IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 December 2021, accepted 19 December 2021, date of publication 23 December 2021,
date of current version 18 January 2022.

Digital Object Identifier 10.1109/ACCESS.2021.3137878

The Performance of Wearable Speech Enhancement System Under Noisy Environment: An Experimental Study

PAVANI CHERUKURU, MUMTAZ BEGUM MUSTAFA, (Member, IEEE), AND HEMA SUBRAMANIAM

Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia

Corresponding author: Mumtaz Begum Mustafa (mumtaz@um.edu.my)

This research was financially supported by Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS/1/2020/ICT09/UM/02/1)

ABSTRACT Wearable speech enhancement can improve the recognition accuracy of the speech signals in stationary noise environments at 0dB to 60dB signal to noise ratio. Beamforming, adaptive noise reduction, and voice activity detection algorithms are used in wearable speech enhancement systems to enhance speech signals. In recent works, a word rate recognition accuracy of 63% for a 0db signal-to-noise ratio is not satisfactory for a robust speech recognition system. This paper discusses the experimental study using fixed beamforming, adaptive noise reduction, and voice activity detection algorithms with the inclusion of −10dB to 20dB signal to noise ratio for different types of noises to test the wearable speech enhancement system's performance in noisy environments. It also compares deep learning-based noise reduction methods as a benchmark for speech enhancement and word recognition for different noise levels. We have obtained an average word rate recognition accuracy of 5.74% at −10dB and 93.79% at 20dB for non-stationary noisy environments. The outcome of the experiments shows that the selected methods perform significantly better in the environment with high noise dB for both stationary and non-stationary noise. We found that there is no significant statistical difference between the stationary and non-stationary noise word recognition and SNRs level. However, the deep learning-based method performs significantly better than the fixed beamforming, adaptive noise reduction, and voice activity detection algorithms in all noisy levels.

INDEX TERMS Wearable speech enhancement, beamforming, adaptive noise reduction, voice activity detection, deep learning.

I. INTRODUCTION

Wearable technology refers to devices users can wear, taking the form of an accessory such as jewelry, sunglasses, watches, etc. [1]. As we can use wearable devices in a noisy environment, the speech application systems such as ASR needs to eliminate the noise to improve their performance and robustness. Wearable devices were used in speech enhancement due to the size and proximity with the mouth that reduces the noise while recording the speech [2], [3]. For example, suppose everybody always wears a personal microphone near his/her mouth; the signal-to-noise ratio can be vastly improved, compared with having a microphone at a greater distance [3]. Most hearing devices, such as hearing aid, implement the recent speech enhancement algorithms [4], [5].

We can classify noisy environments as stationary and non-stationary noisy environments in speech recognition. We typically hear non-stationary background noises in our everyday environment, referring to the background noises that we hear in a real conversation. On the other hand, stationary noises resemble the noise on telephone lines [6].

Speech enhancement deals with the noisy speech signals by reducing the background noises while preventing the alterations in speech features. We use speech enhancement for speech signals processing applications like speech coder, automatic speech recognition, voice over IP, hearing aid, and many more. Speech enhancement systems are frequently used as a pre-processer to enhance speech quality. Generally, algorithms used in speech enhancement systems consist of

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

three types, namely filtering algorithms, spectral restoration algorithms, and speech model-based algorithms [7].

We use Filtering algorithms to filter the unwanted noisy signals that attenuate the noise features to produce a clean speech signal. Filtering algorithms contain time-domain filters, frequency-domain filters [7]–[9], and parametric filters [7].

A spectral restoration algorithm is a function to estimate the performance of noise reductions in the frequency domain to gain a clean speech spectrum from noisy speech spectrums. The spectral restoration algorithms include minimum mean-square error log-spectral amplitude estimator (LSA) [10]–[12], minimum mean square error spectral estimator (MMSE) [8], [10], [13], [14], maximum likelihood spectral amplitude estimator (MLSA) [7], [15] and maximum a posteriori spectral amplitude estimator (MAPA) [7], [16], [17].

The speech model-based algorithms combine the functions of speech reduction and user speech production models to cancel the noise features from noisy speech signals. Well-known speech models used for speech enhancement include the harmonic model [7], [18]– [20], the linear prediction (LP) model [7], [21], and the hidden Markov model (HMM) [7], [10]. There are many filtering algorithms, such as Least Mean Square (LMS), Recursive Least Square (RLS), and Normalized Least Mean Square (NLMS), that have revealed their effectiveness in the reduction of noisy signals and improving the quality of speech. However, they may have limited ability to perform well on low signal-to-noise ratio (SNR) conditions.

The existing speech enhancement algorithms use a single microphone to process the speech signals [22]. However, these algorithms are computationally expensive and ineffective in canceling the noisy speech signals, especially at low signal-to-noise ratio (SNR) levels, i.e., −10dB to 10dB [23]. These noises are difficult to filter as it has different characteristics in different environments. So, speech enhancement is very much required [24].

The wearable microphone array, embedded in textiles, consists of multiple microphones used to record the speech signal that gives the best recognition at 10dB SNR than a single microphone [2]. Microphone array and speech enhancement are the components embedded in wearable speech enhancement (WSE) that process speech signals in multichannel under noisy environments such as the outdoor environments [3]. For example, a spectral statistical filter is applied in wearable hearing devices for handling stationary noise environment (Gaussian noise) and non-stationary noise environments (babble, factory, and car) at −5dB to 20dB [4].

The existing WSE system can filter 0 to 60dB of SNR, which gives 62.5% WRR at 0dB considered low SNR and 83% at 60 dB considered high SNR. However, it was tested only with white Gaussian noise but never environmental noise. For instance, Alessandro *et al.* (2016) did not consider all types of noises in a real-time environment.

Although the existing system developed for WSE in noisy environments was only tested for White Gaussian noise (WGN) at 0 to 60dB SNR (which is stationary noise) [5], [3], they were never tested for non-stationary environmental noises such as Airport noise, babble noise, car noise, Exhibition noise, and Restaurant noise.

Currently, Deep Learning has gained attention in signal processing applications, with significant improvements in various phases of signal processing such as filtering, feature extraction, and recognition [39]. Several of the existing works proposed their techniques based on deep learning such as target speech separation supervised machine learning [40], linear prediction coefficient as DeepLPC [41], which uses deep learning approach to estimate the augmented Kalman filter, deep attention features for speech separation [42] and many more [43].

It is essential to experiment with the WSE under non-stationary environmental noises because wearable devices are used mainly in the outdoor environment. WSE contains beamforming to enhance the speech signal by suppressing the noisy signals, adaptive noise reduction (ANR) to filter the noise signals by using suitable filtering algorithms, and voice activity detection (VAD) to detect the speech signals.

In this paper, we are experimenting with the microphone and multi-channel enhancement-based WSE's with Gaussian noise, babble, airport, car, restaurant, and exhibition noise at low level to high-level SNRs. We take advantage of the existing WSE system developed in [25] to test its word recognition rate in both the stationary Gaussian noise and non-stationary noisy environments at −10dB to 20dB SNR using the AURORA database [25]. We consider Deep Learning-based speech enhancement techniques as a benchmark for comparison.

The rest of the paper is organized as follows. Section II explains the research background of wearable speech enhancement systems. Section III describes the experimental design performed on the existing WSE using the Aurora database modified for a noisy environment. The results are presented in Section IV, and we conclude our work in Section V.

II. WEARABLE SPEECH ENHANCEMENT (WSE) SYSTEM

The existing Wearable Speech Recognition system used a feature extraction method tested in real-time environments at three SNR levels (i.e., 15db, 10db, and 5db) [26]. The existing wearable speech enhancement improves the recognition accuracy at high SNR levels [5].

Wearable speech enhancement (WSE) is developed in the wearable platform, and the output from WSE is applied in computing platforms by using speech to text engine (STT). The WSE consists of a squared array microphone set, a recording component to record the speech signals, noisy signals, beamforming, adaptive noise reduction (ANR), and voice activity detection (VAD).

