Environmental Noise Adaptable Hearing Aid using Deep Learning

Soha A. Nossier
Department of Biomedical Engineering, Medical Research Institute, Alexandria University, Egypt
soha.abdallah.nossier@gmail.com

M. R. M. Rizk
Department of Electrical Engineering, Faculty of Engineering, Alexandria University, Egypt
mrmrizk@ieee.org

Saleh El Shehaby
Department of Biomedical Engineering, Medical Research Institute, Alexandria University, Egypt
shehaby@alexu.edu.eg

Nancy Diaa Moussa
Department of Biomedical Engineering, Medical Research Institute, Alexandria University, Egypt
nancy.diaaeldin.moussa@gmail.com

Abstract: *Speech de-nosing is one of the essential processes done inside hearing aids, and has recently shown a great improvement when applied using deep learning. However, when performing the speech de-noising for hearing aids, adding noise frequency classification stage is of a great importance, because of the different hearing loss types. Patients who suffer from sensorineural hearing loss have lower ability to hear specific range of frequencies over the others, so treating all the noise environments similarly will result in unsatisfying performance. In this paper, the idea of environmental adaptable hearing aid will be introduced. A hearing aid that can be programmed to multiply the background noise by a weight based on its frequency and importance, to match the case and needs of each patient. Furthermore, a more generalized Deep Neural Network (DNN) for speech enhancement will be presented, by training the network on a diversity of languages, instead of only the target language. The results show that the learning process of DNN for speech enhancement is more efficient when training the network using diversity of languages. Moreover, the idea of adaptable hearing aid is shown to be promising and achieved 70% overall accuracy. This accuracy can be improved using a larger environmental noise dataset***.**

Keywords: *Adaptable hearing aid, MFCC, neural networks, noise classification, speech enhancement*.

Received December 1, 2020; accepted December 9, 2021
https://doi.org/10.34028/iajit/19/5/15

1. Introduction

A Hearing aid is an electronic device which improves speech quality for hearing impaired people who suffer from hearing loss. There are three types of hearing loss: conductive hearing loss, sensorineural hearing loss, and mixed hearing loss [13]. Conductive hearing loss is related to any impairment in the outer or the middle ear, while sensorineural hearing loss is related to a deficiency in the inner ear. Mixed hearing loss is a combination of the previous two types. Patients who suffer from sensorineural hearing loss lack the ability to separate desired sound from undesired noise; for that reason, they find it difficult to hear in a noisy environment [24]. As a result, a speech enhancement technique is needed for the hearing aid device to overcome this disability.

Hearing aids can be analog or digital. Analog hearing aids amplifies the sound entering the ear; while, digital hearing aids have other advanced features and preprocessing techniques applied to the input sound before the amplification process [33]. Recently, most available digital hearing aids have an embedded speech enhancement system, that is used to eliminate the noise accompany the speech signal [26]. This system is based on classical speech enhancement approaches. However, recent researches [5, 4, 37, 42] have shown that the speech enhancement process has been greatly improved by using deep learning techniques instead of the classical de-noising approaches. The basic difference between deep learning and the classical approaches for speech enhancement is that these classical approaches are based on statistical assumption of the noise accompanying the speech. On the contrast, deep learning approaches predict the complex nonlinear function, which maps noisy speech to clean speech, without prior knowledge of the statistical relationship between speech and noise. For this reason, deep learning techniques are more powerful and generalized than the classical ones. As a result, deep learning based speech enhancement is a promising technique to be used in future hearing aid devices [22, 32, 35, 40].

The rest of this paper is organized as follows. The related work and the scope of the paper are presented in section 2. The idea of using a diversity of languages in the training of a deep neural network for speech enhancement is suggested in section 3. The noise classification using convolutional neural network is proposed in section 4. The final design for the environmental noise adaptable hearing aid device is described in section 5. The experimental work is explained in section 6. Results and discussion are detailed in section 7. Finally, the paper is concluded in section 8, along with suggestions for future work.

2. Related Work

There are many features nowadays available in hearing aids that make it more adequate for users' needs. Hands-free technology is one of the powerful features that was added to modern hearing aids, which automatically adjusts the device based on the user's listening environment [12], whether the user on the phone, in a restaurant, in a crowd, or a windy area. The device can be programmed to meet personal requirements. However, all these features are still limited, because they are based on classical signal processing techniques [26]. Moreover, every added feature may require a separate electronic circuit to be performed [33].

Many speech enhancement deep neural networks have been proposed recently, and some of them are proved to be very effective and managed to remove all the noise from the speech signal. For these reasons, introducing deep learning signal processing techniques in hearing aids will develop more advanced features that will lead to users' satisfaction; this is the main idea of the work presented in this paper.

The learning procedure of any deep neural network requires a huge amount of data to give a significant improvement in the performance. Regarding the speech enhancement process, used inside all the hearing aids devices, a large dataset of pairs of clean and noisy speech samples are required to learn the mapping function that maps noisy speech to clean speech. There are two common problems that may arise when training a deep neural network, namely: variance problem and bias problem [1]. Variance is the problem of overfitting the training dataset, which means the network is performing very well on the training data, but unable to generalize this good performance on unseen test data. A technique called regularization is used to overcome this problem, and the most common one used nowadays is called Dropout regularization [6]. In dropout, the network randomly drops a certain percentage of the hidden units, in the hidden layers, during the training process. In this way, the learning process becomes more efficient, because dropout regularization prevents the dependence on only some specific features during the training, and in turn makes the network more robust to the changes in the test set. Although this technique negatively affects the network performance on the training set, it improves the network generalization capability.

On the other hand, bias is the problem of underfitting the training dataset. This means that the network is unable to perform the required task. To solve this problem, a more complex network, a deeper network or a network with many hidden units, was used in [1]. This option may not always work, and most importantly, it is not practical to increase the complexity of the network due to hardware restrictions. The other option to solve this issue is to increase the size of the dataset. This was proved to have a positive impact on network performance. However, it will be difficult to collect a large clean speech dataset for each language spoken worldwide. Some researches [17, 38] tried to solve the availability of dataset problem by using what is called transfer learning, applied in [15]. They used the parameters of a pre-trained network on any other language as initialization to their network's parameters, and then fine tune the network using a small dataset of the target language. The problem of this approach is its complexity. Also, it is not granted that any trained model on other languages can be fine-tuned to the target language, and gives good performance without distorting the speech. Moreover, a collection of a small dataset for the target language is still needed for this approach.

