Interspeech 2022
18-22 September 2022, Incheon, Korea

Convolutional Recurrent Smart Speech Enhancement Architecture for Hearing Aids

Soha A. Nossier[1], *Julie Wall*[1], *Mansour Moniri*[1], *Cornelius Glackin*[2], *Nigel Cannings*[2]

[1]Department of Engineering and Computing, University of East London, UK
[2]Intelligent Voice Ltd, UK

soha.abdallah.nossier@gmail.com, j.wall@uel.ac.uk, m.moniri@uel.ac.uk,
neil.glackin@intelligentvoice.com, nigel.cannings@intelligentvoice.com

Abstract

Speech enhancement is the process of removing noise to improve speech quality and intelligibility for applications including hearing aids. Many deep neural networks for speech enhancement have shown great ability in eliminating noise, regardless of its type. In hearing aids, this process may result in removing important noise used in emergency situations, such as fire alarms and car horns. In order to prevent this, a smart speech enhancement architecture is presented in this paper, where a convolution based noise classifier is used to detect emergency noise and activates the speech enhancement model to run in an audio enhancement mode, in which both the emergency noise and the speech are the target system output. The developed speech enhancement model is a deep convolutional recurrent network with several dilated layers to improve feature extraction without increasing network complexity. The results show that the speech enhancement model outperforms state of the art architectures by a 0.22 increase in the PESQ score. Moreover, the smart speech enhancement architecture improves speech and emergency noise quality when evaluated using objective metrics for both normal and hearing-impaired listeners.

Index Terms: Convolutional recurrent network, deep learning, hearing aids, noise classification, speech enhancement

1. Introduction

Speech Enhancement (SE) is the main signal processing technique utilised in Hearing Aids (HAs), which are devices that amplify sounds to improve speech intelligibility and quality for people with hearing disabilities [1]. Deep Neural Networks (DNNs) have shown a promising SE performance, where the DNNs managed to effectively mitigate background noise and generalize to real noise environments [2]. These DNNs include the: Multi-Layer Perceptron (MLP) [3], Convolutional Neural Network (CNN) [4], Recurrent Neural Network (RNN) [5], Convolutional Recurrent Network (CRN) [6], Convolutional Denoising Autoencoder (CDAE) [7], and the Generative Adversarial Network (GAN) [8].

As SE processing mitigates background noise regardless of its type, the hearing-impaired person has to rely on an external alert system to ensure their safety in emergency conditions. These systems detect the emergency sound, such as fire alarms, and use flashing lights or vibrating elements to notify the user [9]. With the advancement in signal processing using DNNs, a smart hearing aid device can be developed which can detect and amplify emergency noise [10].

In this work, we present a smart SE DNN architecture for HAs, which is an integrated SE and alert system. The system is shown in Figure 1, and consists of a CNN classifier and a Deep Convolutional Recurrent Network (DCRN). The DCRN was trained twice: first, it applies SE ($DCRN_{SE}$) to output speech only, while secondly, it performs Audio Enhancement (AE) ($DCRN_{AE}$) to enhance both speech and emergency noises, and removes other noise types. The classifier acts as a switch to change the mode of the DCRN network, either to work on speech or audio enhancement mode, to construct the full Smart SE (SSE) architecture ($DCRN_{SSE}$). The architecture aims to improve the performance of HAs, using this emergency component, while maintaining the performance of the SE module.

Inspired by the promising performance of deep learning for hearing aids [11, 12], the developed architecture ($DCRN_{SSE}$) is a 1 Dimensional (1D) convolution based DCRN network, in which strided and dilated convolution are applied to allow for better feature extraction [13]. The network operates in the time domain, which recently shows promising performance for speech enhancement [14, 7]. The classifier is a CNN based network, which was shown to be efficient in noise classification [15, 16]. It uses Time-Frequency (T-F) features for the input audio to perform the classification, because they lead to better classification accuracy [16].

The contributions of this paper are as follows:

- The development of a new DCRN for SE that uses several 1D dilated convolution layers with increased kernel size to achieve larger receptive fields with decreased complexity.
- Proposes an integrated hearing aid and alert system architecture to improve the functionality of currently available HAs.

