Real-Time Human–Machine Interaction Through Voice Augmentation Using Artificial Intelligence

M. N. Sumaiya, B. V. Sreekanth, U. S. Akash, Aravind Sharma Kala, and G. M. Dharanendra Gowda

Abstract In real time, there is a huge demand to access dynamic, personalized, and adaptive information. Interacting through voice augmented systems yields efficient and faithful access to great deal of knowledge. And the obtained information becomes more meaningful and understandable because it is appropriately customized in the aspect of real-time physical, digital, and virtual interactions. The objective of this project is to display the interactions of voice augmentation functions to achieve human to machine interaction (HMI). The quality of the services can be improved with the help of artificial intelligence through deep learning models to illustrate the possibilities of its potential. There exist a good range of applications which include vending machines, billing counters, home assistant, etc. This work includes audio processing, artificial intelligence, text-to-speech conversion, speech to text, audio conversion, various machine learning algorithms and models, implementation using the Internet of things, and applications in various fields, to evaluate the model.

Keywords Human–machine interaction · Voice augmentation · Artificial intelligence · Voice assistant · Recurrent neural network

1 Introduction

The need for acquiring knowledge and information is increasing day by day. A couple of decades ago, few of the primitive methods to accumulate information were through books, scriptures, and manuscripts. As technology has risen to a level where human can access every data in this world within fraction of seconds, the need to access information has also increased. In this real world, human to machine interactions (HMI) through voice are essential for aged, physically challenged people and people involved in multiple tasks. Instead of using physical movements of their body to interact with machines, easily, they can interact through their voice to get their work

M. N. Sumaiya (✉) · B. V. Sreekanth · U. S. Akash · A. S. Kala · G. M. Dharanendra Gowda
Department of Electronics and Communication Engineering, Dayananda Sagar Academy of Technology and Management, Bangalore 560082, India
e-mail: drsumaiyamn@dsatm.edu.in

M. A. Chaurasia and C.-F. Juang (eds.), *Emerging IT/ICT and AI Technologies Affecting Society*, Lecture Notes in Networks and Systems 478,
https://doi.org/10.1007/978-981-19-2940-3_10

done. It is a very challenging task to bring the machines to communicate on the same level as us. So, this is a step exploring toward that domain, which is the real-time interactive voice augmentation. A device capable of communicating on the same level as a human makes it easily accessible to everyone irrespective of their knowledge of language, age and abilities. The objectives are to build a voice augment mainframe device, capable of carrying real-time conversation with the user using audio processing, artificial intelligence and machine learning algorithm, to establish a database for voice recognition and audio synthesis module, capable of speech-to-text and text-to-speech conversion. To build a machine learning model, by training it with the database and producing the optimum model capable of choosing the optimum solution and to deploy and demonstrate the wide range of potential applications in the field of banking, kiosks, reservation for restaurants, ticket booking vending machines for transportations and public vending machines. For example, Amazon offers Transcribe, an automatic speech recognition (ASR) service that permits developers to feature speech-to-text capability to their applications. Once the voice capability is integrated into the application, users can analyze audio files and, in return, receive a text file of the transcribed speech. Google has made moves in making assistant more universal by opening the software development kit through actions, which allows developers to create voice into their own products that support AI. Another one of Google's speech recognition products is the AI-driven cloud speech-to-text tool which enables developers to convert audio to text through deep learning neural network algorithms.

In [1], a sequence-to-sequence-based voice conversion (VC) method is proposed, which is capable enough of converting voice characteristics and the pitch contour along with duration of the input speech. It enables the use of batch normalization in all the hidden layers. A few drawbacks are restricted only to speaker to identity conversion tasks, and the framework of S2S learning approach has a lot of room to improve in the aspects of accuracy. In [2], recently, neural networks are applied to tackle audio pattern recognition problems. Audio pattern recognition is a vital research topic within the machine learning area and includes several tasks like audio tagging, acoustic scene classification, music classification, speech emotion classification, and sound event detection. It is studied that the system is inspired by conventional cognitive models for memory. A few drawbacks are that the trained data is confined to the audio set's dataset, and PANNs are susceptible to multiple pattern recognition tasks, which is time consuming. Recently, there has been increasing progress in end-to-end automatic speech recognition (ASR) architecture, which transcribes speech to text with none pre-trained alignments [3]. These online systems have the advantages over the offline baselines in both decoding latency and decoding speed. In the application of low latency encoders, the recognition accuracy was observed to be low. The connectionist temporal classification (CTC) architecture is primitive at best, which has to be tuned perfectly to form hybrid CTC. Speech separation is a method to extract the speech data from the ambience noise and background distortions [4]. The latest method in speech separation illustrates as a problem of supervised learning, by training lots of datasets on differential patterns of speech, speakers, and background noise. Implementing deep learning models on supervised speech separation methods

