Hindawi
Journal of Sensors
Volume 2023, Article ID 1911385, 18 pages
https://doi.org/10.1155/2023/1911385

Research Article

IoT-Based Voice-Controlled Smart Homes with Source Separation Based on Deep Learning

Ghalib H. Alshammri

Department of Computer Science, Community College, King Saud University, Riyadh 11437, Saudi Arabia

Correspondence should be addressed to Ghalib H. Alshammri; galshammri@ksu.edu.sa

Received 23 August 2022; Revised 19 September 2022; Accepted 6 October 2022; Published 28 March 2023

Academic Editor: Jaroslav Frnda

The widespread availability of cutting-edge computer technologies has shed light on the relevance of artificial intelligence (AI) applications in almost all sectors of the economy. As a result of the incorporation of voice control processing into many Internet of Things (IoT) devices, many of these IoT devices may be operated using spoken commands. The environment that is controlled by speech may include several devices, each of which may be used for a separate activity; yet, all of the devices may collect and process the same command at the same time. This may be the case if the devices can communicate with one another. Because other devices may choose to ignore orders that are intended for particular devices if those devices are not equipped to deal with those orders, only the device that is designed to carry out the activity and process the command will be able to carry out the activity. This is because only the device that is designed to carry out the activity and process the command will be able to carry out the activity. On the other hand, when all of the voice-controlled devices capture the command through the microphone, there is a greater chance that it will mix with other sounds coming from a variety of sources. This is because the microphone is being used to capture the command from all of the voice-controlled devices. These noises may include those that are emanating from the television, music systems, and other sounds that are created by activities taking on inside the family, among other things. During the identification of instructions via processing, any blending of other sounds that are not the primary command is regarded as noise and has to be deleted. This is because any such blending is deemed to be noise. The direction of arrival (also known as DOA) of the sound waves is given primary consideration by this approach. This is done at the same time as the performance of the system, and the proposal for it are being evaluated. Based on the angle of arrival estimate, a specific room impulse response (RIR) from a collection of defined RIR is identified as a room acoustic characteristic, and source separation is carried out using the technique of independent component analysis (ICA). Following the completion of the analysis of the signals produced by the split command speech, the characteristics of the speech are retrieved from the signals. The Mel-frequency cepstral coefficients (MFCC) approach is used so that the operation of feature extraction may be carried out. This is the goal of the technique. Following that, a support vector machine classifier is used to the data in order to further split these characteristics into a large range of distinct groups. Comparisons are made between the performance of the SVM classifier and the performance of a large number of different classifiers, including decision trees, which are often used in applications that incorporate machine learning (DT). After analyzing its performance, the multiclass SVM classifier is found to have an accuracy of 91%, according to the conclusions of the study. Utilizing a classifier that is based on a probabilistic neural network, which is sometimes referred to as a PNN, is one way in which the accuracy of future classifications may be enhanced. This particular classifier is made up of three layers: one layer of gated recurrent units (GRU), one layer of long short-term memory (LSTM), and one layer that integrates the two of those different kinds of memory. This classification seems to have obtained an accuracy of 94.5 percent, which is higher than the classification accuracy attained by the multiclass SVM classifier.

1. Introduction

The vast majority of unpaid track royalties are mixtures of various sound assets, such as units or human voices. At this point in time, the bad effects and unfavorable effects of specific units are sometimes no longer accessible in distinct amounts. Even taking into consideration the fact that the units are recorded individually, the isolated side effects and side effects do exist, despite the fact that they are often not disseminated at this time [1]. This is the manner in which well-known track manifestations do their business the vast majority of the time. It is not an uncommon procedure for musicians to carry out their job while recording the bulk of the instrument pointers for a select number of orders, some of which include jazz, traditional track, or society. This practice is not rare since it is not an unusual activity. In this particular case, there is not an outstanding amount of disengaged currency owed [2]. In addition, as of today, the cash payment for a single gadget that was due for a variety of distinct vintage track manifestations has been missing. Despite this, getting closer to the precise equipment markers is vital for tracking cash due for a couple of usage seasons. It may be necessary, for example, to exclude the vocals of the original artist from the track that will be used for karaoke in order to obtain the desired effect. In order to make another track piece, a couple of skilled specialists are required to eliminate one instrument from the total amount and then remix it with the given quantity of cash that is due. Craftsmen have the ability to create their own coordinate tracks by withdrawing their tools and supplies from a track on which they still have room to improve their abilities [3]. When it comes to the market for consumer audio products, a combination will often have the standard number of channels included within it. In order to play back a recording on a system that has more channels, one may need to "up mix" the recording and replace the spatial area that was occupied by an instrument. This allows the recording to be played back on the system that has more channels. In addition, a personalized study into monetary commitments is carried out by way of an introduction to the pertinent monetary assets. Examples of models include replicating the singer persona, recreating the instruments or language of a song, and reporting the stanzas or the melodic score of a song [4].

1.1. Blind Source Separation. At the tail end of the 1980s, a strategy of managing solid indications known as the blind stock segment, which is sometimes abbreviated as BSS in certain circles, gained popularity. BSS was what was happening in present compositions in the fields of biological sciences, talk sign communication, image management, geography, and mining for insights from literary content [5]. This was accomplished via the utilization of certified sign management, phoney psyche connections, and record theory. The term "source separation" (SS), which is also known as "signal separation," was coined because the methodology focused on convalescing a gathering of difficult to comprehend supply alarms "(time series, pictures from a gathering of insights (for example expected alarms), which can be combinations of those supply alarms)." In this manner, the term "SS" was portrayed. The BSS arrangement is relevant to the circumstance, but the possible limitations and benefits of the concept are not discussed here. Figure 1 shows the blind source architecture.

Picture spotlight extraction, face confirmation, moving item ID, widespread picture watermarks, picture denoising, picture separation, and picture recovery are typical applications of BSS basically based totally on picture management [6]. The recovery of the support is the cognitive process that is being tested in this experiment. This reenacts an extra mounted literary material that was previously deleted and in some cases rewritten. Figure 1 illustrates the red, green, and blue channels in isolation from one another.

1.2. Machine Monitoring. Signal parcel could be used to protect against potential mechanical frustrations by isolating the acoustic component (in this particular scenario, the sound produced by a mechanical framework all through hurt) [7] from the weather conditions involving a combination of different assets that include various likely normal running parts. This would allow for the acoustic component to be protected from any potential mechanical issues.

1.3. Investigative Procedures in Medicine. It is common practice to use a wide variety of logical tools in order to investigate a particular constellation of adverse effects and adverse effects originating from the human body (as illustration EEG signals, and ECG signals). A signal parcel might be used to separate a sign of relevance that is connected with a selected significant cycle (or) lifts that is likely available. This separation could be accomplished by using a signal parcel. It is possible that this will prevent an erroneous diagnosis brought on by the influence of clatter inside the system [8].

This is a selected piece of programming that may be used globally in order to preserve melodic displays at 0 in just on certain instrument sounds and voices included inside the recordings. This might be specially employed to boost a chosen device, such as a percussion instrument that has to be somewhat more intense than expected [9].

The measurements provided by the Little Array are a very beneficial way of typifying the facts that have been obtained about DNA and protein explanations. Blind stock parcel in multicollector exhibiting depiction may probably be used to seclude the gene [10] significance, to accommodate as an option for delayed progressions, district of periodicity, grouping, and order of characteristics. Blind stock parcel in multicollector exhibits depiction.

The suggested VAD algorithm that is addressed in the proposed work is dedicated to building a deliberate VAD algorithm that is ideal for ASR systems. This is because simplicity and robustness are two of the primary issues that VAD algorithms face when used for ASR systems. Not only is this technique durable in a variety of different noisy settings [11], but it also has a low computing cost. The suggested method includes the incorporation of a noise estimate to characterize the various ambient noise signals. This estimation is then used to update the noise for each frame.

$$y(t) = H_{\cdot} x(t). \tag{1}$$

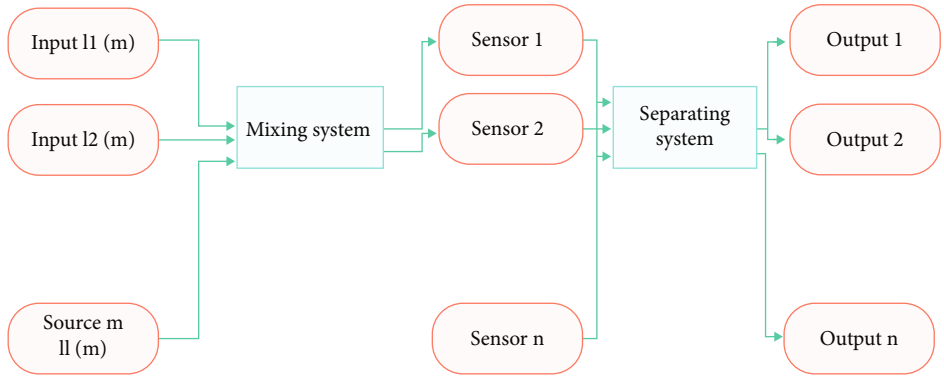

Figure 1: Blind source system architecture.

There have been many different VAD algorithms suggested up to this point. In order for these algorithms to be effective in combating the myriad of external sounds, a variety of discriminating characteristics have been implemented. The energy-based characteristics are the most sought-after of all of them due to the fact that they are both straightforward and efficient. In addition, it is challenging to build an effective combination algorithm that is appropriate for all situations. Traditional characteristics such as zero-crossing rate, pitch-based detection, and weak fricative detection are not used since they are not noise-immune. Other conventional features that are not utilized include weak fricative detection. These distinguishing traits have cheap calculation costs, which makes them excellent for automatic speech recognition because of their applicability.

Popular speech enhancement algorithms have been developed with the primary goal of improving the intelligibility and/or quality of the speech signal. When developing these algorithms, designers did not take into account the impact that their decisions may have on other speech processing systems.

