Two-stage deep learning approach for speech enhancement and reconstruction in the frequency and time domains

Soha A. Nossier
Dept. of Engineering and Computing
University of East London
London, UK
soha.abdallah.nossier@gmail.com

Julie Wall
Dept. of Engineering and Computing
University of East London
London, UK
j.wall@uel.ac.uk

Mansour Moniri
Dept. of Engineering and Computing
University of East London
London, UK
m.moniri@uel.ac.uk

Cornelius Glackin
Intelligent Voice Ltd
London, UK
neil.glackin@intelligentvoice.com

Nigel Cannings
Intelligent Voice Ltd
London, UK
nigel.cannings@intelligentvoice.com

***Abstract*—Deep learning has recently shown promising improvement in the speech enhancement field, due to its effectiveness in eliminating noise. However, a drawback of the denoising process is the introduction of speech distortion, which negatively affects speech quality and intelligibility. In this work, we propose a deep convolutional denoising autoencoder-based speech enhancement network that is designed to have an encoder deeper than the decoder, to improve performance and decrease complexity. Furthermore, we present a two-stage learning approach, in which denoising is performed in the first frequency domain stage using magnitude spectrum as a training target; while, in the second stage, further denoising and speech reconstruction are performed in the time domain. Results show that our architecture achieves 0.22 improvement in the overall predicted mean opinion score (Covl) over state of the art speech enhancement architectures, using the Valentini dataset benchmark. Moreover, the architecture was trained using a larger dataset and tested using a mismatched test corpus, to achieve 0.7 and 6.35% improvement in Perceptual Evaluation of Speech Quality (PESQ) and Short Time Objective Intelligibility (STOI) scores, respectively, compared to the noisy speech.**

***Index Terms*—Deep learning, denoising autoencoders, speech enhancement, speech features, speech reconstruction**

I. Introduction

Speech enhancement is a signal processing technique, which aims to improve speech quality and intelligibility by removing background noise. Applications of speech enhancement include mobile communication systems, hearing aids, and Automatic Speech Recognition (ASR). Classical speech enhancement techniques were all based on statistical assumptions to model the relationship between speech and noise. These techniques include Spectral Subtraction [1], Wiener Filter [2], Signal Subspace [3], and Minimum Mean Square Error (MMSE) estimator [4]. Although some of these techniques managed to partially mitigate the background noise [5], as they are based on statistical assumptions, they fail to generalize for intrusive and non-stationary noise types [6].

With the recent massive increase of data, deep learning has made a breakthrough in this denoising process, because of its ability to remove most of the background noise, regardless of its type and intensity. In this approach, a Deep Neural Network (DNN) is trained in a supervised learning fashion to map from noisy to clean speech, without any statistical assumptions of the relationship between the speech and noise [7]. In deep learning-based supervised speech enhancement, a DNN is trained using pairs of clean and noisy speech signals to minimize a defined loss function, and it finally predicts the clean speech signal [8].

The input signal representation is an important factor that impacts network learning and generalization. For better feature extraction, many speech enhancement DNNs operate in the frequency domain [9]–[11], where a time-frequency (T-F) representation of the noisy speech is used, to estimate a mapping target; mapping directly to a clean speech T-F representation, or a masking target; a mask that classifies every portion of the spectrum as either speech or noise and it is multiplied by the noisy speech to generate the clean speech [12].

After the introduction of Convolutional Neural Networks (CNNs) in audio processing, learning in the time domain becomes more common, because it shows promising performance [13]–[15]. Recent research in the field practically chooses between time and frequency domain learning based on the performance of the proposed DNN in each domain [16], and the deep Convolutional Denoising Autoencoder (CDAE) is one the best-performing DNNs for both frequency and time domain-based speech enhancement [9], [13], [15], [17].

This research is sponsored by University of East London and Intelligent Voice Ltd.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823907 (MENHIR project https://menhir-project.eu)

The following subsection presents the work in the literature that is related to our work, and discusses the research gap and the contribution of this work.

A. Relation to Prior Work

Deep learning-based speech enhancement has shown a massive progression over the last decade. Many DNNs have been proposed for speech enhancement, starting from the simple architectures with few hidden layers [18] to the latest more complex and deeper networks [9], [19].

The T-F representation was previously used in most DNNs for speech enhancement. The one dimensional (1D) time-domain noisy speech signal is converted to the two dimensional (2D) time-frequency form using Short-Time Fourier Transom (STFT) analysis, for the DNN to process only the magnitude spectrogram. The output from the DNN is then converted back to the time-domain representation using Inverse STFT (ISTFT), and the noisy phase was used assuming the phase is not highly affected by the noise [20]. However, in high noise environments, the noise affecting the phase becomes more significant and negatively affects the performance. This leads to the presentation of techniques that process both the magnitude and phase using complex spectral mapping [21] or complex masking approaches [22]. Furthermore, the introduction of CNNs for audio processing opens the door for time-domain waveform-based speech enhancement processing, which is based on a Fully Convolution Neural Network (FCNN) [23]. In this approach, 1D convolution is performed on the time-domain signal, in which both magnitude and phase denoising are considered. Moreover, CDAEs have recently shown promising performance for time domain-based speech enhancement [13], [15], [17].

In DNN based speech enhancement, speech distortion is the main drawback of the speech denoising process, especially at low Signal to Noise Ratio (SNR) levels, in which the DNN removes part of the speech spectrum while trying to remove the background noise. The significance of this issue appears when making a subjective test, where some of the listeners prefer the noisy speech version rather than the clean one because of the distortion, which mainly affects speech intelligibility [10]. Many of the proposed DNNs for speech enhancement are very effective in improving the quality of noisy speech; however, it is still very challenging to avoid the distortion that accompanies the noise removal process [24], [25]. The generalization ability of DNNs is another issue that becomes more significant when testing the network using a mismatched test corpus [26], and poor network generalization also causes speech distortion.

Consequently, recent research is giving more attention to this distortion issue and proposing different techniques and approaches to overcome it. Two approaches have been recently proposed, which are effective in dealing with distortion, both of which are based on a two-stage speech enhancement architecture. The first approach involves two different DNNs, where the first stage network performs the denoising process and the second stage network minimizes speech distortion [25]. The second approach is to process the noisy speech in two stages with the same DNN, but using different features. The work in [19] is based on this approach, as the authors used time then frequency domain cascaded approach. Both magnitude and phase denoising are performed in the first time-domain stage and then further magnitude denoising is applied in the second frequency-domain stage. This is an interesting approach, however, the order by which the two stages should be cascaded is not considered in [19].

