Data Driven Neural Speech Enhancement for Smart Healthcare in Consumer Electronics Applications

Pavan D. Paikrao, Amrit Mukherjee, *Senior Member, IEEE*, Uttam Ghosh, *Senior Member, IEEE*, Pratik Goswami, Milan Novak, Deepak Kumar Jain, Mohammed S. Al-Numay, *Senior Member, IEEE*, and Pradeep Narwade

***Abstract*—This paper presents the practical response and performance-aware development of online speech enhancement from a consumer electronic perspective. To improve the efficiency of human-machine interaction, speech can play a vital role as a transmission medium on the Internet of Medical Things (IoM). However, some intelligent speech recognition systems cannot preserve the confidentiality of speech data. Additionally, the preservation of privacy is onerous, especially for model training and speech recognition in real-time. The recent development of big data-oriented wireless technologies associated with edge computing, interconnected devices of the Internet of Medical Things (IoMT), and big data analytics has great demand for connected human-machine interaction for many applications like automated cars, health monitoring, and consumer personal health care monitoring systems. Although big data-oriented wireless technologies serve these applications, the challenge remains of ignoring emotional care. This paper starts by explaining how to make a neural network-based architecture that can improve the speech of multichannel first-order Ambisonics mixtures and lower the need for human intervention through ambient intelligence (AmI). This will make the system work better overall in medical situations. Second, we demonstrate the effectiveness of different noise estimation techniques on proposed modulation domain processing (MDP) applications in smart hospitals, including electronic medical documentation, disease diagnosis, and evaluation. The proposed approach outperforms the enhancement of the conventional modulation domain in the cortex with several objective evaluation parameters such as Log Likelihood Ratio (LLR), Weighted Spectral Slope (WSS), Perceptual Evaluation of Speech Quality (PESQ), Csig and segmental (SNR seg.) Different noise estimators are used to figure out what effect the system has on different spectral modification parameters, like the over-subtraction factor and the modification domain. The experimental results show that the MDP system achieves better performance in terms of SNRseg. scores (49%) for the state-of-the-art consumer electronics perspective in a health care system. The proposed framework would greatly contribute to personalized communication health monitoring by consumers in a noisy environment.**

***Index Terms*—Consumer healthcare system, speech recognition, ambient intelligence, big data computing.**

Manuscript received 3 July 2023; revised 31 October 2023, 26 December 2023, 4 February 2024, and 5 April 2024; accepted 5 April 2024. Date of publication 11 April 2024; date of current version 29 August 2024. The work of Mohammed S. Al-Numay was supported by the Researchers Supporting Project, King Saud University, Riyadh, Saudi Arabia, under Grant RSP2024R150. *(Corresponding author: Amrit Mukherjee.)*

Pavan D. Paikrao is with the Department of Electronics and Telecommunication Engineering, Dr. D. Y. Patil Institute of Technology, Pune 411018, India (e-mail: pavankumar.paikrao@dypvp.edu.in).

Amrit Mukherjee and Milan Novak are with the Computer Science Department, Faculty of Science, South Bohemia University in Ceske Budejovice, 37005 České Budějovice, Czech Republic (e-mail: amrit1460@ieee.org).

Uttam Ghosh is with the CS & DS Department, Meharry Medical College, Nashville, TN 37208 USA (e-mail: ghosh.uttam@ieee.org).

Pratik Goswami is with the School of Computer Science and Engineering, Yeungnam University, Gyeongsan 38541, South Korea (e-mail: pratikgoswami@ieee.org).

Deepak Kumar Jain is with the Key Laboratory of Intelligent Control and Optimization for Industrial Equipment, Ministry of Education, Dalian University of Technology, Dalian 116024, China, and also with the Symbiosis Institute of Technology, Symbiosis International University, Pune 412115, India (e-mail: dkj@dlut.edu.cn).

Mohammed S. Al-Numay is with the Electrical Engineering Department, College of Engineering, King Saud University, Riyadh 11421, Saudi Arabia (e-mail: alnumay@ksu.edu.sa).

Pradeep Narwade is with the Department of Electronics and Telecommunication Engineering, Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded 431606, India (e-mail: narwade.pradeep@gmail.com).

Digital Object Identifier 10.1109/TCE.2024.3387740

I. Introduction

THE PRESENT trend defines autonomous software that improves decision-making and energy distribution operations that may eventually govern energy demand and supply. Modern machine learning (ML) technologies are essential to improve energy distribution and network decision-making. The development of the Internet of Things (IoT) will depend on ML and data-driven applications soon. The idea of using more is to automate a data-driven system with the least possible human interaction and achieve the best outcome. Massive amounts of data processing can be initiated and accessed anywhere with a strong communication network. Therefore, in the case of the 5G networks exception, the Non-Terrestrial Network (NTN) systems are also able to placate requests to provide connectivity anywhere and anytime, along with data delivery to large numbers of user equipment (UEs). Provides the availability, continuity, and scalability of the service in a wide coverage area [1], [2]. With the fast growth of user equipment, wireless technologies, and IoE services, there is an unprecedented demand for a wide range of applications [3]. Health care [4] is one of the basic pillars of human need, and smart health care is projected to generate several billion dollars in revenue shortly. For example, on healthcare platforms, information about the gesture state of the car driver to initiate user safety [1] is of prime importance. Similarly, the rapid development of wireless devices and appliances has led to a significant increase in the number of interconnected devices for smart healthcare [5] homes and hospitals. The IoT applications

in healthcare are useful because they enable remote patient monitoring [6]. This enables medical personnel to remotely monitor the health status of people with chronic diseases, which can be especially useful [7]. For people with diseases that require ongoing therapy, this remote monitoring can be very helpful in identifying potential health risks early on.

The IoT can be used to improve the effectiveness and efficiency of healthcare services, in addition to remote monitoring. For example, automatically transmitting patient data to healthcare providers using linked medical devices can save time and reduce errors. The IoT can also contribute to better patient outcomes by providing real-time feedback on compliance with prescription plans and treatment schedules. Therefore, connected healthcare, which plays a significant role in the next-generation healthcare system, with connected devices and sensors, generates a massive amount of data sets in terms of volume, value, variety, and velocity that need to be analyzed. This need for big data harms the complexity and multidimensionality of the system of these large wireless data generated from connected healthcare devices [8]. If the required linguistic data resources are available, one possible benefit of data-driven techniques for natural language processing is that they can be adapted to new languages. If models are overfitted to a specific language or linguistic classification scheme, this advantage may be difficult to realize in practice. Next-generation (Nx) neurofuzzy model-based wireless devices coupled with IoE have enormous potential to address issues with low latency and improved resource utilization along with the best accuracy. Similarly, in the recent coronavirus pandemic situation, the field of e-learning also requires the identification of students' emotional states, and making appropriate decisions can enhance the quality of teaching and learning [9]. In this era of smart technology, the metaverse [10] continues to evolve, and its impact on healthcare is poised to be transformative. By navigating the challenges and embracing the opportunities, the healthcare industry can unlock the full potential of the metaverse, creating a future where virtual realms seamlessly enhance the delivery of medical services and contribute to improved patient outcomes. To boost patient satisfaction in hearing health care, address concerns in the "clinical process" through patient-centered strategies, emphasizing effective communication and improving key areas of concern [11]. Blazer et al. [12] encouraged hearing care professionals to adopt a mindset of doing whatever is reasonable and clinically appropriate to meet the needs and desires of the consumer by implementing choice in service, technology, and channel across whichever model of care adopted by a provider. Speech technology typically involves gathering, encoding, transmitting, and handling speech signals. However, speech signals from doctors and patients in public areas of hospitals often come with background noise. In addition, certain patients face challenges in articulating words clearly, either due to illness or dialect variations. These difficulties pose challenges to the effective acquisition and processing of speech signals. One solution is to enhance acquisition equipment to mitigate noise interference, for example, by employing a microphone array for directional noise suppression and targeted speech signal capture [13].