The intelligibility of speech is not improved by the speech enhancement algorithms, not even the most advanced ones. The reason for this is that it does not possess a reliable estimate of the background noise spectrum, which is essential for the execution of the majority of algorithms. Accurate voice-activity detection techniques are necessary in order to accomplish this goal. It has been possible to make significant strides in the development of algorithms for noise estimation and voice activity detection in recent years. These algorithms are now able to constantly monitor the mean of the noise spectrum, at the very least. It is well known that algorithms for estimating noise work admirably in settings with a constant background noise level. As evidence of this, a little increase of 10% in intelligibility was noticed with speech processed in automotive contexts, but this was not the case with speech processed in other situations (e.g., babble). The constancy of the automobile noise, which made it possible to make reliable measurements of the noise, was thought to be the cause of the little improvement [12].

An accurate noise estimate may provide a contribution to increases in intelligibility, but it cannot, on its own, produce significant gains in intelligibility. This is because it is necessary to precisely monitor the spectrum of nonstationary noise. The fact that there is no discernible gain in intelligibility using the currently available speech enhancement algorithms is not just attributable to an inaccurate lack of estimations of the noise spectrum.

The single-channel speech augmentation scenario, in which only the distorted version of the original speech is available for recovery, is the one that poses the most challenge. When input signals are distorted as a result of background noise, the objective is to maximize the performance of voice communication systems while minimizing the influence of noise as much as possible.

A signal is said to be nonstationary if its power spectrum is constantly shifting over a number of distinct frequency bands. It may be difficult to determine if a portion of a nonstationary signal contains speech or noise for that portion of the signal. There are a great number of factors that contribute to the difficulties that often accompany voice recognition.

Natural speech is uninterrupted; in most cases, there are no gaps in between the individual words. As a result of this, among other things, it is more difficult to discern where the borders of the word are. Changes in global or local rates of speech, pronunciations of words within and among speakers, and phonemes in various contexts may all contribute to variations in natural speech. A large vocabulary can lead to misunderstandings. Variability in recorded speech may be attributed to factors such as the acoustics of the recording space, the characteristics of the channel and microphone [13], and the level of ambient noise. The updating of the estimate is only allowed during times when there is no speech when using traditional techniques of noise estimating, which are dependent on the detection of voice activity. In addition, VADs are notoriously difficult to tune, and their dependability is greatly compromised when dealing with speech components that are weak or when the input SNR is low. Therefore, the pace at which the noise estimate is updated is considered to be somewhat slow.

The estimates of the signal-to-noise ratio (SNR) and the noise powers are key difficulties in speech processing. The signal-to-noise ratio (SNR) is an important measure of the speech quality index that is used often in the process of data collection and classification. Additionally, speech improvement systems make use of the estimate of the local noise powers. Estimating the signal-to-noise ratio (SNR) and the local noise power of noisy speech may be very challenging in many situations since neither a clean reference signal nor speech activity is provided.

2. Survey of Related Work

In its most basic form, the procedure involves the processing of sound by use of an analog-to-digital converter (ADC), which converts analog sound signals into digital ones. In order to prepare the sound for subsequent processing, the technology removes noise and distortion. A sound signal is chopped up into smaller pieces at predetermined intervals and given a number. The algorithm that was created compares the produced phoneme to the framed phoneme that is utilized during the training phase. A number of different algorithms are used in order to carry out the pattern matching procedure for phonemes, words, and phrases. It is still a developing topic in terms of study areas [14], since the accomplishment of accuracy and performance is a tough problem. Word error rate, often known as WER, is the metric that is used to evaluate the performance of voice recognition systems. This metric detects words and creates mistakes that occur in the transcription representation. In recent years, there has been a rise in interest in "voice recognition technology," which is another name for the speech recognition system. However, throughout the last 65 years, this scientific field has seen significant change. SRS may be helpful in a variety of different real-time applications. The expansion of SRS will mostly take place in the following three areas: vocabulary size, speaker independence, and processing speed [15]. A method for recognizing single digits under the name "Audrey" was used. In the same year, Gunnar Fant created the source filter model for speech creation, which was the combination of vocal cords and an acoustic filter. This model was used to explain how speech is produced. Applications relating to voice proposed work and speech recognition found the source filter model to be helpful. Speech recognition technology known as "Shoebox" devices were first launched by IBM in 1962. These computers could understand 16 English words. At this point in time, Soviet researchers had constructed an analyzer for a vocabulary of 200 words and discovered the DTW (dynamic time warping) method. [16, 17] saw the beginning of the DARPA SUR (speech understanding research) project, which received funding from the Defense Advanced Research Projects Agency (DARPA) for a period of five years.

In 1987, a doll for kids named "Julie" was introduced, and it has the capability [18] of being taught by kids to reply to what they say. Dragon Dictate was the first consumer speech recognition tool for voice recognition call processing when it was originally made available by Dragon Company in the year 1990. In 1997, after making a few alterations, they came up with the idea for "Dragon NaturallySpeaking." In 2007, Google introduced its first product, GOOG-411, which was a directory that could be accessed through telephone. Siri, a virtual assistant developed by Apple and based on cloud computing, was made available to users in 2011 and responds to voice inquiries using a natural language user interface. Siri is able to answer questions or carry out tasks depending on the context. Google announced the availability of its voice search capabilities for iOS devices [19]. Since 2014 and continuing to the present day, the demand for virtual assistant services such as giving information and carrying out tasks has increased significantly. Amazon Alexa, Google Assistant, Amazon Echo, and Microsoft Cortana are just some of the most well-known examples of virtual assistants. Transcribing conversational telephony speech using several deep learning algorithms was one of the tasks that Microsoft researchers worked on in 2017 with the goal of achieving improved accuracy. Table 1 shows the existing methodology comparison.

The currently available voice recognition systems for Indian languages have reached a sufficient level of development and performance. But every system that is now in use has included an auditory and linguistic model. It was discovered that any system for English that is now available works solely on English characters and words. The researchers have not taken into account the most significant difficulties that pertain to digits, punctuation marks, and the end of utterance (EOU) [27]. It is far from exclusive that recurring components do not replace throughout the course of time while using DFT for the most significant part of the examination. As a consequence, the length of the window would not have an impact on the DFT outcomes, and banner houses hold as soon as from the beginning to the farthest limit of the window. Because it does not provide any measurements taken at the precise moment in time at which a repeat component takes place, a single DFT evaluation is often not sufficient for determining the significance of such markers [28]. STFT, also known as time-dependent Fourier transform (TDFT), is a notion that has been presented in order to govern such points. STFT is generally considered to be an important concept in conversation monitoring applications [29]. The goal of this project is to construct a model for voice recognition for the English language that takes into account English characters, numerals, and punctuation marks. To create a model that can identify the 'end of utterance' in speech, which is notoriously difficult to do. Furthermore, there is room for the development of the model described above, which has the potential to reduce the amount of time needed for the pattern finding process.

3. Methodology of Proposed Work

The following objectives will be met by the suggested works of art for supply primarily based on totally sound separation and request confirmation in IoT primarily based on fully sharp homegrown systems:

(1) To nurture new forms of important learning that are fundamentally founded on comprehensive supply segment computation for use in sound and request attestation

(2) To promote substantial learning primarily based on fully request test computation in voice-oversaw Internet of Things primarily based on utterly outstanding homemade situations

(3) To select how the suggested structure will be presented using a variety of sound combos and styles that will be taken into consideration

Table 1: Existing methodology comparison.

Ref.	Language	Techniques used	Category	Accuracy
[20]	Hindi	LVQ learning	Speech	95%
[21]	Hindi	ANN, RNN	Speech	94.63%
[22]	Assamese	CMU sphinx tool	Speech	93%
[23]	Assamese	HMM	Words	92.56%
[24]	Urdu	MFCC	Words	91.01%
[25]	Urdu	HTK	Speech	90.26%
[26]	Marathi	Sphinx 3 trainer	Speech	86.21%

(4) To dismantle the implementation of the suggested method regarding room acoustic homes

(5) To bring down the curtain on existing structures that are accessible for voice-controlled Internet of Things devices

3.1. Methodology. The improvements to the calculation of the supply sector's sound impact are included in the proposed changes. The sound dividers are mostly reliant on combination designs that count on the greatest possible massive component to fulfill the proposal ascribes of the space acoustics. It is predicted that the room inspiration response combination styles would be developed for the purpose of surveying the precise sound districts of the collector group area. Figure 2 shows the proposed work model.

3.1.1. Datasets. The sound barrier between the source and the receivers is determined by the direction of appearance (DOA) along the area of the receivers located within the room. The room's motivating response has an effect on the receiver's area as well as the sound it catches. Therefore, even when evaluating the framework's performance, such obstacles may be thought of. The suggested paintings have a variety of gestural elements, which are shown in discernment 1. The records and their respective warnings are first merged together, along with the concept of room motivation reaction. Because of the influence that the acoustic properties of the room have, the room pressure response is the most important factor that affects the sound symptoms and symptoms even as detecting and analyzing their individual roles in their individual sources. In spite of the fact that there are a number of different management hubs connected to the management, the original management is carried out according to the arrival direction (DOA). When compared to direct capture, room motivation response primarily based on fully mixing is dependable to develop nonstop acoustic results. This is because of the fact that it follows the guiding fundamental of amplifier display. In addition, the active mixing model, commotion effects, and sometimes the division of sound using the CBSS approach are detailed in this section.