In the work presented in this paper, we demonstrate that processing the noisy speech in the frequency domain followed by time-domain processing outperforms the time then frequency-domain approach presented in [19]. Moreover, we show that feeding the second stage with the denoised speech from the first stage together with the original noisy speech leads to further improvement.

In this work, we propose a new asymmetric CDAE-based architecture for speech enhancement, in which the encoder is designed to be deeper than the decoder. This will take advantage of deep architectures to improve the performance but with reduced complexity. Additionally, we present a new approach to deal with distortion, where the denoising and reconstruction processes are performed separately by training the architecture in a two-stage scheme, first in the frequency and then in the time domain. The proposed architecture uses the denoised, distorted speech estimated by the first stage together with the original noisy speech as an input to the second stage, which is supposed to focus on speech reconstruction, to achieve the best possible performance in terms of speech quality, intelligibility and distortion. Moreover, the proposed architecture was trained using a very large dataset, 1,000 hours, containing a range of different accents and languages, to improve generalization. It was then tested using a mismatched speech corpus that was not used in the training process, corrupted with mismatched noise environments, to fairly assess its generalization ability.

This work makes the following contributions:

- Develops a new asymmetric CDAE based speech enhancement architecture with better performance and less complexity.
- Proposes a two-stage deep learning speech enhancement approach that compromises between speech denoising and reconstruction, and outperforms State Of The Art (SOTA) speech enhancement models.

The rest of this paper is organized as follows. In Section II, the proposed CDAE-based network is explained. Section III presents the two-stage speech enhancement approach. The relevant datasets and experimental setup are described in Section IV. Section V shows the obtained results. Finally, the conclusion is provided in Section VI.

II. The Developed speech enhancement network

The implemented architecture, shown in Figure 1, is a fully 1D CDAE-based DNN. The network accepts an input of size 2,048, and compression is applied through the encoder network until the input reaches a bottleneck layer of size 8. Afterwards, decompression is performed by the decoder

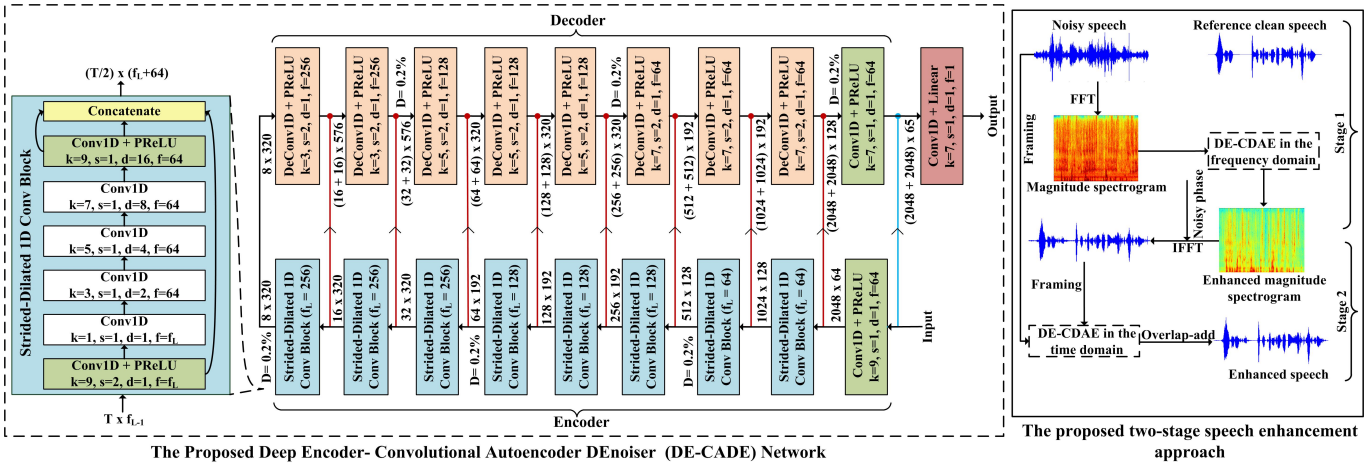

Fig. 1: The proposed two-stage Deep Encoder - Convolutional Autoencoder DEnoiser (DE-CADE) speech enhancement architecture; *k*, *d*, *f*, and *L* represent kernel size, dilation rate, number of convolution channels and layer number respectively; *s* represents stride size in the encoder, and upsampling size in the decoder. *T* is the time samples. The red lines represent skip connections and the blue line shows the shortcut between the input and output.

network to restore the signal. Because the network is very deep, skip connections are added between the encoder and decoder to avoid information loss [23]. These connections pass the information learned by the processing layers of the encoder to the decoder; skip connections are represented by the red lines in Figure 1. We concatenated the unprocessed noisy input with the input to the last layer of the encoder because this was found to decrease distortion and leads to better network generalization; this connection is represented by the blue line in Figure 1.

The encoder combines two techniques: strided and dilated causal convolution, which is proven to improve the overall performance [13]. We used several strided-dilated causal convolution blocks, shown in Figure 1, which enhance the denoising process by compressing the input. Each block has a 1D strided convolution layer of stride size 2, kernel size of 9, and Parametric Rectified Linear Unit (PReLU) activation. This layer is then followed by 5 dilated 1D causal convolution layers of increasing dilation rates and a final PReLU activation. This allows exponential expansion of the receptive field, to decrease distortion without increasing the network's complexity [27]. We also used increasing kernel sizes as the dilation rate increases to decrease sparsity. The strided-dilated convolution block ends with a concatenation layer to combine both the fine and coarse features extracted by these techniques.

The decoder consists of 1D deconvolution layers of upsampling size 2 and PReLU activation. The input to each layer is a concatenation of the output of the previous layer and the output of the corresponding concatenation layer in the encoder network, received by the skip connections. The final layer is a convolution layer with linear activation and a kernel size of 7, and the original noisy input is concatenated with the input to this layer. Further details of the network's hyperparameters are detailed in Figure 1.

Repeating the strided-dilated causal convolution blocks in the decoder does not show significant improvement in the performance, due to the use of skip connections, which send the information gained by the encoder to the decoder. Consequently, the encoder is designed in our architecture to be deeper than the decoder, to decrease network complexity and processing time. Hence, we use the name Deep Encoder - Convolutional Autoencoder DEnoiser (DE-CADE) for this architecture throughout the rest of the paper. The encoder has 74 layers, making a total of 4.2 million parameters; while the decoder has 36 layers, making a total of 2.1 million parameters.