yields effective results. Learning machines, training targets, and acoustic features are the three significant constituents of deep learning-based supervised speech separation. It provides a general overview on the various steps involved in speech separation and provides a comprehensive overview of deep neural network (DNN) based on supervised speech separation. The main drawback is that the DNN-based speech enhancement as described has met the criterion in limited conditions, but not in all conditions, and the DNN-based speech enhancement has many flaws in the aspect of background speech and noise separation. The introduction of smart mobile devices is of great advantage for user interaction as these devices are equipped with numerous sensors, making applications context aware [5, 6].

To further improve user experience, the most mobile operating systems and repair providers are gradually shipping smart devices with voice-controlled intelligent personal assistants, reaching a replacement level of human and technology convergence. It is observed that, to provide defense mechanisms against many of these attacks, the underlying operating system, in this case Android, first needs to decouple voice input and output. Data is only accessible through appropriate authenticated channel, without which the data is secure from malicious treats. Data is susceptible to malware attacks and is open to spyware attacks as the connectivity is branched throughout. Identity security is another aspect where improvement is immediately needed [7]. The literature review also covers differing types of deep architectures, like deep convolution networks, deep residual networks, recurrent neural networks, reinforcement learning, variational auto-encoders, etc. [8–10]. Convolution neural network (CNN) can progressively extract higher representations of the image after each layer and finally recognize the image [11–13]. Numerous research works are based on machine learning which is a method that learns from past experiences and uses gained knowledge to do better in the future [14–16]. Machine learning emphasizes on automatically learning and adapting when exposed to data without the need of human intervention. The reason for that is based on the fact that no system can be described as intelligent if it does not have the ability to learn and adapt [17–19]. In order to exemplify applications of supervised and unsupervised learning, we will offer annotated tips that could be the literature on machine learning for communication systems. Tasks are administered at the sting of the network, that is, at the bottom stations or access points and at the associated computing platforms, from tasks that are instead responsibility of a centralized cloud processor connected to the core network [20–22]. Indeed, for many tasks in communication networks, it is possible to gather or generate training datasets and there is no need to apply sense or to supply detailed explanations for how a decision was made [23–25]. Alternatively, under an algorithm deficit, a physics-based model, if available, can be possibly used to carry out computer simulations and obtain numerical performance guarantees [26–28]. As a solution to unstable gradient values problem, a novel RNN architecture was developed which avoids vanishing and exploding gradients, while it can be trained with conventional RNN learning algorithms. The improved RNN architecture is referred to as long short-term memory neural network (LSTM) [29–31]. The context length exploited when applying the neural network can be longer than the context length considered in training [32]. Supervised speech separation has also been shown to

generalize well given sufficient training data [33–35]. A huge amount of research has gone into finding ways of constraining GMMs to extend their evaluation speed and to optimize the trade-off between their flexibility and therefore the amount of training data required to avoid serious overfitting [36, 37]. Other types of models may work better than GMMs for acoustic modeling if they can more effectively exploit information embedded in a large window of frames [38]. In fact, two decades ago, researchers achieved some success using artificial neural networks with a single layer of nonlinear hidden units to predict HMM states from windows of acoustic coefficients [39]. The application of recurrent networks to speech detection and recognition in clean and noisy environments have been proposed in some early studies [40–42]. As we have come across abundant journal papers which have covered every possible aspect in audio processing, speech-to-text conversions, machine learning models and training them, different ways of communication methods and the IoT aspect of it have better connectivity and accessibility. It is evident that there are a lot of hurdles to overcome to achieve this project. So, the objective of the project is to bring the machines to communicate on the same level as us. Then the purpose of the algorithm is to map the audio signal into textual inputs with a proper conversion method which is capable to discriminate background noise with higher accuracy. Then, the machine learning algorithm must be designed, developed, built and trained, the model must be re-trained with huge number of iterations, since it will be deployed in various applicative environments. To choose an optimum algorithm with the highest accuracy to run the model, which can access the database and interpret the input accurately and with the help of the trained dataset, recognize the situation and analyze the inputs and other constraints which restricts or governs the situation, and then to be capable of arriving at the most appropriate output is the biggest achievement of all. So, this is a small step exploring toward that domain which is the real-time interactive voice augmentation. A device capable of communicating on the same level as a human makes it easily accessible to everyone irrespective of their knowledge of language, age, and abilities.

The chapter is organized as follows: Sect. 2 describes the methodology and materials used for this work. The experimental quantitative and qualitative results are discussed, and results are tabulated in Sect. 3. Finally, conclusion is drawn in Sect. 4.