3.1.2. Preprocessing Stage. The first thing that we do is provide the important rendition of convolutive mixes. A collection of N supply alerts, denoted by the notation $s(t) = (s1(t), \cdots, sN(t))$ are obtained from a set of M sensors at the discrete time interval t. The alerts that were sent out are suggested by the equation $x(t) = (x1(t), \cdots, xM(t))$. It is anticipated that the reassessment will be convolutive (or capably) combined in a few of relevant projects. The convolutive version provides the following connection between the mth contradictory message, the vital stock signals, and a pair of brought substance sensor disturbance $v_m s(t)$:

$$x_m(t) = \sum_{n=1}^{N} y \sum_{k=0}^{K-1} y\, a_{mnk} s_n(t-k) + v_m(t). \tag{2}$$

The going against message is a short sum of filtered assortments of each of the stock signs, and a_{mnk} addresses the related blending channel coefficients. The going against message may be thought of as a counterargument. In any case, for the sake of simplicity, we will assume that the coefficients will have a steady-to-combination variation. All things considered, those coefficients may also in like manner extrude throughout the span of time. In the statute, the channels are most likely of an unknowable length (which is most likely accomplished as IIR systems); despite this, when taken as a whole, it is more than adequate to anticipate that it will be K. It is possible that the convolutive adaptation will be formed as a result of the system structure:

$$x(t) = \sum_{k=0}^{K-1} y\, A_k s(t-k) + v(t), \tag{3}$$

where A_k is a M by N matrix that holds the kth filter coefficients and k is the index number of the matrix. The $M1$ noise vector is denoted by $v(t)$. It is possible to write the convolutive combination as follows in the z-domain:

$$X(z) = A(z)S(z) + V(z), \tag{4}$$

where $A(z)$ is a matrix that has FIR polynomials embedded inside each entry [22]. In special cases, there are certain unique applications of the convolutive mixture that may be simplified using equation (3). Assuming that all of the signals reach the sensors at the same time and are not filtered in any way, the convolutive mixture model may be simplified to

$$x(t) = As(t) + v(t). \tag{5}$$

The immediate or delay-less (straight) combination version is the name given to this variant. In this situation, $A = A0$ refers to a M by N grid that contains the mixing coefficients. In order to address the problem of the short aggregate, a number of computations were developed; for examples, see [15, 23]. In until further notice sources, assuming a resonance loose weather with engendering defers, the mixing version may be moved ahead to account for this.

$$x_m(t) = \sum_{n=1}^{N} a_{mn}\, s_n(t - k_{mn}) + v_m(t), \tag{6}$$

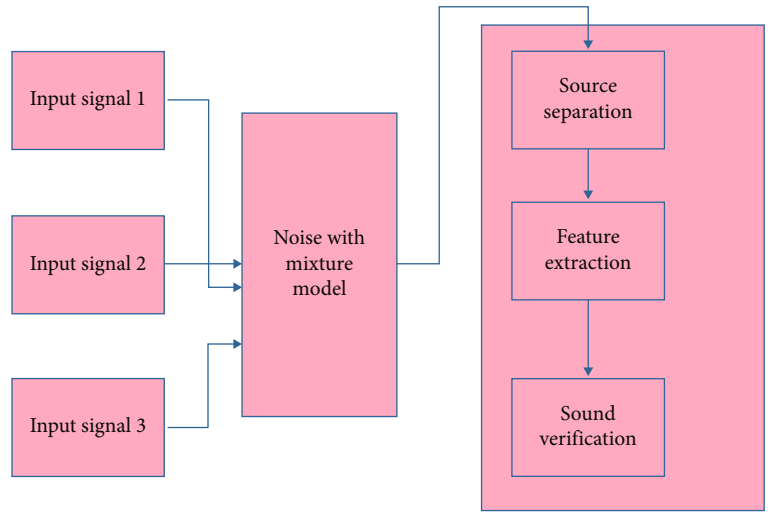

Figure 2: Block diagram of proposed work.

where k_{mn} denotes the time it takes for the signal to travel from source n to sensor m. In the process of deriving various algorithms, it is assumed that the convolutive model is noise-free, which means that there is no background noise in the model:

$$x(t) = \sum_{k=0}^{K-1} y\, A_k s(t-k). \tag{7}$$

In over and undersettled sources, frequently, it is miles expected that the quantity of sensors would draw near (or outperform) the quantity of reassessment, in which case, in an instant, strategies may work to reverse the quick blending. In any case, in the event that how much resets outperforms how many sensors the trouble is currently at this point not permanently set up, and surprisingly, underneath best realities at the blending system, methodologies cannot get well the resets to be composed withinside the repeat region quickly as fair-minded will increment for every repeat region:

$$X(\omega) = A(\omega)S(\omega) + V(\omega). \tag{8}$$

In most cases, the frequency transformation is calculated by using a discrete Fourier transform, often known as a DFT, within a time frame of length T, beginning at some point in time t:

$$X(\omega, t) = \mathrm{DFT}([x(t),\cdots,x(t+T-1)]), \tag{9}$$

and correspondingly for $S(!, t)$ and $V(1, t)$. Often a windowed discrete Fourier transform is used:

$$X(\omega, t) = \sum_{\tau=0}^{T-1} y\, w(\tau)x(t+\tau)e^{-j\omega\tau/T}, \tag{10}$$

where the window function $w(t)$ has been used because of the restricted temporal aperture in order to reduce the amount of band overlap. Because massive time-domain filters are commonly needed in acoustics, the fast Fourier transform (FFT) enables effective convolutions to be done in the discrete Fourier domain. This is crucial because convolutions may be done efficiently in the discrete Fourier domain. In frame blocking, the more generalized form of voice signal that is created by preprocessing and that will be used as input data will be called for. A series of speech frames is created from the input signal $s1(n)$, which represents the next occurrence of $s(n)$. A window will be formed out of the collection of speech frames. The steps involved in blocking and windowing are shown in Figure 3.

3.1.3. Feature Extraction. The decoding approach is used in order to locate the utterance feature vector that provides the greatest possible fit with the acoustic model. The process of decoding makes use of a dynamic programming method that is known as the Viterbi algorithm. The decoding step consists of modeling the acoustic data, the language data, and the pronunciation data. In continuous voice recognition, the system may identify an endless number of different sequences while trying to determine the optimum route of word sequence W for the input signal X. The Viterbi algorithm searches for the answer that is most plausible in light of the given speech. In deciphering looking for the word $W*$, it is possible to define it as in

$$W^* = \mathrm{argmax}_w(p(X|w)p(w)), \tag{11}$$

where $p(w)$ calculated from language model $p(X|w)$ can be calculated from available sequence of phonemes of words available in the dictionary in

$$p(X|w) = \mathrm{argmax}_s \pi\left(p\left(x|s_j\right)p\left(s_j\right)\right). \tag{12}$$

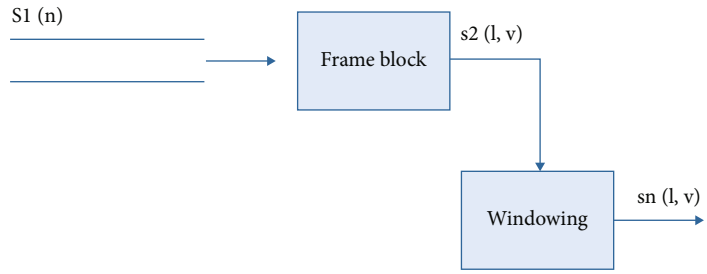

Figure 3: Preprocessing stage.

Figure 4: MFCC Architecture.

In MFCC computation, windowing process is performed by dividing speech signals into different frames. FFT algorithm is applied to calculate power spectrum for each frame. Filter bank processes the power spectrum using Mel-scale. Then, discrete cosine transform (DCT) is applied to speech signal to translate power spectrum to log domain to calculate MFCC coefficient as shown in Figure 4.

$$\text{mel}(f) = 2595x \log_{10}\left(1 + \frac{f}{700}\right), \tag{13}$$

where $\text{mel}(f)$ is the frequency and f represents frequency in (Hz). The MFCC calculation formula is

$$C_n = \sum_{k}^{n=1} (\log S\hat{}k) \cos\left[nk - \frac{1}{2}\frac{\pi}{k}\right]. \tag{14}$$

(i) The speech uttered by speaker in continuous form is entered as an input for the model which is shown in Figure 4

(ii) To differentiate voiced and unvoiced segment, threshold value is defined. If the uttered sound is less than the threshold value, it will be considered as unvoiced

(iii) For word detection process, from the voiced speech sequence of words is generated as $W(N)\bullet$. For each utterance of $W(i)$, feature vector is extracted

(iv) Pattern matching process is performed on $W(i)$, to match from the language model

(i) Word pattern can be words, digits, and punctuation mark. For word utterance, the best matched pattern gets it displayed in text format

(ii) Rule-based approach is implemented with the proposed model for digit and punctuation mark recognition

(iii) Rule-based approach stated that to append every digit pronunciation with "५:" to differentiate uttered digit with word sequences. Same approach can be used for punctuation mark representation that every punctuation mark pronunciation should end with "५:" to represent in symbol form and not in form of character sequence

(iv) If unknown word occurs, it matches with most relevant phoneme for the representation

Figure 5 shows the flowchart of the speech feature extraction model. The recommended model will perform an operation known as pattern matching on the uttered words. This operation is carried out on the spoken words. In the case that the said word does not locate a perfect match for itself, the algorithm will choose the most closely matched occurrence of that word. The model that has been suggested can take in data from a broad range of speakers with a diversity of accents and sound quality. For speech recognition to perform at its most efficient level, a significant quantity of training data is necessary. The gathering and preparation of a considerable quantity of data for training are carried out. There will be a total of eight different persons contributing their voices to the recording. These individuals will range in age, gender, and speaking style. The voice recognition model can accommodate a large number of users

Pseudocode of proposed methodology.
Determine the population size, the dimensions of the issue (NP, D), the crossover rate (c r), and the scaling factor (F).
Initialization: Establish a starting point for the population S i by setting its initial value to s $(1, i)t, s(2, i)t, \cdots, s(D, i)t$ with each individual being equally distributed in the range [s s"low," s"high,"]
Despite the fact that the requirements for termination are not fulfilled
In the population NP, for each unique goal vector, there is a total of NP
Using the following equation for mutation, choose three individuals at random from the population, and build a donor vector called *vit*.
$v(j, i)\hat{}t = s(j, p)\hat{}t + F\, i^{*}([\![\text{""}s]\!](j, r)\hat{}t + s(j, q)\hat{}t)$
Calculate the trial vector for the ith target vector using the following formula: $u(j, i)(t + 1)$:
$u(j, i)\hat{}t = \{\blacksquare(v(i, j)\hat{}t \,\&\text{"}\ \text{if"}\ \&\text{r i} \le \text{c r"}\ \text{or"}\ j = J\text{"}\ \text{rand"}@s(i, j)\hat{}\text{t\&\&"}\ \text{otherwise"})\dashv$
Apply the LSTM classifier as the fitness function f and then assess the *sit* and *uit* values as follows:
If $f(s\, i\hat{}t) \le f(u\, i\hat{}t)$ then $s\, i\hat{}(t + 1) = u\, i\hat{}t$
Else $s\, i\hat{}(t + 1) = s\, i\hat{}t$
Finish For
Finish While