In addition to addressing noise and obtaining high-quality speech signals, current research primarily focuses on advanced AI algorithms for their processing. The motivation for the development of data-driven neural speech enhancement in consumer-based smart healthcare systems is to address the difficulties of speech communication in chaotic and unpredictable environments. In terms of explainable AI (XAI), it is significant, and the proposed work describes its importance. By efficiently eradicating noise and distortions from speech signals, the proposed technique can enhance the quality of speech communication in healthcare applications. In addition, the proposed method can be used to develop remote health monitoring and control systems allowing consumers to remotely monitor and control equipment from a remote location in real time.

The advantages of the proposed method, i.e., data-driven neural speech enhancement in smart healthcare, have several advantages for sustainable consumer electronics applications like transcription, and disease diagnosis interactive control. These benefits can help improve the efficiency and sustainability of consumer healthcare products, ultimately contributing to a cleaner and more sustainable healthcare backbone for the future. Data-driven neural speech enhancement can improve the clarity and intelligibility of speech signals, enabling consumers to communicate more effectively and reducing the risk of errors or accidents. Additionally, by improving speech communication in noisy environments, data-driven neural speech enhancement can help consumers communicate more effectively and respond quickly to any safety hazards.

Saz et al. [14] focused on the development and evaluation of a semi-automated system aimed at providing interactive speech therapy to the growing population of individuals with disabilities, utilizing commercial data-driven neuronal speech enhancement for smart healthcare. The primary goal is to support professional speech therapists in their work. The article discusses the creation and evaluation of a set of interactive therapy tools, along with the underlying speech technologies that enable these tools. These interactive tools are specifically designed to help with the acquisition of language skills, including basic phonatory skills, phonetic articulation, and language comprehension, primarily for children who have neuromuscular disorders such as dysarthria. Successful human-machine interaction in these areas is dependent on the presence of robust speech analysis, speech recognition, and speech verification algorithms that can handle the unique speech variability exhibited by this group of speakers. The paper also includes an experimental study that demonstrates the effectiveness of this interactive system in eliciting speech from a group of impaired children and young speakers between 11 and 21 years of age.

Mental illness encompasses symptoms related to behavioral or psychological patterns that influence various aspects of one's life. Human beings can experience a wide range of pleasant and painful emotions, and it is widely recognized that emotional distress is a universal human experience. Emotions are transient and frequently fluctuate, affecting human existence in both positive and negative ways. However, mental health issues such as stress, sadness, anxiety, and discomfort

can lead to disability and are prevalent among people with chronic diseases, such as cancer. In the proposed method, objective evaluation parameters including PESQ, SDR, and SNR are used to quantify and evaluate the performance of speech enhancement algorithms. These parameters provide a quantitative evaluation of the efficacy of the system and facilitate the selection of the optimal algorithm and configuration for particular healthcare applications [15]. According to the simulation results, these objective evaluation parameters measure the perceptual quality, intelligibility, and suppression of noise and distortions in the enhanced speech signal. The analysis of objective evaluation parameters optimizes the intelligent speech enhancement system in healthcare. The objectives of the proposed work are as follows:

- Develop a speech enhancement neural network, preprocess a large dataset of noisy speech from smart healthcare systems, and integrate the model into intelligent healthcare systems in congested environments.
- Explore transfer learning and domain adaptation for the speech enhancement model, evaluating neural network architectures, and training algorithms for optimal application in smart healthcare.

The workflow of the article is presented below.

Section II discusses related studies and challenges based on the above objectives, while Section III sheds light on the modulation domain recognition approach, which incorporates noise estimation in the designed data framework for smart healthcare. Further, Section IV illustrates the data augmentation using background noise in the consumer's environment using a feedforward neural network approach. Section V discusses the experimental setup along with simulation results, and Section VI incorporates the conclusion and future directions.

II. Related Studies and Challenges

The most prominent and primary mode of communication in this era is speech for personal smart devices. Various research projects have been ongoing on the improvement of accuracy in voice recognition systems in the last few decades. The design of such a system involves concerns like language vocabulary, transducers, speech classes, speech styles, and illness. Berouti et al. [16] presented one of the most popular, simple, and efficient methods of pre-processing that is used to minimize noise from speech signals in the acoustic environment.

Hansen and Clements [17], Hansen [18], and Cairns and Hansen [19] presented reliable indicators of stress like fundamental frequency, spectral energy of voice. Stress has been referred to as speech spoken during one of the following states: task-induced due to the Lombard effect and human emotions (like happiness, anger, fear, or sadness). This recognition has a significant impact on accuracy in a noisy background. Therefore, it needs to be explored in healthcare applications.

The features to relate speech recognition in noise and non-speech recognition in noise are discussed by Vermiglio et al. [20], Bost et al. [21], Prabhakar et al. [22] and Paikrao et al. [23]. In the speech enhancement methods

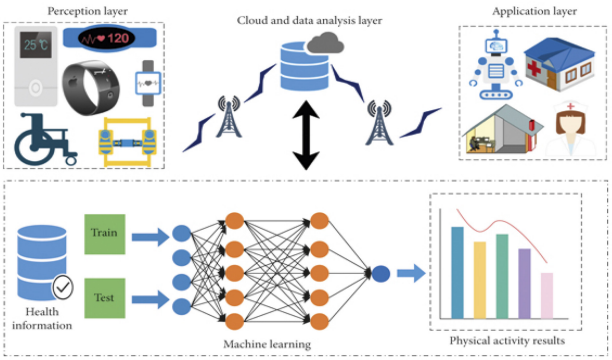

Fig. 1. The Function block architecture consumer-based sustainable healthcare system.

presented by Paliwal and Alsteris [24] and Paliwal et al. [25] (ModSpecSub), improvements in perceptual speech quality metrics over traditional enhancement baselines and noisy data. In this work, the Berouti et al. [16] and Modspecusub [25] are adopted to compare the speech enhancement in the experiments.

Fig. 1 illustrates the function block architecture of a consumer-based sustainable healthcare system. The servers in the real-time environment connect with IoT and IoE servers with transport and network layer protocols. Wu et al. in [26] defines next-generation consumer electronics and presents their state-of-the-art architecture in compliance with IEEE 2668, addressing essential research topics and future opportunities to foster their development securely and reliably. Robust and efficient parsing methods are required to parse unrestricted text. MaltParser is one of the language-independent systems where parsing is data-driven and can generate a parser for a new language from a treebank sample straightforwardly and flexibly. The experimental evaluation confirms that MaltParser is capable of achieving robust, efficient, and accurate parsing for a wide variety of languages without language-specific enhancements and with relatively limited training data. Compared to state-of-the-art parsers for the language in question, the current trends in a data-driven approach to dependency parsing, which has been applied to a variety of languages, consistently produce a dependency accuracy of 80-90 percent. All of these results were obtained without language-specific enhancements and, in most cases, with a relatively limited number of data sources. Several techniques exist in robotics to detect drones, including the Radio-Frequency (RF) technique, the Visual Analysis approach, and Radar approaches, such as a GSM passive coherent location system. These methods are susceptible to other devices in the vicinity. In addition, these recognition techniques are susceptible to confounding similar-appearing objects, such as birds or aircraft, and RF techniques are costly to deploy. To address the current limitations of robotic drone detection systems, an autonomous system is proposed that, in addition to detecting and identifying drones based on their acoustic signatures, requires no human intervention in its implementation.