The architecture of WSE, as shown in Figure 1, consists of two platforms: 1. Wearable platform, where the speech

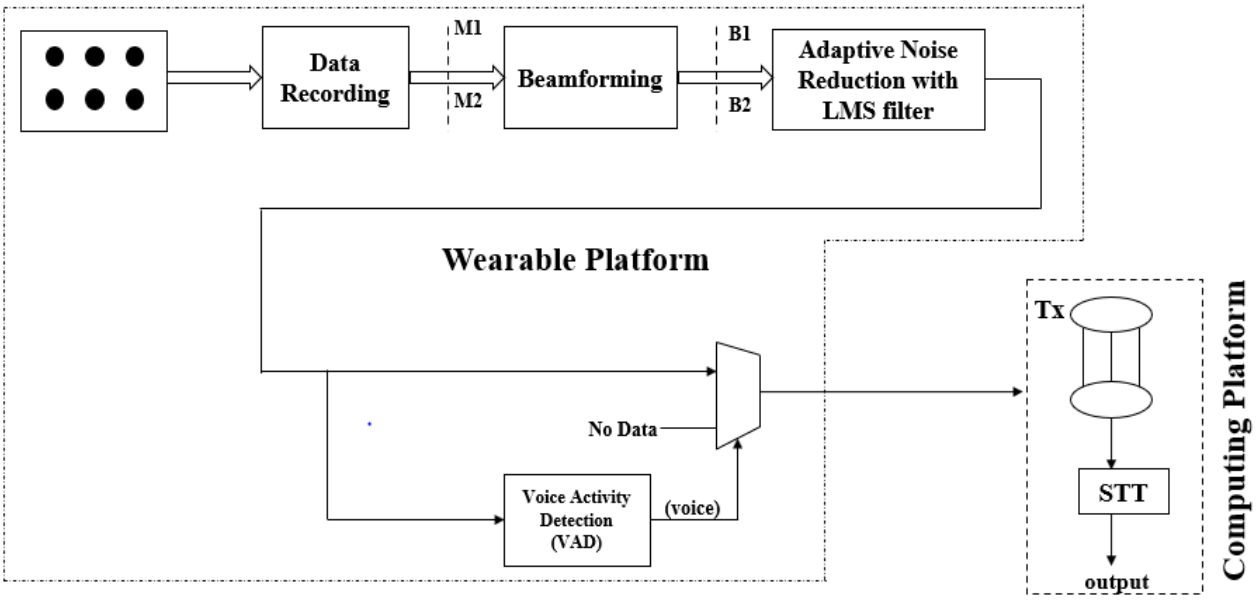

FIGURE 1. The architecture of wearable speech enhancement (WSE) [3].

enhancing process is performed, and 2. Computing platform where the word recognition accuracy process is carried out [3].

A. BEAMFORMING

Beamforming is a signal processor used together with a microphone array to supply the capability of spatial filtering. The microphone array gives the spatial tests of the transmitting signals, which the signal processor controls to produce the output signals of beams [27]. Beamforming can be achieved by filtering the microphone array signals and combining the output signals to extract the desired speech signal and reduce the unwanted noisy signals [28]. In simple terms, beamforming helps to reduce the noisy signal from the speech signal by removing a particular frequency of noisy signals from the speech signal frequency captured from the microphone array.

There are two types of beamforming: adaptive beamforming and fixed beamforming. Adaptive beamforming is where the input signal directivity varies based on the noise signals changes in the environment.

Fixed beamforming is where the input signal directivity is fixed across time, and the distance between the microphones is constant. Fixed beamforming was obtained using the delay and sum beamformer [27]. Based on the microphone array in figure 1, Alessandro *et al.* (2016) developed a fixed beamformer using DMA theory as explained in [3] [5]. The output beam B(t) has been calculated by applying a delay to the microphone M2 and subtracting from M1 as shown in equation (1)

$$B(t) = M1(t) - M2(t\text{-}T)\sum_{0}^{n} x^2(n) \tag{1}$$

The variance in Time of Signal Arrivals between the two microphones is

$$T_0 = \frac{a}{c}\cos\theta \tag{2}$$

where d is the distance between $M1$ and $M2$, c is the sound speed (constant at 20°C) and θ is the input angle. Distance between microphones and input angle is constant.

The first stream is a user beam that contains the highest SNR signals, and another stream is the noise with the lowest SNR signals. These two streams are the input signals to the adaptive noise reduction component. Fixed beamforming is not suitable for outdoor environments as the noises can come from different directions, reducing the capability of noise filtering of WSE.

B. ADAPTIVE NOISE REDUCTION (ANR)

Adaptive noise reduction (ANR) is used as a multi delay block frequency adaptive filter to delete the environmental noises using an LMS filter [32], [33]. User beam and reference noise are the inputs to the ANR. The ANR component filters the noise in the user beam that is consistent with reference noise (speech signal already exists in the user beam as beamforming and not attenuated) [34].

In general, we cannot assume that noise was filtered in a speech signal. In this scenario, the LMS filter used by the ANR partly suppresses and changes the required signal, which depends on the attenuation of the speech signal in the reference beam and user beams. The SNR of the output signal is defined in [3].

$$SNR_{output} = \frac{1}{SNR_{rfn}} \tag{3}$$

where SNR_{rfn} is Signal-to-Noise ratio of reference noise signals.

C. VOICE ACTIVITY DETECTION (VAD)

The Voice Activity Detection (VAD) algorithm separates the user's voice in the user stream [3], which is helpful for two reasons:

(i) Segmentation: the system needs to know the exact boundaries of each word in the spoken utterance.
(ii) Data Reduction: the system only sends data as required and not continuously over the transmission channel.

The VAD is implemented in a time domain, and the recognition is performed every 20 seconds, taking 20 samples per frame to calculate the Zero crossing and Energy characteristics using the angular windows frame function [35], [36].

$$Energy\ Zero\ Crossing = \sum_{0}^{n} sign(x[n] - sign(x[n-1])) \tag{4}$$

where $x[n]$ indicates the number of samples per frame. When Zero Crossing is small and Energy is high, it is categorized as voiced speech signal, otherwise it is deemed to be the unvoiced region of the speech signal.

In discreet time signal processing, zero-crossing occurs if the successive samples of the signal have different algebraic signs. The zero-crossing is a measure of the frequency content of a signal, i.e., the rate at which ZC occurs is the measurement of the frequency content of the input signal. It provides the total count in each time interval that the amplitude of the speech signal passes through the value of zero can be expressed as

$$sign\left[x\left(n\right)\right] = \begin{cases} 1, & x\left(n\right) \geq 0 \\ -1, & x\left(n\right) < 0 \end{cases} \tag{5}$$

D. DEEP LEARNING MODEL

Currently, Deep learning-based schemes such as Recurrent neural networks (RNN), LSTM (Long Short-Term Memory) networks, and convolution neural networks are widely adopted in various applications. These schemes follow the feed-forward learning process, because information flows through the function being evaluated from x, through the intermediate computations used to define f, and finally to the output y. There are no feedback connections in which outputs of the model are fed back into itself [45].

The feature map of these signals is obtained and employed as Mel-filterbank (FBANK) feature. These features are arranged in a two-dimensional array whose size is. These FBANK attributes are considered more robust when compared with the decorrelated features such as Mel-frequency cepstral coefficients (MFCCs).

This model of CNN (Convolutional Neural Network) with the FBANK attribute helps to obtain the learning weights of the hidden layer. Similarly, the convolution layers are based on the weighted sum of the input attributes. The fully

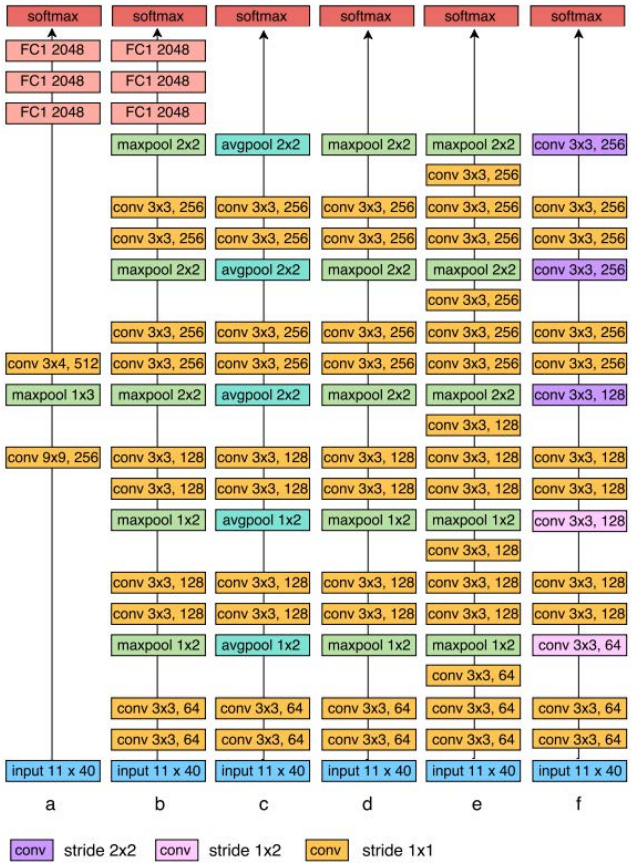

FIGURE 2. Deep learning architecture for speech enhancement and word recognition [44].