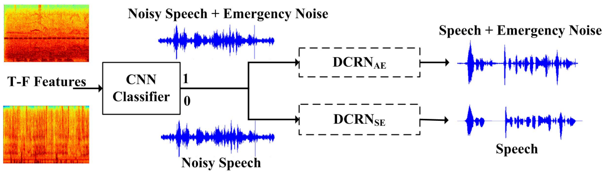


Figure 1: *Smart Speech Enhancement Architecture: An integrated speech enhancement and alert system for hearing aids.*

The organization of this paper is as follows. Section 2 illustrates the problem this paper is trying to solve. Section 3 describes the developed architecture. Section 4 provides the

This research is co-sponsored by Intelligent Voice Ltd.

10.21437/Interspeech.2022-522

used datasets and the experimental setup. Section 5 presents the obtained results. Finally, the conclusion is given in Section 6.

2. Problem Formulation

In SE, the input noisy speech signal can be represented as follows:

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

where *y*, *s*, and *n* are the noisy speech, clean speech and noise, respectively, $\{y, s, n\} \in \mathrm{R}^{K \times 1}$, where K is the total number of samples, and k is the sample index.

In the developed network, audio framing is the only pre-processing performed on the input, and the audio frames are then fed to the DCRN, shown in Figure 2(b). The DCRN will try to enhance the input noisy speech by removing background noise using a Mean Squared Error (MSE) loss function, L_{SE}, to finally generate an estimate to the clean speech signal *ŝ(t)*. This can be described as follows:

$$L_{SE} = \frac{1}{T}\sum_{t=0}^{T}\left[\hat{s}(t) - s(t)\right]^2, \tag{2}$$

where, *t* is the time frame index, and *T* is the total number of frames.

In SSE, we categorize the input noise as emergency (n_e) or unimportant noise (n_u), so in this case Equation 1 can be represented as:

$$\begin{aligned} y(k) &= s(k) + n_e(k) + n_u(k) && (3)\\ &= x(k) + n_u(k), && (4)\end{aligned}$$

where *x(k)* is the clean speech in addition to the emergency noise. The DCRN here will try to enhance both speech and emergency noise while suppressing other unimportant background noise. As a result, the MSE loss function here, L_{SSE}, will be:

$$L_{SSE} = \frac{1}{T}\sum_{t=0}^{T}\left[\hat{x}(t) - x(t)\right]^2. \tag{5}$$

The fact that the network in the case of SSE is trying to enhance emergency noise will decrease its ability to eliminate unimportant noise as well. The role of the CNN classifier is important here, as the classifier will act as a switch to run the network in one of two modes: SE mode; in which the network enhances the speech signal only and discards any other noise environment, or AE mode; in which the model performs SSE and outputs speech with emergency noise while suppressing any other kind of noise.

The classifier accepts five features as input that are useful for audio classification: Mel-Spectrogram (Y_{Mel}), MFCC (Y_{MFCC}), Spectral Contrast (Y_{SC}), Chromagram (Y_{Chroma}), and Tonnetz (Y_T) [17]. Mel-Spectrogram and MFCC are mainly used to model human hearing perception, while Chromagram and Tonnetz model the harmonic structure of speech and noise and shows harmonic relationships. Spectral Contrast is defined as the decibel difference between peaks and valleys in the spectrum. It measures energy variations of frequency at each timestamp and represents the relative spectral characteristics. These features were extracted, averaged and concatenated to form the input vector to the classifier C_i. This is shown below:

$$C_i = \overline{Y}_{MFCC} \oplus \overline{Y}_{Mel} \oplus \overline{Y}_{SC} \oplus \overline{Y}_{Chroma} \oplus \overline{Y}_{T}. \tag{6}$$

The decision of the classifier will be based on the detected noise environments, as it is trained to differentiate between emergency and non-emergency noise through a Binary Cross Entropy (BCE) loss function, detailed below:

$$L_C = \frac{1}{M}\sum_{i=1}^{M}\left[Z_i \log \hat{Z}_i + (1 - Z_i)\log\left(1 - \hat{Z}_i\right)\right], \tag{7}$$

where *M* is the total number of input samples, *i* is the sample index, *Z* is the target binary value (1 if emergency noise is detected and 0 otherwise), and $\hat{Z}$ is the predicted probability generated by the model.

