A Deep Learning Speech Enhancement Architecture Optimised for Speech Recognition and Hearing Aids

Soha A. Nossier
Dept. Computer Science and Digital Technologies
University of East London
London, UK
soha.abdallah.nossier@gmail.com

Julie Wall
Dept. Computer Science and Digital Technologies
University of East London
London, UK
j.wall@uel.ac.uk

Mansour Moniri
Dept. Computer Science and Digital Technologies
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*—With the fast progression of the speech enhancement field after the introduction of deep learning techniques, there is a need to consider the adjustments needed to employ these techniques for real-life applications. In this work, we present an optimised deep learning speech enhancement architecture for automatic speech recognition and hearing aids, two key speech enhancement applications. A speech enhancement architecture with a signal-to-noise ratio switch is presented for automatic speech recognition systems, to avoid denoising artifacts that cause performance degradation in the case of clean or high signal-to-noise speech. Moreover, a smart speech enhancement architecture is presented for hearing aids to retain important emergency noise in the audio signal. The presented work achieved 13.9% reduction in the word error rate of an automatic speech recognition system. Additionally, the smart speech enhancement architecture resulted in 0.18 improvement in HAAQI audio quality metric.**

***Index Terms*—Automatic speech recognition, convolutional classifiers, deep learning, hearing aids, speech enhancement**

I. Introduction

Recent deep learning-based speech enhancement (SE) architectures have shown a great ability to generate estimated clean speech signals with high quality and intelligibility [1]–[3]. This allows these architectures to be employed for real-life SE applications, including Automatic Speech Recognition (ASR) [4], [5] and hearing aids [6], [7]. However, when applying SE architectures to these applications, other factors should be taken into consideration.

On the one hand, SE is not always useful for applications such as ASR, because the artifacts added by the distortion caused by the enhancement networks sometimes result in worse Word Error Rates (WERs) [4], [8]. To solve this issue, two-stage SE networks can be implemented to minimize distortion [3], [9]. Another idea is the joint training of the SE network and ASR system, which also shows better performance [4]. Although these solutions result in improving the WER, the SE network should only be turned on when necessary. This is to keep these distortion artifacts to a minimum and to avoid clean speech processing, which increases processing time without any performance gain. The decision of switching on or off the SE network can be performed based on the Signal-to-Distortion (SDR) ratio of the enhanced speech [10]. Alternatively, a deep neural network (DNN) can be trained to decide whether to perform SE or not, using the WER of the ASR system under testing for the enhanced speech and clean speech [11]. However, these solutions are based on making the decision based on the enhanced speech signal, which means SE processing is always required, and this increases processing time.

On the other hand, it is crucial to enhance emergency noise along with speech, known as Smart Speech Enhancement (SSE) [12], especially for applications such as hearing aids, where the users have reduced hearing ability. As presented in the literature [12], [13], having a SE and alert system in one SSE architecture can replace the need to install separate alert systems [14] for hearing aid users. This idea is based on adding a frontend noise classifier to detect emergency noise and runs an audio enhancement network, trained to output speech and emergency noise. However, based on the reported results, this can negatively affect the performance of the SE module, especially for highly intrusive noise environments, where it is very challenging for one network to perform speech and emergency noise enhancement simultaneously [12].

In this work, we present two improved implementations of the two-stage Deep Encoder-Convolution Autoencoder De-

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

noiser (DE-CADE), proposed in [9] and shown in Fig. 1, one for ASR systems and another for hearing aids. The architectures fill the gaps in the literature through the following contributions.

- A SE architecture is presented for ASR systems that significantly reduces the WER in noisy environments, and minimizes clean speech distortion through the usage of a signal-to-noise (SNR) based switch, which also avoids extra processing time when SE is not required.
- A new SSE architecture is presented, where the emergency noise and speech are enhanced separately using a second stage network, to improve performance.

The following sections are organized as follows. In Section II, the developed DNNs will be explained. Section III defines the problems under investigation. Experimental setup is given in Section IV. Results and discussions are presented in Section V. Finally, Section VI provides the conclusion.