2 Methodology and Materials

A voice augment mainframe device is built which is capable of carrying real-time conversation with the user, using audio processing, artificial intelligence, and machine learning algorithm. Also, a database for voice recognition and audio synthesis module is established, which is capable of speech-to-text and text-to-speech conversion. Now, the main objective of this chapter is to design and build a machine learning model and training it with the database and producing the optimum model capable of choosing the optimum solution. To choose an optimum algorithm with the highest accuracy to run the model, which can access the database and interpret

the input accurately and with the help of the trained dataset, recognize the situation and analyze the inputs and other constraints which restricts or governs the situation, and then to be capable of arriving at the most appropriate output is the biggest achievement of all. The block diagram of the proposed model is shown in Fig. 1.

The most widely used algorithm for natural language processing (NLP) and speech recognition is recurrent neural network (RNN) that can been trained to ingest speech data into small frames. The RNN has three layers: input layer, hidden layer, and the output layer; the hidden layers are the computation layers. At each time step, the non-recurrent layers work on independent data. The hidden layer may be a bidirectional recurrent layer with two hidden unit sets. One set has forward recurrence, while the other has backward recurrence, which requires some memory to perform the recurrence as illustrated in Fig. 2.

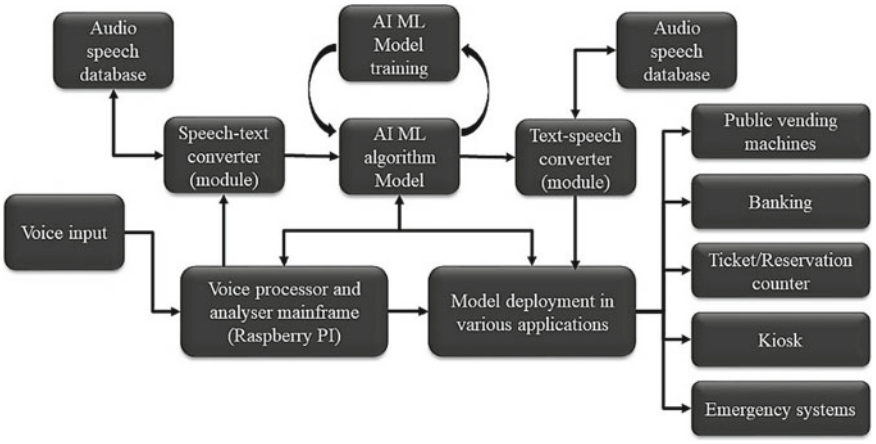


Fig. 1 Block diagram of the proposed system

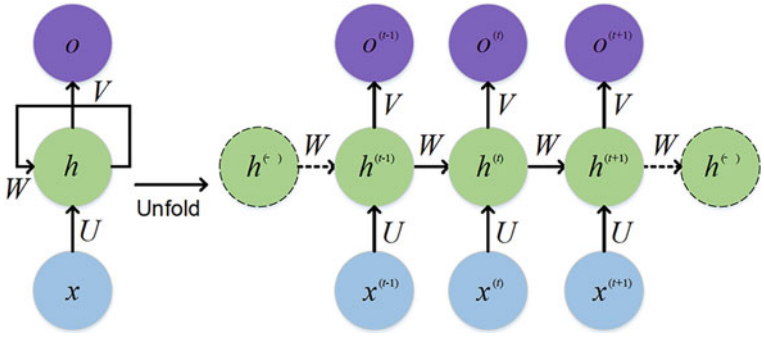


Fig. 2 General block diagram of RNN

One of the prominently used recurrent neural networks for the applications that involve sequential input data processing, such as speech recognition, image recognition, music composition, and handwriting recognition, is long short-term memory recurrent neural network (LSTM-RNN). First, considering the input speech data which are nothing but sequence of words, every word will be encrypted to unique binary vectors, then the neurons are initialized with random values coined as weights. During the training, when the binary vector of the input sequence is applied to the input layer, the nodes of the layer perform the respective logical function or sigma function on the weights to get an arbitrary value, and with the help of activation function, this value is fed to the successive layers. In application such as speech recognition, recognizing only the current input word from the speech data is not the aim, but to generate a value by co-relating the weight and the binary vector along with the previous word vector of the input sequence and then storing the value arbitrarily to co-relate with the successive vectors of the input sequence.