3. Model Architecture

3.1. The Convolutional Classifier

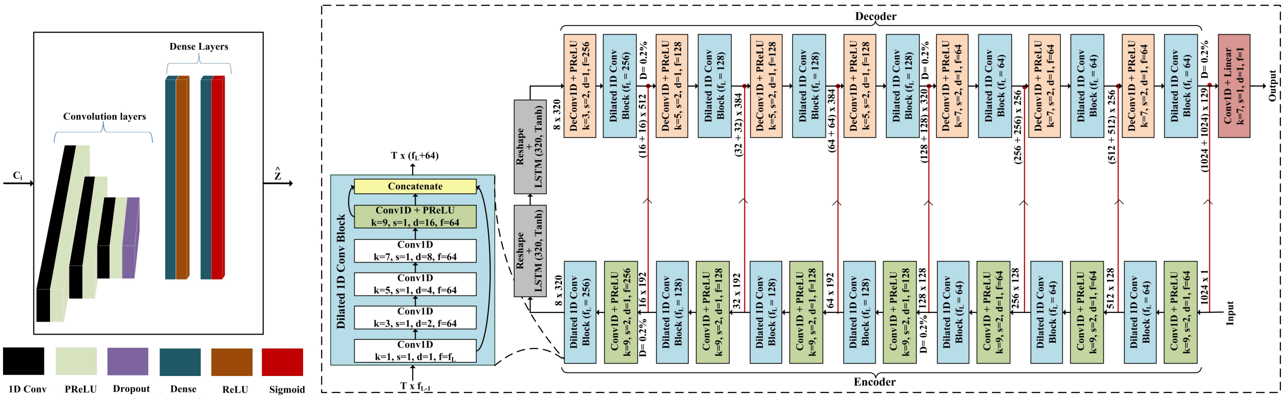
The classifier, shown in Figure 2(a), consists of three 1D CNN layers with Parametric Rectified Linear Unit (PReLU) activations. A stride of size 2 is applied for each CNN layer to compress the input feature vector. We used 64, 128, and 256 filter sizes for the first, second and third layers, respectively, and a kernel of size 10 for all layers. The output from the CNN layers is further processed by two dense layers. The first layer has 512 units and Rectified Linear Unit (ReLU) activation, while the second layer is an output layer with sigmoid activation. The classifier will output 1 if it detects emergency noise and 0 otherwise. Afterwards, it will feed this output to the DCRN network to run in one of the two modes.

3.2. The Deep Convolutional Recurrent Network (DCRN)

The developed network, shown in Figure 2(b), accepts a time frame of 1,024 size, and it is divided into three parts: the encoder, two Long Short-Term Memory (LSTM) layers, and the decoder. A block of dilated convolutions was added across the encoder and decoder networks. These dilated convolution layers help the network to better extract useful features by increasing the receptive field of the convolution operation. This can be illustrated in Figure 3, in which the receptive field and kernel size increase across hidden layers. This setup avoids information loss that might occur by network processing in deep hidden layers, and decreases network complexity by using different kernel sizes instead of large kernels. Strided convolution was applied in the encoder to decrease the size from 1,024 to 8, while upsampling is applied in the decoder to reconstruct the audio back to its original size. PReLU is the activation function used in both the encoder and the decoder, and a dropout of 0.2% is used after every three dilated convolution blocks. The two LSTM layers were added in the middle before feeding the signal to the decoder network, each has 320 units with Tanh activation. The role of these layers is to process the compressed bottleneck features to consider temporal dynamics of speech. Further details for other hyperparameters used for each layer are provided in Figure 2.