Algorithm 1

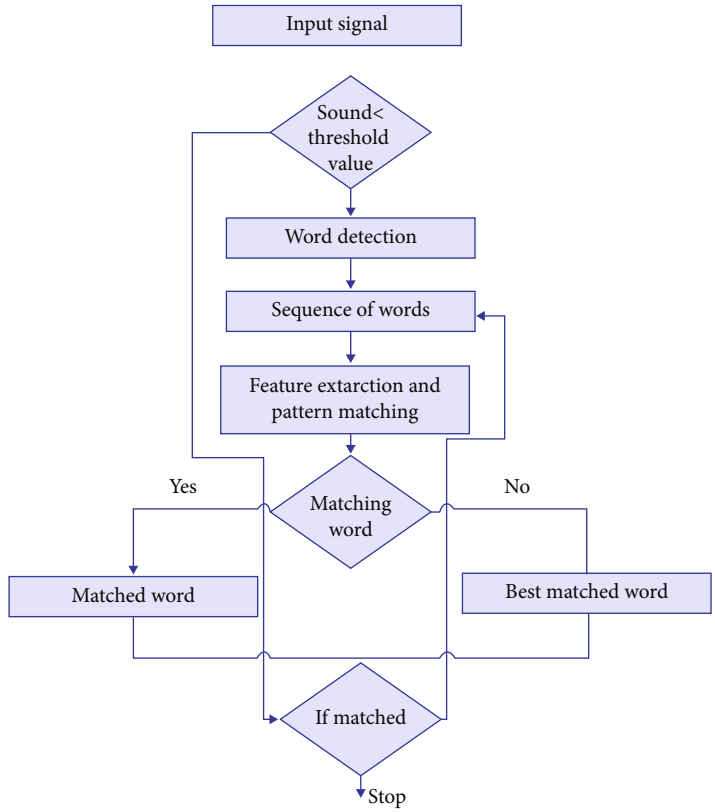

Figure 5: Flowchart of speech feature extraction.

simultaneously. The model has a threshold value that decides whether or not it will receive sound from a speaker. This value is determined by how loud the sound is. If the amount of voice in the sound is lower than the threshold value, then the sound will be classified as one that does not include any voice. The framework may be used in a different language by making certain fundamental adjustments to the way language is employed in its components. The model utilizes a method that is based on rules in order to properly represent numerals and punctuation marks. The framework may be used in a different language by making certain fundamental adjustments to the way language is employed in its components. The framework may be used in a different language by making certain fundamental adjustments to the way language is employed in its components. The graphical user interface that was built for the process of speech recognition is both user pleasant and includes all of the key interface functions. This section displays the textual

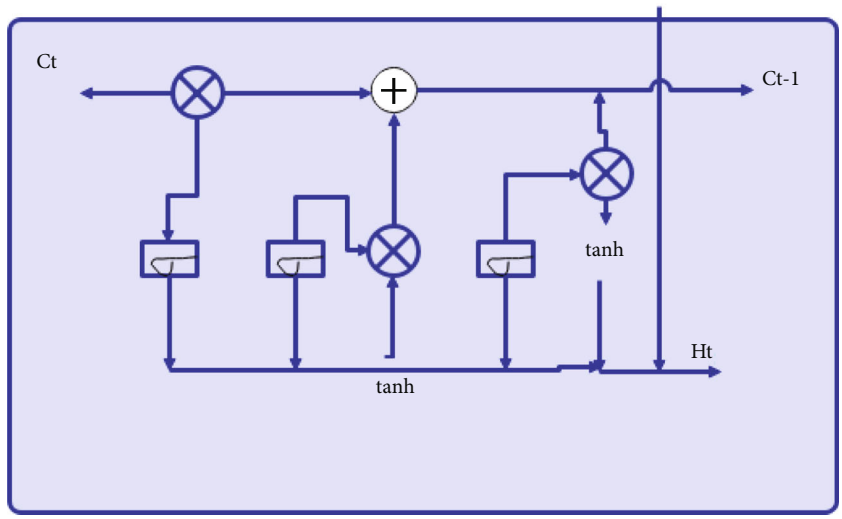

Figure 6: LSTM with GRU unit.

representation of the output that the recognizer generated. It is possible to generate a wide variety of output-oriented reports for the purpose of conducting further research and monitoring the recognition ratio and accuracy of the system. Even though the output is in the form of simple text, it is still possible to perform a number of text-based operations on it.

3.2. Classification. Candidate layer:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i). \tag{15}$$

Input gate:

$$C\sim_t = \tan h(W_C \cdot [h_{t-1}, x_t] + b_c). \tag{16}$$

Output gate:

$$C_t = f_t * C_{t-1} + i_t * C\sim_t. \tag{17}$$

Hidden state:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o). \tag{18}$$

Memory state:

$$h_t = o_t * \tan h(C_t), \tag{19}$$

where W and b each represent a weight vector for the forget gate (f), the candidate (C), the input gate (I) and the output gate (o), respectively. The symbol "$*$" indicates element-by-element multiplication, while the "σ" symbol symbolizes the sigmoid function. The LSTM's internal structure is shown in Figure 6, which depicts the state of the system at time step t.

Even though LSTM solves the issue of vanishing gradients, gated recurrent unit (GRU), a generalized version of LSTM, was developed. GRU is a generalized variation of LSTM [4]. The GRU is a kind of unit that, like the LSTM unit, includes gating units that influence the flow of information inside the unit. However, unlike the LSTM unit, the GRU does not have distinct memory cells. The gated recurrent unit, or GRU, is responsible for calculating two gates referred to as the update gate and the reset gate. These gates are responsible for controlling the flow of information through each hidden unit. The following equations are used to determine the value of each hidden state at each time step t:

Update gate:

$$z_t = \sigma(W_z \cdot [h_{l-1}, x_t]). \tag{20}$$

Reset gate:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]). \tag{21}$$

New memory:

$$h\sim_t = \tan h(W \cdot [r_t * h_{t-1}, x_t]). \tag{22}$$

Final memory:

$$h_t = (1 - z_t) * h_{t-1} + z_t * h\sim_t, \tag{23}$$

where W denotes weight vector, * denotes element wise multiplication, and σ is the sigmoid function. Figure 6 demonstrates the internal structure of LSTM at time step t. Figure 6 shows the LSTM with GR unit.

LSTM calculates the hidden states by a set of equation as follows:

$$i = \sigma\left(x_t U^i + s_{t-1} W^i\right), \tag{24}$$

$$f = \sigma\left(x_t U^f + s_{t-1} W^f\right), \tag{25}$$

$$o = \sigma(x_t U^o + s_{t-1} W^o), \tag{26}$$

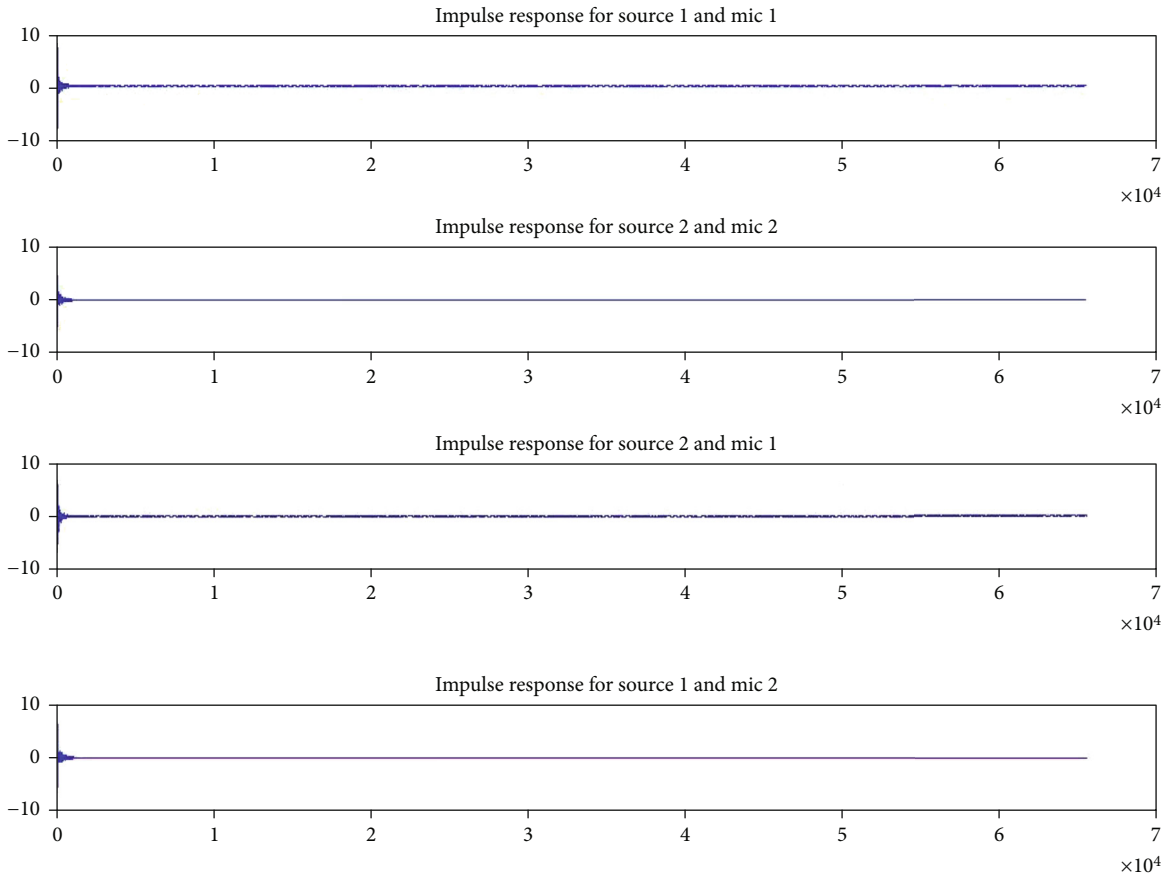

Figure 7: Room impulse response for node 1.

$$g = \tan h \tan h \left(x_t U^g + s_{t-1} W^g\right), \tag{27}$$

$$c_t = c_{t-1} \circ f + g \circ i. \tag{28}$$

Under these circumstances, the realities entrance, forget about entryway, provide doorway, and self-drawn-out independently are related with the letters I, f, and o, respectively, as seen in the previous sentence. Due to the fact that they provide the sense of being confused, the LSTM conditions might be discussed in more detail. I demonstrate how a large quantity of the spic and span measurements might possibly be supported by the memory mobileular. It is necessary for f to take responsibility for the insights, which need to be removed from mobile memory.

c.t. may be referred to on the grounds that it is the internal memory of the memory mobileular, which is the way a truckload element wise broadening of previous inward memory country through method of method for the brush to the side entrance and elementwise duplication of self-inauspicious country with enter entryway. At the end of the day, sh*t is related with the spine chiller country, which is now as of now not really chosen using the technique of approach for part proper duplication of the internal memory with the end-product entry. In addition, the end product that was not consumed will ultimately become part of the atmosphere as a result of the process outlined in Equation (27), which is comparable to Equation (28).