At the output layer, the difference of the predicted vector value to the actual vector value is calculated using a loss function; according to the obtained loss value, using a gradient descent function, the weights of the neurons are calibrated. This process is iterated as many times as required to train the model with minimum loss. Gradient descent function calculates the next optimum value from the current value of the weight and reduces by one step after scaling to the learning rate, and it is subtracted from the present weight because we want to minimize the loss function value. The learning rate controls the step size of the gradient descent which effects strongly on the performance of the model. Larger the value of learning rate, larger will be the step size, which will usually miss the optimum weight value. Smaller the value of learning rate, higher will be the possibility of attaining the optimum weight value, in larger number of steps of small size.

One more major task is to determine optimum parameters for the neural network. In speech recognition, since the number of words in each sentence might vary for every instance, determining the number of nodes in input layer is quite difficult. The hyperparameters determined for this system are determined as follows:

- The number of nodes in input and output layer or also called as batch size is equal to 10.
- The number of hidden layers is equal to 3.
- Learning rate for the gradient descent function is equal to 0.001.
- The number of epochs to determine the number of iterations to train the dataset is equal to 1000.

The sigma function or the logical function of the nodes is a simple linear transformation function $Y = a_1 * T * (v_1) + a_2 * T * (v_2)\ldots$, where Y is the arbitrary value of the node, T is the transformation function, a_1 and a_2 are the weights of the nodes, and v_1 and v_2 are the input vectors. The activation function used is called as rectified linear unit function (RELU), which is a piecewise linear function that will output the input data. For loss detection of the predicted value, cross-entropy loss function is implemented for calibrating the weights.

PyTorch library is used to build and deploy the proposed work as it provides the required framework to design the machine learning algorithm. After designing all the mentioned modules, this project will be implemented to perform some of the real-time machines like ATM, vending machine, ticket reservation machine, and kiosk machine, where all these machines can be controlled using voice rather than using buttons, with the help of software development toolkit to design the graphical user interface.

For the mainframe device, we have used Raspberry Pi 4 which features a quad-core ARM Cortex-A72 processor, dual video output, and a good selection of other interfaces. It also requires a microphone module compatible with Raspberry Pi for audio recognition, a speaker module for the output, and a Bluetooth module to enable the Raspberry Pi to be capable of conveying output via Bluetooth speakers. The main software, Raspbian OS, is used to operate and simulate Raspberry Pi.

Since the system is being deployed in different environment, the dataset required to train the model in various environment will vary. Hence, the datasets were obtained separately for personal home assistant, vending machine, and digital assistant. About ten thousand words in each domain of application were collected to train the model.

First, the dataset required to train the model is preprocessed before using it to train the data. For example, for any given sentence, the very first step is to perform segmentation of the sentence into smaller words, and this process is called as stemming. Then, the stemmed words will be of any form according to the grammar of the sentence in which it is used; hence, it is essential to filter down to its root form. This process is called as tokenization. These tokenized words are mapped into its corresponding binary symbols with unique value into a binary array.

This preprocessed dataset is used for the model training. Now, we have to build an appropriate neural network model with appropriate batch size, training parameters, and learning parameters. The model based on recurrent neural network algorithm is built, since it is the most suitable algorithm to develop natural language processing. The dataset is divided into small batches, and each batch is called as epoch. This is done so that after each epoch training, the model adjusts the weight of the neuron by calculating the cross-entropy loss. Cross-entropy loss is the measurement of the amount of difference from the predicted values to the actual value. By rigorous training, the model predicts the exact words, with loss ranging maximum up to 0.09%.

Now, the trained model saves the data into a path file, which consists of the trained data, and this is used imported in the functionality program, where we have to map each and every functionality to the predicted words and sentence. Various library functions along with the path file are imported, in order to accommodate as many as functionalities as possible. Then, in order to present the voice augmented system, a graphical user interface program is devised which enables the user to operate by audio and visual feedbacks.

3 Experimental Results

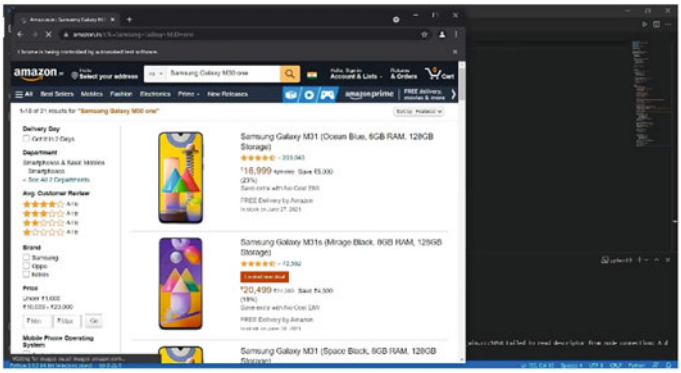
The experimental results, at the progressing stage of the project, yielded successful outcomes in several implemented applications such as voice operated interactive personal assistant in digital systems, voice operated interactive personal assistant in home automation, voice operated vending machine, and voice operated interactive e-commerce systems, as illustrated in Fig. 3a–c.