II. The Developed Architectures

For both ASR systems and hearing aids, the developed architecture consists of a frontend binary classifier and a two-stage SE network, as shown in Fig. 2.

For ASR, the classifier was trained to classify low and high SNRs, as shown in Fig. 2(a). In our implementation, SNR values that are less than or equal to 15 dB belong to class 1 (low SNR), while SNR values that are greater than 15 dB belong to class 0 (high SNR). If the noisy speech is of low SNR, the classifier switches on the two-stage SE network, and then the enhanced speech by the network is fed to the ASR system to generate the transcription. While for high SNR, the unprocessed speech is directly fed to the ASR system, without performing SE, to avoid the artifacts generated by the SE network.

For hearing aids, the frontend classifier was trained to detect emergency noise in noisy speech, as shown in Fig. 2(b). The classifier outputs 1 if emergency noise accompanies the speech signal and 0 otherwise. If emergency noise is detected, two second stage DE-CADE architectures will run, one performs speech reconstruction, and the other performs emergency noise enhancement. Otherwise, the standard SE procedure is applied, in which the network enhances the speech signal only.

The following subsections describe the implementation of the classifier and SE network separately.

A. The CNN Classifier

A binary one-dimensional (1D) convolution-based classifier was implemented [12] that consists of three convolution layers with Parametric Rectified Linear Unit (PReLU) activations, stride of size 2, and kernel of size 10. The convolution layers are followed by two dense layers for prediction. The first dense layer has 512 units and Rectified Linear Units (ReLU), while the second dense layer generates the output using a Sigmoid activation.

The classifier accepts five frequency domain-based features that were proven to be effective for audio classification [15], [16]. These features are then concatenated together to generate an input feature vector that is fed to the classifier network, C_i. This feature vector can be represented as in Equation 1:

$$C_i = \overline{y}_{MFCC} \oplus \overline{y}_{Mel} \oplus \overline{y}_{SC} \oplus \overline{y}_{Chroma} \oplus \overline{y}_{T}, \quad (1)$$

where y_{Mel} is the Mel-Spectrogram, y_{MFCC} is the Mel-Frequency Cepstral Coefficients (MFCCs), Y_{SC} is the Spectral Contrast, Y_{Chroma} is the Chromagram, and Y_T is the Tonnetz [16].

B. The Deep Encoder-Convolution Autoencoder Denoiser (DE-CADE)

The SE DE-CADE [9], shown in Fig. 1, performs a first denoising stage in the frequency domain and a second reconstruction stage in the time domain. The architecture is a fully convolution encoder/decoder-based implementation, where the encoder is deeper than the decoder, to improve performance and minimize network complexity. The noisy speech is first processed by the frequency domain-based DE-CADE, where aggressive noise removal is performed, generating a highly denoised but distorted speech. The output of the first stage is then processed by the second stage time domain-based DE-CADE, which performs speech reconstruction and outputs the enhanced speech.

When implementing the architecture to perform SSE for hearing aids, the second reconstruction stage was trained twice: the first to reconstruct speech and the second to enhance emergency noise. This will result in having an enhanced speech and emergency noise signals as outputs, as shown in Fig. 2(b).

III. Problem Definition

As this work presents an improved SE architecture for two different applications: ASR systems and hearing aids, the description of the problem for each application will be presented separately in the following two subsections. It is also represented in Fig. 2

A. Speech Enhancement for ASR

The time domain noisy speech signal can be expressed as in Equation 2:

$$y(k) = s(k) + n(k), \quad (2)$$

where, *y* is the noisy speech, *s* and *n* are the speech and noise signals, and *k* is the time index.