4. Results and Discussion

Vocabulary size A speech recognition system's vocabulary has an effect on its performance, which includes its accuracy and the amount of time it takes to complete recognition jobs. The needs of the system are what guide the development of the vocabulary in its entirety. One has the option of picking from one of four different groups, and this decision is based on how the system is being utilized.

The suggested system has undergone preliminary testing and error checking, both of which were carried out with the help of the Python programming language and its execution. The reaction that is produced as a result of the inspiration provided by the space is what establishes the genuine direction of the glance evaluation. Regarding the mouthpieces located at node 2 and the receivers located in center 1, the room inspiration response has been applied. As a consequence, for exploratory purposes and to account for the possibility of mistake, we have utilized combinations of the key inspiration response data. These centers observed for random causes in a room may have varying reactions to the power of the room. The plot of the room's inspiration vs. reaction may be seen in Figures 4 and 5, respectively.

(1) A vocabulary that is very restricted, consisting of little more than tens of words at most

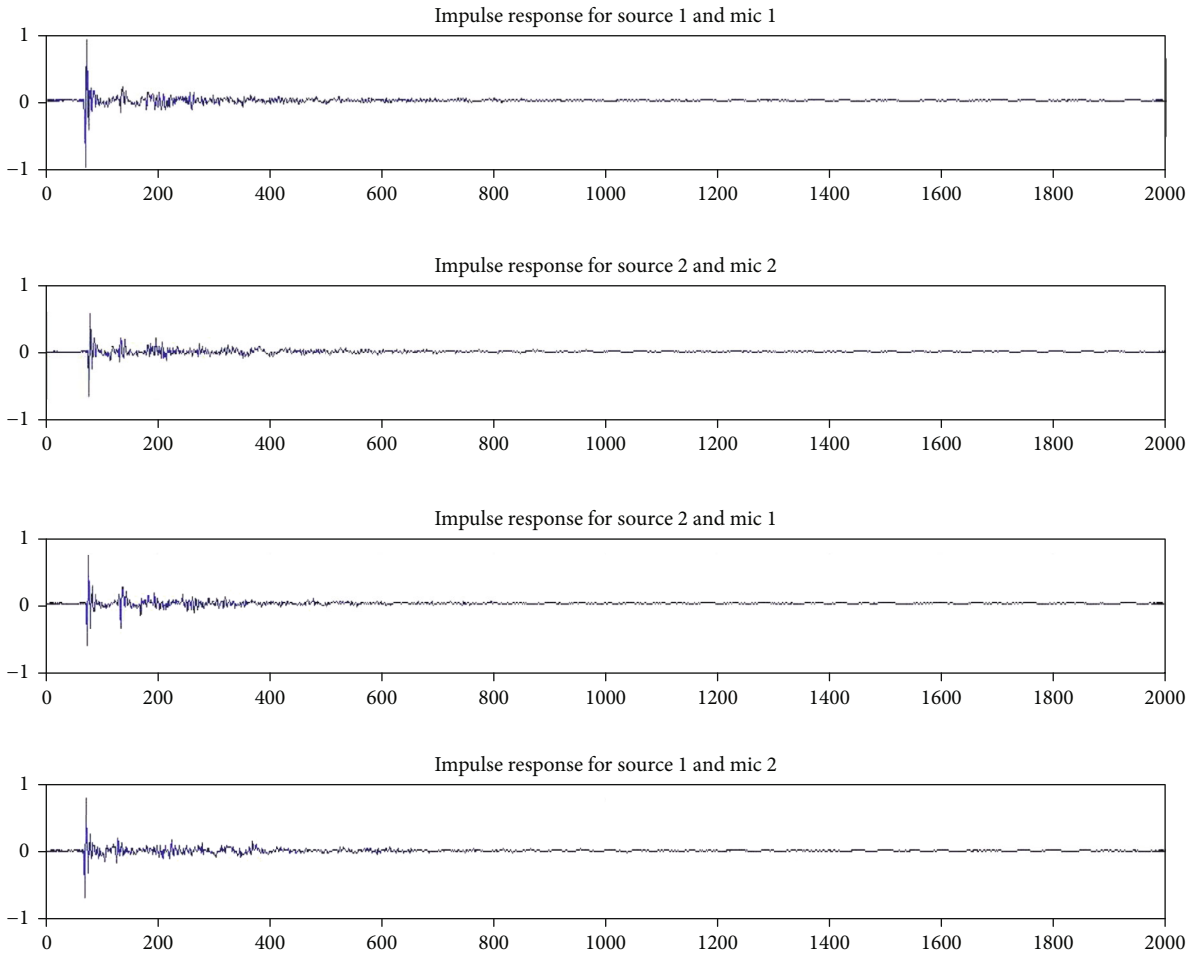

Figure 8: Room impulse response for node 2.

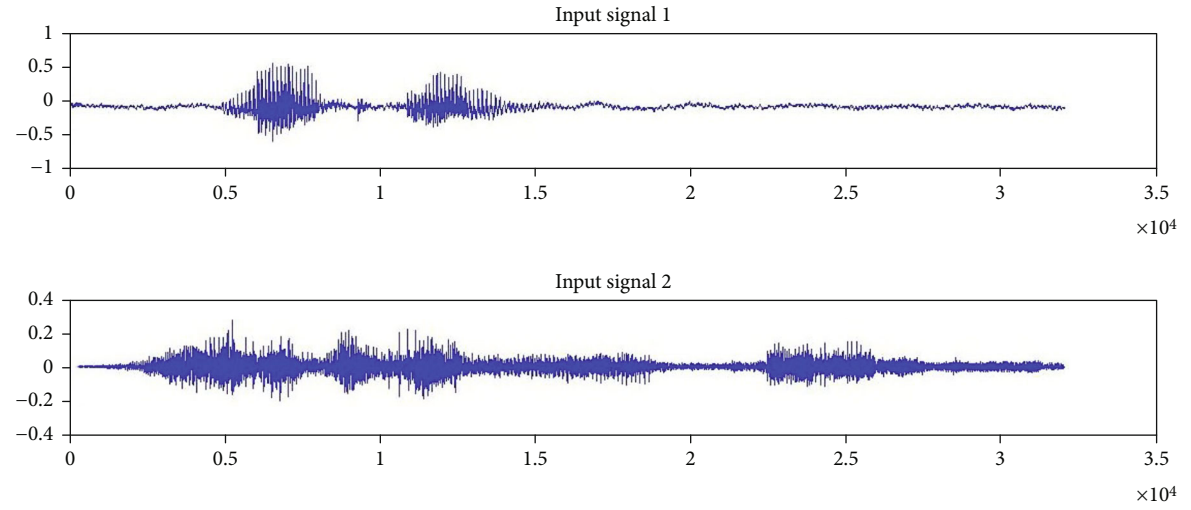

Figure 9: Input speech signals.

(2) A vocabulary of an intermediate level, comprised of hundreds of different terms

(3) A large vocabulary consisting of hundreds of different words

(4) A very extensive vocabulary that includes tens of thousands of words from a variety of sources Words that have the potential to be mixed up There are certain words in each language that, although having entirely distinct meanings, have the same sound.

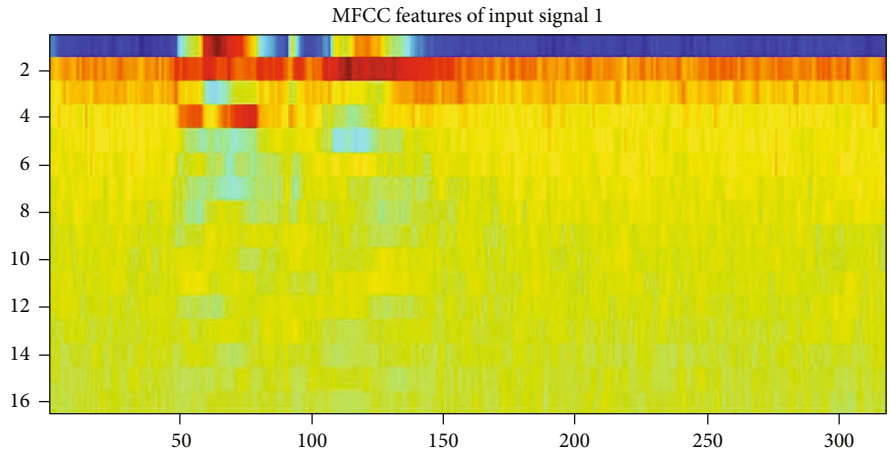

Figure 10: MFCC coefficient.

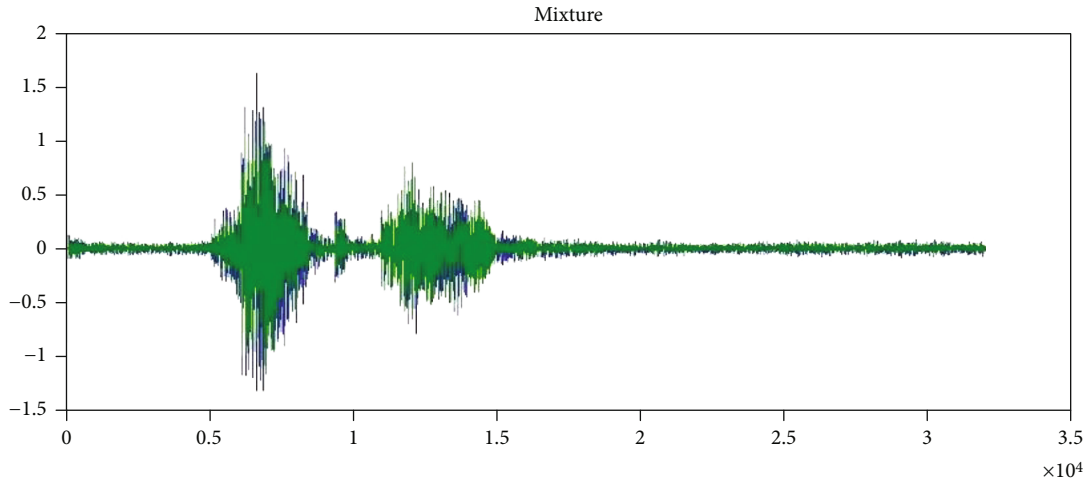

Figure 11: Mixed signal with respect to node 1 room impulse response.