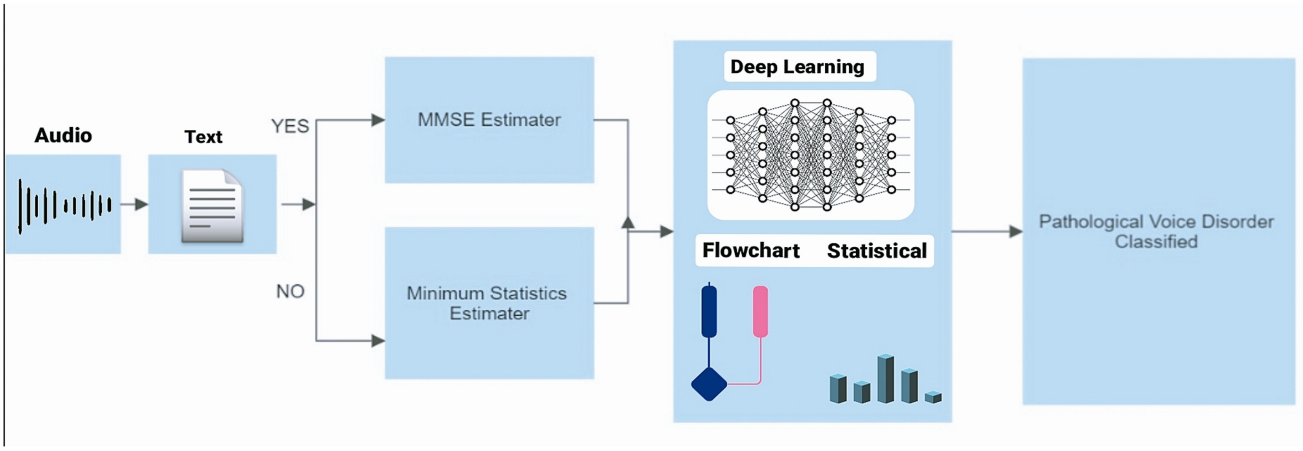

Fig. 2. Neural network-based data-driven architecture system design heuristic to associate speech using deep generative models for the consumer-based sustainable healthcare system.

The lack of large, diverse training datasets that are representative of the population being served is one of the extant issues with data-driven neural speech enhancement in smart healthcare. This can restrict the applicability and efficacy of the models in actual healthcare settings. Variables such as age, gender, and medical conditions can influence the quality and characteristics of speech signals in distinct patient populations. If the training dataset is insufficiently diverse to capture these differences, the resulting model may not perform well for certain patient populations, resulting in inaccurate or unreliable speech enhancement results. A second issue is the requirement for real-time processing of speech signals, especially in healthcare contexts where timely and accurate diagnosis and treatment are essential. However, data-driven neural speech enhancement models can be computationally intensive, demanding significant processing power and memory, which can hinder their real-time performance. Subsequently, there is a need for interpretable models that can provide insight into the decision-making process of the neural network, especially in the healthcare context where accountability and transparency are essential. However, data-driven neural network models can be difficult to interpret, making it difficult to comprehend how they make decisions or which speech enhancement features are most essential. To address the issue of real-time processing in these existing works, the proposed method could employ techniques such as model compression, quantization, or pruning to reduce the computational complexity of the neural network and enable faster inference on low-power devices such as smartphones or wearable devices.

The authors developed and evaluated numerous machine learning and deep learning algorithms before concluding that RNN outperformed the other two algorithms in terms of F1 score. To solve the problem of being able to understand, the suggested method might use attention mechanisms or saliency maps to show how different features or inputs affect the neural network's decision-making process. A neural network-based data-driven architecture system design heuristic to connect speech using deep learning generative models is shown in Fig. 2. The generative data will help detect dynamic human activity, such as speech or audio. Automatic voice pathology detection systems can effectively help clinicians by enabling objective assessment and diagnosis in the early stages of voice pathologies. The deep neural networks trained on the voice samples were enhanced by the MMSE and min statistics estimators.

This proposed Algorithm 1 describes a data-driven neural speech enhancement algorithm, where a deep neural network is trained on a dataset of noisy and clean speech signals. The trained network is then used to enhance new noisy speech signals. The performance of the algorithm is evaluated using objective metrics such as SNR and PESQ.

Algorithm 1 Data-Driven Neural Speech Enhancement Algorithm

1: **Input**: Noisy speech signal $\mathbf{x}$, Clean speech signal $\mathbf{s}$
2: **Output**: Enhanced speech signal $\hat{\mathbf{s}}$
3: **Initialization**: Training dataset $\mathcal{D} = (\mathbf{x}_i, \mathbf{s}i)i = 1^N$
4: **Training**:
5: Initialize a deep neural network modulation frame duration with parameters M and domain o enhance
6: train the network by minimizing the mean squared error loss:

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} |\mathbf{s}i - f\theta(\mathbf{x}_i)|_2^2 \tag{1}$$

7: **Testing**:
8: For a new noisy speech signal $\mathbf{x}$, obtain the enhanced speech signal as:

$$\hat{\mathbf{s}} = f_\theta(\mathbf{x}) \tag{2}$$

9: **Evaluation**:
10: Evaluate the performance of the algorithm using objective metrics such as signal-to-noise ratio (SNR) and perceptual evaluation of speech quality (PESQ).

Algorithm 2 Extension of Algorithm 1 for Healthcare Applications

Require: Noisy speech signal y, clean speech signal x, training dataset $\mathcal{D} = (y_1, x_1), \ldots, (y_N, x_N)$
Ensure: Enhanced speech signal $\hat{x}$
function TRAINMODEL($\mathcal{D}$)
    Initialize parameters of neural network model θ
    Define loss function $\mathcal{L}(\theta, \mathcal{D})$
    Optimize model parameters using stochastic gradient descent or other optimization algorithm
end function
function ENHANCESPEECH(y, θ)
    Apply short-time Fourier transform to y to obtain noisy speech spectrum Y
    Feed Y into neural network model with parameters θ to obtain estimated clean speech spectrum $\hat{X}$
    Apply inverse short-time Fourier transform to $\hat{X}$ to obtain enhanced speech signal $\hat{x}$
    return $\hat{x}$
end function

As mentioned above, Algorithm 2 consists of two main functions: "TrainModel" and "EnhanceSpeech". "TrainModel" takes a training dataset $\mathcal{D}$ and learns the parameters θ of a neural network model for speech enhancement. "EnhanceSpeech" takes a noisy speech signal y and the learned parameters θ, and applies the neural network model to obtain an enhanced speech signal $\hat{x}$. The algorithm assumes that the short-time Fourier transform is used to convert the speech signals into the frequency domain for processing.

III. Modulation Domain Recognition Approach

Improvement in modulation domain: To minimize the shortcomings of the methods discussed In the literature, the proposed methods used the minimum mean squared error (MMSE) estimator and the minimum statistics estimator at varying input SNR, as discussed below.

A dataset of speech signals corrupted with varying levels of noise and input SNRs is compiled.