connected layers are connected to the sub-region of the input feature map and share the weight across the local receptive fields to extract the same feature across the input representation. The feature map hidden unit value of CNN model can be presented as h_{ij}

$$\begin{aligned} h_{i,j} &= \sum_{k=0}^{m-1} \sum_{l=0}^{m-1} f\left(w_{k,l} x_{i+k,j+l} + b_{i,j}\right) \\ &= f\left(W \otimes X + b_{i,j}\right) \end{aligned} \tag{6}$$

where f denotes the activation function, $w_{k,l}$ denotes the weights of weight matrix W, $b_{i,j}$ denotes the shared bias and $x_{i+k,j+l}$ denotes the input value at position $i + k$ and $j + l$, and $\otimes$ is the cross-correlation field. Based on this model, we have considered 6 different architectures of the convolution model. Figure 2 depicts the architectural model proposed by Rownicka *et al.* [44].

The architecture depicts the (a) CNN model, (b) Very Deep Convolutional Neural Network (VDCNN) with max pooling and fully connected layer, (c) VDCNN with average pooling layer (VDCNN-avg), (d) VDCNN with max pooling (VDCNN-max), (e) VDCNN with max pooling and convolution (VDCNN-max-conv) and (f) VDCNN with convolution layer (VDCNN-conv).

III. METHOD

This research aims to conduct a noise reduction experiment for a wearable speech enhancement (WSE) system in stationary and non-stationary noisy environments at different SNR levels of speech signals. We will perform several enhancement experiments using the existing wearable speech enhancement methods including the Fixed Beamforming, ANR, and VAD algorithms. The experiment intends to examine the WSE system's performance for stationary and non-stationary environmental noises. Our experiment considers the low-level SNRs to high-level SNRs (−10dB to 20dB) using white Gaussian noise, airport noise, babble noise, car noise, exhibition noise, and restaurant noise. The following sections describe the speech database, the experimental setup, and the evaluation method.

A. SPEECH DATABASE

In this research, we have used the Aurora [25] speech database to experiment with the wearable speech enhancement system under noisy environments at different SNR levels of speech signals. We have selected 25 utterances from the Aurora noisy dataset consisting of 13 unique male voices and 16 unique numbers of female voices. While the numbers of trials are different for different noise levels, we have ensured a minimum of 25 samples for each dB level.

The selected noisy speech utterances incorporate five non-stationary environmental noise types: Airport. Babble, Car, Exhibition and Restaurant and one stationary noise type which is the White Gaussian noise at seven different SNRs: −10dB, −5 dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. For −10dB noisy speech signals. We also selected 25 utterances from the AURORA clean training dataset, where we theatrically mixed −10dB noisy signals with the clean training dataset. We have prepared 42 different conditions for every 25 utterances.

For Deep learning methods, we have used the Aurora 4 dataset that contains 7138 utterances from 83 speakers. The clean dataset has an audio duration of about 15 hours. We grouped the test set of this dataset into four different subsets as 330 clean signals selected from SI-84 WSJ corpus (Set A), 330 sets with five types of different noises varying from −10 to 20 dB SNR (Set B), 330 sets of recording with different microphone (set C), and 330 sets of recording with different microphone with added white Gaussian noise (set D). The overall duration of all these test sets is about 9 hours.

B. EXPERIMENTAL DESIGN

The procedure of this experiment begins with the development of wearable speech enhancement (WSE) based on Beamforming, Adaptive Noise Reduction, and Voice Activity Detection algorithms, as shown in Table 1, with the detailed procedure is explained in the section below.

1) BEAMFORMING

In this experiment, we used the directional microphone array (DMA) in the squared array position to capture the signals recorded and stored in the recording component. These recorded signals become the input signals to the beamforming component [3].

TABLE 1. Experimenting the wearable speech enhancement (WSE) using different types of environmental noises at different SNR levels.

Environment	Types of noise	SNR levels (dB)
Stationary Noises	White Gaussian Noise	-10dB,-5dB, 0dB, 5dB, 10dB, 15 dB, 20 dB
Non-Stationary Noises	Airport Babble Car Exhibition Restaurant	-10dB,-5dB, 0dB, 5dB, 10dB, 15 dB, 20 dB

The beamforming device was placed at 10mm to 15mm from the left channel speaker of the sample utterance drive, while the noise generator driver was at the left corner. Python was used to write the code for mixing the speech and noise samples. On the other hand, we used the C programming language to write the beamformer code.

This research only tested the existing fixed beamforming algorithms in the wearable speech enhancement by using stationary and non-stationary environmental noises, as adaptive beamforming was never implemented on WSE. As such, it provides the opportunity for us to improve the WRR accuracy rate in non-stationary environments using adaptive beamforming.

2) ADAPTIVE NOISE REDUCTION (ANR)

The output signals of the beamforming device are the input signals of adaptive noise reduction. We apply the LMS algorithm written in python to filter the noise signals using the MATLAB simulations and ARM processor, which filter noise in real-time.

3) VOICE ACTIVITY DETECTION (VAD)

The output audio signals from the adaptive noise reduction will be the input signals for voice activity detection (VAD). The VAD algorithm, written in python language, identifies the presence or absence of speech in audio signals.

4) DEEP LEARNING APPROACH

The input audio signals are processed through Deep learning methods adopted from [44], which are Convolutional Neural Networks (CNN), Deep Neural networks (DNN), Very deep convolutional neural networks-Fully Connected (VDCNN-FC), VDCNN-avg, VDCNN-max, VDCNN-max-conv, and VDCNN-conv [44]. In VDCNN-avg, the max-pooling layers were replaced with average pooling using the same kernel sizes and strides as in the baseline VDCNN model [44].

In the first phase, all audio signals are trained, where the convolution layers are activation functions are used to learn the weight parameters and transfer these weights.

The fully connected layers are used to learn the non-linear functions in the considered space. Finally, the softmax layer is used to provide the outcome of the Deep learning model.

C. EXPERIMENTAL SETUP

1) DEVICE CONFIGURATION

(a) Controller: stm32f103CBT6
(b) Main clock: 72MHz
(c) Memory: 128KB ROM/ 20KB RAM
(d) External storage: Transcend 8GB class 4 memory card.
(e) Transducers: capacitive electret microphones
(f) Servos: 9G servo

2) SAMPLING SETUP

Two transducers output was pre-amplified, then fed to a single-stage bandpass filter(80Hz-16KHz), then gain adjusted, level shifted to 1.75V, then fed to individual ADC's (analog to digital converter).

We configure ADCs at 12bit vertical resolution and 16000 Samples per second (+/−50 due to clock stability). Data written to SD card via Conversion complete interrupt linked to DMA channel which writes the value in SD card and a copy in Buffer variable defined in RAM. Both ADC sampling times were synchronized. Amplifiers used were based on LM358 general purpose Opamp.

3) VARIABILITY SETUP

Timer 1 PWM channels connect the Two 9G servos at 16bit resolution (Effective usable steps were around 30000 per servo due to higher ARM deflection of Servos). The distance between each microphone is fixed at 10 mm.

4) NOISE AND SAMPLE UTTERANCE SYSTEM SETUP

The primary noise driver is Edifier 2.0 channel speaker. The speech is varied at the amplifier and the noise samples were continuously looped and fed to the amplifier from the BeagleBone Black Single Board.

The speech samples were driven with only left channel speakers of Logitech USB speakers and the BeagleBone Black single-board computer feds the samples.

5) SNR SETUP

The desired SNR (−10dB, −5dB, 0dB, 5dB, 10dB, 15dB and 20dB) was achieved by individually tuning the noise sound amplifier gain control and the sample utterance amplifier gain control by measuring individual Sound Pressure levels (SPL) to calculated values.

D. EVALUATION METHODS

In this research, we used the spectrogram analysis and the Word Recognition Rate (WRR) to evaluate the performance of the wearable speech enhancement system in a noisy environment (stationary and non-stationary noise).