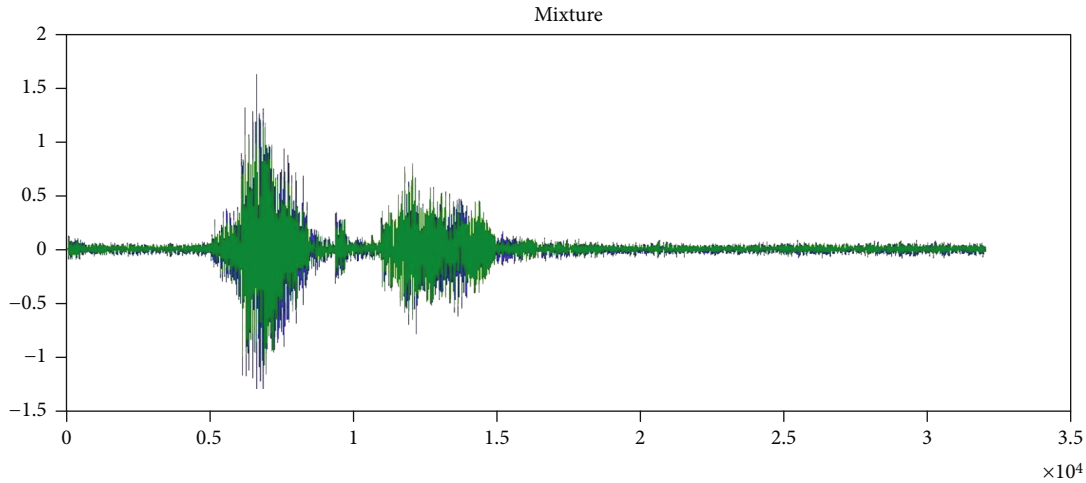

Figure 12: Mixed signal with respect to node 2 room impulse response.

These words might be misleading since they sound the same. The accuracy of the system suffers as a direct result of the confusion that is brought on by the usage of terminology of this sort. As an example, you may use ""and "", as well as ""and "", and so on. Speaker-dependent vs. speaker-independent: when

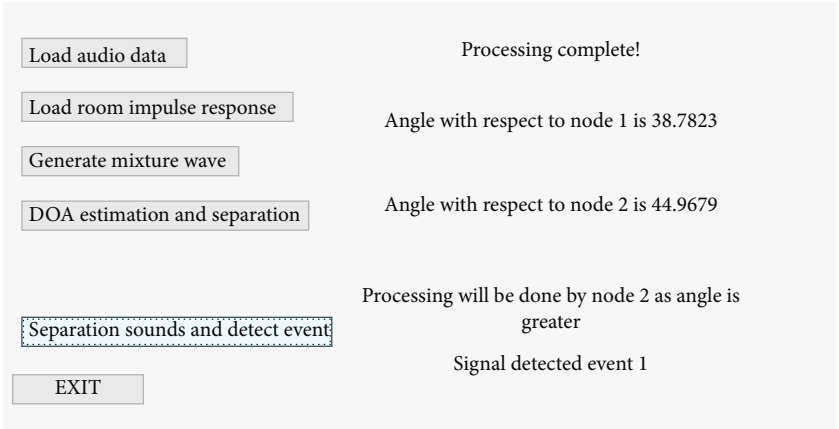

Figure 13: Result obtained for DOA and SVM-based event detection.

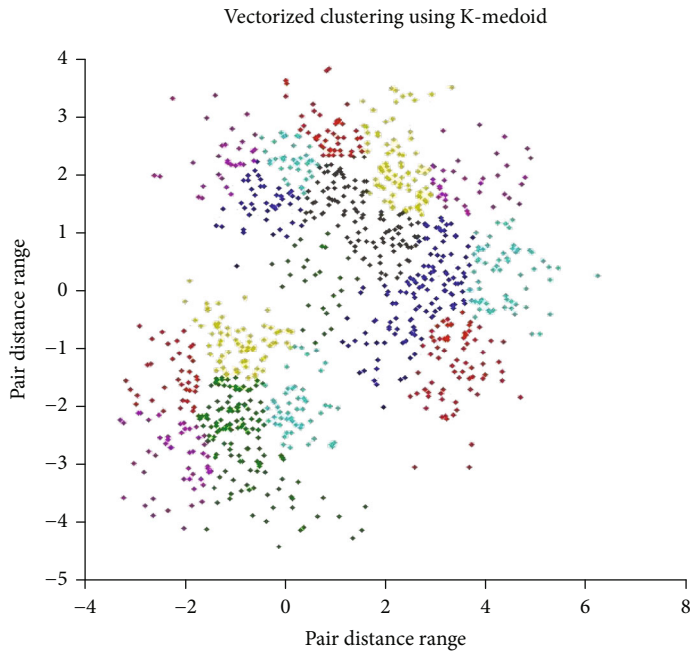

Figure 14: *K*-medoid clustering.

Table 2: Accuracy using machine learning methods.

Method	Accuracy	Specificity	Sensitivity
SVM	95.6%	0.91	0.89
KNN	93.2%	0.86	0.81
Naive Bayes (NB)	94.12%	0.89	0.82
Decision tree (DT)	94.05%	0.89	0.82

employing speaker-dependent systems, only certain speakers with constrained vocabularies are taken into consideration. A training phase must first be finished in order to save the speaker's utterances before the speech recognition system can be put into use. This phase must be finished before the system can be deployed. In comparison to the speaker-independent system, it has a greater degree of accuracy. A system that is independent of the speakers using it makes it possible for a large vocabulary to be utilized concurrently by several users. When compared to systems that are dependent on the speaker, the quantity of training data that is required here is far reduced. Isolated speech in addition to continuous and sporadic speech were both included. A better level of precision may be achieved by the use of individual words in the process of speech recognition when isolated speech is used. An example of discontinuous speech would be a string of words that are interspersed by silences in the course of a discussion. This kind of speech is easy to recognize since its boundaries are distinct. Continuous speech creates considerable difficulties in terms of real-time response as well as speech overlap, both of which have an impact on accuracy. These issues must be

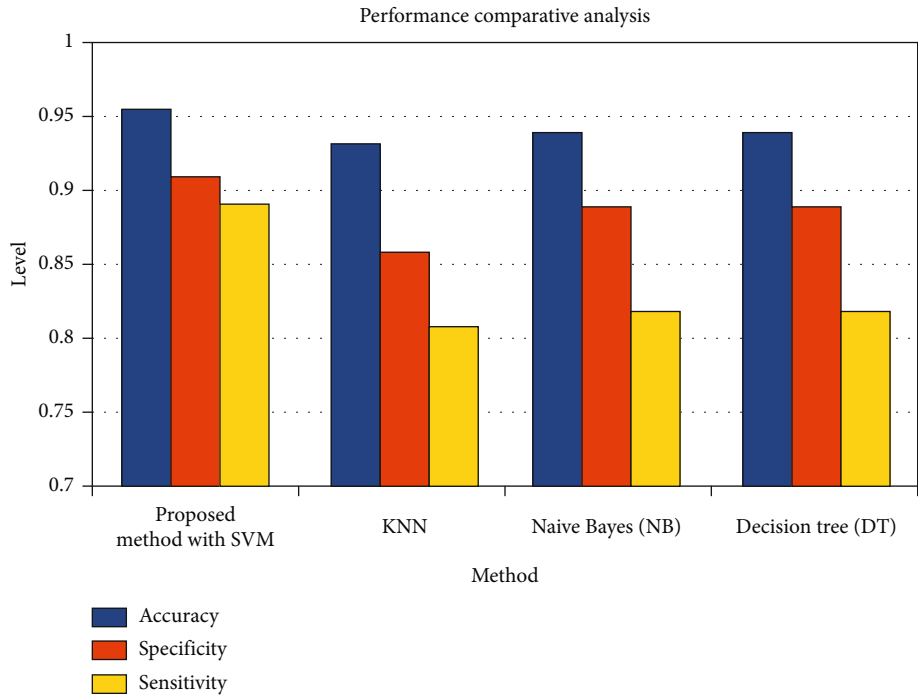

Figure 15: Graphical analysis of comparative methods.

Table 3: Accuracy using clustering and machine learning methods.

Method	Accuracy	Specificity	Sensitivity
SVM	96.80%	0.93	0.89
KNN	94.50%	0.89	0.82
Naive Bayes (NB)	94.12%	0.89	0.82
Decision tree (DT)	94.05%	0.89	0.82

Table 4: Deep learning-based model analysis.

Method	Accuracy	Specificity	Sensitivity
Model1	96.50%	0.91	0.9
Model2	96.80%	0.91	0.88
Model3	97.00%	0.89	0.89
Proposed model	98.50%	0.91	0.92

overcome. Taking notes on the circumstances, the circumstances under which the recording was made have an impact on the precision of the recognition. A loud setting, a broad channel bandwidth, the quality of the microphone, the speaker's dialect, and the speaker's speaking manner may all have a detrimental influence on the performance of a recognition system. Speech mistake: if the person speaking does not have a sufficient comprehension of the language or the topic at hand, then she or he will make mistakes in the pronunciation of specific words that are foreign to the person listening to them. The mental state of the speaker has an effect not only on the speaker's manner of speech but also on the data of their speech. Because of their worry, nervousness, and lack of self-confidence, the speaker uses a lot of unnecessary filler phrases and terminology that are hard to grasp, which brings the performance to a lower level

(i) Word Error Rate (WER)

The performance of speech recognition systems may be effectively tested using the standard measurement method, rate of errors in words (WER) [8]. The WER is an equation that indicates the system's degree of accuracy.

$$\text{WER} = \frac{S + D + I}{N} * 100, \tag{29}$$

where S is the number of substitutions, D is the number of deletion, I is the number of insertion, and N is the number of words in the reference.