In addition, sensorineural hearing loss affects the hair cells of the inner ear, which have the functionality of analysing the frequencies in the received sound. As a result, all the frequencies in the received sound, that are supposed to be analyzed by the affected hair cells, will not be heard properly [9], and patients may suffer from lower ability to receive high, mid, or low frequency sounds. While other patients may experience something called recruitment, which means the ear is unable to hear low intensity sound (below 50dB). Furthermore, they find high intensity sounds, higher than 80dB, unbearably loud. Some patients also may have a lower ability to hear any kind of speech, which is known as profound hearing loss. As a conclusion, the hearing impairment differs from one patient to another due to the reason of the deficiency in the ear functionality and the severity of the hearing loss [28]. Based on this fact, hearing aids should be adapted according to the type of hearing loss and its level. The Audiologist can use tests, such as audiometry test [25], to detect the ability of a patient to hear different sounds, and recognize which ranges of frequencies are not received appropriately. The work in [32, 40] applied this idea using classification approach.

The above survey reveals the contribution of deep learning in the speech enhancement field. However, available speech enhancement techniques, employing deep neural networks, suffer from the problems listed below.

a) They require a large dataset of clean and noisy speech for training, which is difficult for each language worldwide.
b) They may suffer from the variance and bias problems mentioned previously.
c) They lack the ability to classify the sound frequency

range, which is an important feature that should be added to hearing aid devices, as discussed earlier. Consequently, having a hearing aid that can classify the type and frequency of a certain noise environment will be more satisfactory to the patient's needs.

This paper attempts to solve the above mentioned three problems. It develops an environmental adaptable hearing aid using deep neural networks by making an integrated speech enhancement and noise frequency classification system to be embedded in the hearing aid device. Where, the hearing aid should treat different frequency sounds according to the case of each patient, and this can be achieved by adding the frequency classification stage in the hearing aid device.

The proposed system first enhances the speech signal by using deep neural network by removing any accompanying noise. Afterwards, it classifies the frequency and the type of the noise environment using two cascaded classifiers based on convolutional neural network. Finally, to output an improved version of the noisy speech, the noise is multiplied by certain weights based on each patient's needs and the importance of the noise.

In order to avoid the problem of collecting a huge dataset for the target language, the deep neural network used in this work was trained using an online available dataset with a diversity of languages. These proposed ideas are presented in the following sections.

3. The Suggested Speech Enhancement DNN using Diversity of Languages

Speech enhancement is one of the most challenging tasks in the signal processing field. Recently, many researchers used deep learning to solve this challenging task. The training of the deep neural network is based on feeding the network with a large number of pairs of clean and noisy speech for the network to learn the mapping function that maps noisy speech to clean speech [11]. Since neural networks learn similarly as human brains, the idea that attracts our attention in this research is that the brain can distinguish between noise and speech of any language even if this language is unknown to the brain. This means there are specific features for any kind of speech that make it different from noise. Based on this idea, a deep neural network for speech enhancement was trained in this work three times, using different speech corpora as an input to address the effect of using a diversity of languages in the training process.

The deep neural network used in this work is based on multi-layer perceptron architecture, as the commonly used architecture for speech enhancement in recent studies [35, 37, 39]. The network has three fully connected hidden layers with ReLU activation function, and a final output fully connected layer with linear activation for prediction. Dropout is the regularization technique used, which randomly drops percentage of the hidden units during the training process to make the network more robust to the changes in the test set, and hence avoid overfitting by overcoming the variance problem, and hence make the network more generalized. A 20% dropout was used in the three hidden units. Adam optimizer is the algorithm used to decrease the mean square error between the clean speech signal and the estimated clean speech signal. Spectrogram mapping is the feature extraction method used, and the features were normalized to zero mean and unit variance before entering the network. The phase angle of the noisy speech was separated to be added at the end after the enhancement process. It is assumed in this work that our ear is not sensitive to the changes that happened in the phase angle because of the noisy environment [34].

This network was trained with three different datasets. The first dataset used English speech corpus. The second dataset, half of it is composed of English speech and the other half has 175 other languages. The third dataset employed 175 languages, excluding English. A comparison was then made by testing the three trained networks on English speech samples as the target language, to figure out the effect of using other languages rather than the target language in the training process and to solve problem (a), mentioned in section 2.

Furthermore, this proposed approach is implemented as a solution for both bias and variance problems (problem (b) in section 2). This is due to the fact that using this huge online available dataset will lessen the effort of searching or collecting a large dataset of the target language, leading to a reduction in bias. Furthermore, exposing the network to various languages during the training will prevent the network from overfitting the training dataset, thus reducing the variance.

4. The Proposed Noise Classifier using Convolutional Neural Network

Recently, Convolutional Neural Networks (CNNs) proves to be very effective in audio classification [20]. CNN firstly made for image classification to work with the huge amount of image’s parameters. The power of CNN is that it depends on the idea of convolution, which leads to the usage of less parameter because of two reasons: parameter sharing and sparsity of connections. Parameter sharing means that a feature takes the advantage of other features in a certain part of the image, and uses it in another stage in the network training. While sparsity of connections means that the output value in each layer does not depend on all inputs of the previous layer, but instead depends only on a small number of inputs [41]. The basic convolutional neural network consists of three main layers: convolution layer in which the convolution operation is done to the input features; pooling layer that is used to further reduce the dimensionality of the convolution layer output by keeping only important information and discarding the others; and the last layer is fully connected layer that is responsible for the classification process and generating the output [11].

When dealing with the audio signal, there are varieties of features to represent the data. These acoustic features represent the important information in any sound, which leads to a compact representation of the sound [10]. In this paper, Mel Frequency Cepstrum Coefficient (MFCC) is the audio feature used for the classification purpose.