In deep learning-based SE, the noisy speech *y* is processed by a DNN to generate an estimate to the clean speech signal, $\hat{s}$. As proved in [4], [8], the SE process adds some unwanted artifacts that negatively affects the enhanced speech signal, as it causes speech distortion. Considering the effect of these artifacts, the time domain enhanced speech signal, $\hat{s}_2$, that is generated by the second stage DE-CADE network, shown in Fig. 1, can be defined as in Equation 3:

$$\hat{s}_2(k) = s(k) + \alpha n(k) + z(k), \quad (3)$$

where α is a scaling factor to the noise signal, describing the decrease in noise intensity due to the noise removal process,

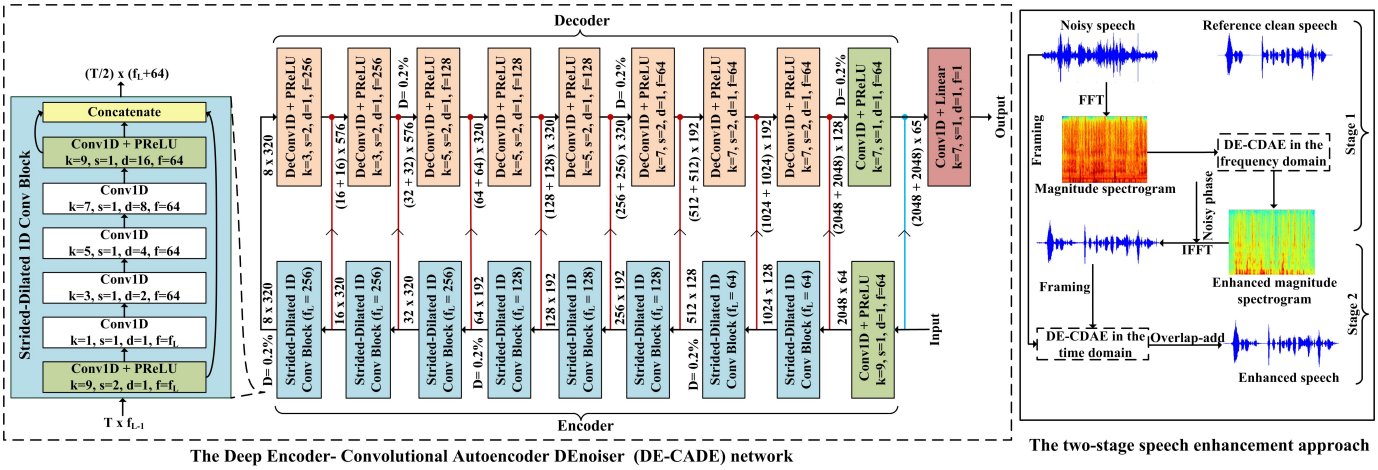

Fig. 1. The two-stage Deep Encoder-Convolution Autoencoder Denoiser (DE-CADE) [9]

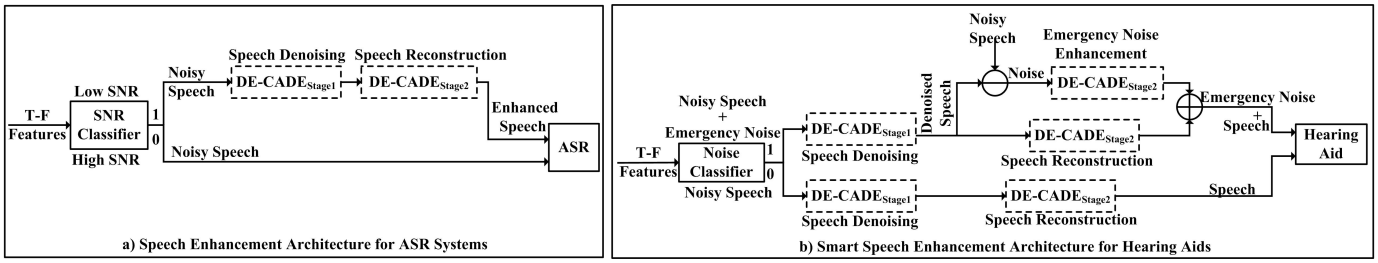

Fig. 2. The proposed architectures: speech enhancement for ASR (Sub-figure a) and smart speech enhancement for hearing aids (Sub-figure b)

and z is the signal that represents the added artifacts by the DNN.