4. Experimental Setup

In training, speech and noise data were taken from the Microsoft Deep Noise Suppression (DNS) challenge dataset [18]. The dataset consists of more than 500 hours of speech and 181 hours of noise data. For emergency noise, a total of 1,478 audio samples were collected for 5 noise types: 118 alarm audio samples, including fire alarms, door bells, and alarm clocks; 440 car horn audio samples; 440 car siren audio samples; 440 baby crying


Dense Layers
Convolution layers
C_i
$\hat{Z}$
1D Conv
PReLU
Dropout
Dense
ReLU
Sigmoid
Dilated 1D Conv Block
Concatenate
Conv1D + PReLU k=9, s=1, d=16, f=64
Conv1D k=7, s=1, d=8, f=64
Conv1D k=5, s=1, d=4, f=64
Conv1D k=3, s=1, d=2, f=64
Conv1D k=1, s=1, d=1, f=f_L
T x (f_L+64)
T x f_L-1
Reshape + LSTM (320, Tanh)
Reshape + LSTM (320, Tanh)
Decoder
8 x 320
DeConv1D + PReLU k=3, s=2, d=1, f=256
Dilated 1D Conv Block (f_L = 256)
D= 0.2%
(16 + 16) x 512
DeConv1D + PReLU k=5, s=2, d=1, f=128
Dilated 1D Conv Block (f_L = 128)
(32 + 32) x 384
DeConv1D + PReLU k=5, s=2, d=1, f=128
Dilated 1D Conv Block (f_L = 128)
(64 + 64) x 384
DeConv1D + PReLU k=5, s=2, d=1, f=128
Dilated 1D Conv Block (f_L = 128)
D= 0.2%
(128 + 128) x 320
DeConv1D + PReLU k=7, s=2, d=1, f=64
Dilated 1D Conv Block (f_L = 64)
(256 + 256) x 256
DeConv1D + PReLU k=7, s=2, d=1, f=64
Dilated 1D Conv Block (f_L = 64)
(512 + 512) x 256
DeConv1D + PReLU k=7, s=2, d=1, f=64
Dilated 1D Conv Block (f_L = 64)
D= 0.2%
(1024 + 1024) x 129
Conv1D + Linear k=7, s=1, d=1, f=1
Output
8 x 320
Dilated 1D Conv Block (f_L = 256)
Conv1D + PReLU k=9, s=2, d=1, f=256
D= 0.2%
16 x 192
Dilated 1D Conv Block (f_L = 128)
Conv1D + PReLU k=9, s=2, d=1, f=128
32 x 192
Dilated 1D Conv Block (f_L = 128)
Conv1D + PReLU k=9, s=2, d=1, f=128
64 x 192
Dilated 1D Conv Block (f_L = 128)
Conv1D + PReLU k=9, s=2, d=1, f=128
D= 0.2%
128 x 128
Dilated 1D Conv Block (f_L = 64)
Conv1D + PReLU k=9, s=2, d=1, f=64
256 x 128
Dilated 1D Conv Block (f_L = 64)
Conv1D + PReLU k=9, s=2, d=1, f=64
512 x 128
Dilated 1D Conv Block (f_L = 64)
Conv1D + PReLU k=9, s=2, d=1, f=64
1024 x 1
Input
Encoder

a) CNN Classifier

b) The Proposed Deep Convolutional Recurrent Network (DCRN)

Figure 2: *The proposed Smart Speech Enhancement Architecture for Hearing Aids, a) the CNN classifier; C_i is the input feature vector and $\hat{Z}$ is the predicted class, b)the Deep Convolutional Recurrent Network (DCRN); 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.*