III. The Proposed Speech Enhancement Approach

A. Problem Definition

The noisy speech signal can be represented as follows:

$$y(m) = s(m) + n(m), \tag{1}$$

where *y* represents the noisy speech, *s* and *n* are the speech and additive noise signals, respectively, and $\{y,\ s,\ n\} \in \mathrm{R}^{M \times 1}$, where M is the total number of samples in the signals, and m is the time sample index. If we also considered reverberate noise affecting the speech signal, this equation can be redefined as follows:

$$y(m) = x(m) + n(m), \tag{2}$$

where,

$$x(m) = s(m) * r(m) = \sum_{j=0}^{M-1} r[j]s[m-j], \tag{3}$$

where $*$ denotes the convolution operator, *x* denotes the reverberant speech, *r* represents the Room Impulse Response (RIR), *j* is the discrete RIR sample, and *m* is sample point of the discrete signals, *x* and *s*. Because reverberation is a special type of noise affecting the speech signal, and that recently dereverberation is applied as a separate stage to

improve performance [28], we trained the speech enhancement network to only suppress additive noise, without applying dereverberation, and to map noisy, reverberant speech to the reverberant target speech, which is proven to improve speech intelligibility [29].

A main requirement of the denoising procedure is to have a good estimate of the relationship between speech and additive noise, to be able to predict the clean speech signal. In deep learning-based speech enhancement, the noisy speech is fed to a DNN that performs some linear and non-linear functions to generate an estimate of the clean speech. Based on the fact that most recent DNNs in the literature managed to generate a good prediction of the clean speech, we hypothesize that feeding the output of the DNN, $\hat{x}$, with the original noisy speech, y, to another second stage DNN will result in a better learning process and prediction for the second stage DNN.

For the two stages to perform differently, we need to either implement two different DNNs for each stage, or apply two different approaches using the same architecture. In this work, we applied the latter idea by using a first stage DNN that operates in the frequency domain to estimate the magnitude spectrogram of the clean speech. The output from this stage is then fed to a second stage DNN running in the time domain to perform both magnitude and phase denoising. This will allow a different estimation of the clean speech using the time domain representation.

For the first stage network, time-frequency features were extracted from the noisy speech by applying STFT, which can be calculated as described below:

$$Y(t,f) = \sum_{f=0}^{F-1} y(m+t)h(m)e^{-j2\pi fm/F}, \tag{4}$$

where *Y(t,f)* is the STFT of the noisy signal, f is the frequency bin index; $\{f = 0, 1, ..., F\text{-}1\}$ and F is the total number of frequency bins, t is the time frame, $\{t = 0, 1, ..., T\text{-}1\}$ and T is the total number of frames, m is the input signal time sample, h denotes the applied window function, which is a Hamming window in our implementation. The time frame size was set to 256 with 50% overlap. After applying STFT and taking the magnitude of the signal to obtain the spectrograms, the frequency domain representation of Equation (2) can be expressed as:

$$|Y(t,f)| = |X(t,f)| + |N(t,f)|, \tag{5}$$

where, $|Y(t,f)|$, $|N(t,f)|$, and $|X(t,f)|$ are the magnitude spectrograms of the noisy speech, noise and speech signals, respectively. We then trained the proposed network, described in Section II, in the frequency domain, DE-CADE(F), using the clean speech magnitude spectrogram, $|X(t,f)|$, as a training target, because masking targets fail to generalize for CDAEs [30]. The noisy phase was stored to be added to the final estimated clean speech, assuming that the phase component is not highly affected by noise, compared to the magnitude [20]. When processing the noisy speech by the DE-CADE(F), every layer of the network will apply a 1D dilated causal convolution operation [31], which can be expressed as follows:

$$B(u,v) = \sum_{c} \sum_{w+d*q=v} A(c,w) * weight(u,c,q), \tag{6}$$

where, *B(u,v)* is the output of the 1D dilated causal convolution layer, *A(c,w)* is the layer input, *weight(u,c,q)* is the filter applied to the input, u is the number of applied convolution channels, v is the output width, c is the number of input channels, w is the input width, q is the filter width and d is the dilation rate.

Each convolution layer is followed by a nonlinear function, PReLU in our case, so the output, G, from the non-linearity layer will be:

$$G(u,v) = PReLU(B(u,v)), \tag{7}$$

where,

$$PReLU(B(u,v)) = \begin{cases} \text{B(u,v)}, & \text{if } B > 0, \\ \alpha \text{ B(u,v)}, & \text{otherwise}, \end{cases} \tag{8}$$

where α is a variable parameter that changes based on the model during training. Mean Square Error (MSE) is the loss function used with the Adam optimizer, learning rate = 0.0001, β_1 = 0.1, β_2 = 0.999. The DE-CADE(F) will minimize the frequency domain MSE loss, given below, to estimate the speech magnitude spectrum.

$$L_F = \frac{1}{TM} \sum_{t=0}^{T} \sum_{f=0}^{F} \left[|\hat{X}(t,f)| - |X(t,f)| \right]^2, \tag{9}$$

where L_F is the loss function for the frequency network, DE-CADE(F), T is the total number of frames, and F is the number of frequency bins. After processing the noisy speech using several convolution and non-linearity functions, the estimated clean speech STFT, $\hat{X}$*(t,f)*, can be reconstructed using the estimated speech magnitude spectrogram, $|\hat{X}(t,f)|$, and the STFT phase of the noisy speech, $\angle$*Y(t,f)*. This can be expressed as follows:

$$\hat{X}(t,f) = \sqrt{|\hat{X}(t,f)|} \otimes e^{j\angle Y(t,f)}, \tag{10}$$

where $\otimes$ denotes element-wise multiplication. Finally, the time domain estimated speech signal from the first stage, $\hat{x}$, can be generated using the ISTFT.

$$\hat{x}_1(m) = ISTFT(\hat{X}(t,f)). \tag{11}$$

For the second stage, both the noisy speech, y, and the estimated clean speech by the first stage, $\hat{x}_1$, are concatenated on two different channels, and then fed to a similar second stage network but operating in the time domain, DE-CADE(F-T). Framing is the only preprocessing operation applied to the inputs using a frame size of 2,048 with 50% overlap. The input concatenated time frames, y_2*(t)*, to the second stage network, DE-CADE(F-T), can be represented as follows:

$$y_2(t) = (y(t), \hat{x}_1(t)), \tag{12}$$

where, t is the time frame, *y(t)* and $\hat{x}_1$*(t)* are the framed noisy

and estimated speech, respectively. The network here will try to enhance both the magnitude and phase, given the time-domain representation of the noisy speech and the denoised speech from the first stage. This will allow different learning and enhancement processes from that of the first stage. MSE is the loss function used for the second stage, as an optimum choice to reduce the time domain prediction error [32]. This can be expressed as given below.

$$L_T = \frac{1}{T}\sum_{t=0}^{T}\left[\hat{x}_2(t) - x(t)\right]^2, \tag{13}$$

where L_T is the loss function of the second enhancement stage and $\hat{x}_2(t)$ is the estimated clean speech frame from the second stage. We finally apply overlap-add procedure to obtain the final estimated clean speech, $\hat{x}_2(m)$.

