Detecting Persian Speaker-Independent Voice Commands based on LSTM and Ontology in Communicating with the Smart Home Appliances

Leila Safarpoor Kalkhoran; Shima Tabibian; Elaheh Homayounvala

Abstract

Nowadays, various interfaces are used to control smart home appliances. The human and smart home appliances interaction may be based on input devices such as a mouse, keyboard, microphone, or webcam. The interaction between humans and machines can be established via speech using a microphone as one of the input modes. The Speech-based human and machine interaction is a more natural way of communication in comparison to other types of interfaces. Existing speech-based interfaces in the smart home domain suffer from some problems such as limiting the users to use a fixed set of pre-defined commands, not supporting indirect commands, requiring a large training set, or depending on some specific speakers. To solve these challenges, we proposed several approaches in this paper. We exploited ontology as a knowledge base to support indirect commands and remove user restrictions on expressing a specific set of commands. Moreover, Long Short-Term Memory (LSTM) has been exploited for detecting spoken commands more accurately. Additionally, due to the lack of Persian voice commands for interacting with smart home appliances, a dataset of speaker-independent Persian voice commands for communicating with TV, media player, and lighting system has been designed, recorded, and evaluated in this research. The experimental results show that the LSTM-based voice command detection system performed almost 1.5% and 13% more accurately than the Hidden Markov Model (HMM)-based one, in scenarios 'with' and 'without ontology', respectively. Furthermore, using ontology in the LSTM-based method has improved the system performance by about 40%.

Keywords: Voice commands detection, ontology, smart home appliances, Long Short-Term Memory, accessibility.

1 Introduction

1.1 Speech as an interaction mode

Speech is a natural way of communication between humans. Thus, speech-based user interfaces have many advantages in comparison to other types of user interfaces. Speech-based user interfaces

AMM copy, Published in Artificial Intelligence Review 2022, http://dx.doi.org/10.1007/s10462-022-10326-x

make multi-tasking easier because users’ hands are free in this mode of communication and eye concentration on the interface is not necessary either. It is also a faster way of communication than other kinds of user interfaces. Speech as an input mode is a better choice for tasks in which hands and eyes are busy doing other tasks such as controlling other devices (Bird et al. 1997) with a satisfactory speech recognition performance (Këpuska 2011).

1.2 Speech recognition

Speech recognition refers to converting a spoken utterance into its textual equivalent. Speech recognition systems are developed to improve voice-based human-computer interaction, to solve hard problems such as translating spoken utterances from one language to another, and developing smart systems that can understand spoken utterances as efficiently as human beings (Bird et al. 1997). Moreover, one of the major advantages of voice modality for interaction is accessibility for visually impaired people, the elderly, or people with some kinds of mobility disabilities.

1.3 Paper main idea and contributions

This paper will address human-device interaction (HDI) and advanced intelligent assistance while using smart home appliances with a focus on speech-based technology. Users' communication with smart home appliances (TVs, lamps, Media players) is occurred by ordering a set of general voice commands. In order to model voice commands, we have to choose suitable machine-learning methods. Hidden Markov Model (HMM)-based method is the most popular classic machine learning method which models the time-varying nature of speech signals with high performance (Rabiner and Juang 1986). On the other hand, among different neural network-based methods especially deep learning, the most suitable ones to model the time-varying nature of the speech signals are those which exploit the time concept in their structures such as recurrent neural networks (RNNs) and time delay neural networks (TDNNs). Long Short Term Memory (LSTM) is one of the most popular RNN-based methods used successfully for modeling time series data such as speech signals (Eramo et al. 2020; Eramo et al. 2021; Hochreiter and Schmidhuber 1997). Moreover, it has solved the problem of vanishing gradient in RNN-based approaches. Therefore, HMM-based and LSTM-based methods, as two of the most successful approaches in the field of speech processing, have been used in this paper for modeling voice commands. Sometimes users’ commands do not have a predefined structure and only include users' needs. In this case, ontology using a defined knowledge base can identify users' needs and provide a suitable answer. The other advantage of exploiting ontology is the ability to infer indirect or ciphered commands. This advantage help reduce the users' privacy concerns about these kinds of smart speakers with voice assistants due to their continuously listening microphones (Huxohl et al. 2019). Therefore, in this paper, our approach aims to improve the performance of speech-based human-to-machine communication and to solve the challenge of forcing the end-users to follow a fixed set of pre-defined commands utilizing ontology. According to the research studies in the field of smart home appliances, a limited number of datasets have been collected in the Persian language for controlling these appliances. Most of these datasets have been prepared by audio and speech processing laboratories for personal use. Therefore, in this work, we pay attention to the Persian language to compensate for the mentioned lack in this language. In summary, our work has the following contributions:

1- Creating a dataset of Persian voice commands for communicating with the smart home appliances

2- Exploiting the ontology for alleviating the necessity for users to learn or remember the appliances' specific words/commands and allowing them the opportunity to communicate with their devices by uttering their favorite commands.