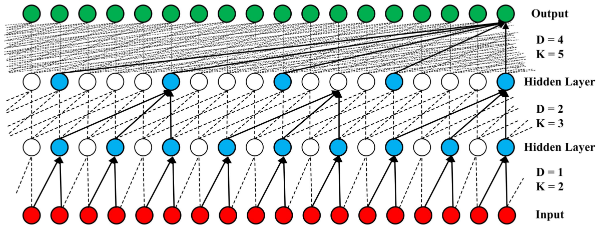


Figure 3: *Illustration of Dilated Causal Convolution with Increased Kernel Size*

audio samples; and 40 footstep audio samples. A total of 240 emergency noise audio samples were taken from the ESC-50 dataset [19], 800 from UrbanSound8K database [20], 400 from Donate-a-Cry corpus [21], and 38 from Mixkit website [22]. The speech and noise data was divided, 90% for training and 10% for validation. For SSE training, we mixed the speech and emergency noise at 0 dB SNR; to help the network dealing with them similarly during training, and then corrupted this mixture with undesired noise from the DNS challenge dataset at different SNR levels, from 0 dB to 20 dB with a step of 1, to form a total of 65,000 noisy utterances. For SE training, the speech utterances were only mixed with the undesired noise from the DNS challenge dataset at the same SNR levels (0-20 dB).

In testing, 100 speech utterances for 5 male and 5 female speakers were randomly selected from the Librispeech corpus [23]. On the other hand, we collected five emergency noise types, unseen during training, from the Mixkit website. For the undesired noise, we used 10 unseen noise environments from the 100 Nonspeech Environmental Sounds dataset [24]: 9 crowd noises, including babble noise, and Additive White Gaussian Noise (AWGN). For SSE model testing, we created the *AE test set* by mixing the speech with the emergency noises at 0 dB, and then added to the mixture the undesired noises at -5 dB, 0 dB, and 5 dB, where -5 dB is unseen SNR during training. For SE testing, we created the *SE test set* by corrupting the speech audio samples with the undesired noise only at -5 dB, 0 dB, and 5 dB SNRs. Both the classifier and the DCRN are trained and tested using the same dataset.

We sampled the input audio to 16 kHz sampling frequency, and it was normalized to zero mean and unit variance. For the classifier, we extracted frequency features and used BCE loss function. The network was then trained for 200 epochs. While for the DCRN, audio framing was performed with 1,024 frame size and 50% overlap. We used MSE loss function and Adam optimizer, learning rate = 0.0001, $\beta_1 = 0.1$, $\beta_2 = 0.999$. The used batch size is 4, and the number of epochs is 20, which was found to be sufficient for the network to converge. The final weights were finally taken based on the validation data.

5. Results and Discussion

We used the Perceptual Evaluation of Speech Quality (PESQ) [25] (from -0.5 to 4.5) and the Short Time Objective Intelligibility (STOI) [26] (from 0 to 100) to evaluate speech quality and intelligibility, respectively, for normal hearing listeners. The composite speech quality measures are also used: Csig (from 1 to 5), speech signal quality measure; Cbak (from 1 to 5), noise intrusiveness measure; and Covl (from 1 to 5), overall quality of the enhanced speech [27]. For hearing loss, we used the Hearing-Aid Speech Quality Index (HASQI) [28] (from 0 to 1) and the Hearing-Aid Speech Perception Index (HASPI) [29] (from 0 to 1) to measure speech quality and intelligibility, respectively. The Hearing-Aid Audio Quality Index (HAAQI) [30] (from 0 to 1) is also used to measure the quality of the output speech with emergency noise audio.

5.1. Smart Speech Enhancement Architecture Performance for Normal and Hearing-Impaired Listeners

5.1.1. Speech Enhancement Evaluation

In this experiment, we evaluated the architecture performance for improving the noisy speech when undesired noise only accompanies the speech signal. We used the *SE test set*, described in Section 4, and the presented results are the average of the three SNRs: -5 dB, 0 dB, and 5 dB. The evaluation is shown in Table 1 for the quality and intelligibility of the output speech using PESQ and STOI scores for normal hearing listeners, and the HASQI and HASPI scores for hearing impaired. For the HASQI and HASPI, two hearing loss degrees were used:

- HL1: Mild hearing loss, in which the person has difficulty hearing soft sounds in noisy environments.

- HL2: Moderate hearing loss, in which the person has difficulty hearing conversational speech, especially in noisy environments.

The values of the hearing loss degree were taken from the real Occupational Hearing Loss (OHL) Worker Surveillance Data [31], which is a dataset used to estimate the prevalence of hearing loss among U.S. industries. We took the hearing degree values for 100 workers, 50 males and 50 females, for mild and moderate hearing loss cases, 50 values for each.