B. Frequency versus Time Domain Learning

The reason for first applying speech enhancement in the frequency domain is to achieve a better denoising process, as the network will focus only on enhancing the magnitude spectrogram. Phase denoising is considered in the second stage, where the denoising process becomes less challenging to the network when adding the estimated clean speech of the first stage. The degradation in the denoising ability of the network when trying to enhance both magnitude and phase components can be demonstrated by comparing the output of the frequency network, DE-CADE(F), with the output of the DE-CADE when trained as a single stage in the time domain only, DE-CADE(T).

Figure 2 shows the evaluation of the quality of the output speech using the Cbak score [33], which evaluates the quality of speech based on background noise intrusiveness (the higher the score the better the quality), and the Log Spectral Distortion (LSD, in dB) [34], which measures speech distortion; low value indicates less distortion, and it can be calculated as follows:

$$LSD = \frac{1}{T}\sum_{t=0}^{T-1}\left\{\frac{1}{F/2+1}\sum_{f=0}^{F/2}\left[10log\frac{P(X(f))}{P(\hat{X}(f))}\right]^2\right\}^{\frac{1}{2}}, \tag{14}$$

where, *P* is the clipped power spectrum such that the dynamic range of the log-spectrum is limited to about 50 dB. The function *P* for a signal z can be expressed as:

$$P(z(f)) = max\left[|z(f)|^2, 10^{-50/10}(|z(f)|^2)\right]. \tag{15}$$

This evaluation is based on testing the network using 200 speech audios from a mismatched test corpus, the Librispeech corpus [35], corrupted with five mismatched noise environments: Babble, Factory, Engine, HF radio channel and Operating Room; taken from the the NOISEX-92 dataset [36] and not seen during training. Details of the experimental setup is presented in Section IV.

In Figure 2, it is clear from the Cbak results that the denoising ability of the frequency network is much better, except at a very low SNR, -5 dB, where the denoising of frequency and time networks are approximately the same. In this case, the effect of the noisy phase becomes more significant and negatively affects the performance of the frequency network. On the other hand, the LSD results show less distortion for the time network, especially at low SNRs; -5, 0, 5 dB, where aggressive noise removal of the frequency network results in high distortion.

The trade-off between speech denoising and reconstruction can also be justified using spectrograms, shown in Figure 3, which represent clean and noisy speech at three SNRs: -5 dB crowd noise, 0 dB tooth brushing noise and 5 dB shower noise, and their corresponding estimated output from the frequency and time domain single-stage DE-CADE. It is clear that at all SNR levels, the frequency domain network can effectively remove background noise. However, the output speech experiences high distortion due to the spectrum representation, which gives more attention to the fundamental frequencies when reconstructing the estimated speech. On the other hand, the time domain network shows less denoising ability, but with better speech reconstruction, especially for the high-frequency components.

Consequently, by applying the two-stage frequency and time training scheme, the frequency domain stage can be considered as a denoising stage, in which aggressive noise removal is performed; while the second time domain stage is a reconstruction stage that is fed an estimate of the clean speech as additional information together with the noisy speech, to apply both magnitude and phase denoising while compromising between noise removal and speech distortion.

IV. Experimental Setup

A. Verification Dataset

To compare with the SOTA speech enhancement models, we used the noisy speech from the Valentini training and test dataset [37]. The dataset is a subset of the Voice Bank corpus [38], it has a total of 30 speakers, 28 for training and 2 for testing. The speakers are native English, reading about 400 English sentences. The noisy speech training set was created by mixing speech audios with 10 noise environments: 8 from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) dataset and two artificial noise, at four SNRs: 0, 5, 10 and 15dB, to make 11,572 training samples. While the noisy test set was created by corrupting the test speech audios with 5 unseen noise environments from the DEMAND dataset, to make 824 test samples. We used 10% of the noisy training data for validation and trained the two-stage DE-CADE architecture model for 100 epochs in each domain. Afterwards, we evaluated the architecture using the Valentini test data and compared the performance with the reported results of other SOTA networks. This evaluation is presented in Table III.

B. Large Scale Dataset

In this experiment, we trained the architecture using a very large noisy speech dataset of 1,000 hours. The clean speech data includes 800 hours of English speech, and 200

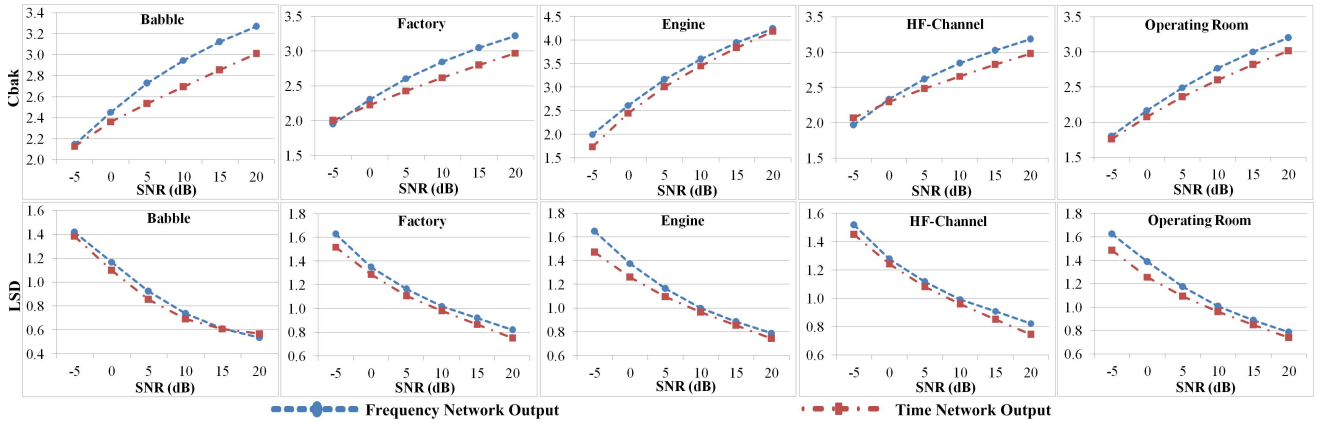

Fig. 2: The Cbak and LSD results for the proposed network, DE-CADE, when operating as a single stage in the frequency and time domain, tested on mismatched babble, factory, engine, HF-channel, and operating room noises.