By observing the performance of the ASR system used for testing using different noisy speech data at many SNR levels, we noticed that the ASR system can generate a transcription for the noisy speech of SNR value greater than 15 dB with lower WER without performing SE than the case of transcribing the enhanced speech. This means that the artifacts generated by the enhancement network in this case outweigh the advantages of the denoising process. Therefore, the negative effect of z is greater than n in the case of high SNR values (more than 15 dB in our system), leading to higher WERs if SE is applied before ASR. As a result, 15 dB was set as the threshold value of the SNR classifier in our implementation, shown in Fig. 2(a), which means that the classifier will activate the SE network only if 15 dB or less SNR value was detected.

B. Smart Speech Enhancement for Hearing Aids

When applying SSE, as proposed in [12], the noise signal in Equation 2 is redefined to represent emergency noise and any other unimportant noise as two separate signals. This can be represented by Equation 4:

$$y(k) = s(k) + n_e(k) + n_u(k), \tag{4}$$

where, n_e and n_u represent emergency and unimportant noise, respectively. To retain the emergency noise while maintaining a good SE performance, the idea proposed in this work is based on restoring an estimate to the noise signals, $\hat{n}$, from the first stage enhanced speech by the DE-CADE, $\hat{s}_1$, and the original noisy speech signal, y, using a subtraction process. This can be described by Equation 5:

$$\hat{n}(k) = y(k) - \hat{s}_1(k). \tag{5}$$

Each of the estimated noise and speech signals are then fed separately to the second stage DE-CADE network for signal reconstruction, as the second DE-CADE stage in our implementation was trained twice. Once to perform speech reconstruction to the enhanced speech by the first DE-CADE network, and a second time to perform emergency noise enhancement, where the unimportant noise is removed and the emergency noise is reconstructed. The output from the second stage DE-CADE networks will be then added to generate the final speech and emergency noise signal, x. This is shown in Fig. 2(b) and can be represented as in Equation 6:

$$\hat{x}(k) = \hat{n}_e(k) + \hat{s}_2(k), \tag{6}$$

where, $\hat{n}_e$ is the estimated emergency noise by the second stage DE-CADE. By applying this idea, the negative effect of the audio enhancement mode, described in [12], on the SE process will be minimized, as here the emergency noise and speech are processed using two independent networks. Moreover, this processing separation facilitates the generation of speech and emergency noise with better quality.

IV. Experimental Setup

The Deep Noise Suppression (DNS) challenge dataset was used in the training process. This dataset consists of 500 hours of clean speech and 181 hours of unimportant noise data. For the SE network for ASR, all the clean speech utterances were corrupted by the noise data at different SNR levels (0 dB to 20 dB with a step of 1), to form 65,000 noisy utterances that were used to train the SE network and the SNR classifier.

While for the SSE network for hearing aids, besides the above-described speech and unimportant noise data, emergency noise utterances were collected from different sources: 240 from the ESC-50 dataset [17], 800 from UrbanSound8K database [18], 400 from Donate-a-Cry corpus [19], and 38 from Mixkit website [20]. This makes a total of 1,478 audio samples for five emergency noises: 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 audio samples; and 40 footstep audio samples. The emergency noise data was first randomly mixed with 50% of the clean speech data at 0 dB SNR, as using 0 dB SNR value for the two target signals was found to help in network training. The speech and emergency noise mixture and the remaining 50% of clean speech data were then corrupted by the unimportant noise data at a range of SNR levels (0-20 dB). This makes a total of 65,000 noisy utterances, which are used to train the SSE network and the noise classifier.

In all training procedures, the data was divided into 90% for training and 10% for validation.

For testing, the Librispeech corpus [21] was used, where 100 clean speech utterances for 5 male and 5 female speakers were corrupted by 10 unseen unimportant noise environments from the 100 Nonspeech Environmental Sounds dataset [22]: 9 crowd noises, including babble noise, and an Additive White Gaussian Noise (AWGN). Four testing SNRs were used: 0 dB (low SNR), 5 dB (low SNR), 15 dB (classifier threshold), and 20 dB (high SNR). This data forms the test set for the SE network for ASR; denoted by *ASR Test Set*.