Spectrogram analysis is used to analyze the amplitude of speech signals [37]. We perform the spectrogram analysis for both stationary (White Gaussian Noise) and non-stationary environmental noises (Babble, Airport, Car, Exhibition, and Restaurant) on time-domain using MATLAB.

We test the voiced speech signal received after the voice activity detection with the ASR speech to text engine to determine the word error rate (WER). We calculate the Word error rate to evaluate the performance of the wearable speech enhancement systems. WER is computed as follows:

$$WER = \frac{S + D + H}{N} \tag{7}$$

where N is the total number of words/letters in the sentence, S is the number of substitutions of other words with, D is the number of deletions I is the number of insertions in a sentence.

By calculating the WER, we determine the word recognition rate (WRR) as:

$$WRR = 1 - WER \tag{8}$$

WRR measures the performance accuracy of wearable speech enhancement system.

IV. RESULTS

A. SPECTOGRAM ANALYSIS

A spectrogram analysis shows the spectral illustrations of a time-varying signal [38]. Figure 3 represents the clean speech signal, while Figures 4 to 15 show spectrograms analysis: (a) noisy speech at −5 dB SNR and (b) enhanced speech using adaptive beamforming, ANR, and VAD algorithms.

We have considered different noise types for this analysis as white Gaussian noise, airport noise, Exhibition, Restaurant, Babble, and car noise. This experiment adds the 5dB noise to the original signal and processes it through the considered speech enhancement model.

The filtered signal is also known as the reconstructed signal, which can be used to analyze the performance of the existing approach.

Figure 4 depicts the spectrogram for unfiltered white Gaussian noise in speech signal at −5db, while figure 5 shows filtered white Gaussian noise in speech signal at −5db. From the spectrogram, we can see that speech distortion due to noise has been rectified through speech reconstruction with a suitable filter. Figure 6 shows the unfiltered airport noise in speech signals at −5db, while figure 7 shows the filtered airport noise at −5db.

Figure 8 depicts the unfiltered babble noise in speech signal at −5db, while figure 9 shows the filtered babble noise at −5db. Figure 10 shows the unfiltered car noise in speech signal at −5db, and figure 11 depicts filtered car noise in speech signal at −5db. Figures 12 and 13 depict the unfiltered exhibition noise in speech signal at −5db and filtered exhibition noise at −5db, respectively. Finally, figure 14 depicts the unfiltered restaurant noise in speech signal at −5db, while figure 15 shows the spectrogram for filtered restaurant noise in speech signal at −5db.

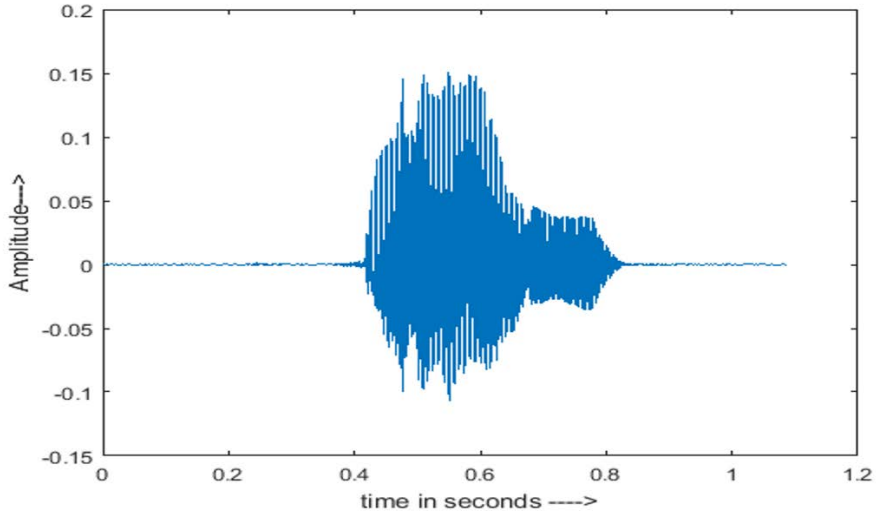

FIGURE 3. Clean speech signal.

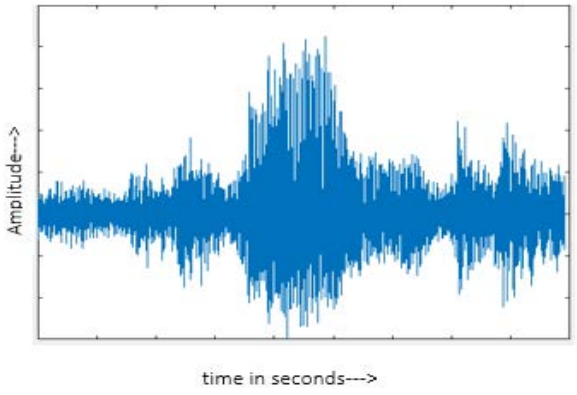

FIGURE 4. The spectrogram for unfiltered white Gaussian noise in speech signal at −5db.

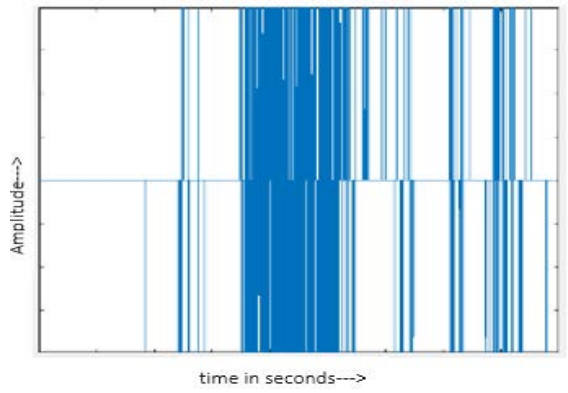

FIGURE 5. The filtered white Gaussian noise in speech signal at −5db.

On the other hand, Figures 16 and 17 show the unfiltered and the reconstruction quality of the deep learning scheme at 10db noise for the airport noise. Figures 18 and 19 show the unfiltered and the reconstruction quality of the deep learning

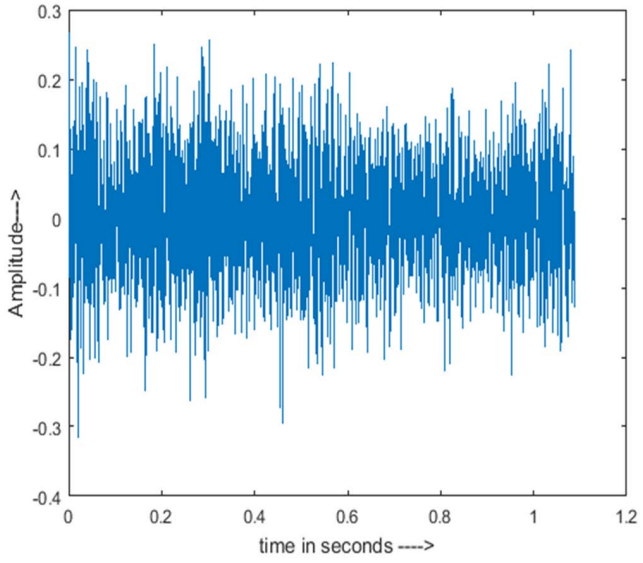

FIGURE 6. The unfiltered Airport noise in speech signal at −5dB.

scheme at 10db noise for white Gaussian noise (WGN), while Figures 20 and 21 show the unfiltered and the reconstruction quality of the deep learning scheme at 10db noise for Babble noise.

B. WORD RECOGNITION RATE

Table 2 and Table 3 show the evaluation outcomes of the experiments using the Beamforming, ANR, and VAD algorithms in WSE at different levels of SNRs under stationary and non-stationary noisy environments.

From Table 2, the WSE with fixed beamforming, ANR, and VAD is very effective at 20dB SNR, but the recognition accuracy gradually decreases at 15dB, 10dB, 5dB, 0dB, −5dB, −10dB, respectively for the Stationary white Gaussian noise.