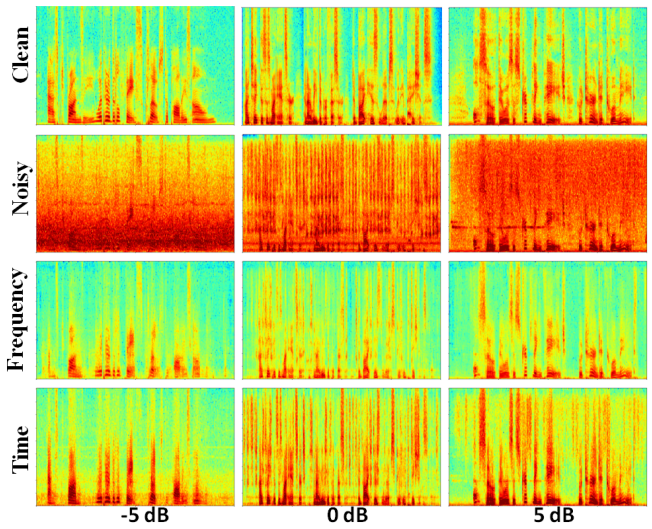

Fig. 3: The spectrograms of the clean, noisy, and estimated speech from the single-stage frequency and time domain DE-CADE at -5 dB crowd noise, 0 dB tooth brushing noise and 5 dB shower noise.

hours of an additional 175 languages [39]. The 800 hours of English speech was collected from the Microsoft Deep Noise Suppression (DNS) challenge dataset [40], the CSTR VCTK Corpus [41] and the Reverberant speech dataset [42]. These make a total of 267,841 different speech utterances. The noise environments were taken from the DNS noise dataset, which is about 181 hours of noise data, and makes a total of 60,000 different noise clips [40]. We first randomly selected 10% of the speech and noise data to create the validation dataset, and then we used the rest of the data for training. To create noisy speech for both training and validation, the speech utterances were randomly mixed with the noise environments at a wide range of SNRs from -5 to 15 with a step of 1.

For testing, we randomly selected 200 speech audio files of 20 male and 20 female speakers from the Librispeech corpus [35] and corrupted them with 20 unseen noise environments from 100 Nonspeech Environmental Sounds [43]: 9 crowd noise, an Additive White Gaussian Noise (AWGN), 2 human yawn noises, human cry, shower, tooth brushing, 2 footsteps, door moving, and 2 phone dialling. We mixed the test speech and noise at six SNRs from -5 dB to 20 dB with a step of 5 dB. The results presented in Section V are based on the average of these six SNRs, and we will use the term mismatched noise environments for this test set. We also mixed these speech files with unseen Babble, Factory, Engine, HF radio channel and Operating Room noises from the NOISEX-92 dataset [36], to perform the analysis presented in Figure 2 and discussed in Section III.

To assess the network's generalization, the architecture was also tested using a matched test data that is seen during the training process. This data was used to compare the network's performance for seen and unseen noisy speech, to evaluate the network's generalization ability. We used 200 speech audios from the DNS dataset, seen in the training, and of similar length as the mismatched Librispeech speech audios. These audios were corrupted with 20 seen noise environments, randomly selected from the training DNS noise dataset. These noises include: church bell, sweeping sound, motorcycle, train, music, cry, water, crowd, wind, sea waves, siren, hummer, kitchen machine, piano, and birds. We mixed them at the same 6 SNRs of the mismatched test set, to obtain similar conditions.

C. Training Hyperparameters

Regarding network hyperparameters, the chosen values and approaches are based on best practices. Training is based on a 16 kHz sampling frequency and a wide range of SNRs from -5 to 15 with a step of 1. The input is normalized to zero mean and unit variance. In the frequency domain, the training is based on magnitude spectral mapping, with STFT of frame size 256 and 50% overlap, and the noisy phase was added to the output spectrogram before transforming back to the time domain using ISTFT. In the time domain, framing was performed with frame size 2,048 and 50% overlap, and

the traditional overlap-add method was applied to the output frames.

In both stages, MSE is the loss function used with the Adam optimizer, learning rate = 0.0001, β_1 = 0.1, β_2 = 0.999. We used a batch size = 2. For the first stage, the network was trained till convergence for 18 epochs in the case of large scale training, and for 100 epochs for the verification experiment using the Valentini dataset. For the second stage, the network was trained for 50 epochs in the case of large scale training, and for 100 epochs when using the small scale Valentini dataset. Training and validation curves of the second stage for both verification and large scale training are presented in Figures 4 and 5. It is clear that in the case of large scale training using 1,000 hours of data, the network converges quickly, approximately at the 30th epoch. By the 30th epoch, the network was exposed to 30,000 hours of data, which is enough for convergence. The decrease afterwards in the validation loss is not significant. On the other hand, the network takes longer to converge in the case of the small scale Valentini dataset. Approximately at the 50th epoch, the validation curve saturates, and the network starts to overfit the training data. The best network's weights were taken based on the validation data in both experiments, to avoid overfitting.

V. Results and Discussion

To evaluate the performance, we used the well-known speech enhancement metrics, described below:

- Perceptual Evaluation of Speech Quality (PESQ) [44] score to assess speech quality (from -0.5 to 4.5); the higher the score, the better the speech quality.
- Short-Time Objective Intelligibility (STOI) [45] score to evaluate speech intelligibility (from 0 to 1, presented in %); the higher the score, the better the speech intelligibility.
- Log Spectral Distortion (LSD) [34] to measure speech distortion; low value indicates low distortion.
- Csig [33], Mean Opinion Score (MOS) prediction of the signal distortion, considering speech signal only (from 0 to 5); high value indicates low speech distortion.
- Cbak [33], MOS prediction of background noise intrusiveness (from 0 to 5); high value indicates less background noise.
- Covl [33], MOS prediction of the overall quality of the enhanced speech (from 0 to 5); high value indicates better overall speech quality.