In order to create the test set for the SSE network for hearing aids, we randomly selected five audio samples, one for each of the emergency noise types used in the training. These five audio samples were collected from the Mixkit website and they were not seen during training. These emergency noise audios were first mixed with the 100 testing clean speech utterances, and then the mixture was corrupted by the 10 unimportant testing noise environments, described above, at -5 dB, 0 dB, and 5dB SNRs; this test set will be denoted by *SSE Test Set*. The clean speech utterances were also corrupted by the unimportant testing noise environments at the same SNR levels, to create the *SE Test Set*, which is used to evaluate the SE performance of the network. This test setting is similar to the one used in this work [12], to make a fair comparison.

Regarding training hyperparameters, 16 kHz is the sampling frequency used. Binary Cross Entropy (BCE) loss function was used to train the classifiers, while Mean Square Error (MSE) is the loss function used for the first and second stage DE-CADE. The Adam optimizer was used, learning rate = 0.0001, β_1 = 0.1, β_2 = 0.999. The networks were trained until convergence and the best weights were taken based on the validation data.

V. Results and Discussion

A. Automatic Speech Recognition Performance

The performance of the SE architecture for ASR was evaluated using the WER for the *ASR Test Set*, and these results are presented in Table I. The results show the WER of the unprocessed audio; WER_{Unproc}, first SE stage output; WER_{SE1}, second SE stage output; WER_{SE2}, and the output of the two-stage SE network with the frontend classifier, WER_{C+SE}.

It is clear from the results in Table I that the SE networks significantly improves ASR performance for low SNR values, where 13.9%, 11%, and 7% reduction in WER were shown for 0 dB, 5 dB, and 15 dB SNRs, respectively. The negative effect added by the SE network artifacts starts to be clear at high SNR and for clean speech, where 0.4% and 0.5% degradation in the WER were caused after processing the 20 dB and clean speech audio with the SE network, respectively. The addition of the frontend SNR classifier reduces this negative effect, where 0.2% and 0.4% reduction in the WER were shown for 20 dB and clean speech, respectively, compared to the case of the SE network only without the SNR classifier. The classifier accuracy is 100% for both 0 dB and 5 dB noisy speech utterances, 96.5% and 93% for clean speech and 20 dB noisy speech utterances, respectively, and 88% for 15 dB noisy speech utterances. It should be mentioned here that the low accuracy for 15 dB SNR is due to the fact this is the most challenging SNR value for the classifier, as it was chosen as the classification boundary that differentiates between high and low SNR values. This explains the slight increase in the WER at 15 dB SNR after adding the classifier.

B. Smart Speech Enhancement Performance

The performance of the SSE network for hearing aids was evaluated using different speech quality metrics. For normal listeners, we used the Perceptual Evaluation of Speech Quality (PESQ) [23] (from -0.5 to 4.5) and the Short Time Objective Intelligibility (STOI) [24] (from 0 to 100), to assess speech quality and intelligibility, respectively. For hearing-impaired listeners, we used the Hearing-Aid Speech Quality Index (HASQI) [25] (from 0 to 1) and the Hearing-Aid Speech Perception Index (HASPI) [26] (from 0 to 1) to measure speech quality and intelligibility, respectively. While to measure the quality of the speech with emergency noise audio, we used the Hearing-Aid Audio Quality Index (HAAQI) [27] (from 0 to 1). These measures are presented for two hearing loss degrees: Mild hearing loss (HL1) and Moderate hearing loss (HL2). 50 values for each hearing loss degree were taken for 100 workers, 50 males and 50 females, from the real Occupational Hearing Loss (OHL) Worker Surveillance Data [28].

The SE performance of the proposed architecture was first tested using the *SE Test Set*, which contains unimportant noise only, to show the effect of adding the SSE processing. The

frontend noise classifier accuracy for this test set is 90%. These results are shown in Table II, where comparison was made with another architecture in the literature, $DCRN$ [12]. The subscripts *SE* and *SSE* denote the performance when the network enhances speech only and when enhancing speech and emergency noise, respectively. The presented architecture shows better SE performance compared to the *DCRN*, in terms of all the evaluation metrics. Moreover, the negative effect of adding the emergency noise enhancement processing is also decreased for our network, as the difference between SE network, *DE-CADE*$_{SE}$, and the SSE network, *DE-CADE*$_{SSE}$ is less than that of the *DCRN* architecture.