We evaluated the performance of the SE network only, $DCRN_{SE}$, and the full SSE architecture with the classifier, $DCRN_{SSE}$. The classifier accuracy for the used test set is 90%. The results show that both networks improve the quality and intelligibility of the speech for both normal and hearing-impaired listeners. The full architecture, $DCRN_{SSE}$, gives slightly worse performance than the SE only model, $DCRN_{SE}$. The reason for this is the failure of the classifier to classify some challenging undesired crowd noise; and as a result, the noisy speech will be processed with the AE network in this case, which is trained to output emergency noise and in turn, this negatively affects the SE capability of the network.

The spectrograms in Figure 4(a) show the performance of the SE model when tested using speech corrupted with undesired babble noise at 0 dB SNR. It is clear that the model managed to remove most of the challenging babble noise.

5.1.2. *Speech and Emergency Noise Enhancement Evaluation*

In this evaluation, we used the HAAQI to assess the quality of the enhanced speech and emergency noise audio samples generated by the network, by testing the network using the *AE test set*, described in Section 4. The results of this evaluation are shown in Table 2, which are the average scores at three SNRs: -5 dB, 0 dB, and 5 dB. The results show that the architecture is able to improve the quality of the audio in comparison to the unprocessed audio for both mild and moderate hearing loss degrees, which proves the applicability of the presented architecture.

The spectrograms in Figure 4(b) show the performance of the AE model when tested using speech with fire alarm emergency noise corrupted with undesired challenging babble noise at 0 dB SNR. It is clear that the model is trying to generate both speech and fire alarm noise and mitigate babble noise.

5.2. Speech Enhancement Model Comparison to Baselines

In this experiment, we trained the DCRN for SE only using the Valentini Voice Bank dataset benchmark [32], to compare its performance with other networks in the literature. The results for the evaluation on the Valentini test set are shown in Table 3, in which our network outperforms with respect to the PESQ and Covl scores. Additionally, the architecture shows good performance for the other evaluation metrics in comparison to other state of the art DNNs. It should be noted that STOI results were not reported by the authors of Wave U-Net [33], Metric-GAN [34], SEGAN-D [8], Koizumi et al. [6], and T-GSA [35].

6. Conclusions

In this paper, we present an integrated speech enhancement and alert system architecture, designed especially for hearing aid applications. The architecture performs smart speech enhancement, which is to enhance both speech and emergency noise to ensure the safety of people with hearing disabilities in the case of emergency. This is achieved by operating the architecture in speech or audio enhancement mode based on the decision of a classifier network, which detects emergency noise. We tested the architecture using several evaluation metrics for normal hearing and different types of hearing loss, and the results show that the idea is applicable and can be used to develop hearing aid devices in the future. Further work is needed to improve the classification accuracy and network performance.

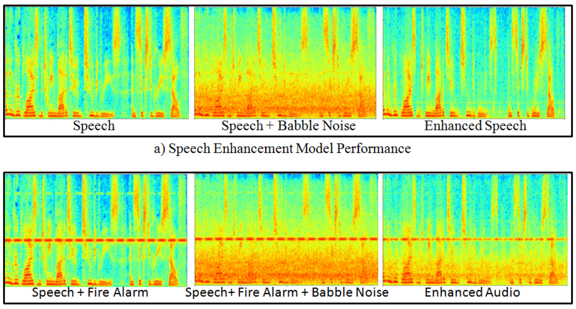


Figure 4: *The Performance of the Proposed Architecture, a) speech enhancement mode, b) audio enhancement mode*

Table 1: *Speech Enhancement Performance of the Architecture for Normal and Hearing-Impaired Listeners*

Metric	Normal Hearing		Hearing Loss			
	PESQ	STOI	HASQI		HASPI	
			HL1	HL2	HL1	HL2
Unprocessed	1.57	70	0.37	0.24	70	65
$DCRN_{SE}$	2.12	77	0.57	0.38	76	70
$DCRN_{SSE}$	2.00	76	0.56	0.36	75	68