The MMSE estimator and minimum statistics estimator are used to perform speech enhancement and denoising, respectively, on noise speech signals. This will result in improved speech signals at various SNRs. Our proposed method is evaluated as part of the preprocessing backbone Fig. 1, where the input to the network is the magnitude spectrum of the noisy speech.

$$X(n,k,m) = \sum_{l=-\infty}^{\infty} |x(l,k)| Wm(n-l) \times e^{\frac{-j\times 2\times \pi \times z\times l}{L}} \tag{3}$$

Here n is known as the discrete-time frame number of the acoustic domain, k refers to a discrete frequency index, and m refers to a discrete modulation frequency index. L is modulation frame duration in terms of acoustic frame, and $Wm(n)$ is modulation window.

Improvement in the modulation domain is represented as in Eq. (4).

$$\hat{S}(n,k,m) = \begin{cases} (X_R(n,k,m)|^\gamma - \alpha |N(n,k,m)|^\gamma)^{1/\gamma}, \\ if X_R(n,k,m)|^\gamma - \alpha |N(n,k,m)|^\gamma \quad \geq B \\ \beta |N(n,k,m)|^\gamma \, otherwise \end{cases} \tag{4}$$

where $B = \beta |N(n,k,m)|^\gamma$

This is expressed as:

$$X(n,k,m) = X_R(n,k,m) + iX_I(n,k,m) \tag{5}$$

where m is the index of a modulation frame. k denotes the index of acoustic frequency. After that, the real part of the complex modulation spectrum $X(n,k,m)$ is derived.

In general, the polar form [0] is shown in Eq. (6).

$$X(n,k) = |X(n,k)| ej \angle X(n,k) \tag{6}$$

where $|X(n,k)|$ and $\angle X(n,k)$ denotes an acoustic magnitude and phase spectrum.

In order to analyze, evaluate, categorize, and classify a given dataset, CNN is a form of deep learning technique that is implemented based on supervised learning. The following are the two characteristics that separate CNN: (1) CNN's patterns are said to be translation-invariant. (2) Learning spatial hierarchies of patterns. In the proposed model, data enhancement is accomplished by applying synthetic noise to clean speech signals. This noise can be generated using various noise estimators, such as the MMSE and minimum noise estimators. Adding these noise estimators to the data changes the PESQ objective evaluation parameter in many ways, depending on things like the type and amount of noise added and how complicated the speech enhancement model is. In general, however, the MMSE noise estimator is anticipated to produce superior PESQ scores than the minimum noise estimator. This is because the MMSE noise estimator can more precisely estimate the statistical properties of the noise and use this information to suppress the noise in the speech signal more effectively. The minimum noise estimator, [27] on the other hand, posits a simple model for the noise and may not be able to represent the complex statistical properties of the noise in real-world environments.

A. Estimation of Noise

Conventional estimation of noise with initial silence frames of input noisy speech signal: The most crucial part of the speech enhancement technique is appropriate noise estimation. Here, the minimum statistics for noise identification [27] in NTN applications are presented, and the MMSE model for noise estimation [28] is implemented in the proposed method using a neuro-speech computing approach. For better noise estimation, various properties of speech signals must be exploited. The periodicity of the signal can be used to design VAD because, as in the case of speech signals, most of the noise signals are periodic. The disadvantage related to the use of periodicity methods is that VAD does not run for periodic noises [29]. Many statistical-based speech enhancement systems employ simple noise estimation, whose reference

implementation is given in [30]. The noise estimate is first computed by initial silence frames, and then the noise estimate is updated using the recursive averaging rule [25], [30], [31] as shown below.

$$W_{n,k} = \psi |W_{n-1,k}|^2 + (1-\psi)|Z_{n-1,k}|^2 \tag{7}$$

where ψ is the forgetting factor depending on the stationarity of noise. The forgetting factor is also called a smoothing constant. The smoothing constant helps reduce the variance in the estimated noise power spectrum.

The noise estimated through initial silence frames was updated using VAD in the conventional ModSpecSub [30]. This reduces speech quality scores in the non-stationary environment, and computational load surges on the system. This is an estimation method in which the mean of squared error between the actual and the estimated error is minimized.

Minimum Mean Squared Error (MMSE) Estimator: Assume that we want to estimate some parameter θ from the available set of observations, $x = [x[0], x[1], \ldots, x[N-1]]$. For example, assuming the observed data comes from the following process,

$x[n] = A + w[n]$, w[n] is additive noise with a normal distribution. In this example, we would be trying to estimate the parameter A (in real scenarios, the actual value of A will not be available) from N observations. The most simple method for estimating A (this is the *theta* for this problem) is the sample mean of the observed data.

$$\hat{A} = fracx[0] + x[1] + \cdots + x[N-1]N$$

This is the best estimate of A we can have. Well, that depends on how the user is judging an estimator function.

$\bar{A} = f(x[0], x[1], \ldots, x[N-1])$

$$MSE\left(\hat{A}\right) = E\left[\left(A - \hat{A}\right)^2\right] \tag{8}$$

Thus, a minimum mean squared error, or MMSE estimate, would be $\hat{A}$, which minimizes the value of $MSE(\hat{A})$.

B. Trade Off Between Alpha and Gamma Values in Spectral Subtraction With SNR Dependency

Spectral modification enhancement methods enhance noise-carrying speech by removing the short-time spectral amplitude of the estimated noise from the background/interfering noise signal. To minimize this, the noise flooring B_n is applied to compensate for over-subtraction. The modified spectrum is given by Eq. (9)

$$\hat{S}(n,k) = |X(n,k)|^\gamma - \alpha |N(n,k)|^\gamma \tag{9}$$

where γ refers to domain of spectral subtraction, α refers to the over-subtraction parameter.

noise floor $Bfloor_n$ represented as

$$Bfloor_n = \left(\beta (N(n,k)^\gamma)\right)^{\frac{1}{\gamma}} \tag{10}$$

$\alpha \geq 1$ and $0 \leq \beta \leq 1$. where γ is the spectral subtraction domain, n is known as the discrete-time frame number of the acoustic domain, k refers to a discrete frequency index, m refers to a discrete modulation frequency index.

TABLE I
PESQ SCORES FOR THE PROPOSED METHOD USING MINIMUM STATISTIC AND MMSE ESTIMATOR OF NOISE AT 0 DB INPUT SNR

	Spectral Subtraction Type	Minimum Statistics Noise Estimation [27]	MMSE Noise Estimation [32]
PESQ	Noisy	1.773	1.773
	Magnitude spectral subtraction (gamma =1)	1.9749	1.996
	Power spectral subtraction (gamma = 2)	2.704 (PESQ significant improvement)	2.436

To evaluate the recognition scores for the Minimum Mean Squared Error (MMSE) Estimator and the minimum statistics estimator at varying input SNR, the following method is proposed:

- A dataset of speech signals corrupted with varying levels of noise and input SNRs is compiled.
- The MMSE Estimator and minimum statistics estimator are used to perform speech enhancement and denoising, respectively, on noise speech signals. This will result in improved speech signals at various SNRs.
- The speech recognition system is used to assess the efficacy of the MMSE and minimal statistics estimator on enhanced speech signals with varying SNRs. Word error rate (WER) and phoneme error rate (PER) can be used by the recognition system to measure the accuracy of the recognition.
- The data-driven neural network model is trained on chaotic speech signals and their corresponding clean speech signals in order to discover a mapping between the input SNR and the corresponding recognition score for the MMSE Estimator and minimal statistics estimator.
- The trained neural network model is utilized to predict the recognition score for the MMSE Estimator and minimum statistics estimator at different input SNRs. This will enable the system to evaluate the efficacy of speech enhancement methods without costly speech recognition experiments at each SNR.
- The predicted recognition scores for the MMSE Estimator and minimum statistics estimator are contrasted to the actual recognition scores derived from speech recognition experiments for the MMSE Estimator and minimum statistics estimator. This will enable the system to validate the precision and efficacy of the neural network model for predicting recognition scores.