MFCC is one of the widely used acoustic features in today's research [2]. The Mel scale maps the normal frequency representation to a representation more similar to the way our auditory system deals with different frequencies. The relation between standard frequency and frequency on the mel scale is given in Equation (1) as follows:

$$F_{\mathrm{mel}} = \frac{1000}{\log(2)}\left(1 + \frac{F_{\mathrm{Hz}}}{1000}\right), \qquad (1)$$

Where F_{mel} is the mel frequency and F_{Hz} is the normal frequency. While, Cepstrum *c(n)* is defined as the inverse Fourier transform of the log spectrum of a time signal *s(n)*, and this is defined in Equation (2) as follows:

$$c(n) = F^{-1}\{\log|F\{s(n)\}|\}. \qquad (2)$$

When calculating the MFCCs, the Mel-scale and the cepstrum are combined, to give perceptual meaningful acoustic features. The steps of extracting the MFCC features are summarized in Figure 1, shown below:

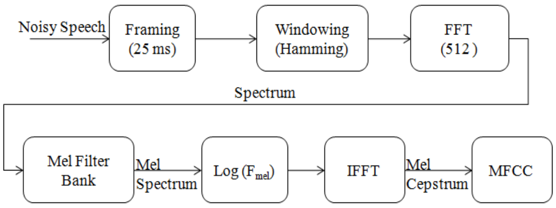


Figure 1. MFCC feature extraction method.

After extracting the MFCC features, the features are fed to two deep convolutional neural network cascaded classifiers. Each of the two classifiers consists of two convolutional layers with ReLU activation function, followed by max pooling layer for dimensionality reduction [20]. The final two layers are fully connected: the first with 500 activation units and ReLU activation function, and the second is the output classification layer with linear activation function.

It may be observed that two cascaded classifiers are suggested in this paper. The first classifier is designed to classify the frequency of the background noise as one of three classes: low, mid, or high frequency, and then a second classification stage is carried out to classify the environmental noise itself. This proposed two stages noise classification solves problem (c), discussed in section 2, and makes the hearing aid more adaptable to each user case and needs, as will be illustrated in the results section. The full block diagram of the proposed adaptable hearing aid is depicted in Fig. 2, and will be discussed in details in the following section.

5. The Proposed Environmental Noise Adaptable Hearing Aid

This section compiles the approaches proposed in sections 3, and 4, to obtain the proposed design of the environmental noise adaptable hearing aid as illustrated in Figure 2. The deep neural network, explained in section 3, is first used to perform the speech enhancement process to produce clean speech. Following the speech enhancement process, the clean speech is subtracted from the noisy speech resulting in the background noise. This background noise signal is then passed through two classification stages. The first stage classifies the noise as one of three broad categories based on its frequency range, as explained below.

a) Natural noise sounds, which are of low frequency range, less than 500 Hz
b) Animal noise sounds, which are of medium frequency range, between 500-1500 Hz.
c) Urban noise sounds, which are of high frequency range, higher than 1500 Hz.

The average dominant frequency [18, 21] of each sound is stated in Table 1, and the spectrograms of the sounds are shown in Figures 3, 4, and 5.

The second classification stage consists of three classifiers; each categorizes the type of the environmental noise (Figure 2). Then, this noise is multiplied by a weight according to its frequency and type. Finally, this weighted noise is added to the output clean speech from the speech enhancement stage to have an improved version of the noisy speech. This weight should be adapted to each user's case, and based on the importance of the noise environment. The noise could be directly multiplied by a weight and added to the clean speech signal based on the first classification stage; however, the second classification stage is used to satisfy the needs of users who will be interested in hearing certain sounds over others. Additionally, the third category in the first classification stage, namely the urban noise, serves as an alerting device as it detects emergency important noise. This adds to the hearing aid device a smart feature [14]. These important kinds of noises will be multiplied by higher weight to act as an integrated alerting circuit. Moreover, the first classification stage represents a deep learning implementation of Hands-free technology that is available in nowadays hearing aids. Based on this classification, the listening environment can be detected, and accordingly, the device can be automatically adjusted. Consequently, a deep learning based environmental noise adaptable smart hearing aid can be developed using the proposed design. The details of the experimental work are presented in section 6.