Table 2: *Performance of the Architecture for Speech and Emergency Noise Enhancement*

Metric	HAAQI	
	HL1	HL2
Unprocessed	0.21	0.16
$DCRN_{SSE}$	0.44	0.34

Table 3: *Performance Comparison with State-of-the-Art Speech Enhancement Models using the Valentini Voice Bank dataset benchmark [32].*

Metric	PESQ	STOI	Csig	Cbak	Covl
Noisy	1.97	91.5	3.35	2.44	2.63
Wiener [36]	2.22	92	3.23	2.68	2.67
SEGAN [37]	2.16	93	3.48	2.94	2.80
Wave U-Net [33]	2.40	-	3.52	3.24	2.96
MMSE-GAN [38]	2.53	93	3.80	3.12	3.14
Deep Xi-ResLSTM [39]	2.65	91	4.01	3.25	3.34
Metric-GAN [34]	2.86	-	3.99	3.18	3.42
SEGAN-D [8]	2.39	-	3.46	3.11	3.50
DEMUCS [7]	3.07	**95**	4.14	3.21	3.54
Koizumi et al. [6]	2.99	-	4.15	3.42	3.57
T-GSA [35]	3.06	-	4.18	**3.59**	3.62
Deep MMSE [40]	2.95	94	**4.28**	3.46	3.64
$DCRN_{SE}$	**3.29**	93.5	4.18	2.96	**3.76**

7. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[2] 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, 2020.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[4] S. Fu, T. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE Transactions on audio, speech, and language processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[5] 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.

[6] 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.

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

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

[9] F. Beritelli, S. Casale, A. Russo, and S. Serrano, "An automatic emergency signal recognition system for the hearing impaired," in *12th Digital Signal Processing Workshop & 4th IEEE Signal Processing Education Workshop*. IEEE, 2006, pp. 179–182.

[10] S. A. Nossier, M. Rizk, N. D. Moussa, and S. el Shehaby, "Enhanced smart hearing aid using deep neural networks," *Alexandria Engineering Journal*, vol. 58, no. 2, pp. 539–550, 2019.

[11] H. Schröter, T. Rosenkranz, A. N. Escalante-B, M. Aubreville, and A. Maier, "Clcnet: Deep learning-based noise reduction for hearing aids using complex linear coding," in *ICASSP*. IEEE, 2020, pp. 6949–6953.

[12] D. Wang, "Deep learning reinvents the hearing aid," *IEEE spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[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. 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.

[15] G. Park and S. Lee, "Environmental noise classification using convolutional neural networks with input transform for hearing aids," *International journal of environmental research and public health*, vol. 17, no. 7, p. 2270, 2020.

[16] Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," *Applied Acoustics*, vol. 167, p. 107389, 2020.

[17] F. Alías, J. C. Socoró, and X. Sevillano, "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds," *Applied Sciences*, vol. 6, no. 5, p. 143, 2016.

[18] 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.

[19] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[20] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1041–1044.

[21] G. Veres. (2015) Donate-a-cry corpus. [Online]. Available: https://github.com/gveres/donateacry-corpus

[22] E. Elements. (2019) Mixkit. [Online]. Available: https://mixkit.co/

[23] 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.

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

[25] 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.

[26] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on audio, speech, and language processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[27] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.

[28] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (hasqi)," *Journal of the Audio Engineering Society*, vol. 58, no. 5, pp. 363–381, 2010.

[29] ——, "The hearing-aid speech perception index (haspi) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[30] ——, "The hearing-aid audio quality index (haaqi)," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 2, pp. 354–365, 2015.

[31] E. A. Masterson, S. Tak, C. L. Themann, D. K. Wall, M. R. Groenewold, J. A. Deddens, and G. M. Calvert, "Prevalence of hearing loss in the united states by industry," *American journal of industrial medicine*, vol. 56, no. 6, pp. 670–681, 2013.

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

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

[34] 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 *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[35] 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.

[36] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.

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

[38] 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.

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

[40] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1404–1415, 2020.