Table I shows PESQ scores for the proposed method using Minimum statistic and MMSE estimator of noise at 0 dB input SNR. The comparison between perceptual evaluation metrics shows that the minimum statistic estimator outperforms the MMSE estimator where a significant improvement in the mean PESQ score of 2.703 has been observed. Again this improvement is observed in magnitude spectral subtraction for gamma = 2.

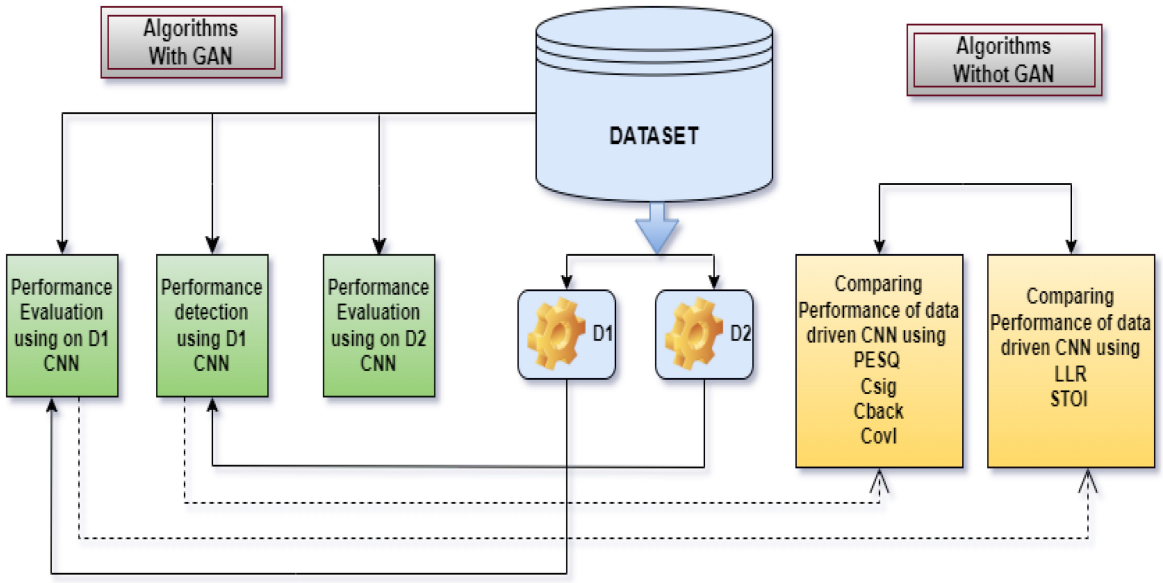

Fig. 3. High-level design of the data-driven framework for consumer-based sustainable healthcare system.

IV. Data Augmentation Using Background Noise in Consumer Environment

We have approached the issue by introducing a method of augmentation, in which a real-life background noise is overlapped with the drone audio without any modification to the actual audio features, such as the amplitude or frequency of the audio clip, because the application of the drone detection and recognition could be deployed in areas with a variety of background noises. Fig. 3 shows the High-level design of the data-driven framework for a consumer-based sustainable healthcare system. In this after augmentation for background noise, the SNR varies from 32 dB to 3 dB. To ensure the uniformity of the audio files, it was crucial to reformat the audio snippets obtained from these datasets, as detailed in the section on data preprocessing. This mechanism allowed us to simulate real-world situations. The proposed method uses MMSE in combination with a data-driven neural network, which is performing well in enhancing speech quality and reducing noise in the input signal in smart healthcare. The effectiveness of the proposed technique analyses various factors, including the quality and size of the training dataset, the architecture of the neural network, and the specific techniques used in the speech enhancement pipeline. In addition to MMSE, other performance metrics are also considered when evaluating the effectiveness of data-driven neural speech enhancement techniques. These include error rate, accuracy, computational efficiency, and perceptual quality, among others. The optimal set of metrics depends on the specific renewable energy application and requirements of the system.

V. Experimental Setup

The training and testing of the proposed framework were developed. The Figure 2 shows the procedure for obtaining the modulation domain frame $X(n, k, l)$. The L acoustic frames constitute one modulation frame, and one acoustic frame is composed of T time domain speech samples. Therefore, TL time-domain samples construct one modulation frame. The increase in the acoustic frame determines the modulation domain sampling frequency. In our proposed work, the speech stimuli with a sampling frequency of 8000 Hz are considered to have a modulation frequency of $8000/M$. Now, the noise estimator is used on each modulation frame of the noisy speech to get an estimate of the clean speech. This is used to make a spectral envelope using the overlap-add synthesis method.

A. Database Used

For the validation of the proposed methodology, the NOIZEUS speech corpus dataset is implemented [33], [34]. The NOIZEUS dataset consists of 30 sentences spoken by six speakers: three males and three females. The audio stimuli in the dataset are sampled at 8 kHz along with the nonstationary noises at various input signal-to-noise ratios. In objective validation, the noisy stimuli obtained by degrading clean stimuli with babble noise, AWGN, at different input signal-to-noise ratios, by Paliwal et al. [25], can be accessed publicly.

Stimuli Enhanced speech stimuli are derived by applying procedures to input, as shown in Fig. 2.

B. Model Setup Using Feed Forward Neural Network

A feed-forward-based model has been performed using the collected data. As discussed, Fig. 3 shows a high-level dataset incorporating GAN for a consumer-based sustainable healthcare system. The feed-forward type of neural network helps in the faster formation of the model when compared with other variants of neural networks. The main reason is that there is no retraining of weights and biases back to the previous nodes/layers. The information is transferred to the next layer by optimizing the weights and biases to map the output data to the input data efficiently. Before training the

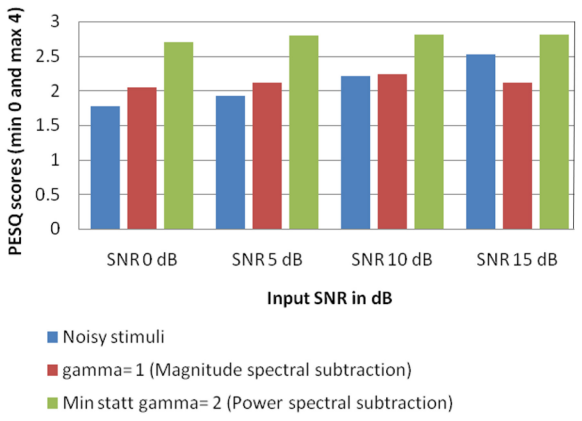

Fig. 4. PESQ scores for Noisy stimuli, proposed stimuli using Minimum statistics method for gamma = 1 (Magnitude spectral subtraction) and gamma = 2 (Power spectral subtraction).

neural network model, the data obtained needs to be pre-processed. Pre-processing of data helps in the removal of unwanted data elements that are not in the range of the objective defined. In this part, we may be reducing the data size that needs to be implemented in the network. However, it will help in improvising the fitness function formation, and hence, better mapping can be performed between the input and output parameters. The complete dataset after pre-processing needs to be normalized. In this normalization method, 0 is taken as the lower limit and 1 is taken as the higher limit for all the input and output variables. It can be observed that variables like SNR, PESQ, WSS, and others have different minimum and maximum values. This normalization helps with synchronization and better mapping between the variables. Then the normalized data is categorized into 3 levels, i.e., training data, validation data, and testing data. These data sets are randomly categorized from the normalized data sets. Most importantly, all data sets are different in all three categories. Also, redundant data sets have been removed. The categorization has been done in a manner where the training data set is comprised of 70%, validation data is comprised of 10% and the rest of the data is categorized into testing data. There are different training algorithms available. However, after repeated runs, it has been observed that the Levenberg Marquardt (LM) training method has provided better performance. After repeated iteration, the best and most accurate model generated is fixed and saved for further characterization of speech estimation. The output datasets are denormalized before implementation in the next stage.