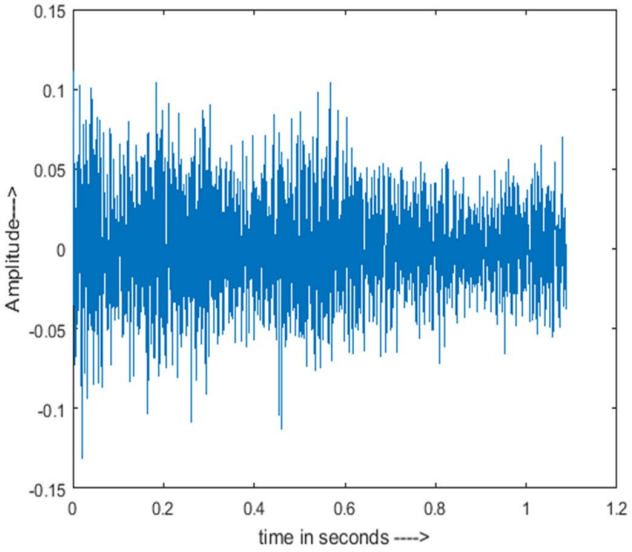

FIGURE 7. The filtered Airport noise in speech signal at −5dB.

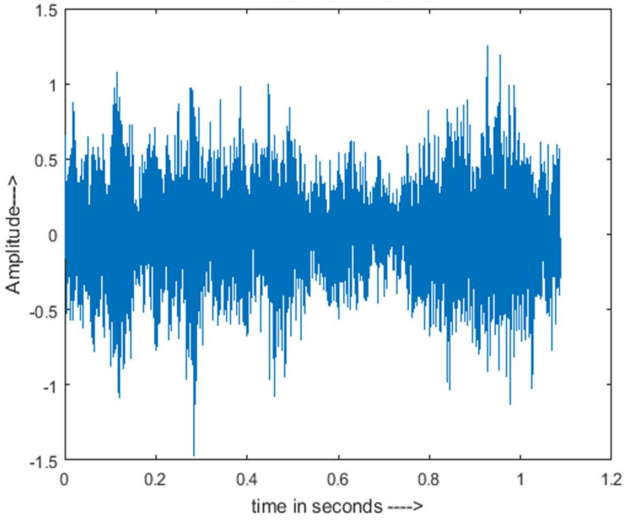

FIGURE 8. The unfiltered Babble noise in speech signal at −5dB.

TABLE 2. Word recognition rate (WRR) for wearable speech enhancement (WSE) under stationary environmental noise.

	SNR/dB	WRR (WSE)	WRR (CNN)
Stationary Noise White Gaussian noise	-10dB	42.6	56.21
	-5dB	58.3	62.20
	0dB	63.4	65.80
	5dB	68.5	71.23
	10dB	72.6	78.55
	15dB	74.8	79.26
	20dB	89.7	92.25

The WRR was lower for the negative dB than the positive dB. We also found that the performance of CNN is better than the WSE for all noise levels.

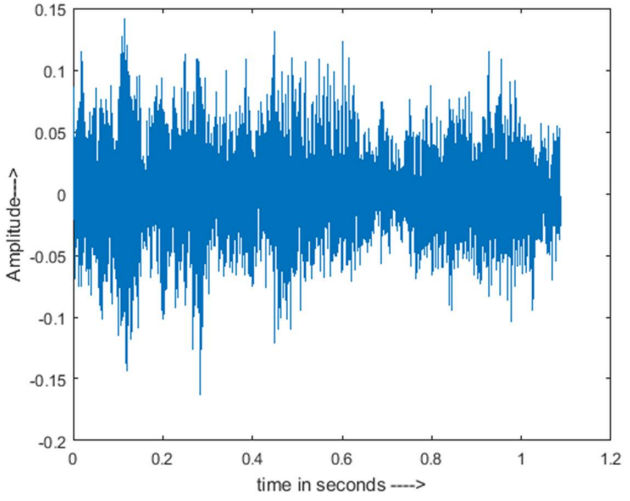

FIGURE 9. The filtered Babble noise in speech signal at −5dB.

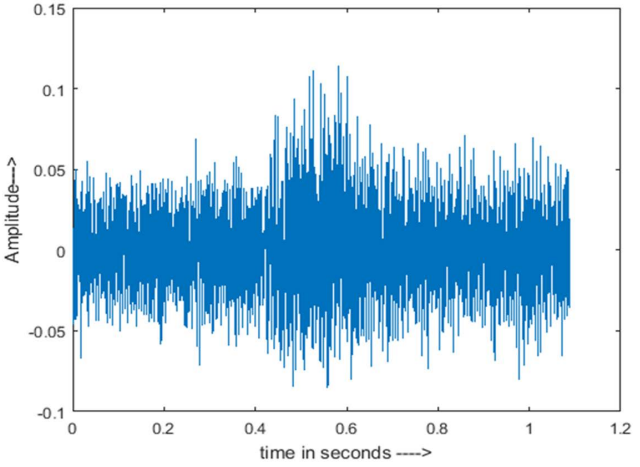

FIGURE 10. The unfiltered Car noise in speech signal at −5dB.

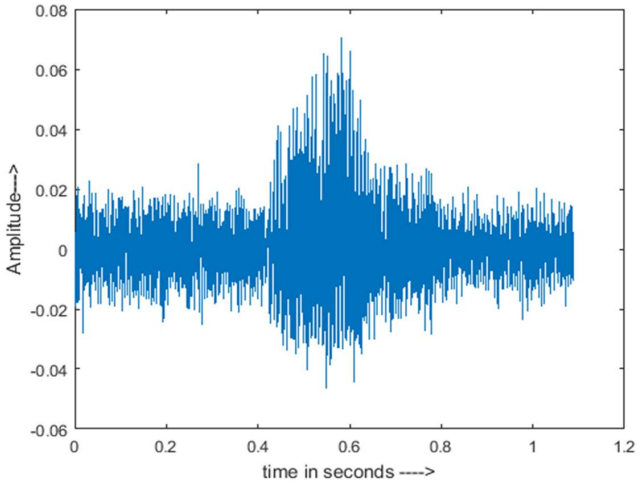

FIGURE 11. The filtered Car noise in speech signal at −5dB.

The scatter diagram shown in Figure 22 shows a linear relationship between the SNR level and WRR (p-value < 0.001).

Similarly, the result in Table 3 also shows that WSE with fixed beamforming, ANR, and VAD is very effective at 20dB

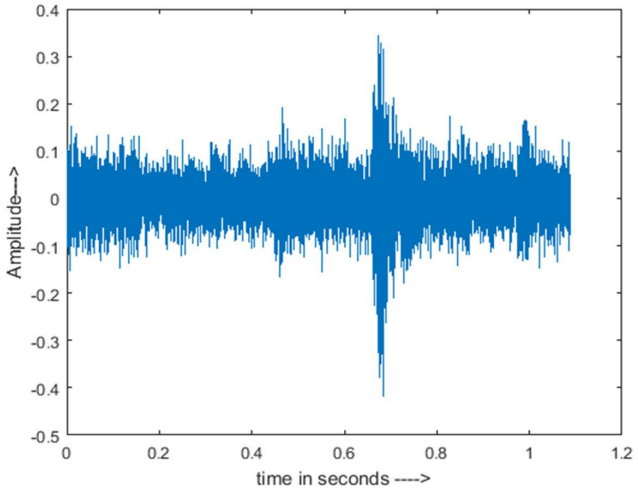

FIGURE 12. The unfiltered exhibition noise in speech signal at −5db.

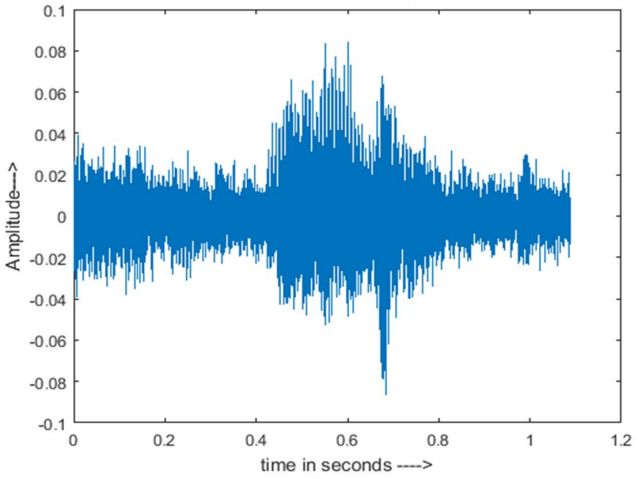

FIGURE 13. The filtered exhibition noise in speech signal at −5db.