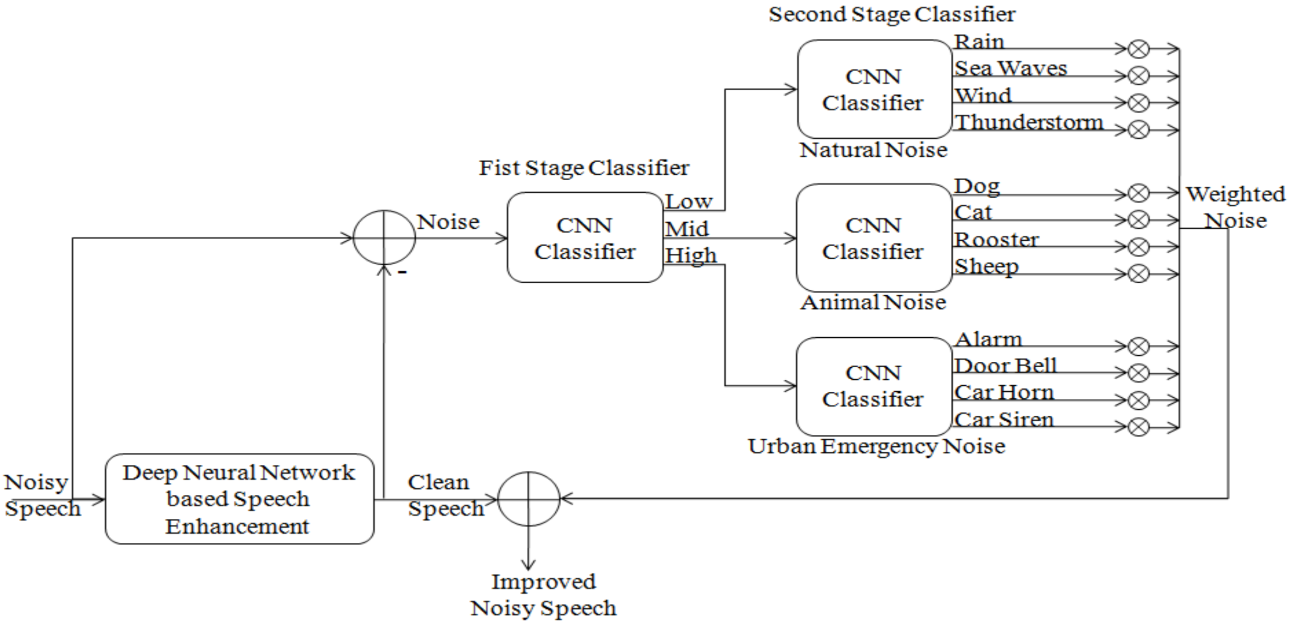
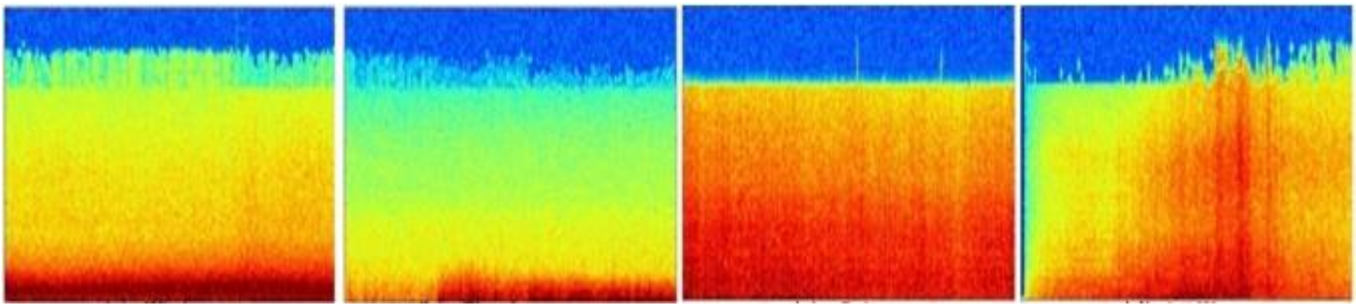
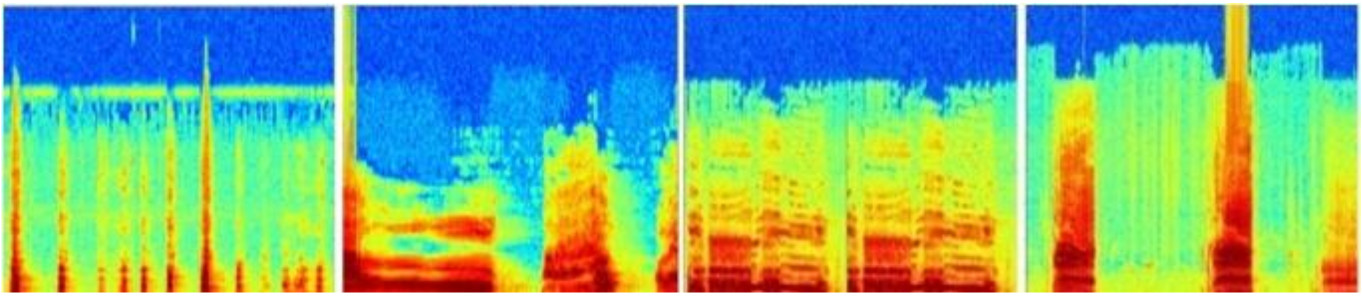


Figure 2. A block diagram of the proposed system.



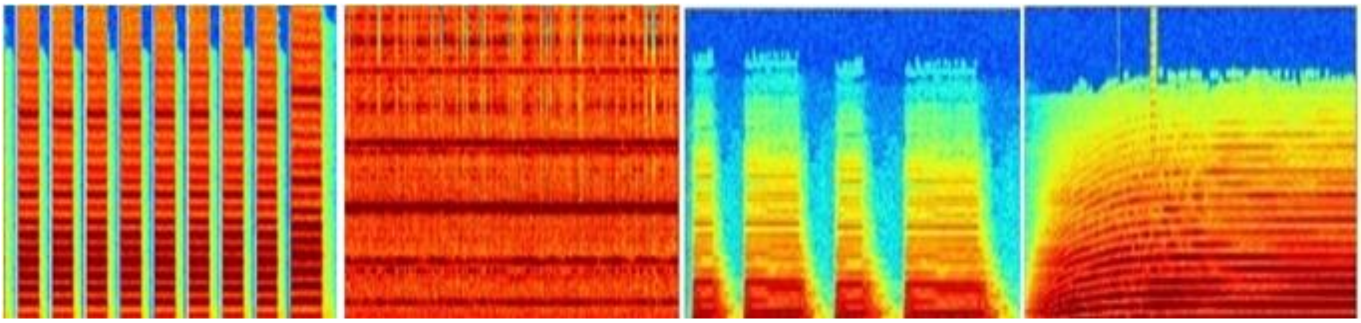
a) Wind. b) Thunderstorm. c) Rain. d) Sea waves.

Figure 3. Spectrograms of nature low frequency sounds: classified by the first classifier.



a) Dog. b) Cat. c) Rooster. d) Sheep.

Figure 4. Spectrograms of animal mid frequency sounds: classified by the second classifier.



a) Alarm. b) Door bell. c) Car horn. d) Car siren.

Figure 5. Spectrograms of urban noise high frequency sounds: classified by the third classifier.

Table 1. Average dominant frequency of each sound.

Sound	Dominant Frequency (Hz)
Dog	765.5
Cat	1438.4
Rooster	1350.1
Sheep	789.2
Rain	423.1
Sea Waves	377.9
Wind	278.9
Thunderstorm	114.1
Alarm	2540.9
Door Bell	1843.8
Car Horn	1500
Car Siren	1687.5

6. Experimental Work

The clean English speech data were collected by randomly selecting audio samples of about 3 hours from the Voice Bank Corpus [31], this corpus consists of 400 English sentences for each of 28 English speakers, 14 male and 14 female, and another 56 different accent speakers, 28 male and 28 female, from Scotland and United States.