Fig. 4 illustrate PESQ scores for noisy stimuli, proposed stimuli using Minimum statistics method for gamma = 1 (Magnitude spectral subtraction) and gamma = 2 (Power spectral subtraction)

C. Effect of VAD and Noise Updating

In this section by experimental evaluation, the work illustrates the effect of voice activity detection [31] on noise updating. As noise is non-stationary, we need to consider the speech frame as very short in length for spectral modification and analysis. But over this short frame duration, the updating of noise has a very negligible effect. From Table II improvement in the PESQ score at an input SNR of 5 dB is observed for noisy and enhanced speech stimuli.

TABLE II
SNRSEG, WSS AND PESQ SCORES FOR THE PROPOSED METHOD USING VOICE ACTIVITY DETECTOR AND NOISE UPDATING AS NOISE ESTIMATOR AT 0, 5, 10 AND 15 DB INPUT SNR

	Object. Metric	Input SNR in dB			
		0 dB	5 dB	10 dB	15 dB
Noisy	WSS	58.08	52.23	43.31	33.75
	PESQ	1.77	1.92	2.21	2.03
alpha = 1; gamma = 1	SNRseg	-1.62	-0.13	0.63	0.91
	WSS	66.77	55.12	49.14	43.81
	PESQ	1.95	2.17	2.29	2.28
alpha = 0.1; gamma = 1	SNRseg	-1.94	0.026	1.02	1.39
	WSS	65.65	55.83	55.25	51.29
	PESQ	1.96	2.16	2.24	2.17

TABLE III
COMPUTATIONAL COMPLEXITY OF PROPOSED METHOD WITH CONVENTIONAL

	ModSpecSub [25]		Proposed	
Normalized processing time	2.657		1	
Modulation frame duration M in ms				
64	63.12		12.87	
100	30.55		5.92	
128	30.35		6.45	
192	23.03		4.32	
256	14.95		6.11	
Sub Frames in Call	**Calls**	**Total Time**	**Calls**	**Total Time**
Hamming window	38916	7.856 s	2	0.04 s
angle	38401	0.375 s	2	0.21 s
specsub_frame	38400	13.68 s	2048	0.17 s
berouti [16]	38400	0.44 s	2048	0.05 s
specsub	512	14.94 s	-	-
repmat	2050	0.12 s	2	0.03 s
Noise estimate	512	10.94	1	3.34 s

Table III shows the improved computational complexity of the proposed method compared to the conventional method. The normalized processing time has increased drastically.

D. Objective Evaluation: Effect of Over Subtraction Factor-Alpha in Modulation Domain

Figure 2 shows a methodology for modification of the magnitude spectrum in modulation domain processing. It will be useful to find the optimized parameters in modulation domain processing for both minimum statistics and the MMSE estimator. Therefore, in this section, we further continue to investigate the effect of over-subtraction factor-alpha in modulation domain processing. In this process, how speech quality varies with different parameters of the acoustic and modulation domains is studied more closely. In both the acoustic and modulation domains, the hamming window is applied until and unless stated. The frame duration of 20–40 ms is applied in speech processing.

TABLE IV
LLR AND SNRSEG., SCORES FOR THE PROPOSED METHOD USING MINIMUM STATISTIC ESTIMATOR OF NOISE AT 0, 5, 10 AND 15 DB INPUT SNR

alpha2 and gamma2 dependency	Object. metric	Input SNR in dB			
		0	5	10	15
Noisy	LLR	3.63	3.40	3.18	2.99
	SNRseg	-4.27	-2.01	0.41	0.61
alpha2 = 1; gamma2 = 1	LLR	3.484	3.40	3.44	3.42
	SNRseg	**-1.74**	-2.01	0.21	0.98
alpha2 = 2; gamma2 = 1	LLR	3.47	3.63	3.48	3.40
	SNRseg	-1.55	-0.33	0.21	0.57
alpha2 = 3; gamma2 = 1	LLR	3.46	3.43	3.40	3.31
	SNRseg	-1.46	-0.33	0.56	0.33
alpha2 = SNR dependency equation ; gamma2 = 1	LLR	3.51	3.44	3.43	3.28
	SNRseg	-1.46	-0.33	0.57	0.77
alpha2 = 0.1; gamma2 = 1; (clear voice with background noise)	LLR	3.46	3.43	3.31	3.28
	SNRseg	-1.46	-0.38	0.78	0.82

TABLE V
WSS AND PESQ SCORES FOR THE PROPOSED METHOD USING MINIMUM STATISTIC ESTIMATOR OF NOISE AT INPUT SNR 0, 5, 10, AND 15 DB SHOWING ALPHA2 AND GAMMA2 DEPENDENCY

Dependency	Object. Metric	Input SNR in dB			
		0 dB	5 dB	10 dB	15 dB
Noisy	WSS	58.08	52.24	43.31	33.75
	PESQ	1.77	1.92	2.21	2.53
alpha2 = 1; gamma2 = 1	WSS	64.93	53.12	59.78	53.88
	PESQ	1.97	1.91	2.11	2.13
alpha2 = 2; gamma2 = 1	WSS	65.44	58.39	60.21	55.26
	PESQ	2.00	2.05	2.14	2.21
alpha2 = 3; gamma2 = 1	WSS	67.91	59.37	59.78	56.34
	PESQ	2.04	2.16	2.15	2.11
alpha2 = SNR dependency equation ;	WSS	67.91	60.12	57.69	54.31
	PESQ	2.04	2.11	2.16	2.11
alpha2 = 0.1; gamma2 = 1	WSS	65.11	59.37	54.31	52.11
	PESQ	2.04	2.16	2.11	2.18

Fig. 5. Segmental SNR improvement for various input SNR at alpha = 0.1 and alpha = 1 for minimum statistic noise estimate with gamma =1 (magnitude spectral subtraction).

TABLE VI
SNR SEG., WSS AND PESQ SCORES FOR THE PROPOSED METHOD USING MMSE ESTIMATOR OF NOISE AT 0, 5, 10 AND 15 DB INPUT SNR SHOWING ALPHA AND GAMMA DEPENDENCY FOR MMSE NOISE ESTIMATOR

alpha and gamma dependency for MMSE Noise Estimator	Object. metric	Input SNR in dB			
		0	5	10	15
Noisy	SNRseg.	-4.27	-2.01	0.41	0.61
	WSS	58.08	52.23	43.21	33.74
	PESQ	1.77	1.92	2.29	2.52
gamma = 1; alpha = 1	SNRseg.	-2.69	-2.04	-1.70	-1.48
	WSS	90.06	91.54	98.17	91.32
	PESQ	1.86	2.05	2.16	2.30
gamma = 2; alpha = 0.1	SNRseg.	-4.6	-4.78	-4.84	-4.74
	WSS	95.53	78.99	78.01	74.23
	PESQ	2.69	2.78	2.81	2.80
gamma = 2; alpha = 1	SNRseg.	-4.03	-3.70	-3.69	-3.67
	WSS	140.93	131.36	129.93	124.17
	PESQ	2.43	2.46	2.53	2.81

However, negligible changes in PESQ scores for values of alpha = 1 and alpha = 0.1 are reported.