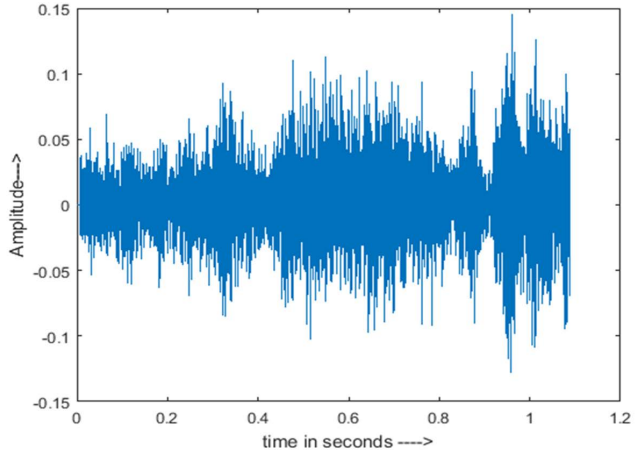

FIGURE 14. The unfiltered restaurant noise in speech signal at −5db.

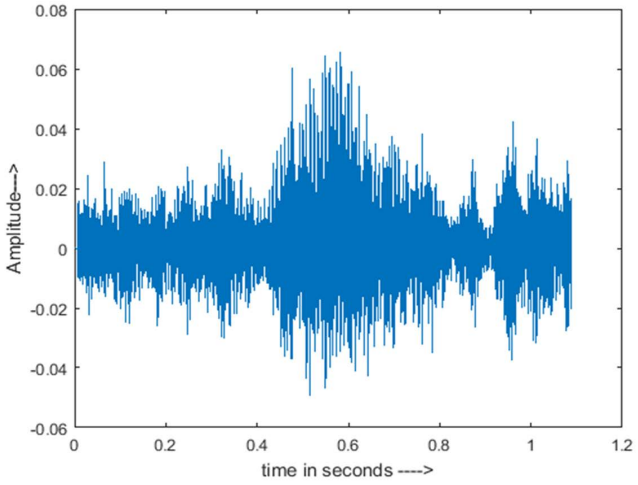

FIGURE 15. The spectrogram for filtered restaurant noise in speech signal at −5db.

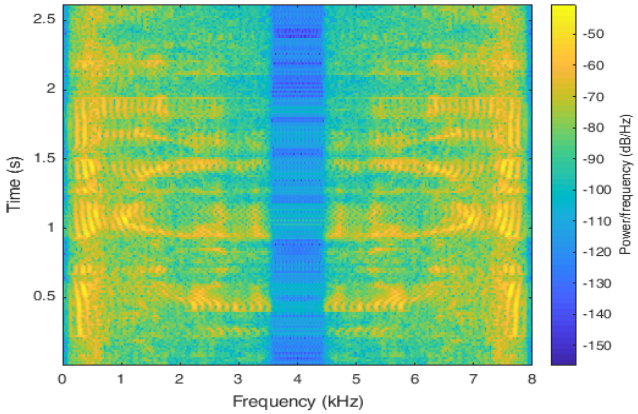

FIGURE 16. The unfiltered Airport noise in speech signal at 10dB.

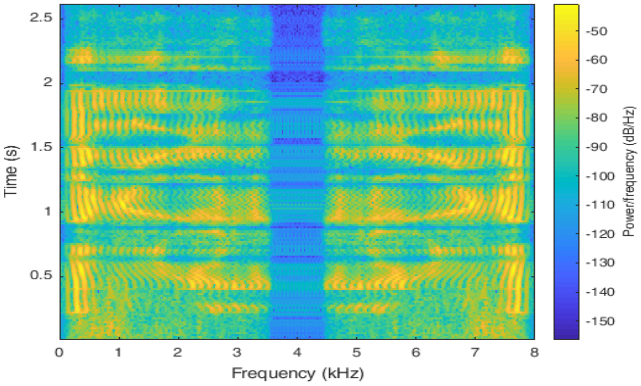

FIGURE 17. The reconstruction quality of the deep learning scheme for Airport noise in speech signal at 10dB.

SNR, but the recognition accuracy gradually decreases at 15dB, 10dB, 5dB, 0dB, −5dB, −10dB for non-stationary noise. The results show that selected methods for WSE can better deal with environmental noise issues at 20dB, 15dB, and 10dB SNRs and can be more suitable for speech recognition applications, especially in the outdoor environment.

From Table 3, the average WRR for non-stationary noises was the highest for exhibition and the lowest for restaurants

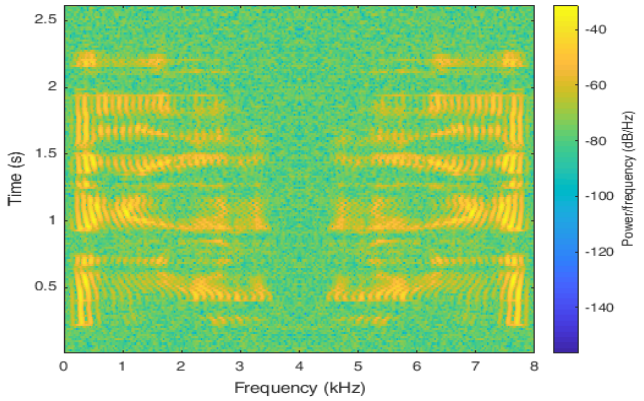

FIGURE 18. The unfiltered white Gaussian noise in speech signal at 10dB.

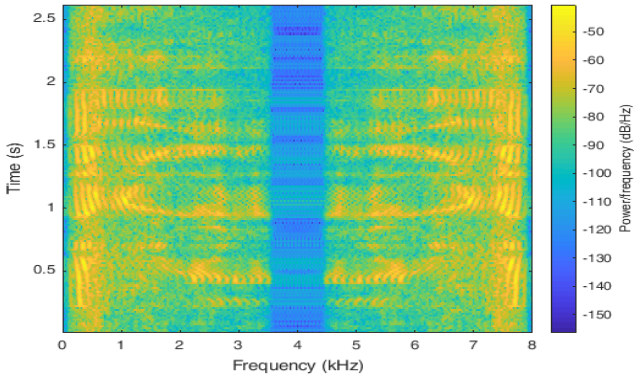

FIGURE 19. The reconstruction quality of the deep learning scheme for white Gaussian noise in speech signal at 10dB.

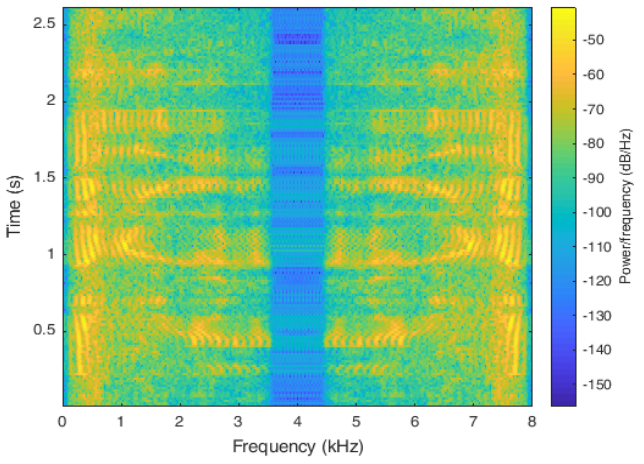

FIGURE 20. The unfiltered Babble noise in speech signal at 10dB.

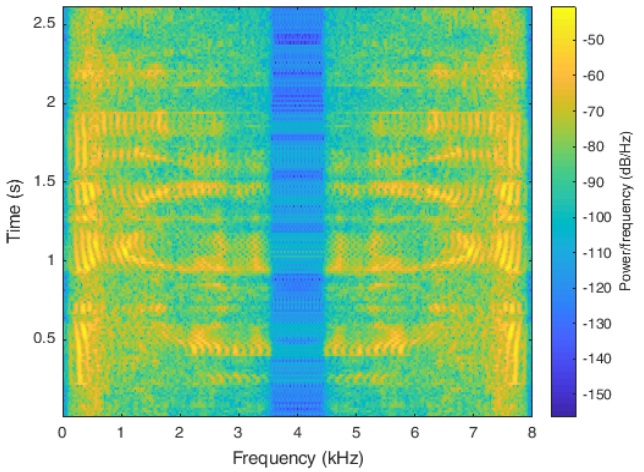

FIGURE 21. The reconstruction quality of the deep learning scheme for Babble noise in speech signal at 10dB.