A. Architecture Performance

Figure 6 shows a comparison between the performance of the proposed DE-CADE network, as a single stage in the frequency and time domain, and the two-stage frequency then time approach. The higher denoising ability of the frequency network is clear when looking at the Cbak scores. Conversely, although the time domain network shows the least denoising ability, it generates speech with better intelligibility and lower distortion, especially for babble noise, which is similar to speech, and this leads to high distortion in the case of the

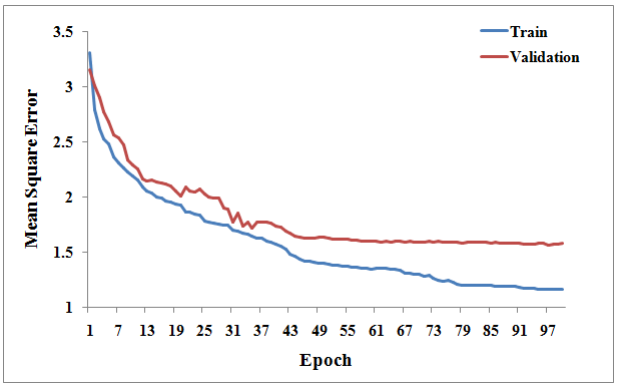

Fig. 4: The training and validation curves for small scale training with the Valentini dataset benchmark.

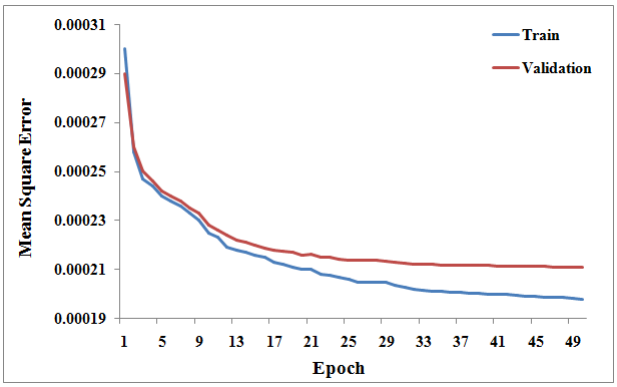

Fig. 5: The training and validation curves for large scale training with 1,000 hours of noisy speech.

frequency network. The compromise between speech denoising and reconstruction is achieved by the two-stage approach, as the network sometimes increases the noise level, compared to the frequency domain stage, as shown in the Cbak graphs, which leads to better performance for all the other evaluation metrics.

Table I shows the comparison between the performance of the network for matched and mismatched conditions, described in Section IV. The difference between PESQ and STOI results is acceptable, considering the fact that the matched data is seen during training, and the mismatched data is highly challenging and considers 40 people and 20 noise environments, all unseen during the training process. Consequently, the network experiences low variance and good generalization to highly mismatched data. Moreover, the ability of the proposed approach in avoiding distortion is proven by the Cbak score, which is lower for matched data, due to the reconstruction process of the second stage, which negatively affects the denoising process, to achieve overall better speech quality and intelligibility.

B. Comparison to Cascaded Approach

Several experiments were conducted using the DE-CADE architecture, described in Section II, to compare the proposed

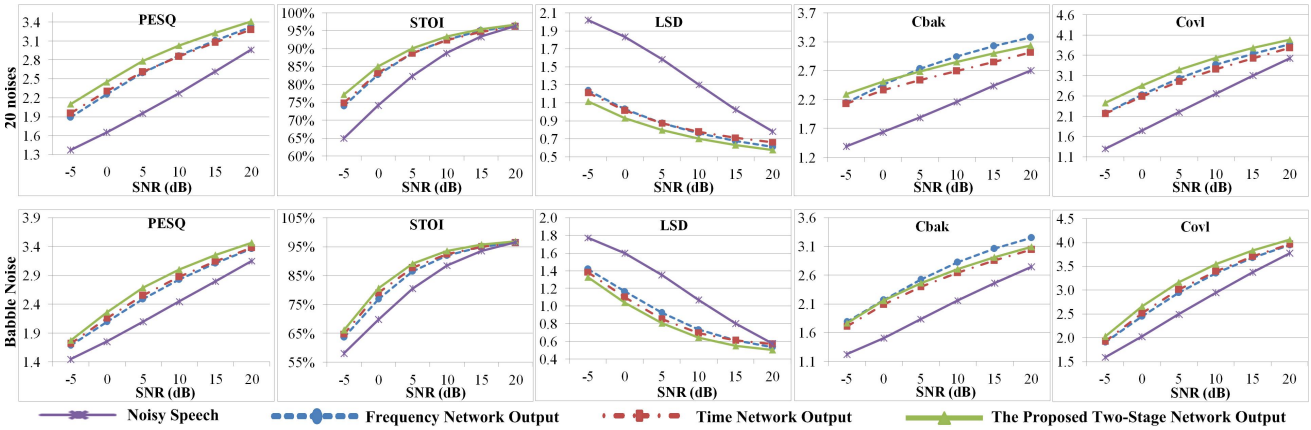

Fig. 6: The PESQ, STOI, LSD, Cbak, and Covl of the proposed DE-CADE architecture, trained in the frequency domain, blue line; time domain, red line; the proposed two-stage approach in frequency then time domain, green line; and the reference noisy speech, purple line, for the 20 mismatched noise environments in Figure 5 and babble noise.

two-stage frequency then time approach to other cascaded approaches. In all approaches, the estimated output of the first stage, $\hat{x}$, is fed to the second stage; while, in the proposed approach, both the noisy speech, y, and the estimation of the first stage $\hat{x}$ are concatenated and fed to the second stage. Table II shows this comparison, and the description of each approach is defined below:

- T(y)-T($\hat{x}$): two-stage DE-CADE, in which the first and second stages are operating in the time domain.
- F(y)-F($\hat{x}$): two-stage DE-CADE, in which the first and second stages are operating in the frequency domain.
- T(y)-F($\hat{x}$): two-stage DE-CADE, in which the first stage is operating in the time domain and the second stage in the frequency domain.
- F(y)-T($\hat{x}$): two-stage DE-CADE, in which the first stage is operating in the frequency domain and the second stage in the time domain.
- F(y)-T($\hat{x}$,y): the proposed two-stage DE-CADE with first frequency domain stage and second time domain stage, and the noisy speech is taken through to the second stage along with the output of the first stage.

The evaluations show that the proposed approach, (F(y)-T($\hat{x}$,y)), outperforms other cascaded approaches for all evaluation metrics, except the Cbak results that measure the denoising ability. The cascaded frequency-frequency approach, (F(y)-F($\hat{x}$)), shows the best noise removal performance; however, all the other evaluation metrics are negatively affected. This is more evidence that the frequency network has better denoising ability, as discussed in Section III.