Another clean speech corpus [29] was used, which has 66,176 files, each of them contains approximately 10 second of speech recorded in 1 of 176 possible languages spoken worldwide. A total of 3 hours audio samples were also randomly selected from this dataset, which were chosen to be a collection of the available languages, excluding English.

For test purposes, we have chosen 75 speech utterances from LibriSpeech ASR corpus [16], which is a 1000 hours corpus of clean read English speech.

On the other hand, 128 noise environments were collected from three different datasets which are: the 100 environmental noise dataset [7], the USTC-made 15 noise types dataset [36], and the NOISEX-92 corpus [30], from which we selected another 13 noise types. The first two datasets had been used for the training procedure, while the last dataset for testing purpose.

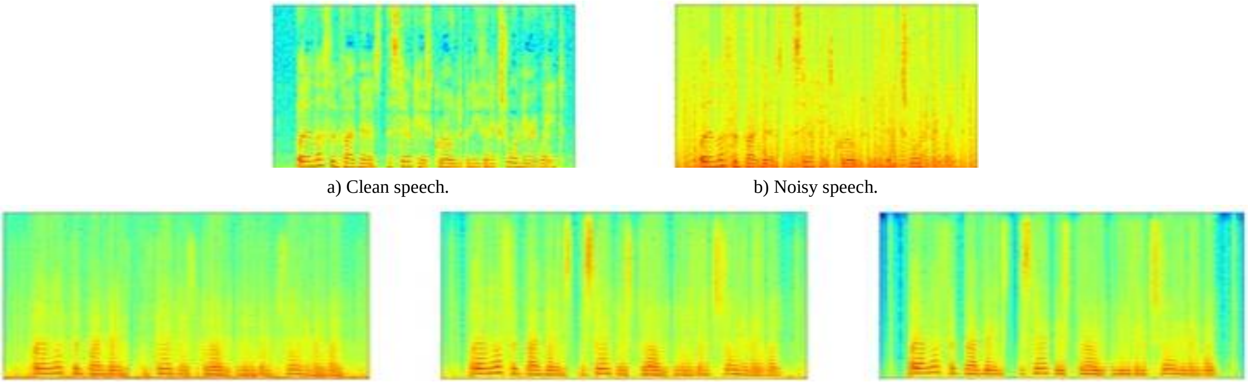
All the collected audio files were truncated to be 3 seconds in length, and re-sampled at 16 KHz. In all cases, the deep neural network was tested on the unseen 75 English speech audio samples, taken from LibriSpeech ASR corpus, and each of them was corrupted with unseen 3 noise types from the NOISEX-92 corpus, to form a total of 225 noisy speech samples, used in the test stage. Concerning the training process, the deep neural network was first trained on 4000 pairs of clean and noisy speech audio samples, which were created by mixing each of 2000 speech audios with two randomly selected noise environments from the 115 noises used in the training process. After that, we doubled the training dataset and trained the network a second time. In both cases, the deep neural network was trained three times, using three different training sets. In the first time, the network was trained on English only dataset, taken from the Voice Bank English corpus. In the second time, the training set was split into two halves, half of clean English speech samples, taken from the voice bank English corpus, and the other half was from the 176 different languages corpus. In the third time, it was trained on speech audio samples, which all were taken from the 176 different languages corpus, excluding English Language.

The noise environments used to train the classifiers were collected from ESC-50 dataset [19] and from urban noise dataset [23] to make a total of 400 samples of each broad category in the first stage, and 40 sample of each class in the second stage. The dataset was divided to 60% for training set, 20% for validation, and 20% for testing. The obtained results are discussed in the following section.

7. Results and Discussion

7.1. Speech Enhancement Stage Results

The speech quality was evaluated based on the three well known scores: Perceptual Evaluation of Speech Quality (PESQ) [8], Short Time Objective Intelligibility (STOI) [27], and Log Spectral Distortion (LSD) [3]. All these measurements were done using six values of Signal to Noise Ratios (SNRs) ranged from -5 to 20 with a step of 5. The speech audios were tested on three unseen noise environments which are: Machine gun, Volvo 340, and HF channel, and then the average was computed. In all the spectrogram graphs illustrated in Figure 6, the x-axis represents the frequency and the y-axis represents the power spectral density.



a) Clean speech.

b) Noisy speech.

c) English only trained network enhanced speech.

d) Half/ half trained network enhanced speech.

e) Language diversity trained network enhanced speech.

Figure 6. Spectrograms of clean speech utterance (a) and its noisy version (b) with unseen Volvo 340 noise, tested at 0 dB SNR, when the network was trained on 3 hours of speech, in the three cases of the training datasets, English only (c), half English and half different languages (d), and 176 different languages (e).

PESQ is an objective method of measuring speech quality, its score ranges from -0.5 to 4.5 and the higher the score, the better the speech quality. It can assess the speech quality by making a comparison between the original and degraded speech. PESQ results are shown in Table 2 for the three training sets used in the training procedure. It is found that using a training set of half English speech and half of the other languages gives the highest PESQ value. Moreover, training the network with diversity of languages, excluding the target language, produces slightly higher PESQ values than the network trained on the English dataset only.

Table 2. Average PESQ comparison between the three used datasets in the training process, in case of 1.5 and 3 hours training, at different SNRs using three unseen noise environments.

SNR	1.5 Hours Training			3 Hours Training		
	English	Half/Half	Mixture	English	Half/Half	Mixture
SNR 20	2.3586	2.4530	2.3839	2.4258	2.5146	2.3976
SNR 15	2.2982	2.3685	2.3059	2.3606	2.4306	2.3266
SNR 10	2.1746	2.2567	2.2007	2.3070	2.3204	2.2105
SNR 5	1.9791	2.1131	2.0602	2.0155	2.1638	2.0569
SNR 0	1.7212	1.9078	1.8488	1.7300	1.9462	1.8537
SNR -5	1.4376	1.6192	1.5757	1.4068	1.6993	1.6060
Ave	**1.9949**	**2.1197**	**2.0625**	**2.0410**	**2.1792**	**2.0752**
STDEV (σ)	**0.3584**	**0.3121**	**0.3047**	**0.4051**	**0.3099**	**0.3017**

STOI is another measure that evaluates the intelligibility of the enhanced speech after removing the noise, which means how many words could be interpreted from this processed speech. It ranges from 0 to 1, and the higher the value, the better the speech intelligibility. Table 3 reveals the results of the STOI, in which again the network that was trained on half English speech and half of the other languages gave the best STOI values, and the network that was trained on diversity of languages resulted in slightly higher STOI values than the English only trained network, so these results matched with the PESQ results, given previously.