3- Incorporating indirect commands as well as direct ones for interacting with the smart home context.

4- Using the LSTM method to increase the accuracy of detecting the Persian voice commands in smart homes.

5- Providing high accessibility for visually impaired users, elderly persons, and people with mobility disabilities via a speech-based interface.

1.4 Paper outline

The structure of the paper is presented in the following. The second section of the paper presents the related works. Our proposed dataset of Persian voice commands for interacting with smart home appliances has been presented in the third section. Then, in the fourth section, we show how to exploit LSTM and Ontology for detecting Persian voice commands. The experimental results have been presented in the fifth section. Finally, the paper is concluded in the sixth section.

2 Related works

The interaction between human and machine needs to meet two goals; the machine has to figure out the human purpose and the human has to find out the operation of the machine (Mehrabani et al.

2015). To make human purposes understandable for a device requires translating the human commands into the device control commands. Many works have been conducted on human-device interaction and its challenges in the form of research papers, chapter books, thesis, etc. (Preece et al. 2015; Rubio-Drosdov et al. 2015; Weng et al. 2016).

In recent years, improvements in information and communication technology have led to the smart home's adventure. Smart homes provide new chances to increase the in-home comfortability of people by providing different kinds of human-device interactions. One of the most important applications of smart home domains is assisting people with different types of disabilities, especially elderly persons. The main basis of most research studies in the field of smart homes is how human interacts with different appliances such as lighting system, audio and video equipment, and security system. Human communication with devices can be via modes such as mouse, keyboard, and webcam. However, these modes can sometimes be confusing and inefficient. Thus, they may lead to user error. The interaction between human and machines can be established via a microphone (based on speech) to provide a more natural, convenient, and efficient speech-based communication (Chuang et al. 2017; Munir et al. 2019). Since a smart home is made up of connected smart objects, a controller should be embedded for each object of the smart home such as light, sound, and door. For the elderly and disabled people, a speech recognition system is needed to work with these controllers (Chenxuan 2021). According to the research in (Lau et al. 2018), the most important interactive properties of personal voice assistants such as authentication and authorization, activity-based interaction, explainability and transparency can be used to improve the quality of interaction with smart home appliances. Moreover, an accurate voice command detection or speech recognition method is needed to achieve better performance in detecting users' tasks while using voice mode. In (Wang 2020), an acoustic model based on speech feature recognition and adopted Deep Neural Network (DNN)- Hidden Markov Model (HMM) is proposed. Additionally, the nonlinear activation function has been replaced with the linear unit (ReJUs) to obtain better performance. The Deep Belief Network (DBN) and the Deep Auto-encoder model are used as DNNs for feature training and hence improving the speech recognition accuracy (Wang 2020).

The speech-based communications have several challenges. One of these challenges is that existing speech-based interfaces in the smart home field limit the users to express a fixed set of pre-defined commands and do not support indirect commands. Thus, users have to memorize all the commands. If the user commands do not match the device control ones, the device cannot reply to the user queries and hence it makes the user dissatisfied.

There are several approaches to solving the mentioned challenge. In (Rubio-Drosdov et al. 2017) the machine-to-machine interaction has been used instead of the human-to-device interaction. In this approach, the user inputs his/her orders to the central device. Then, the devices communicate with each other, automatically, to respond to the user's needs. Although in this approach the human-device interaction is easy, all the devices must interact with each other, properly. In (Bajpai and Radha 2019) all devices can also be operated by only a single remote controller rather than one controller per device. This universal controller can be implemented easily and cost-effectively using an existing smartphone and an Arduino microcontroller via Bluetooth transmission. Thus, smartphone-based controller eliminates the need for carrying many different remote controllers. In addition, speech-based device controlling avoids looking for various buttons/options while operating the devices.

In (Zhang et al. 2012), the main contribution is based on combining speech recognition and natural language understanding methods to improve human interaction with smart home applications.