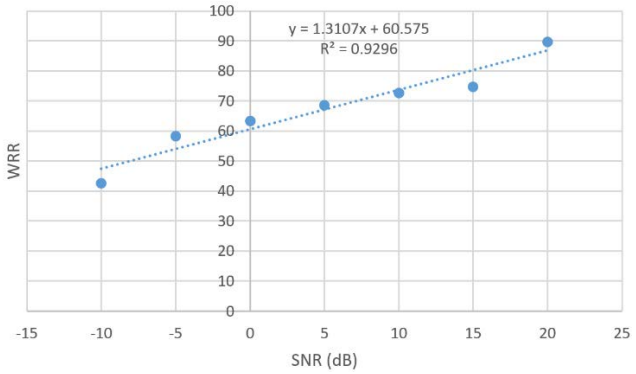

FIGURE 22. The SNR and WRR linear relationship for stationary noise.

TABLE 3. Word recognition rate (WRR) for wearable speech enhancement (WSE) under Non-stationary environmental noises.

WRR for Non-Stationary environmental noises (%)						
SNR/dB	Airport	Babble	Car	Exhibition	Restaurant	Average
-10dB	5.82	4.04	7.26	6.54	4.54	5.64
-5dB	12.32	7.12	13.26	11.54	6.38	10.12
0dB	19.06	17.56	16.55	20.23	12.12	17.10
5dB	36.14	35	35.16	44.66	38.24	37.84
10dB	67.26	74.18	67.2	77.72	55.56	68.38
15dB	88.88	90.64	92.02	91.46	75.2	87.64
20db	*93.6*	*91.28*	*97.92*	*94.78*	91.28	93.77
Average	*46.15*	*45.69*	*47.05*	*49.56*	*40.47*	

(a difference of about 9%). The low WRR for the restaurant is due to the mix-up of many speeches and non-speech noises.

By referring to the scatter diagram, there is a linear relationship between the SNR level and WRR. Figure 23 depicts the linear relationship of SNR and WRR for non-stationary noise (p-value < 0.001).

When we compare the linear relationship of SNR and WRR for both the stationary and non-stationary noise, we found that the gradient for the latter was higher than the former. It indicates that the performance of selected methods

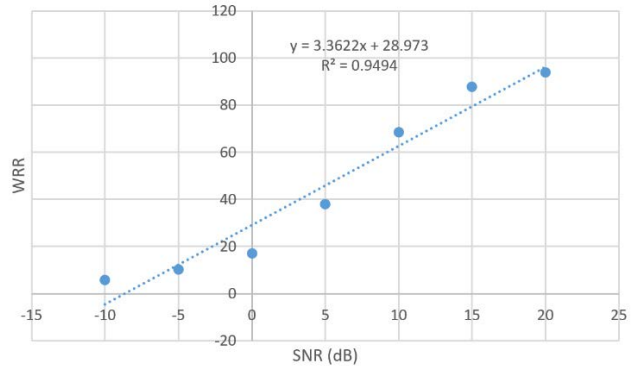

FIGURE 23. The SNR and WRR linear relationship for non-stationary noise.

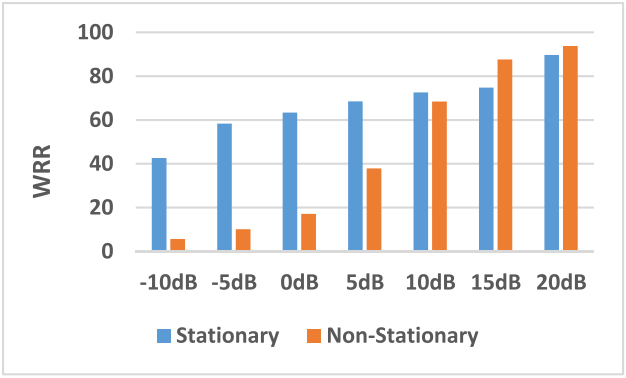

FIGURE 24. WRR for both the stationary and non-stationary noises.

for WSE improves at a higher rate when SNR increases for non-stationary noise.

By comparing the result in Tables 2 and 3, we found that the selected methods for WSE was better for stationary noise at low dB but was more effective for non-stationary noise at high dB noise. Figure 24 depicts the differences in the WRR for both the stationary and non-stationary noises for the WSE based on fixed beamforming, ANR, and VAD. The selected methods work well for stationary noise at −10dB to 10dB. For 15dB and 20dB, the WSE is very effective for recognizing speech in non-stationary noises.

Finally, to determine whether the result for stationary and non-stationary noise was significantly different, we have conducted the Analysis of Variance (ANOVA), and the result is shown in Figure 25 below. It was found that the linear relationship of SNR and WRR was not significantly different between the stationary and non-stationary noise. As such, the performance of the selected methods for both the stationary and non-stationary noise was statistically similar at a 95% confidence level.

We have also compared the performance of several Deep learning-based methods for word recognition. Table 4 depicts the average word recognition rate for all levels of SNR from −10dB to 20dB.

Among the Deep learning methods, VDCNN-conv reported the highest WRR at 90.45%, while DNN shows

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1594.525	1	1594.525	1.987332	0.184008	4.747225
Within Groups	9628.135	12	802.3446			
Total	11222.66	13				

FIGURE 25. Result of ANOVA.

TABLE 4. Word recognition rate for the deep learning methods.

Model	A	B	C	D	Avg.
CNN	96.67	82.48	93.41	93.11	88.66
DNN	96.53	80.27	92.15	92.33	87.45
VDCNN-FC	97.57	83.74	94.26	94.08	89.91
VDCNN-avg	97.25	83.54	92.73	94.12	89.71
VDCNN-max	97.44	84.6	94.64	94.22	90.36
VDCNN-max-conv	97.5	83.65	93.05	93.98	89.74
VDCNN-conv	97.68	84.64	94.62	94.55	90.45
Avg.	97.23	83.27	93.55	93.77	89.47

the lowest WRR at 87.45%. In terms of the speech datasets, Dataset A has the highest WRR, while dataset B has the lowest WRR.

When we compare the result of dataset B results and the selected methods for WSE, the performance in terms of WRR is lower than Deep Learning methods. The WRR for most of the Deep Learning methods is twice as much. It shows that the selected methods for WSE are inferior to the Deep Learning methods.

V. CONCLUSION

In this study, we experiment on a wearable speech enhancement system based on fixed beamforming, ANR, and VAD algorithms and their recognition accuracy. We conducted the spectral analysis and word recognition rate evaluations on the WSE. The word recognition rate evaluation had confirmed that the selected methods for WSE could not perform effectively under low SNR conditions. However, it performed better in a noisy stationary environment than in non-stationary noisy environments.

We also found that the selected methods for WSE perform effectively at high SNR in stationary and non-stationary noisy environments. The linear relationship between the SNR and WRR has proven that the current WSE successfully filters noise at higher SNR but fails at lower SNR. The strength of the noise is too small for the selected methods for WSE to filter the noise away. As such, more work is needed to increase the ability of the WSE to filter noise with low SNR, which can be a promising future direction in WSE research.

We have used the Benchmark Deep Leaning methods to compare the selected methods for WSE in terms of WRR. Among the Deep learning methods, VDCNN-conv reported the highest WRR, while DNN shows the lowest WRR. The

performance of selected methods for WSE in terms of WRR is lower than Deep Learning methods. The WRR for most Deep Learning methods is twice as much as the WSE. It shows that the performance using fixed beamforming, ANR, and VAD is inferior to the Deep Learning methods. It suggests that the existing methods for noise reduction in WSE, such as fixed beamforming, ANR, and VAD, are not adequate. New methods for noise reduction in WSE are needed to improve the performance of the WSE in a noisy environment.