LSD is used to measure the amount of distortion happened after processing the speech signal, so lower values indicate low distortion, and higher values point out to high distortion. In Table 4, the results of the LSD for the three datasets are given. It is observed that the dataset of half English and half other languages makes the network outputs a speech with the lowest level of distortion, and this distortion decreases when increasing the size of the dataset. On the other hand, the network trained on a diversity of languages outputs speech with a bit higher distortion than the other two.

Table 3. Average STOI comparison between the three used datasets in the training process, in case of 1.5 and 3 hours training, at different SNRs using three unseen noise environments.

SNR	1.5 Hours Training			3 Hours Training		
	English	Half/Half	Mixture	English	Half/Half	Mixture
SNR 20	0.8087	0.8392	0.8341	0.8214	0.8552	0.8423
SNR 15	0.8058	0.8281	0.8212	0.8156	0.8458	0.8298
SNR 10	0.7967	0.8086	0.8022	0.8043	0.8286	0.8111
SNR 5	0.7772	0.7785	0.7737	0.7790	0.7971	0.7814
SNR 0	0.7343	0.7328	0.7264	0.7301	0.7492	0.7339
SNR -5	0.6576	0.6684	0.6506	0.6560	0.6765	0.6581
Ave	**0.7634**	**0.7759**	**0.7680**	**0.7677**	**0.7921**	**0.7761**
STDEV (σ)	**0.0586**	**0.0651**	**0.0692**	**0.0641**	**0.0684**	**0.0696**

Table 4. Average LSD comparison between the three used datasets in the training process, in case of 1.5 and 3 hours training, at different SNRs using three unseen noise environments.

SNR	1.5 Hours Training			3 Hours Training		
	English	Half/Half	Mixture	English	Half/Half	Mixture
SNR 20	1.0190	0.9697	1.1080	1.0076	0.8417	1.0210
SNR 15	1.0311	0.9839	1.1096	1.0333	0.8657	1.0404
SNR 10	1.0540	1.0422	1.1337	1.0536	0.9007	1.1213
SNR 5	1.0910	1.1230	1.1484	1.0753	0.9573	1.1754
SNR 0	1.2666	1.1641	1.2329	1.2131	1.1174	1.3237
SNR -5	1.4644	1.2124	1.4435	1.3772	1.3482	1.4646
Ave	**1.1544**	**1.0826**	**1.1960**	**1.1267**	**1.0052**	**1.1911**
STDEV (σ)	**0.1768**	**0.0992**	**0.1296**	**0.1422**	**0.1948**	**0.1727**

The results of the three metrics: PESQ, STOI, and LSD prove that using a diversity of language can serve as an alternative way of using a huge database of the target language. This will save much time for researchers to search for a huge database of the target language or try to collect large number of clean speech audio, because the results are very close for the three used datasets. The results also support the assumption that using diversity of language may enhance the network performance and the generalization ability, especially when increasing the dataset size.

The spectrograms of the trained network with the three datasets are shown in Figure 6, in the case of 3 hours training. These spectrograms also prove that exposing the network to diversity of languages enables it to better eliminate the noise accompanied with the speech.

From these results, it may be interpreted that the speech possesses unique features, regardless of the language, that are captured by the network. Training the network with the target language and some other languages seems to render the network robustness to any changes in the dataset. Moreover, this allows the network to learn all the speech features by re-tuning its parameters to better differentiate between speech and noise, instead of overfitting some specific features in the target language. This will solve the variance problem, and increase the generalization capacity of the network.

Furthermore, training the network with diversity of languages, without the target language, did not greatly affect the network performance. On contrast, as the results show, it gave better performance with respect to some evaluation metrics. Consequently, the proposed method of training the deep neural network for speech enhancement by using diversity of languages can be implemented as a substitute to any unavailable target language dataset (problem (a) in section 2) and as a solution for both variance and bias problems (problem (b) in section 2).

7.2. Noise Classification Stage Results

The accuracy of the first stage classifier depicted in Figure 2, which classifies the broad category of the sound, was found to be 83% for unseen sound environment. The second classification stage consists of three classifiers. The accuracies of the first classifier, which classifies four nature sounds; the second classifier, which classifies four animal sounds; and the third classifier, which classifies four urban noise environments, were 75%, 83% and 80%, respectively. All these three classifiers' accuracies were evaluated on unseen noise environments. The overall accuracy of the system was measured by running the whole system several times, using unseen noise environments, and it was found to be 70%. These accuracies are all summarized in Table 5.

Table 5. Accuracy of each classification stage.

	Accuracy	
First Stage	**83%**	
Second Stage	First Classifier	**75%**
	Second Classifier	**83%**
	Third Classifier	**80%**
Whole System	**70%**	

The accuracy was not high enough in the second stage because of the small dataset used. This is due to the fact that only 60% of the 40 audio files of each sound contributed in the training process. This means that only 24 audio files, each of 5 seconds in length, were used, leading to about 2 minutes of training or less. These audio files contain silent periods that were deducted before the training. Thus, the actual training was even less than 2 minutes. As a result, this training data was not enough, and lead to this low accuracy. Moreover, the test set is of 10 randomly chosen sound files, so if the system fails to classify only 2 audios, this will lead to 80% output accuracy.

To increase the data size, manual data were collected, and the network was retrained. However, the accuracy remained approximately the same due to data redundancy, because for example there is no variety in the rain sounds online available. As a result, this data scarcity problem affected the overall performance of the system. However, these results can be improved by collecting larger amount of non-redundant data, as proved by any deep learning approach. The results also depend on the sounds required to be classified, because as these sounds are of similar features, the classification process becomes more challenging. This problem can be clearly observed when studying the spectrograms, of the sounds in each broad category, in Figures 3, 4, and 5. It is noticed that the first category, which is classifying four sounds of nature, has almost similar random representation. For this reason, the classifier for this category resulted in the least accuracy. This compares favourably to the work done in [36], which considered the whole ESC 50 dataset, and resulted in an accuracy of 82%.