Ontology has been utilized in a lot of research studies to improve the performance of speech recognition or synthesis (Elsayed and Fathy 2020; Huang et al. 2015; Reedoy et al. 2021). Moreover, the results of using ontology in speech-based user interfaces have been very promising. According to studies in the field of smart homes, the ontology has been mostly used in the field of energy management (Saba et al. 2018; Saba et al. 2021) and in designing smart home models (Khan and Ndubuaku 2018; Reedoy et al. 2021). In (Milward and Beveridge 2003), the speech recognition complexity in breast cancer detection tools has been decreased by exploiting the ontology as a question/answer system. In this way, the users express their needs and the system understands the purpose of the users by asking general and partial questions, and finally, it acts, accordingly.

In (Al-Osaimi and Karim 2017), the device-to-device interaction is based on ontology. In this paper, the speech recognition system language models are based on ontology. Moreover, the device-to-device communication is knowledge-based (KB). As an example, suppose that the user has to enter the room; he/she orders the door to open. The door infers that the room lamp has to be turned on. Therefore, it orders the lighting system to turn the lamp on.

In (Kalkhoran et al. 2020), ontology has been exploited as a KB in a platform to examine the possibility of improving the performance of an HMM-based voice command detection system for controlling smart home appliances. The experimental results confirmed that the accuracy of detecting the voice commands while exploiting the ontology is higher than that in the “without ontology” case.

The main purpose of the current study is to increase user satisfaction while interacting with smart home appliances. In this regard, an ontology-based approach has been proposed which enables users to establish two-way communication with their selected devices without the necessity to use the device’s predefined commands. Moreover, the possibility of using indirect or ciphered commands has been realized. Since there is very limited research in the field of voice command detection in the smart home in the Persian language, the main focus of this work is the Persian language. The next section presents the proposed Persian dataset gathered for interacting with the smart home appliances mentioned in our work.

3 The proposed Persian dataset

Different research studies in the field of speech-based interaction with smart home appliances confirm the limited work in the Persian language.Therefore, in this work, we decided to collect a Persian dataset of speaker-independent voice commands for interacting with smart home appliances. We named our database PVC-SHA database which stands for **P**ersian **V**oice **C**ommands for interacting with **S**mart **H**ome **A**ppliances. For this purpose, a population of students (30 persons aged 17-19 and about 10 persons aged 24-28) was requested to declare all the conceptual commands they preferred or needed when instructing lamps and audio-visual equipment such as televisions and media players. The English-translated form of the PVC-SHA dataset direct Persian commands for TV and their keywords and all the indirect commands have been presented in Table 1. The direct commands for other appliances such as media players, lamps, and lighting system are similar to the TV commands for turning them on and off and setting their volume or light. For example, “Media player! Turn on”, “Media player! Play it” and “Media player! Get started” commands are the English translations of the Persian commands “ضبط صوت! روشن شو”, “ضبط صوت! بخوان” and “ ضبط صوت! شروع کن”, respectively.

Table 1 The English-translated form of some of the Persian Commands and Keywords of the PVC-SHA Dataset

Commands	Keywords
TV! Shut down (Cut it/ It is enough/ Shut up/ Stop it)	TV, shut down (cut/ enough/ shut up/ stop)
TV! Get started (Open up/ Show/ Play it/ Turn on)	TV, started (open/ show/ play/ turn on)
TV! Reduce the volume	TV, reduce, volume
TV! Increase the volume	TV, increase, volume
TV! Make the volume normal (medium)	TV, volume, normal (medium)
TV! Increase the screen light	TV, increase, screen light
TV! Reduce the screen light	TV, reduce, screen light
TV! Make the screen light normal (medium)	TV, normal (medium), screen light
TV! The volume is weak	TV, volume
TV! The volume is very weak	TV, volume, weak, very
TV! The volume is loud	TV, volume, loud
TV! The volume is very loud	TV, volume, loud, very
Lamp! It is dark	Lamp, dark
Lamp! It is very dark	Lamp, dark, very

As can be seen in Table 1, the PVC-SHA database includes 32 keywords, six garbage words, contain auxiliary verbs, conjunctions, etc., and 63 voice direct and indirect commands. In direct commands, the users instruct the device directly by expressing their requests explicitly, that include commands such as: “TV! Shut down!” and “Lamp! Reduce the light”. The direct commands have been recorded using 30 speakers, including 15 men and 15 women with an age range of 13 to 50 years and with different educational levels. The whole duration of all recorded direct commands is about one hour, 42 minutes, and 30 seconds. The age distribution of the 30 speakers is as follows: two women and four men from 10 to 20 years old, nine women and 13 men from 21 to 30 years old, one woman from 31 to 40 years old, and one woman from 41 to 50 years old. Moreover, the speakers have different Persian accents such as Azari (two men and five women), Kurdish (three men), Tehrani (three men), Arabic (two women and one man), Khorasani (one woman), Semnani (two women and one man), Golestani (one woman), Ghomi (one woman and one man), Mazandarani (two men), Kashani (one man), Araki (one man and one woman) and Shirazi (one woman).