C. Baselines Comparison

1) Comparing with SOTA Models: The comparison with SOTA speech enhancement models is presented in Table III, using the three predictions of the MOS score, Csig, Cbak, and Covl, reported in the literature. For comparison, we used the classical Wiener filter approach [2], and the SOTA DNN-based speech enhancement architectures: SEGAN [46], Wave U-Net [47], WaveNet [23], MMSE-GAN [48], Deep Feature Loss [49], Deep Xi-ResLSTM [50], Metric-GAN [51], SEGAN-D [52], DEMUCS [17], Koizumi et al. [53], T-GSA [54], and Deep MMSE [55].

TABLE I: The performance of the proposed two-stage network for matched and mismatched test data. The results are averaged over 6 SNRs, from -5 to 20 with 5 dB step.

Metric	PESQ	STOI	LSD	Csig	Cbak	Covl
matched	**2.848**	**91.44**	**0.773**	**3.865**	2.635	**3.341**
mismatched	2.782	89.64	0.791	3.862	**2.744**	3.305

TABLE II: Performance comparison of the proposed two-stage approach to the cascaded approach, using the mismatched test data. The results are averaged over 6 SNRs, from -5 to 20 with 5 dB step.

Metric	PESQ	STOI	LSD	Csig	Cbak	Covl
Noisy	2.086	83.28	1.422	2.856	2.037	2.421
T(y)-T($\hat{x}$)	2.591	88.01	0.981	3.566	2.581	3.050
F(y)-F($\hat{x}$)	2.609	87.52	0.892	3.557	**2.794**	3.066
T(y)-F($\hat{x}$)	2.643	87.63	0.906	3.580	2.785	3.094
F(y)-T($\hat{x}$)	2.680	87.92	0.884	3.712	2.665	3.176
F(y)-T($\hat{x}$,y)	**2.797**	**89.69**	**0.792**	**3.865**	2.762	**3.315**

The models are ordered based on the overall predicted MOS score, Covl. Our two-stage architecture, DE-CADE(F-T), outperforms all the STOA models in terms of speech signal quality, Csig, and the overall predicted MOS score, Covl. Moreover, the first-stage frequency domain network, DE-CADE(F) achieves better performance in comparison to most of the models. Other models show better denoising ability; however, the overall performance is negatively affected due to the speech distortion issue, which our architecture is designed to solve.

2) Large Scale Training: Our first stage frequency network, DE-CADE(F), and the two-stage architecture, DE-CADE, after 18 and 50 training epochs, DE-CADE(18^{th}) and DE-

TABLE III: Performance comparison of SOTA speech enhancement models using the Valentini Voice Bank dataset benchmark [37].

Metric	Csig	Cbak	Covl
Noisy	3.35	2.44	2.63
Wiener [2]	3.23	2.68	2.67
SEGAN [46]	3.48	2.94	2.80
Wave U-Net [47]	3.52	3.24	2.96
WaveNet [23]	3.62	3.23	2.98
MMSE-GAN [48]	3.80	3.12	3.14
Deep Feature Loss [49]	3.86	3.33	3.22
Deep Xi-ResLSTM [50]	4.01	3.25	3.34
Metric-GAN [51]	3.99	3.18	3.42
SEGAN-D [52]	3.46	3.11	3.50
DEMUCS [17]	4.14	3.21	3.54
Koizumi et al. [53]	4.15	3.42	3.57
DE-CADE(F)	4.00	3.11	3.60
T-GSA [54]	4.18	**3.59**	3.62
Deep MMSE [55]	4.28	3.46	3.64
DE-CADE	**4.36**	3.01	**3.86**

CADE(50th), were compared to other best performing CDAE based speech enhancement architectures in the literature. We tested these architectures using the mismatched test data, described in Section IV. The comparison includes the standard CDAE network, trained in the time domain [15], CDAE-T, and the same network was also trained in the frequency domain, CDAE-F. All the architectures were trained using the same dataset size for a fair evaluation.

Table IV shows this comparison, where both the first stage DE-CADE(F) and the two-stage architecture DE-CADE show better performance than the traditional CDAEs in the frequency and time domain, CDAE(F) and CDAE(T). The two-stage DE-CADE outperforms in terms of all the evaluation metrics, except the Cbak score, where the first stage DE-CADE(F) outperforms but at the expense of all the other evaluation metrics. It is also clear that the improvement is not significant from the 18^{18} to the 50th epoch, which proves that the number of training epochs is enough for the network to converge.

3) Complexity Analysis: Figure 7 shows the number of parameters of the first stage, DE-CADE(F) and the two-stage version, DE-CADE(F-T), of the proposed architecture, highlighted in red, in comparison with other SOTA speech enhancement models. It should be noted that in this analysis, we included only the architectures whose number of parameters were reported by the authors. Our single-stage network, DE-CADE(F), shows a comparable number of parameters to other architectures, such as Wavnet and CDAE-T, but it shows better performance based on the evaluation in Table III and IV. The two-stage architecture, DE-CADE, is more complex, but it significantly improves signal quality and the overall performance as shown in Tables III and IV. Moreover, it is of remarkably less complexity compared to GAN architectures.

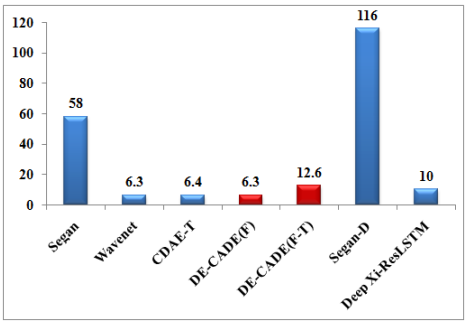

Fig. 7: A comparison between the number of parameters for the first stage of our architecture, DE-CADE(F), its two-stage version, DE-CADE, and SOTA speech enhancement models.

TABLE IV: Performance comparison of the architecture to other speech enhancement networks, using the mismatched test data. The results are averaged over 6 SNRs, from -5 to 20 with 5 dB step.

Metric	PESQ	STOI	LSD	Csig	Cbak	Covl
Noisy	2.086	83.28	1.422	2.856	2.037	2.421
CDAE-F [15]	2.622	86.78	1.285	3.438	2.687	3.009
CDAE-T [15]	2.556	87.33	0.936	3.543	2.588	3.016
DE-CADE(F)	2.623	88.21	0.862	3.658	**2.777**	3.120
DE-CADE(18th)	2.782	**89.63**	**0.791**	**3.862**	2.744	**3.305**
DE-CADE(50th)	2.797	**89.69**	**0.790**	**3.865**	2.762	**3.315**

VI. Conclusion

In this paper, a two-stage DNN architecture for speech enhancement is proposed, which is based on a new approach that takes advantage of the denoising capability of the CDAEs in the frequency domain as a first enhancement stage, followed by the reconstruction capability of the CDAEs in the time domain as a second enhancement stage. This work shows that the cascaded frequency then time approach is effective in decreasing speech distortion, which leads to overall better performance when compared to best performing speech enhancement models in the literature. Moreover, the proposed architecture shows promising results for improving speech intelligibility and quality when tested using challenging mismatched noisy speech. Future work is needed to investigate the combinations of other DNNs as a first and second stage, to further improve performance and decrease the complexity.