Table IV shows LLR and SNRseg scores for the proposed method using the Minimum statistic estimator of noise at 0, 5, 10, and 15 dB input SNR. For alpha2 = 0.1 gamma2 = 1 clear voice with background noise; prominent speech spectra on the spectrogram are observed. Table V shows WSS and PESQ scores for the proposed method using Minimum statistic estimator of noise at input SNR 0, 5, 10, and 15 dB showing alpha2 and gamma2 dependency.

As shown in Figure 5 a significant improvement is observed to segmental SNR (SNR Seg.) for NTN applications. It is improved for the over-subtraction factor alpha = 0.1 compared to alpha = 1. The average improvement is about 48% reported.

We evaluated the improvement in the Segmental SNR for various input SNRs at alpha = 0.1 and alpha = 1 for the minimum estimate of statistic noise with gamma = 1 (magnitude spectral subtraction). A significant improvement is observed in terms of segmental SNR (SNR Seg.). It is improved for the over subtraction factor alpha = 0.1 compared to alpha = 1. The average improvement is about 48% is reported.

Table VI shows the SNR Seg WSS and PESQ scores for the proposed method using the MMSE estimator of noise at 0 5 10 and 15 dB input SNR. For power spectral subtraction (gamma= 2) significant improvement in PESQ scores is observed. This improvement is specifically for the over-subtraction alpha = 0.1. Table VII Overall speech signal quality Covl, SNR Seg., WSS and PESQ scores for the proposed method, MMSE method [28] and Berouti method [16] at 0, 5, 10 and 15 dB input SNR.

TABLE VII
OVERALL SPEECH SIGNAL QUALITY COVL, SNR SEG., WSS AND PESQ SCORES FOR THE PROPOSED METHOD, MMSE METHOD AND BEROUTI METHOD AT 0, 5, 10 AND 15 DB INPUT SNR

Method	Object. Metric	Input SNR in dB			
		0 dB	5 dB	10 dB	15 dB
Noisy	LLR	3.64	3.40	3.18	3.00
	SNR seg	-4.27	-2.01	0.41	3.61
	WSS	58.08	52.24	43.31	33.75
	PESQ	1.77	1.92	2.21	2.53
	Csig	-0.10	0.28	0.76	1.23
	Cback	1.81	2.06	2.41	2.83
	Covl	0.75	1.04	1.44	1.86
	STOI	0.67	0.77	0.87	0.93
Proposed	LLR	3.60	3.29	3.01	2.87
	SNR seg	-1.12	1.20	3.54	6.06
	WSS	69.53	56.19	48.05	36.88
	PESQ	1.98	2.40	2.79	3.13
	Csig	0.15	0.60	1.07	1.51
	Cback	1.88	2.19	2.58	2.98
	Covl	0.91	1.27	1.70	2.09
	STOI	0.69	0.78	0.88	0.94
MMSE [28]	LLR	3.28	3.09	2.86	2.74
	SNR seg	-1.26	0.68	2.79	5.65
	WSS	52.30	43.47	36.56	25.70
	PESQ	2.11	2.43	2.76	3.02
	Csig	0.52	0.99	1.48	1.87
	Cback	2.20	2.53	2.87	3.25
	Covl	1.25	1.66	2.09	2.44
	STOI	0.67	0.77	0.87	0.93
Spectral Berouti [16]	LLR	3.68	3.46	3.10	2.84
	SNR seg	-3.83	-1.83	0.55	5.53
	WSS	70.23	62.74	53.92	38.63
	PESQ	1.83	2.00	2.39	3.02
	Csig	-0.22	0.18	0.86	1.64
	Cback	1.77	2.03	2.43	3.16
	Covl	0.69	0.99	1.55	2.30
	STOI	0.62	0.75	0.88	0.93

Using data-driven techniques, the proposed method addresses computational complexity in the acoustic and modulation domains. Specifically, it uses neural networks and other machine learning algorithms to learn the underlying patterns and relationships in the acoustic and modulation domains and then uses this information to facilitate data processing and analysis. The proposed method can use neural networks to execute tasks such as speech enhancement, noise reduction, and source separation in the acoustic domain. By training these models on large datasets of speech and noise signals, they can learn to extract the relevant features from the data and perform complex computations that would be challenging or impossible to perform manually. This can considerably reduce the computational complexity of the processing and analysis while simultaneously enhancing the precision and robustness of the results. Similarly, in the modulation domain, the proposed method can employ machine learning algorithms to discover the patterns and relationships between various modulation schemes and the signals they generate. This allows the system to detect and classify various modulation schemes autonomously, without the need for complex signal processing techniques or domain-specific knowledge.

VI. CONCLUSION

This study first provides a sustainable perspective for consumer healthcare systems including electronic medical document transcription, pathological voice recognition, and medical process optimization through consumer-healthcare interaction. We discuss how a speech-based healthcare system facilitates the early recognition, rehabilitation assistance, and intelligent assessment of consumers. The existing studies on speech enhancement focuses on the source of one noise present or not. The main work in real-world situations deals with the presence of stationary and nonstationary noise sources in speech and tries to improve speech quality and intelligibility. This type of problem can be solved through DNN. Here feed-forward-based neural network is implemented. The neural network-based architecture provides mask prediction by estimating roughly speech and noise amplitude spectra, with the assumption of known directions of arrival. The study demonstrates the effectiveness of different noise estimation techniques on proposed Modulation domain processing (MDP). The proposed approach outperforms the conventional modulation domain enhancement in the cortex of several objective evaluation parameters such as LLR, WSS, PESQ, and SNR seg. The effect of the system on different parameters of spectral modification like over-subtraction factor and modification domain by various noise estimators is evaluated. The experimental results show that the MDP system achieves better performance in terms of segmental SNR scores (49%). The proposed framework would greatly contribute to consumer electronic personalized healthcare in a noisy environment.

An extensive experimental evaluation is done to investigate the potential of the proposed method in modulation domain processing. The proposed enhancement method is evaluated for speech corruption in various noisy environments.

Experimentally, we determined the following key points:

i) A mean PESQ score for the estimation of minimum statistic noise estimation [27] is found to be 21% for various SNRo between 0-15 dB.

ii) Segmental SNR improvement for various input SNR at alpha = 0.1 and alpha = 1 for minimum statistic noise estimate with gamma = 1 (magnitude spectral subtraction). As shown in the figure, a significant improvement is observed in terms of segmental SNR (SNR Seg.). It is improved for over subtraction factor alpha = 0.1 as compared with alpha = 1. An average improvement is about 48% is reported.

iii) In this context, personal healthcare systems with advanced preprocessing transmissions have gained great attention owing to their ability to boost the overall system performance by incorporating the speech quality of noisy speech signals in different noises such as babble, and AWGN noise. With the help of ML models, the healthcare sector transitions from a safe, low-risk, regulation-driven environment to a cutting-edge, more complex, and uncertain consumer market. Changes would be necessary, along with the employment of new technology and commercial structures in the healthcare systems.

ACKNOWLEDGMENT

The authors thank Dr. Rudolf Vohnout for the database and computing resources in the Department of Computer Science, Faculty of Sciences, University of South Bohemia, Czech Republic.

REFERENCES

[1] F. Rinaldi et al., "Non-terrestrial networks in 5G & beyond: A survey," *IEEE Access*, vol. 8, pp. 165178–165200, 2020, doi: 10.1109/ACCESS.2020.302298.

