kotler, A.-L., Leeper, H., and Doyle, P. (1999). “Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition accuracy,” Augmentative Altern. Commun. 14(1), 51–56.

Valentin-Botinhao, C., Wang, X., Takaki, S., and Yamagishi, J. (2016). “Investigating RNN-based speech enhancement methods for noise-robust text-to-speech,” in 9th ISCA Speech Synthesis Workshop, pp. 146–152.

Vary, P., and Martin, R. (2009). Digital Speech Transmission: Enhancement, Coding and Error Concealment (Wiley, Hoboken, NJ).

Veaux, C., Yamagishi, J., and King, S. (2013). “The voice bank corpus: Design, collection and data analysis of a large regional accent speech database,” in 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), pp. 1–4.

Veaux, C., Yamagishi, J., and MacDonald, K. (2017). “CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit,” Technical report (The University of Edinburgh).

Vincent, E., Griboval, R., and Favrotte, C. (2006). “Performance measurement in blind audio source separation,” IEEE Trans. Audio, Speech, Lang. Process. 14(4), 1462–1469.

Wang, D. (2005). “On ideal binary mask as the computational goal of auditory scene analysis,” in Speech Separation by Humans and Machines, edited by P. Divenyi (Springer US, Boston, MA), pp. 181–197.

Wang, D., and Brown, G. J. (2006). Computational Auditory Scene Analysis: Principles, Algorithms, and Applications (Wiley-IEEE, Hoboken, NJ).

Wang, Y., Mohamed, A., Le, D., Liu, C., Xiao, A., Mahadeokar, J., Huang, H., Tjandra, A., Zhang, X., Zhang, F., Fuegen, C., Zweig, G., and Seltzer, M. L. (2020). “Transformer-based acoustic modeling for hybrid speech recognition,” in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6874–6878.

Wang, Y., Narayanan, A., and Wang, D. (2014). “On training targets for supervised speech separation,” IEEE/ACM Trans. Audio, Speech, Lang. Process. 22(12), 1849–1858.

Wang, Y., and Wang, D. (2013). “Towards scaling up classification-based speech separation,” IEEE Trans. Audio, Speech, Lang. Process. 21(7), 1381–1390.

Wang, Z., Roux, J. L., and Hershey, J. R. (2018). “Alternative objective functions for deep clustering,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 686–690.