3.1 Quantitative Measures

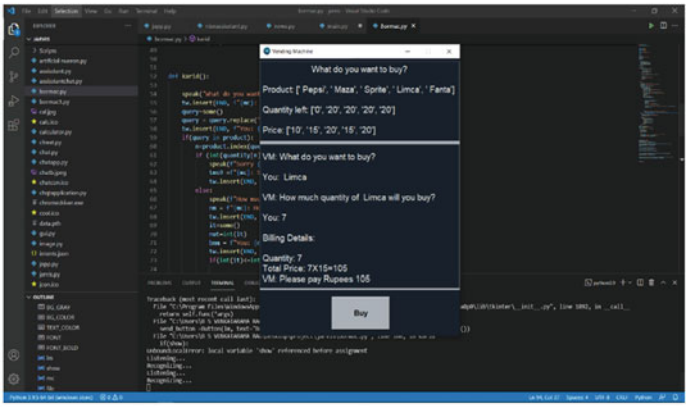
Dataset consists of 1000 words which was taken from Google Speech Command dataset. As the model is trained gradually with considerable number of datasets, the initial accuracy rate of the recurrent neural network algorithm is achieved 88.54%. In order to have a comprehensive understanding of this result, as given in Table 1, the same number of datasets was trained in different algorithms such as linear regression and polynomial regression algorithms which yielded accuracy rate of 53.61% and 72.89%, respectively. Hence, we could set a benchmark for the accuracy rate of the model, trained in recurrent neural network algorithm. After the model training was carried out further with some more datasets, the accuracy rate of 99.56% was achieved, and these results are only due to the primitive datasets trained to the model.

From Table 1, it is evident that recurrent neural network is the optimal algorithm for training the datasets with maximum accuracy. The major difference between the recurrent neural network and the rest of the supervised algorithms is that it is more suitable for natural language processing and audio processing, which is essential in human to machine interaction (HMI). In the case of regression algorithms, linear or polynomial, it determines an arbitrary equation through which it can determine the features in the datasets and produce or predict the optimum result during deployment. This is highly incompatible for language processing, because there are many situations in the datasets, where the equations that determine the features are incapable to recognize multiple possibilities out of a single input data. Since we have discussed the process of tokenization of datasets in the methodology section, where the input word is filtered to its root form, there are many homophones present in English that are read the same but have different meanings, this feature is highly difficult to be recognized using the arbitrary equations of regression algorithms, hence the accuracy level reduces, and we might find a little improvement in the accuracy if the epoch number is increased, but not as good as expected.

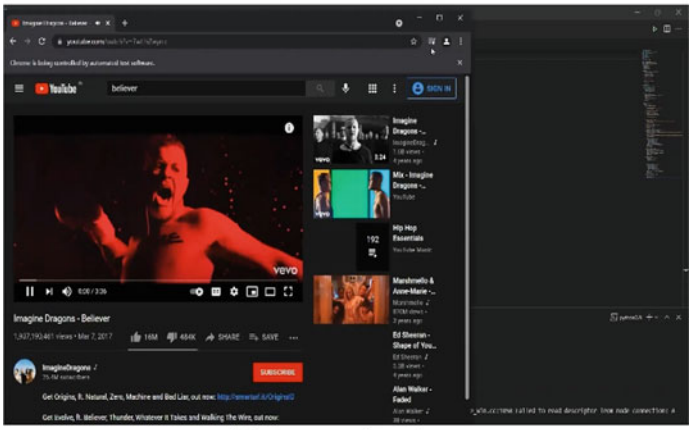
In the case of recurrent neural network, after the preprocessing of dataset, each word which is converted into its binary counterpart with unique binary value is fed into different neurons of different layers. A neuron can be any complicated mathematical function developed according to the application. Basically, a neural network consists of three layers: input, hidden, and output layers. The batch size or the number of input neuron, number of hidden layers, and number of output neurons depend



(a) e-commerce systems



(b) vending machine



(c) Personal Assistant

Fig. 3 Voice operated interactive systems **a** e-commerce systems, **b** vending machine, **c** personal assistant

Table 1 Accuracy of various algorithms

Algorithms	Accuracy (%)
Recurrent neural network	88.54
Linear regression	53.61
Polynomial regression	72.89

on the application, for a given input sentence of minimum number of words in the sentence, each word with its unique binary value is fed into each neuron, and after the mathematical function is performed, an activation function passes the output to the next neuron in the hidden layer. In this layer according to the neural network developed, it performs forward recurrence or backward recurrence, i.e., when a particular binary value of a homophonic word is obtained, it stores the obtained mathematical output of the function in a temporary memory, and this algorithm is called as recurrent neural network long short-term memory (RNN-LSTM). Backward recurrence is performed with the homophonic binary value with the next binary value of the successive word in the input sentence. Hence, the predicted output value will be more accurate as every possibility of the homophonic binary value is calculated by performing recurrence. Therefore, the accuracy is considerably more, with more epoch numbers, the accuracy increases much more.