It should also be mentioned that the second classification stage, which classifies the type of the noise environment, can be deducted from the system in order to decrease the complexity. However, the second classification stage at the same time adds a more advanced feature to the hearing aid device, which is the ability to detect specific noise environments, and multiply them by a weight based on each user needs.

8. Conclusions

In this paper, the idea of environmental adaptable hearing aid using deep neural networks was introduced. A study was conducted first to show how the performance of a deep neural network is affected by using many languages in the training process for speech enhancement. The network was trained with three different datasets: one with English speech only; one with half of English and half of other languages; and the last with a mixture of many languages, excluding English. The results illustrated that the network better detects the clean speech when exposed to a diversity of languages during the training process. This approach is supposed to be used to make the network more generalized by overcoming the variance problem. Furthermore, this could be a proposed solution for the scarcity problem of datasets for some target languages. The removed noise from the speech enhancement network was then analyzed using two cascaded convolutional neural network classifiers. The first classifier classifies the noise frequency, and the second classifier detects the type of noise. Based on this two stages classification, the background noise is multiplied by a weight that mainly depends on each user's needs and the importance of the noise environment. Consequently, the final output of the system is an improved version of the noisy speech. By implementing this design, the hearing aid device can be adaptable to each patient. In addition, it will be smart as it is able to detect emergency noise and make it audible. Finally, this will result in a more satisfying performance for the hearing aid users. The results show that the proposed system is applicable, and further improvements may be achieved by increasing the size of the environmental noise training dataset.

As a future work, collecting large dataset for noise environments is very crucial to be able to apply the idea of adaptable hearing aid with better performance. This dataset should be also of a variety of conditions to avoid data redundancy. Moreover, convolutional neural network based noise classifiers are proved to give good performance in hearing aid. It is suggested to compare the obtained results to those obtained when classical classifiers are employed. Furthermore, most of DNN based speech enhancement networks managed to map noisy speech to clean speech; however, due to hardware implementation restriction, complexity will be the main issue. The speech enhancement process is used for applications like, mobile communication and hearing aids, so the algorithm must fit a certain amount of memory, and run on a small processor that can be attached to the product. As a result, researchers should find a solution to decrease deep neural network complexity.

References

[1] Charniak E., *Introduction to Deep Learning*, MIT Press, 2019.

[2] Dave N., “Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition,” *International Journal for Advance Research in Engineering and Technology*, vol. 1, pp. 1-4, 2013.

[3] Du J. and Huo Q., “A Speech Enhancement Approach Using Piecewise Linear Approximation of an Explicit Model of Environmental Distortions,” *in Proceeding of the 9th Annual Conference of the International Speech Communication Association*, Brisbane, 2008.

[4] Fu S., Tsao Y., Lu X., and Kawai H., “Raw Waveform-based Speech Enhancement by Fully Convolutional Networks,” *in Proceeding of the Asia-pacific Signal and Information Processing Association Annual Summit and Conference*, Kuala Lumpur, pp. 006-012, 2017.

[5] Grais E. and Plumbley M., “Single Channel Audio Source Separation Using Convolutional Denoising Autoencoders,” *in Proceeding of the IEEE Global Conference on Signal and Information Processing*, Montreal, pp. 1265-1269, 2017.

[6] Hinton G., Srivastava N., Krizhevsky A., Sutskever I., and Salakhutdinov R., “Improving Neural Networks By Preventing Co-adaptation of Feature Detectors,” *arXiv*, 2012.

[7] Hu G., “100 Nonspeech Environmental Sounds,” *The Ohio State University, Department of Computer Science and Engineering*, 2004.

[8] Itu R., “Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs,” *Recommendation ITU-T P. 862*, 2001.

[9] Johnsson L. and Hawkins J., “Sensory and Neural Degeneration with Aging, As Seen in Microdissections of the Human Inner Ear,” *Annals of Otology, Rhinology and Laryngology*, vol. 81, no. 2, pp. 179-193, 1972.

[10] Leaver A. and Rauschecker J., “Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object Category,” *Journal of Neuroscience*, vol. 30, no. 22, pp. 7604-7612, 2010.

[11] LeCun Y., Bengio Y., and Hinton G., “Deep Learning,” *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[12] Meyer W., “Programmable Hearing Aid with Automatic Adaption to Auditory Conditions,” ed: Google Patents, 1997.

[13] Nadol J. and Joseph B., “Hearing Loss,” *New England Journal of Medicine*, vol. 329, no. 15, pp. 1092-1102, 1993.

[14] Nossier S., Rizk M., Moussa N., and Shehaby S., “Enhanced Smart Hearing Aid Using Deep Neural Networks,” *Alexandria Engineering Journal*, vol. 58, no. 2, pp. 539-550, 2019.

[15] Pan S. and Yang Q., “A Survey on Transfer Learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, 2009.

[16] Panayotov V., Chen G., Povey D., and Khudanpur S., “Librispeech: an ASR Corpus Based on Public Domain Audio Books,” *in Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, pp. 5206-5210, 2015.

[17] Pascual S., Park M., Serrà J., Bonafonte A., and Ahn K., “Language and Noise Transfer in Speech Enhancement Generative Adversarial Network,” *in Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, pp. 5019-5023, 2018.

[18] Peters G., Baum L., Peters M., and Tonkin-Leyhausen B., “Spectral Characteristics of Intense Mew Calls in Cat Species of the Genus Felis

(Mammalia: Carnivora: Felidae)," *Journal of ethology*, vol. 27, no. 2, p. 221-237, 2009.

[19] Piczak K., "ESC: Dataset for Environmental Sound Classification," *in Proceedings of the 23rd ACM international conference on Multimedia*, Brisbane, pp. 1015-1018, 2015.

[20] Piczak K., "Environmental Sound Classification with Convolutional Neural Networks," in *Proceeding of the IEEE 25th International Workshop on Machine Learning for Signal Processing*, Boston, pp. 1-6, 2015.