(ii) Real-Time Factor (RTF)

The real-time factor (RTF) is used to measure velocity parameter of speech recognition system. It is defined by the equation:

$$\text{RTF} = \frac{P}{I}, \tag{30}$$

where P is the amount of time it takes to process input and I is the amount of time.

(iii) F-measure

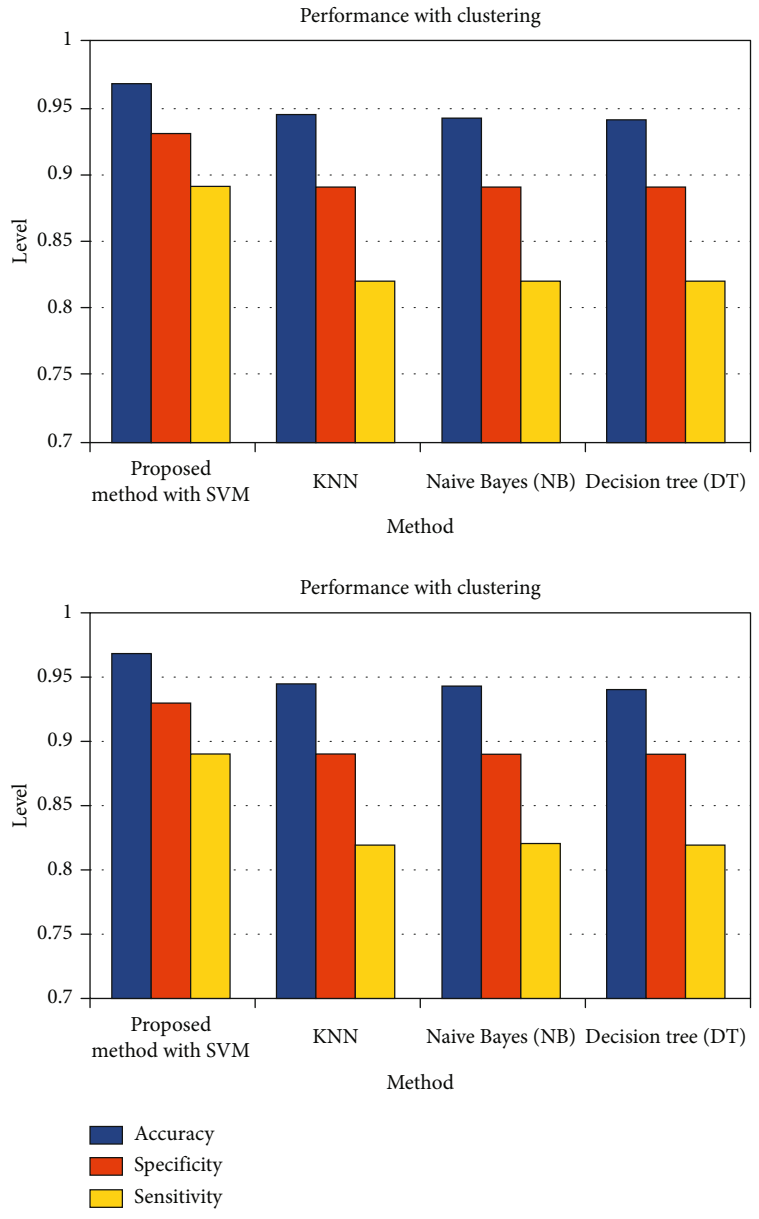

Figure 16: Graphical analysis of comparative methods.

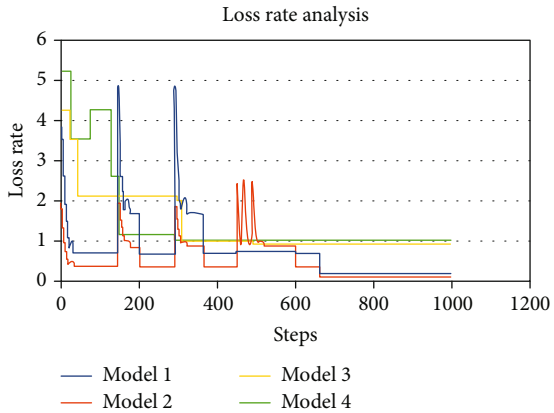

Figure 17: Loss rate analysis during training of DL models.

The F measure serves as a measure parameter in order to carry out performance analysis on the system. The F measure is often referred to as the $F1$ score or simply the F score. The weighted harmonic mean of the accuracy and recall of the test is what is used to define it.

The term "precision" refers to a measurement that compares the total number of words that were properly identified to the sum of the total numbers of correctly recognized words, substitutions, and insertions.

$$\text{Precision} = \frac{C}{(C+S+I)}. \tag{31}$$

Recall is a measure that depicts the total correctly recognized words to the sum of total numbers of correctly recognized words, substitutions, and deletions.

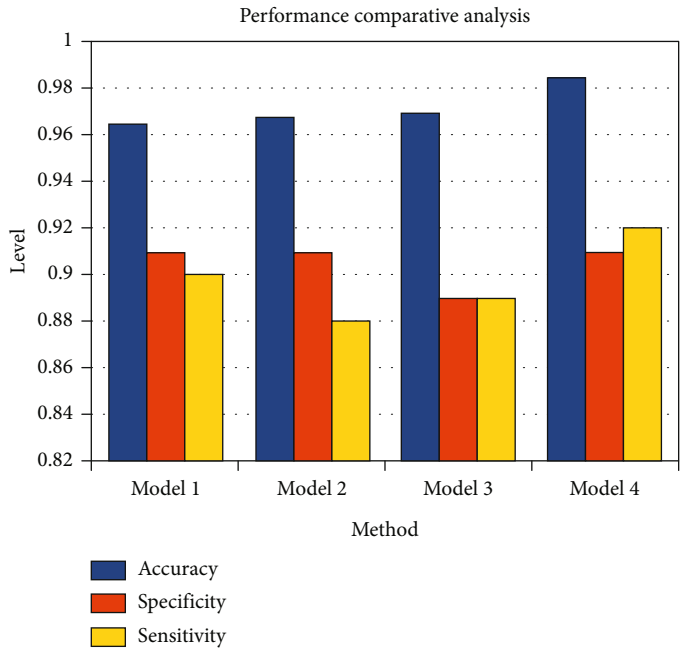

Figure 18: Comparative graph of performance parameters for deep learning models.

Table 5: Comparison of machine learning and deep learning methods.

Method	Accuracy	Specificity	Sensitivity
Model4 (proposed)	0.985	0.91	0.92
SVM	0.956	0.91	0.89

Table 6: Comparison of machine learning and deep learning methods.

Method	Accuracy	Specificity	Sensitivity
Model4	0.985	0.91	0.92
Clustering+model 4 (proposed)	0.99	0.95	0.92

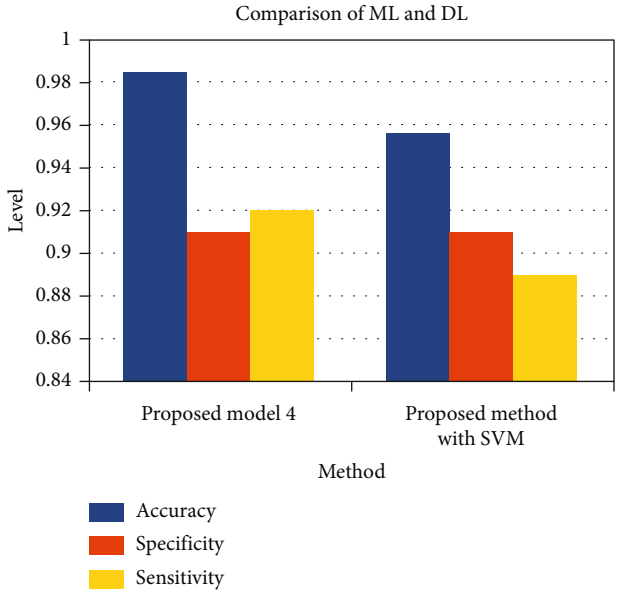

Figure 19: Comparison of ML and DL methods.

4.1. Signal Mixing and Feature Extraction. Using execution with the Python program, the preliminary testing and error checking for the proposed system has been completed. The true orientation of the glance assessment is determined by the response that is caused by the room's inspiration. Regarding the receivers of center 1 and the mouthpieces of node 2, the room inspiration response is implemented. These centers seen for random reasons in a room might have various reactions to the room's power, and as a result, for preliminary purposes and to account for potential errors, we have used mixtures of the essential inspiration reaction data. The room inspiration response plot is shown in Figures 7–9.

Figure 6 shows the input speech signals used for experimentation. The MFCC coefficient are presented in the spectrum graph as shown in Figure 7. The mixed signal is obtained based on room impulse response as shown in Figure 8.

The DOA estimation and SVM-based event detection is shown in Figure 9. As the angle with respect to node 2 is greater, further processing will be done by node 2. The sound separation using CBSS based on DOA is used to detect the estimate using MFCC coefficients. The resulting event is shown in the result (Figure 10). K-medoid clustering is applied on dataset features. The clustered data is plotted in a scatterplot as shown in Figure 11. Figure 12 shows the node 2 room impulse response, and Figure 13 shows the results obtained from the SVM classifier.

Figure 14 shows the K-means clustering scattered plot. Table 2 shows the accuracy comparison with the proposed and existing machine learning algorithms.

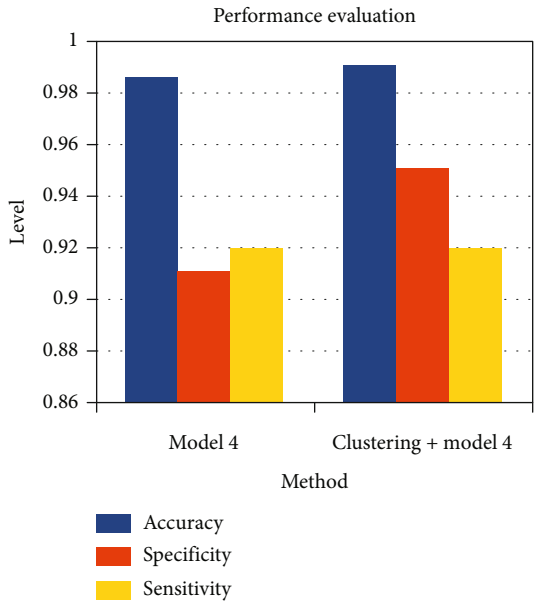

Figure 20: Comparison of clustering and without clustering methods.