3.2 Qualitative Measures

The performance of the proposed algorithm is measured by using the error rate, training accuracy and comprehensive accuracy at different stages of the proposed model. Quality metrics formulas are listed below in Eqs. (1)–(2).

$$\text{Accuracy} = \left(\left(\sum V_i / N\right)\right) * 100 \tag{1}$$

where V_i is the predicted value of input words in each instance and N is the number of words in that epoch.

$$\text{Cross entropy loss } = H(P, Q) = H(P) + \Delta(P||Q) \tag{2}$$

where P is the predicted value, Q is the actual value, $\Delta(P||Q)$ is the divergence from Q to P, and $H(P)$ is entropy of P. In the accuracy of recognizing homophones in every situation, there will be similarly sounding words whose meaning will depend on the context it is being used. It was calculated by a built-in function of the PyTorch library.

Cross-entropy loss, model training accuracy, and the comprehension accuracy at the various stages of the proposed algorithm are given in Table 2. It is observed that the mode training accuracy approaches 99.56% at the final stage due to recurrent neural

Table 2 Quality metrics of the proposed algorithm

Quality metrics	Initial stage (%)	Intermediate stage (%)	Final stage (%)
Cross-entropy loss	10.63	01.80	00.03
Model training accuracy	88.54	91.28	99.56
Comprehension accuracy	85.54	81.36	83.56

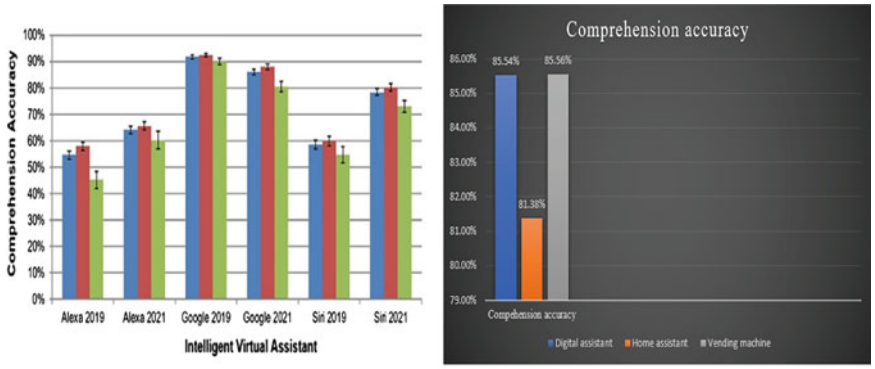


Fig. 4 Comprehension accuracy **a** intelligent virtual assistants **b** proposed system

network model. The comprehension accuracy of other intelligent virtual systems and the proposed system, which were trained on various datasets which included thousands of words in many languages and class of demographic each year, is shown below in Fig. 4.

The results show that the average comprehension accuracy is 57.3%, 89.64%, and 68.39% for Alexa, Google, and Siri, respectively. These results are due to the huge number of datasets monitored in various regions around the world, taking accents and language into consideration. The proposed algorithm yields an average accuracy about 83%, which is attained by considerably less and primitive dataset from English language alone, as shown in Fig. 4b.

The interfacing program successfully could recognize the voice input and appropriately open the respective windows and applications. When the user wishes to open YouTube, the model asks the user's desired video to watch in YouTube and open appropriate video in YouTube browser. If the user wishes to open Amazon shopping website, it will revert back to user with what does the user wish to buy and open appropriate section in the Amazon Web site. Similarly, it can access every application installed in the system and open them at the request of the user. If the user wishes to buy any given condiment from the vending machine, it will display the available stock of the condiments. When the user wishes to buy, it will enquire the number of items the user wishes to buy and display the final amount to be paid. After the transaction, it will update the stock and display the updated stock.

4 Conclusion

Voice-controlled digital assistants are becoming more natural, as it is integrated into everyday devices. Also, the emulations of human conversations will become much more natural. The main purpose of this chapter is achieved one step closer towards having a very intelligent system. In this chapter, a successful description of the terms used to describe actions makes the devices user friendly which is described in detail.