[21] [21] Pongrácz P., Molnár C., and Miklósi Á., "Acoustic Parameters of Dog Barks Carry Emotional Information for Humans," *Applied Animal Behaviour Science*, vol. 100, no. 3-4, pp. 228-240, 2006.

[22] Sajid S., Javed A., and Irtaza A., "An Effective Framework for Speech and Music Segregation," *The International Arab Journal of Information Technology*, vol. 17, no. 4, pp. 507-514, 2020.

[23] Salamon J., Jacoby C., and Bello J., "A Dataset and Taxonomy for Urban Sound Research," in *Proceedings of the 22nd ACM international conference on Multimedia*, New York, pp. 1041-1044, 2014.

[24] Schreiber B., Agrup C., Haskard D., and Luxon L., "Sudden Sensorineural Hearing Loss," *The Lancet*, vol. 375, no. 9721, pp. 1203-1211, 2010.

[25] Stelmachowicz P., Beauchaine K., Kalberer A., Kelly W., and Jesteadt W., "High-frequency Audiometry: Test Reliability and Procedural Considerations," *The Journal of the Acoustical Society of America*, vol. 85, pp. 879-887, 1989.

[26] Stockham J. and Chabries D., "Hearing Aid Device Incorporating Signal Processing Techniques," ed: Google Patents, 1996.

[27] Taal C., Hendriks R., Heusdens R., and Jensen J., "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.

[28] Tonndorf J., "Acute Cochlear Disorders: the Combination of Hearing Loss, Recruitment, Poor Speech Discrimination, and Tinnitus," *Annals of Otology, Rhinology and Laryngology*, vol. 89, no. 4, pp. 353-358, 1980.

[29] TopCoder. Possible Languages Spoken, http://www.topcoder.com/contest/problem/SpokenLanguages2/trainingdata.zip. Last Visited, 2020.

[30] Varga A. and Steeneken H., "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 communication*, vol. 12, no. 3, pp. 247-251, 1993.

[31] Veaux C., Yamagishi J., and King S., "The Voice Bank Corpus: Design, Collection and Data Analysis of A Large Regional Accent Speech Database," *in Proceeding of the International Conference Oriental COCOSDA Held Jointly with Conference on Asian Spoken Language Research and Evaluation*, Gurgaon, pp. 1-4, 2013.

[32] Vivek V., Vidhya S., and Madhanmohan P., "Acoustic Scene Classification in Hearing aid Using Deep Learning," *in Proceeding of the International Conference on Communication and Signal Processing*, Chennai, pp. 695-699, 2020.

[33] Walden B., Surr R., Cord M., Edwards B., and Olson L., "Comparison of Benefits Provided by Different Hearing Aid Technologies," *Journal of the American Academy of Audiology*, vol. 11, no. 10, pp. 540-560, 2000.

[34] Wang D. and Lim J., "The Unimportance of Phase in Speech Enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679-681, 1982.

[35] Wang D., "Deep Learning Reinvents the Hearing Aid," *IEEE spectrum*, vol. 54, no. 3, pp. 32-37, 2017.

[36] Xu Y., USTC-made 15 Noise Types, https://github.com/yongxuUSTC/DNN-for-speech-enhancement, Last Visited, 2021.

[37] Xu Y., Du J., Dai L., and Lee C., "A Regression Approach to Speech Enhancement Based On Deep Neural Networks," *Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7-19, 2015.

[38] Xu Y., Du J., Dai L., and Lee C., "Cross-Language Transfer Learning for Deep Neural Network based Speech Enhancement," *in Proceeding of The 9th International Symposium on Chinese Spoken Language Processing*, Singapore, pp. 336-340, 2014.

[39] Xu Y., Du J., Dai L., and Lee C., "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65-68, 2013.

[40] Zaman K., Sah M., and Direkoğlu C., "Classification of Harmful Noise Signals for Hearing Aid Applications using Spectrogram Images and Convolutional Neural Networks," *in Proceeding of 4th International Symposium on Multidisciplinary Studies and Innovative* Technologies, Istanbul, pp. 1-9, 2020.

[41] Zeiler M. and Fergus R., "Visualizing and Understanding Convolutional Networks," in Proceedings of *European conference on computer vision*, Zurich, pp. 818-833, 2014.

[42] Zhao H., Zarar S., Tashev I., and Lee C., "Convolutional-Recurrent Neural Networks for Speech Enhancement," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, pp. 2401-2405, 2018.

Soha A. Nossier is currently a PhD student at the University of East London, London, UK and an Assistant Lecturer of Biomedical Engineering at Medical Research Institute, Alexandria University, Alexandria, Egypt. She received a B.Sc. degree in Electrical Engineering in 2014 and a M.Sc. degree in Biomedical Devices in 2019, both from Alexandria University, Alexandria, Egypt. She is interested in speech enhancement and deep learning.

M. R. M. Rizk is an Associate Professor of Electrical Engineering, Faculty of Engineering, Alexandria University, Alexandria, Egypt. He received a B.Sc. degree in Electrical Engineering in 1971, from Alexandria University, Alexandria, Egypt, and a M.Sc. and PhD degrees in Electrical Engineering in 1975 and 1979, both from McMaster University, Ontario, Canada. His area of expertise includes Signal, Image and Video Processing and Neural Networks, and he has more than 100 publications

Saleh El Shehaby is the Head of the Biomedical Engineering Department, Medical Research Institute, Alexandria University, Alexandria, Egypt. He received a B.Sc. degree in Electrical Engineering in 1973, and a M.Sc. and PhD degrees in Computer Engineering, all from Faculty of Engineering, Alexandria University, Alexandria, Egypt. His area of expertise includes pattern recognition and artificial intelligence, and he has many publications in this area.

Nancy Diaa Moussa is an Associate Professor of Biomedical Engineering at Medical Research Institute, Alexandria University, Alexandria, Egypt. She received a B.Sc. degree in 2002, M.Sc. degree in 2007, and PhD degree in 2013, in Electrical Engineering, all from Faculty of Engineering, Alexandria University, Alexandria, Egypt. Her area of expertise includes signal processing and machine learning, and she has many publications in this area.