Figure 15 shows the graphical analysis of the existing and proposed methodologies. Table 3 represents the performance metrics comparison.

K-nearest neighbor, naive Bayes (NB), and decision tree (DT) classifiers are considered for comparison. The values of accuracy, specificity, and sensitivity are obtained using formulae shown in Table 4. Figure 5 shows the graph of comparative analysis from which SVM results can be seen as better in terms of accuracy. The improved convergence for SVM model for different types of commands is the main reason for improved performance of the SVM model.

The training is done for 100 epochs with 10 steps per epoch using the dataset [12]. The loss rate for the training steps is tracked, and a loss analysis graph is plotted in Figure 16.

The results of the accuracy of classification for the identification of flawless instructions are studied in terms of accuracy, specificity, and sensitivity. Table 4 compares the scores of four different models, and Figure 17 depicts a graphical representation of the performance parameters. The accuracy, specificity, and sensitivity values are calculated using the formulas in Table 4.

Figure 18 shows the performance metrics of the deep learning models.

The performance obtained using machine learning and deep learning are compared as shown in Table 5 and graph in Figure 19. The comparative of use of clustering is also considered for comparison as shown in Table 6 and Figure 20. The performance of deep learning method is seen well.

5. Conclusion

Voice commands may be used to operate a wide range of devices that are connected to the internet of things network. It is possible to correctly identify orders coming from a single source in the absence of any other sounds that would be considered noise. When there are a large number of sound-producing sources in the surroundings around you, it is possible that the ambient noise will get muddled with the command voices. A number of different processing processes are required in order to separate speech instructions from a mixed sound stream. This study effort covers the separation and classification of command from mixed sound signal for the purpose of ensuring that the proper activity may be done through voice-controlled devices. The process consists of developing a database of mixed sound signals, which may contain a variety of verbal instructions as well as the noises of dogs, cats, or even television shows such as the news or other entertainment options. The evaluation of additional potential sources of noises can call for a large amount of the collection. The study work that was done demonstrates the noises that are usually regarded to originate from a variety of sources as well as their mixing in relation to the acoustic qualities of the space. The model that is able to accurately detect the command while simultaneously extracting the appropriate features. The initial step of the task consists of preparing a mixed sound dataset in such a way that the instructions that fall under a certain class may be appropriately identified. In the system that has been presented, the database that has been produced already has a number of instructions that have been gathered and segregated for the supervised training of the models. The second step of the work is comprised of the traditional blind source separation approach, which focuses its attention on the direction of arrival (DOA) of the sound signals. This is done while the system is being performed and proposed. Based on the angle of arrival estimate, a specific room impulse response (RIR) is regarded from a set of specified RIR as a room acoustic feature, and source separation is carried out using the independent component analysis (ICA) approach. The processed signals of the split command speech are then subjected to feature extraction. The Mel-frequency cepstral coefficient (MFCC) approach is used in order to carry out the process of feature extraction. Comparisons are made between the performance of the SVM classifier and that of other classifiers that are often used in machine learning applications, such as decision trees (DT). The effectiveness of the SVM classifier is evaluated, and the results reveal that it has an accuracy of 91%. Utilizing a classifier that is based on a recurrent neural network (RNN) helps to increase the accuracy of further categorization. This particular classifier is made up of three layers: one layer of gated recurrent units (GRU), one layer of long short-term memory (LSTM), and one layer of both. This classification seems to have an accuracy of 94.5 percent, which is higher than the SVM classifier.

Data Availability

The data that support the findings of this study are available on request from the corresponding author.

Conflicts of Interest

The author declare that he/she has no conflicts of interest to report regarding the present study.

References

[1] T. Ammari, J. Kaye, J. Y. Tsai, and F. Bentley, "Music, search, and IoT," *ACM Transactions on Computer-Human Interaction*, vol. 26, no. 3, pp. 1–28, 2019.

[2] D. Pal, C. Arpnikanondt, S. Funilkul, and W. Chutimaskul, "The adoption analysis of voice-based smart IoT products," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 10852–10867, 2020.

[3] G. Muhammad, S. M. M. Rahman, A. Alelaiwi, and A. Alamri, "Smart health solution integrating IoT and cloud: a case study of voice pathology monitoring," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 69–73, 2017.

[4] P. Ni, Y. Li, G. Li, and V. Chang, "Natural language understanding approaches based on joint task of intent detection and slot filling for IoT voice interaction," *Neural Computing and Applications*, vol. 32, no. 20, pp. 16149–16166, 2020.

[5] G. Alexakis, S. Panagiotakis, A. Fragkakis, E. Markakis, and K. Vassilakis, "Control of smart home operations using natural language processing, voice recognition and IoT technologies in a multi-tier architecture," *Designs*, vol. 3, no. 3, p. 32, 2019.

[6] A. Valera Román, D. Pato Martínez, Á. Lozano Murciego, D. M. Jiménez-Bravo, and J. F. de Paz, "Voice assistant application for avoiding sedentarism in elderly people based on IoT technologies," *Electronics*, vol. 10, no. 8, p. 980, 2021.

[7] A. F. Klaib, N. O. Alsrehin, W. Y. Melhem, and H. O. Bashtawi, "IoT smart home using eye tracking and voice interfaces for elderly and special needs people," *The Journal of Communication*, vol. 14, no. 7, pp. 614–621, 2019.

[8] S. Uma, R. Eswari, R. Bhuvanya, and G. S. Kumar, "IoT based voice/text controlled home appliances," *Procedia Computer Science*, vol. 165, pp. 232–238, 2019.

[9] M. Ali, "Developing applications for voice enabled IoT devices to improve classroom activities," in *In 2018 21st International Conference of Computer and Information Technology (ICCIT)*, pp. 1–4, IEEE, Dhaka, Bangladesh, (2018, December).

[10] H. Isyanto, A. S. Arifin, and M. Suryanegara, "Design and implementation of IoT-based smart home voice commands for disabled people using Google Assistant," in *In 2020 International Conference on Smart Technology and Applications (ICoSTA)*, pp. 1–6, IEEE, Surabaya, Indonesia, 2020.

[11] K. M. Malik, A. Javed, H. Malik, and A. Irtaza, "A light-weight replay detection framework for voice controlled IoT devices," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 982–996, 2020.

[12] Y. Meng, H. Zhu, J. Li, J. Li, and Y. Liu, "Liveness detection for voice user interface via wireless signals in IoT environment," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 6, pp. 1–3011, 2020.

[13] P. J. Rani, J. Bakthakumar, B. P. Kumaar, U. P. Kumaar, and S. Kumar, "Voice controlled home automation system using natural language processing (NLP) and internet of things (IoT)," in *In 2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM)*, pp. 368–373, IEEE, Chennai, India, 2017, March.

[14] CNET, "How to bring Alexa into every room of your home," 2017, https://www.cnet.com/how-to/how-to-install-alexa-in-every-room-of-your-home/.

[15] E. M. Mugler, M. C. Tate, K. Livescu, J. W. Templer, M. A. Goldrick, and M. W. Slutzky, "Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri," *Journal of Neuroscience*, vol. 38, no. 46, pp. 9803–9813, 2018.

[16] C. P. Browman and L. Goldstein, "Articulatory phonology: an overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[17] X. Lei, G. Tu, A. X. Liu, C. Li, and T. Xie, "The insecurity of home digital voice assistants - Amazon Alexa as a case study," 2017, Available: http://arxiv.org/abs/1712.03327.

[18] X. Lei, G. Tu, A. X. Liu, C. Li, and T. Xie, "The insecurity of home digital voice assistants - vulnerabilities, attacks and countermeasures," in *2018 IEEE Conference on Communications and Network Security (CNS)*, pp. 1–9, Beijing, China, May 2018.

[19] J. Tan, C. Nguyen, and X. Wang, "SilentTalk: lip reading through ultrasonic sensing on mobile phones," in *IEEE INFOCOM 2017- IEEE Conference on Computer Communications*, pp. 1–9, Atlanta, GA, USA, May 2017.

[20] Y. Meng, W. Zhang, H. Zhu, and X. S. Shen, "Securing consumer IoT in the smart home: architecture, challenges, and countermeasures," *IEEE Wireless Communications*, vol. 25, no. 6, pp. 53–59, 2018.

[21] W. Zhang, Y. Meng, Y. Liu, X. Zhang, Y. Zhang, and H. Zhu, "HoMonit: monitoring smart home apps from encrypted traffic," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1074–1088, New York, NY, United States, 2018.

[22] Y. Zhang, R. Deng, D. Zheng, J. Li, P. Wu, and J. Cao, "Efficient and robust certificateless signature for data crowdsensing in cloudassisted industrial IoT," *IEEE Transactions on Industrial Informatics*, vol. 15, pp. 1–9, 2019.

[23] X. Yuan, Y. Chen, Y. Zhao et al., "CommanderSong: a systematic approach for practical adversarial voice recognition," in *Proceedings of 27th USENIX Security Symposium (USENIX Security)*, pp. 49–64, Baltimore, MD, 2018.

[24] Y. Chen, J. Sun, R. Zhang, and Y. Zhang, "Your song your way: rhythm-based two-factor authentication for multi-touch mobile devices," in *Proceedings of IEEE Conference on Computer Communications (INFOCOM)*, pp. 2686–2694, Hong Kong, China, April 2015.

[25] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proceedings of IEEE Symposium on Security and Privacy (S & P)*, pp. 582–597, San Jose, CA, USA, 2016.

[26] K. Ali, A. X. Liu, W. Wang, and M. Shahzad, "Keystroke recognition using WiFi signals," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (ACM MobiCom)*, pp. 90–102, New York, NY, United States, 2015.

[27] Y. Liu, "WiVo: enhancing the security of voice control system via wireless signal in IoT environment," in *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, pp. 81–90, New York, NY, United States, 2018.

[28] P. Associates, "Top 10 consumer iot trends in 2017," 2017, http://www.parksassociates.com/whitepapers/top10-2017.

[29] Amazon, "Amazon alexa developer," 2019, https://developer.amazon.com/alexa.