[2] N. Fares, R. S. Sherratt, and I. H. Elhajj "Directing and orienting ICT healthcare solutions to address the needs of the aging population," *Healthcare*, vol. 9, no. 2, p. 147, 2021, doi: 10.3390/healthcare9020147.

[3] H. R. Chi, M. de Fátima Domingues, H. Zhu, C. Li, K. Kojima, and A. Radwan, "Healthcare 5.0: In the perspective of consumer Internet-of-Things-based fog/cloud computing," *IEEE Trans. Consum. Electron.*, vol. 69, no. 4, pp. 745–755, Nov. 2023, doi: 10.1109/TCE.2023.329399.

[4] A. Alelaiwi, "Multimodal patient satisfaction recognition for smart healthcare," *IEEE Access*, vol. 7, pp. 174219–174226, 2019, doi: 10.1109/ACCESS.2019.295608.

[5] N. Shoji, J. Motomura, N. Kokubu, H. Fuse, T. Namba, and K. Abe, "Proposal of a wearable personal concierge system with healthcare using speech dialogue technology," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, 2021, pp. 1–5, doi: 10.1109/ICCE50685.2021.942776.

[6] T. A. Suleiman and A. Adinoyi "Telemedicine, and smart healthcare—the role of artificial intelligence, 5G, cloud services, and other enabling technologies," *Int. J. Commun., Netw. Syst. Sci.*, vol. 16, no. 3, pp. 31–51, 2023, doi: 10.4236/ijcns.2023.163003.

[7] C. Wang, T. He, H. Zhou, Z. Zhang, and C. Lee, "Artificial intelligence enhanced sensors—Enabling technologies to next-generation healthcare and biomedical platform," *Bioelectron. Med.*, vol. 9, no. 1, p. 17, 2023, doi: 10.1186/s42234-023-00118-1.

[8] P. E. Idoga, M. Toycan, H. Nadiri, and E. Çelebi, "Factors affecting the successful adoption of e-health cloud based health system from healthcare consumers' perspective," *IEEE Access*, vol. 6, pp. 71216–71228, 2018, doi: 10.1109/ACCESS.2018.288148.

[9] T. A. Shaikh and R. Ali, *Fog-IoT Environment in Smart Healthcare: A Case Study for Student Stress Monitoring*. Cham, Switzerland: Springer Int. Publ., 2021, pp. 211–250.

[10] A. Musamih et al., "Metaverse in healthcare: Applications, challenges, and future directions," *IEEE Consum. Electron. Mag.*, vol. 12, no. 4, pp. 33–46, Jul. 2023, doi: 10.1109/MCE.2022.322352.

[11] V. Manchaiah, R. J. Bennett, P. Ratinaud, and D. W. Swanepoel, "Experiences with hearing health care services: What can we learn from online consumer reviews?" *Am. J. Audiol.*, vol. 30, no. 3, pp. 745–754, 2021, doi: 10.1044/2021_AJA-21-00041.

[12] D. G. Blazer, S. Domnitz, and C. T. Liverman, Eds., "Hearing health care services: Improving access and quality," *Hearing Health Care for Adults: Priorities for Improving Access and Affordability*. Washington, DC, USA: Nat. Acad. Press, 2016, doi: 10.17226/23446.

[13] M. Chen et al., "Neural-free attention for monaural speech enhancement toward voice user interface for consumer electronics," *IEEE Trans. Consum. Electron.*, vol. 69, no. 4, pp. 765–774, Nov. 2023, doi: 10.1109/TCE.2023.325450.

[14] O. Saz, S. C. Yin, E. Lleida, R. Rose, C. Vaquero, and W. R. Rodríguez, "Tools and technologies for computer-aided speech and language therapy," *Speech Commun.*, vol. 51, no. 10, pp. 948–967, 2009, doi: 10.1016/j.specom.2009.04.006.

[15] T. M. Ghazal et al., "Iot for smart cities: Machine learning approaches in smart healthcare—A review," *Future Internet*, vol. 13, no. 8, p. 218, 2021, doi: 10.3390/fi13080218.

[16] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1979, pp. 208–211, doi: 10.1109/ICASSP.1979.1170788.

[17] J. H. L. Hansen and M. A. Clements, "Evaluation of speech under stress and emotional conditions," *J. Acoust. Soc. Am.*, vol. 82, no. S1, pp. S17–S18, 1987, doi: 10.1121/1.2024686.

[18] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*, vol. 20, nos. 1–2, pp. 151–173, 1996, doi: 10.1016/S0167-6393(96)00050-7.

[19] D. A. Cairns and J. H. L. Hansen, "Nonlinear analysis and classification of speech under stressed conditions," *J. Acoust. Soc. Am.*, vol. 96, no. 6, pp. 3392–3400, 1994.

[20] A. J. Vermiglio et al., "The relationship between speech recognition in noise and non-speech recognition in noise test performances: Implications for central auditory processing disorders testing," *J. Commun. Disord.*, vol. 77, pp. 31–43, Jan./Feb. 2019.

[21] T. J. M. Bost, N. J. Versfeld, and S. T. Goverts, "Effect of audibility and suprathreshold deficits on speech recognition for listeners with unilateral hearing loss," *Ear Hear.*, vol. 40, no. 4, pp. 757–765, 2019.

[22] G. A. Prabhakar, B. Basel, A. Dutta, and C. V. Rama Rao, "Multichannel CNN-BLSTM architecture for speech emotion recognition system by fusion of magnitude and phase spectral features using DCCA for consumer applications," *IEEE Trans. Consum. Electron.x*, vol. 69, no. 2, pp. 226–235, May 2023, doi: 10.1109/TCE.2023.323697.

[23] P. Paikrao, S. Routray, A. Mukherjee, A. R. Khan, and R. Vohnout, "Consumer personalized gesture recognition in UAV-based industry 5.0 applications," *IEEE Trans. Consum. Electron.*, vol. 69, no. 4, pp. 842–849, Nov. 2023, doi: 10.1109/TCE.2023.33082.

[24] K. K. Paliwal and L. D. Alsteris. "On the usefulness of STFT phase spectrum in human listening tests," *Speech Commun.*, vol. 45, no. 2, pp. 153–170, 2005, doi: 10.1016/j.specom.2004.08.00.

[25] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain." *Speech Commun.*, vol. 52, no. 5, pp. 450–475, 2010, doi: 10.1016/j.specom.2010.02.004.

[26] C. K. Wu, C.-T. Cheng, Y. Uwate, G. Chen, S. Mumtaz, and K. F. Tsang, "State-of-the-art and research opportunities for next-generation consumer electronics," *IEEE Trans. Consum. Electron.*, vol. 69, no. 4, pp. 937–948, Nov. 2023, doi: 10.1109/TCE.2022.323247.

[27] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[28] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012, doi: 10.1109/TASL.2011.218089.

[29] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 478–482, Jul. 2000, doi: 10.1109/89.84822.

[30] P. C. Loizou, *Speech Enhancement: Theory And Practice*. Boca Raton, FL, USA: CRC Press, 2007, doi: 10.1201/9781420015836.

[31] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 2, pp. 126–137, Mar. 1999, doi: 10.1109/89.74811.

[32] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985, doi: 10.1109/TASSP.1985.116455.

[33] P. Loizou and Y. Hu, "NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, nos. 7–8, pp. 588–601, 2017.

[34] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011, doi: 10.1109/TASL.2011.2114881.