This chapter has its implementation in various domains, right from being personal assistant in a household, to being a tool to deploy sophisticated operations and functions in various industries. Moreover, this chapter has the advantage of being flexible in the aspect of implementation, it is not only restricted in industries and household, but it has the scope to be implemented in field of public, government, and defense. In other words, both scope and scale of this project are vast. In the future, a much more interactive experience environment through every digital channel is possible. Voice technology is becoming increasingly accessible to developers. And as consumers are getting increasingly easier and reliant upon using voice to speak to their phones, cars, smart home devices, etc., voice technology will become a primary interface to the digital world and with it, expertise for voice interface design and voice app development are going to be in greater demand. The future possibilities of advancements in the field of voice augmented systems are enormous. To build a strong speech recognition experience, the AI behind it is to become better at handling challenges like accents and ambience noise.

Acknowledgements This work is funded and supported by VGST-KFIST L1, Karnataka, India, with the grant number GRD 786.

References

1. Kameoka H, Tanaka K, Kwaśny D, Kaneko T, Hojo N (2020) ConvS2S-VC: fully convolutional sequence-to-sequence voice conversion. In: IEEE/ACM Transactions on audio, speech, and language processing, vol 28, pp 1849–1863
2. Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD (2020) PANNs: large-scale pretrained audio neural networks for audio pattern recognition. In: IEEE/ACM Transactions on audio, speech, and language processing, vol 28, pp 2880–2894
3. Miao H, Cheng G, Zhang P, Yan Y (2020) Online hybrid CTC/attention end-to-end automatic speech recognition architecture. IEEE/ACM Trans Audio Speech Lang Process 28:1452–1465
4. Wang D, Chen J (2018) Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans Audio Speech Lang Process 26(10):1702–1726
5. Alepis E, Patsakis C (2017) Monkey says, Monkey does: security and privacy on voice assistants. IEEE Access 5:17841–17851
6. Rubio-Drosdov E, Díaz-Sánchez D, Almenárez F, Arias-Cabarcos P, Marín A (2017) Seamless human-device interaction in the internet of things. IEEE Trans Consum Electron 63(4):490–498
7. Teo JH, Cheng S, Alioto M (2020) Low-energy voice activity detection via energy-quality scaling from data conversion to machine learning. IEEE Trans Circ Syst I Regul Pap 67(4):1378–1388

8. Shrestha A, Mahmood A (2019) Review of deep learning algorithms and architectures. IEEE Access 7:53040–53065
9. Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K (2019) Speech recognition using deep neural networks: a systematic review. IEEE Access 7:19143–19165
10. Simeone O (2018) A very brief introduction to machine learning with applications to communication systems. IEEE Trans Cogn Commun Network 4(4):648–664
11. Delang K, Todtermuschke M, Schmidt PA, Bdiwi M, Putz M (2019) Enhanced service modelling for flexible demand-driven implementation of human–robot interaction in manufacturing. IET Collab Intell Manuf 1(1):20–27
12. Saini S, Sahula V (2020) Cognitive architecture for natural language comprehension. Cogn Comput Syst 2(1):23–31
13. L'Heureux A, Grolinger K, Elyamany HF, Capretz MAM (2017) Machine learning with big data: challenges and approaches. IEEE Access 5:7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365
14. Yang C, Zeng C, Liang P, Li Z, Li R, Su C (2018) Interface design of a physical human-robot interaction system for human impedance adaptive skill transfer. IEEE Trans Autom Sci Eng 15(1):329–340. https://doi.org/10.1109/TASE.2017.2743000
15. Amarú L, Gaillardon P, De Micheli G (2014) Biconditional binary decision diagrams: a novel canonical logic representation form. IEEE J Emerg Sel Top Circ Syst 4(4):487–500. https://doi.org/10.1109/JETCAS.2014.2361058
16. Vorm ES (2020) Computer-centered humans: why human-AI interaction research will be critical to successful AI integration in the DoD. IEEE Intell Syst 35(4):112–116. https://doi.org/10.1109/MIS.2020.3013133
17. du Boulay B (2016) Artificial intelligence as an effective classroom assistant. IEEE Intell Syst 31(6):76–81. https://doi.org/10.1109/MIS.2016.93
18. Toda T, Black AW, Tokuda K (2007) Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans Audio Speech Lang Process 15(8):2222–2235. https://doi.org/10.1109/TASL.2007.907344
19. Deng L, Li X (2013) Machine learning paradigms for speech recognition: an overview. IEEE Trans Audio Speech Lang Process 21(5):1060–1089. https://doi.org/10.1109/TASL.2013.2244083
20. Le H, Oparin I, Allauzen A, Gauvain J, Yvon F (2013) Structured output layer neural network language models for speech recognition. IEEE Trans Audio Speech Lang Process 21(1):197–206. https://doi.org/10.1109/TASL.2012.2215599
21. Erro D, Moreno A, Bonafonte A (2010) INCA algorithm for training voice conversion systems from nonparallel corpora. IEEE Trans Audio Speech Lang Process 18(5):944–953. https://doi.org/10.1109/TASL.2009.2038669
22. Dahl GE, Yu D, Deng L, Acero A (2012) Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang Process 20(1):30–42. https://doi.org/10.1109/TASL.2011.2134090
23. Louridas P, Ebert C (2016) Machine learning. IEEE Softw 33(5):110–115. https://doi.org/10.1109/MS.2016.114
24. Sheth A, Yip HY, Shekarpour S (2019) Extending patient-Chatbot experience with internet-of-things and background knowledge: case studies with healthcare applications. IEEE Intell Syst 34(4):24–30. https://doi.org/10.1109/MIS.2019.2905748
25. Makishima N et al (2019) Independent deeply learned matrix analysis for determined audio source separation. IEEE/ACM Trans Audio Speech Lang Process 27(10):1601–1615. https://doi.org/10.1109/TASLP.2019.2925450
26. Tu Y, Du J, Lee C (2019) Speech enhancement based on teacher-student deep learning using improved speech presence probability for noise-robust speech recognition. IEEE/ACM Trans Audio Speech Lang Process 27(12):2080–2091. https://doi.org/10.1109/TASLP.2019.2940662
27. Cui X, Goel V, Kingsbury B (2015) Data augmentation for deep neural network acoustic modeling. IEEE/ACM Trans Audio Speech Lang Process 23(9):1469–1477. https://doi.org/10.1109/TASLP.2015.2438544