REFERENCES

[1] T. Page, "A forecast of the adoption of wearable technology," *Int. J. Technol. Diffusion*, vol. 6, no. 2, pp. 12–29, Apr. 2015.
[2] Y. Xu, M. Yang, Y. Yan, and J. Chen, "Wearable microphone array as user interface," in *Proc. 5th Conf. Australas. User Interface*, vol. 28. Darlinghurst, NSW, Australia: Australian Computer Society, 2004, pp. 1–4.
[3] A. Palla, L. Fanucci, R. Sannino, and M. Settin, "Wearable speech enhancement system for motor impaired people," in *Proc. Int. Conf. Appl. Electron. Pervading Ind., Environ. Soc.* Cham, Switzerland: Springer, 2017, pp. 159–165.
[4] S. M. Kim, "Wearable hearing device spectral enhancement driven by non-negative sparse coding-based residual noise reduction," *Sensors*, vol. 20, no. 20, p. 5751, Oct. 2020.
[5] A. Palla, L. Fanucci, R. Sannino, and M. Settin, "Wearable speech enhancement system based on MEMS microphone array for disabled people," in *Proc. 10th Int. Conf. Design Technol. Integr. Syst. Nanosc. Era (DTIS)*, Apr. 2015, pp. 1–5.
[6] A. Leman, J. Faure, and E. Parizet, "Influence of informational content of background noise on speech quality evaluation for VoIP application," *J. Acoust. Soc. Amer.*, vol. 123, no. 5, p. 3066, 2008.
[7] J. Chen, "Fundamentals of noise reduction," in *Spring Handbook of Speech Processing*. NJ, USA: Springer, 2008.
[8] P. Scalart and J. V. Filho, "Speech enhancement based on *a priori* signal to noise estimation," in *Proc. ICASSP*, 1996, pp. 629–632.
[9] E. Hansler and G. Schmidt, *Topics in Acoustic Echo and Noise Control*. Springer, 2006, ch. 9.
[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, pp. 1109–1121, 1984.
[11] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. ICASSP*, 1999, pp. 789–792.
[12] A. Das and J. H. L. Hansen, "Constrained iterative speech enhancement using phonetic classes," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 6, pp. 1869–1883, Aug. 2012.
[13] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.
[14] J. H. L. Hansen, V. Radhakrishnan, and K. H. Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, pp. 2049–2063, Nov. 2006.
[15] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.
[16] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, Dec. 2005.
[17] S. Suhadi, C. Last, and T. Fingscheidt, "A data-driven approach to *a priori* SNR estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 186–195, Jan. 2011.
[18] R. Frazier, S. Samsam, L. Braida, and A. Oppenheim, "Enhancement of speech by adaptive filtering," in *Proc. ICASSP*, 1976, pp. 251–253.
[19] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.
[20] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Signal Process.*, vol. 40, no. 3, pp. 497–510, Mar. 1992.
[21] B. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 3, pp. 247–254, Jun. 1979.
[22] G. P. Surender, "A review of different approaches of spectral subtraction algorithms for speech enhancement," *Current Res. Eng., Sci. Technol. J.*, 2013.
[23] C. Ken, H. Wei, and W. Min, "Wearable support system for intelligent workshop application," in *Proc. Int. Conf. Comput. Problem-Solving (ICCP)*, Oct. 2012, pp. 80–83, doi: 10.1109/ICCPS.2012.6384259.
[24] M. A. Akhaee, A. Ameri, and F. A. Marvasti, "Speech enhancement by adaptive noise cancellation in the wavelet domain," in *Proc. 5th Int. Conf. Inf. Commun. Signal Process.*, 2005, pp. 719–723.
[25] *ETSI/AURORA Project, Aurora 2 Database Documentation*, Ericsson, EuroLab, Berlin, Germany, Jan. 2000.
[26] M. Yağanoğlu, "Real time wearable speech recognition system for deaf persons," *Comput. Electr. Eng.*, vol. 91, May 2021, Art. no. 107026.
[27] G. R. Babu, "Speech enhancement using beamforming," *Int. J. Eng. Comput. Sci.*, vol. 4, no. 4, pp. 11143–11147, 2015.
[28] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
[29] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, vol. 1. Springer, 2008.
[30] F. L. Luo, J. Yang, C. Pavlovic, and A. Nehorai, "Adaptive null-forming scheme in digital hearing aids," *IEEE Trans. Signal Process.*, vol. 50, no. 7, pp. 1583–1590, Jul. 2002.
[31] I. McCowan, "Microphone arrays: A tutorial," Queensland Univ., Brisbane, QLD, Australia, Tech. Rep., 2001, pp. 1–38.
[32] J.-S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, pp. 373–376, Feb. 1990, doi: 10.1109/29.103078.
[33] J.-M. Valin, "On adjusting the learning rate in frequency domain echo cancellation with double-talk," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1030–1034, Mar. 2007.
[34] B. Widrow, J. R. Glover, Jr., J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, Jr., and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, Dec. 1975.
[35] R. V. Prasad, A. Sangwan, H. S. Jamadagni, M. C. Chiranth, R. Sah, and V. Gaurav, "Comparison of voice activity detection algorithms for VoIP," in *Proc. 7th Int. Symp. Comput. Commun. (ISCC)*, 2002, pp. 530–535.
[36] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *Proc. Amer. Soc. Eng. Educ. (ASEE) Zone Conf.*, 2008, pp. 1–7.
[37] S. Haykin, *Advances in Spectrum Analysis and Array Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1991.
[38] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. Springer-Verlag, 1972.
[39] J. M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process. (MMSP)*, CA, USA, Aug. 2018, pp. 1–5.
[40] R. Gu, S.-X. Zhang, Y. Zou, and D. Yu, "Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain," *IEEE Signal Process. Lett.*, vol. 28, pp. 1370–1374, 2021.
[41] S. K. Roy, A. Nicolson, and K. K. Paliwal, "DeepLPC: A deep learning approach to augmented Kalman filter-based single-channel speech enhancement," *IEEE Access*, vol. 9, pp. 64524–64538, 2021.
[42] C. Fan, J. Tao, B. Liu, J. Yi, Z. Wen, and X. Liu, "End-to-end post-filter for speech separation with deep attention fusion features," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1303–1314, 2020.
[43] S. Dong, P. Wang, and K. Abbas, "A survey on deep learning and its applications," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100379.
[44] J. Rownicka, S. Renals, and P. Bell, "Simplifying very deep convolutional neural network architectures for robust speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2017, pp. 236–243.
[45] H. A. Babri and Y. Tong, "Deep feedforward networks: Application to pattern recognition," in *Proc. Int. Conf. Neural Netw. (ICNN)*, vol. 3, Jun. 1996, pp. 1422–1426.

PAVANI CHERUKURU received the B.Tech. degree in electrical and electronic engineering from AITS affiliated to JNTU University, in 2011, and the M.Tech. degree in information technology from the Vellore Institute of Technology (VIT) University, in 2011 and 2014, respectively. She is currently pursuing the Ph.D. degree with the Universiti Malaya. She worked as a Research and Teaching Assistant for different projects of multimodal engagement for children with communication disabilities, a Malay text-to-speech synthesis system for enhancing user interaction with technologies devices and big data technologies. She has published articles on e-learning, network security areas in international conferences. Her research interests include e-learning, game-based learning, speech signal process, wearable learning, and network security. Her research interests include speech signal processing for ASR systems at wearable platform. She holds the Best Merit Student Award from bachelor's degree and the Best Paper Award from the 5th International Conference on Science, Engineering and Technology.

MUMTAZ BEGUM MUSTAFA (Member, IEEE) received the B.Sc. degree in software engineering from the Universiti Putra Malaysia (UPM), in 2002, and the M.Sc. and Ph.D. degrees in computer science from the Universiti Malaya (UM), in 2006 and 2012, respectively. She is currently an Associate Professor at the Universiti Malaya. She has undertaken several speech synthesis research and holds grants from the Ministry of Higher Education. Her research and development of the HMM-based Malay speech synthesis system has won the Most Prestigious Award (MPA) for Excellent Research 2012 from MIMOS Bhd., and the National Research and Development Centre in ICT, and have won gold medals for several national level competitions. She has published several papers in prestigious speech conferences and journals. She has established network with a number of International Speech Research Laboratory in Japan and Singapore. She supervises a group of Ph.D. and master's students working on speech synthesis, speech recognition, and speech signal processing. She has published her work in many of prestigious international journals. Her research interests include emotional speech synthesis and speech assistive tools for disabled individuals.

HEMA SUBRAMANIAM received the M.Sc. degree in software engineering from the Universiti Selangor (UNISEL), and the Ph.D. degree in software engineering from the Universiti Putra Malaysia (UPM), in 2016. She is currently a Senior Lecturer at the Universiti Malaya. She has undertaken several emotions analytics related research and holds a grant from the Ministry of Higher Education on the mood analytics model via social media postings. She has published several papers in software engineering conferences and journals. She supervises a group of master's students working on software maintainability, software management, and mood detection solution development. Her research interest includes providing IT-based solutions for improving mental wellbeing among students, including special needs students.