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Trans. Acoust. Speech Sig. Proc.*, vol. 27, no. 2, pp. 113–120, 1979.

[2] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *ICASSP*, vol. 2. IEEE, 1996, pp. 629–632.

[3] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *Trans. Speech Audio Proc.*, vol. 3, no. 4, pp. 251–266, 1995.

[4] D. Malah and Y. Ephraim, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Trans. Acoust. Speech Sig. Proc.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[5] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[6] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Comm.*, vol. 49, no. 7-8, pp. 588–601, 2007.

[7] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "An experimental analysis of deep learning architectures for supervised speech enhancement," *Electronics*, vol. 10, no. 1, p. 17, 2021.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[9] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *Trans. Audio Speech Lang. Proc.*, vol. 28, pp. 1778–1787, 2020.

[10] Y. Xia, S. Braun, C. K. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *ICASSP*. IEEE, 2020, pp. 871–875.

[11] X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, and W. Chen, "Masking and inpainting: A two-stage speech enhancement approach for low SNR and non-stationary noise," in *ICASSP*. IEEE, 2020, pp. 6959–6963.

[12] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *Trans. Audio Speech Lang. Proc.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[13] A. Pandey and D. Wang, "Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain," in *ICASSP*. IEEE, 2020, pp. 6629–6633.

[14] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *Trans. Audio Speech Lang. Proc.*, vol. 26, no. 9, pp. 1570–1584, 2018.

[15] A. Pandey and D. Wang, "A new framework for supervised speech enhancement in the time domain." in *INTERSPEECH*, 2018, pp. 1136–1140.

[16] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "A comparative study of time and frequency domain approaches to deep learning based speech enhancement," in *IJCNN*. IEEE, 2020, pp. 1–8.

[17] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," p. 3291–3295, 2020.

[18] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *ICASSP*. IEEE, 2015, pp. 4390–4394.

[19] A. A. Nair and K. Koishida, "Cascaded time+ time-frequency Unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps," in *ICASSP*. IEEE, 2021, pp. 7153–7157.

[20] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *Trans. Acoust. Speech Sig. Proc.*, vol. 30, no. 4, pp. 679–681, 1982.

[21] Z. Ouyang, H. Yu, W. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *ICASSP*. IEEE, 2019, pp. 5756–5760.

[22] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *Trans. Audio Speech Lang. Proc.*, vol. 24, no. 3, pp. 483–492, 2015.

[23] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *ICASSP*. IEEE, 2018, pp. 5069–5073.

[24] P. Wang, K. Tan *et al.*, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *Trans. Audio Speech Lang. Proc.*, vol. 28, pp. 39–48, 2019.

[25] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages," in *WASPAA*. IEEE, 2019, pp. 239–243.

[26] A. Pandey and D. Wang, "On cross-corpus generalization of deep learning based speech enhancement," *Trans. Audio Speech Lang. Proc.*, vol. 28, pp. 2489–2499, 2020.

[27] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Int. Conf. Learn. Rep. (ICLR)*, 2016, pp. 1–9.

[28] Y. Zhao, Z.-Q. Wang, and D. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *Trans. Audio Speech Lang. Proc.*, vol. 27, no. 1, pp. 53–62, 2018.

[29] Y. Zhao, D. Wang, I. Merks, and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in *ICASSP*. IEEE, 2016, pp. 6525–6529.

[30] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "Mapping and masking targets comparison using different deep learning based speech enhancement architectures," in *IJCNN*. IEEE, 2020, pp. 1–8.

[31] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mech. Syst. Sig. Proc.*, vol. 151, p. 107398, 2021.

[32] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *Trans. Audio Speech Lang. Proc.*, vol. 28, pp. 825–838, 2020.

[33] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Trans. Audio Speech Lang. Proc.*, vol. 16, no. 1, pp. 229–238, 2007.

[34] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Conf. Int. Speech Comm. Assoc.*, 2008.

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[36] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Comm.*, vol. 12, no. 3, pp. 247–251, 1993.

[37] C. Valentini-Botinhao *et al.*, "Noisy speech database for training speech enhancement algorithms and tts models," *University of Edinburgh*, 2017.

[38] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *O-COCOSDA/CASLRE*. IEEE, 2013, pp. 1–4.

[39] Topcoder, "176 spoken languages." [Online]. Available: http://www.topcoder.com/contest/problem/SpokenLanguages2/training-data.zip, 2017.

[40] C. K. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Interspeech 2021 deep noise suppression challenge," *arXiv*, 2021.

[41] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning Toolkit (version 0.92)," *University of Edinburgh*, 2019.

[42] C. Valentini-Botinhao *et al.*, "Reverberant speech database for training speech dereverberation algorithms and tts models," *University of Edinburgh*, 2016.

[43] G. Hu, "100 nonspeech environmental sounds." [Online]. Available: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html, 2014.

[44] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation, p. 862.*, 2001.

[45] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *Trans. Audio Speech Lang. Proc.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[46] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *INTERSPEECH*, 2017, pp. 3642–3646.

[47] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv*, 2018.

[48] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *ICASSP*. IEEE, 2018, pp. 5039–5043.

[49] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *INTERSPEECH*, 2019, pp. 2723–2727.

[50] A. Nicolson and K. K. Paliwal, "Deep learning for minimum mean-square error approaches to speech enhancement," *Speech Comm.*, vol. 111, pp. 44–55, 2019.

[51] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Int. Conf. Mach. Learn.* PMLR, 2019, pp. 2031–2041.

[52] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving GANs for speech enhancement," *Sig. Proc. Lett.*, vol. 27, pp. 1700–1704, 2020.

[53] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *ICASSP*. IEEE, 2020, pp. 181–185.

[54] J. Kim, M. El-Khamy, and J. Lee, "T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement," in *ICASSP*. IEEE, 2020, pp. 6649–6653.

[55] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "Deepmmse: A deep learning approach to mmse-based noise power spectral density estimation," *Trans. Audio Speech Lang. Proc.*, vol. 28, pp. 1404–1415, 2020.