28. Nakashika T, Takiguchi T, Ariki Y (2015) Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines. IEEE/ACM Trans Audio Speech Lang Process 23(3):580–587. https://doi.org/10.1109/TASLP.2014.2379589
29. Sundermeyer M, Ney H, Schlüter R (2015) From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Trans Audio Speech Lang Process 23(3):517–529. https://doi.org/10.1109/TASLP.2015.2400218
30. Receveur S, Weiß R, Fingscheidt T (2016) Turbo automatic speech recognition. IEEE/ACM Trans Audio Speech Lang Process 24(5):846–862. https://doi.org/10.1109/TASLP.2016.2520364
31. Nakashika T, Takiguchi T, Minami Y (2016) Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine. IEEE/ACM Trans Audio Speech Lang Process 24(11):2032–2045. https://doi.org/10.1109/TASLP.2016.2593263
32. Gannot S, Vincent E, Markovich-Golan S, Ozerov A (2017) A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Trans Audio Speech Lang Process 25(4):692–730. https://doi.org/10.1109/TASLP.2016.2647702
33. Wang Y, Narayanan A, Wang D (2014) On training targets for supervised speech separation. IEEE/ACM Trans Audio Speech Lang Process 22(12):1849–1858. https://doi.org/10.1109/TASLP.2014.2352935
34. Abdel-Hamid O, Mohamed A, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 22(10):1533–1545. https://doi.org/10.1109/TASLP.2014.2339736
35. Truong H, Dustdar S (2015) Principles for engineering IoT cloud systems. IEEE Cloud Comput 2(2):68–76. https://doi.org/10.1109/MCC.2015.23
36. Guo Y, Stolyar AL, Walid A (2020) Online VM auto-scaling algorithms for application hosting in a cloud. IEEE Trans Cloud Comput 8(3):889–898. https://doi.org/10.1109/TCC.2018.2830793
37. Akata Z et al (2020) A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer 53(8):18–28. https://doi.org/10.1109/MC.2020.2996587
38. Kucherbaev P, Bozzon A, Houben G (2018) Human-aided bots. IEEE Internet Comput 22(6):36–43. https://doi.org/10.1109/MIC.2018.252095348
39. Hinton G et al (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag 29(6):82–97. https://doi.org/10.1109/MSP.2012.2205597
40. Juang CF, Cheng CN, Chen TM (2009) Speech detection in noisy environments by wavelet energy-based recurrent neural fuzzy network. Expert Syst Appl 36(1):321–332
41. Juang CF, Lai CL, Tu CC (2009) Dynamic programming prediction errors of recurrent neural fuzzy networks for speech recognition. Expert Syst Appl 36(3P2):6368–6374
42. Tu CC et al (2012) Recurrent type-2 fuzzy neural network using Haar wavelet energy and entropy features for speech detection in noisy environments. Expert Syst Appl 39(3):2479–2488