Table III shows the performance of the SSE network using the *SSE Test Set*, which contains emergency noise. The frontend noise classifier accuracy for this test set is 97%. The proposed architecture generated speech and emergency noise with better audio quality for the two hearing loss degrees, compared to the SSE *DCRN* architecture in the literature.

TABLE I
AUTOMATIC SPEECH RECOGNITION PERFORMANCE USING THE ASR TEST SET

SNR	Clean	20 dB	15 dB	5 dB	0 dB	Ave
WER$_{Unproc.}$	25.3	26.7	32.5	46.5	58	37.8
WER$_{SE1}$	25.8	27.2	27.5	40.3	50.9	34.3
WER$_{SE2}$	25.8	27.1	25.5	35.5	44.1	31.6
WER$_{C+SE}$	25.4	26.9	25.6	35.5	44.1	**31.5**

TABLE II
SPEECH ENHANCEMENT AND SMART SPEECH ENHANCEMENT PERFORMANCE COMPARISON FOR NORMAL AND HEARING-IMPAIRED LISTENERS USING THE SE TEST SET

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
DE-CADE$_{SE}$	**2.36**	**79**	**0.67**	**0.48**	**77**	**69**
DCRN$_{SSE}$	2.00	76	0.56	0.36	75	68
DE-CADE$_{SSE}$	**2.34**	**78.8**	**0.66**	**0.47**	**76.6**	**68.7**

TABLE III
SMART SPEECH ENHANCEMENT PERFORMANCE USING THE SSE TEST SET

Metric	HAAQI	
	HL1	HL2
Unprocessed	0.21	0.16
DCRN$_{SSE}$	0.44	0.34
DE-CADE$_{SSE}$	**0.62**	**0.55**

VI. CONCLUSIONS

This paper presents two improved SE architectures for ASR and hearing aids applications. For ASR, the architecture minimizes the negative effect of the denoising artifacts by applying SE only when required, based on the decision of an SNR classifier. While a SSE architecture was designed for hearing aids, to perform speech and emergency noise enhancement. The results show better SE performance after the adjustments made for each application, and in comparison to other works in the literature. Future work is needed to improve the accuracy of the SNR classifier for ASR at the boundary SNR value.

REFERENCES

[1] 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 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6959–6963.

[2] A. A. Nair and K. Koishida, "Cascaded time+ time-frequency unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7153–7157.

[3] 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 *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 239–243.

[4] P. Wang, K. Tan *et al.*, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 39–48, 2019.

[5] Z.-Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust asr," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 28, pp. 1778–1787, 2020.

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

[7] G. Park, W. Cho, K.-S. Kim, and S. Lee, "Speech enhancement for hearing aids with deep learning on environmental noises," *Applied Sciences*, vol. 10, no. 17, p. 6077, 2020.

[8] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 826–835, 2014.

[9] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "Two-stage deep learning approach for speech enhancement and reconstruction in the frequency and time domains," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–10.

[10] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, T. Moriya, and N. Kamo, "Should we always separate?: Switching between enhanced and observed signals for overlapping speech recognition," in *INTERSPEECH*, 2021, pp. 1149–1153.

[11] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, N. Kamo, and T. Moriya, "Learning to enhance or not: Neural network-based switching of enhanced and observed signals for overlapping speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6287–6291.

[12] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "Convolutional recurrent smart speech enhancement architecture for hearing aids," *INTERSPEECH 2022*, 2022.

[13] 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.

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

[15] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*. IEEE, 2015, pp. 1–6.

[16] 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.

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

[18] 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.
[19] G. Veres. (2015) Donate-a-cry corpus. [Online]. Available: https://github.com/gveres/donateacry-corpus
[20] E. Elements. (2019) Mixkit. [Online]. Available: https://mixkit.co/
[21] 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.
[22] G. Hu. (2014) 100 nonspeech environmental sounds. [Online]. Available: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html
[23] 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.
[24] 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.
[25] 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.
[26] ——, "The hearing-aid speech perception index (haspi) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.
[27] ——, "The hearing-aid audio quality index (haaqi)," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 2, pp. 354–365, 2015.
[28] 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.