The users' requirements are not mentioned, explicitly, in the indirect commands. As an example the command "Media player! The volume is loud." is an indirect command to express the user's need to reduce the player volume. Ontology has been exploited besides the voice command detection system in order to understand the users' indirect queries. The indirect commands have been recorded in 16 minutes and 20 seconds using 20 speakers, including 10 men and 10 women with an age range from 13 to 55 years old and with different educational levels. Moreover, the speakers have different accents such as Azari (two men and two women), Kurdish (one woman and two men), Tehrani (one man), Lori (one woman), Khorasani (one woman), Yazdi (one woman and one man), Golestani (one man), Mazandarani (two men and one woman), Kermani (one man and one woman) and Shirazi (one woman).

Each wave file in the dataset has a unique name. This name contains the abbreviation form of the corresponding voice command (For example, the abbreviation form of the voice command "Media player turn on" is "Mto"), sp (stands for the speaker), and the ID of speakers that is a number between 1 and 50. For example, Mtosp50 is the name of the wave file for the command (Media player turn on) of the speaker with ID 50. The PVC-SHA dataset has been recorded in wave format, mono, and sample rate of 16 kHz using the Voice Media player application (an Android app).

It is necessary to mention that the voice commands have been recorded in a clean (noise-free) condition. Therefore, it can be claimed that all the wave files were collected without any significant background noise signal.

Labeling of the proposed dataset is done manually at the word level using the Cool Edit Pro. Software. For silent and non-keyword parts, the assigned label is "sil" and "filler", respectively. Other words have been labeled with the corresponding keyword. Therefore, the PVC-SHA dataset contains 2737 wave files and 2737 word-level label files. The wave files duration is between 2 seconds (for the shortest command) and 4 seconds (for the longest command).

As it is obvious from Table 1, different commands could be used for doing one task. As an example, "cut", "enough" and "stop" have the same meaning as "shut down". The purpose of recording different commands for each special task is to solve the challenge of understanding the commands via their meanings instead of matching their exact structures. Numerous datasets in the smart home appliances domain have been published in the English language. Table 2 shows the comparison between the PVC-SHA and other datasets in this field.

Table 2 Comparing the PVC-SHA dataset with other datasets in the smart home appliances domain

Reference	Commands	language	Appliances
Current paper	Direct and indirect commands (Table 1)	Persian	TV, lamps Media players, Lighting system
(Mittal et al. 2015)	“turn on/off”, “turn up/down”, “increase/decrease”, “play/pause”, “start/stop”, “resume”, “check”, “read”, “change”, “move next/previous”	English	“light”, “light bulb”, “TV”, “music player”, “air conditioner”, “fridge”, “clock refrigerator”, “washing”, “washing machine”
(Han et al. 2016)	“open”, “close”, “guest”, “silence”, “help”, “on”, “off”	English	light group, fan group, safety group, access group, utility group
(Wang 2020)	“turning off the light”, “turning on the light”, “closing the door”, “opening the door”, “opening the window”, and “closing the window”	Not reported	light, door, window

As Table 2 shows, the PVC-SHA dataset includes both direct and indirect commands for instructing the home appliances. While in the previous works (Han et al. 2016; Mittal et al. 2015; Wang 2020) only direct commands have been proposed. Additionally, the PVC-SHA is the only available Persian-language dataset that can be used in research studies and industrial projects with the need to detect Persian commands in smart homes. Moreover, according to the variety of words with the same meaning in the Persian language, the PVC-SHA dataset includes different versions of expressing a command to convey a specific concept. Thus, for each command, the dataset includes more than just one sentence. Different datasets support different appliances. For example, the PVC-SHA does not support doors and windows, as investigated in Wang’s study (Wang 2020), or air conditioner, fridge and washing machines, which were the subjects of study in (Mittal et al. 2015), or access and safety groups appliances studied by Han and colleagues (Han et al. 2016). However, it supports the main common appliances such as light and TV and media player groups appliances. In addition, in almost all the related references, the research studies refrain from full introduction of the dataset. Therefore, information such as number of speakers, gender distribution, dialect variety, etc. are not available. However, the PVC-SHA dataset has been completely introduced in this paper.

4 The proposed Persian speaker-independent voice command detection system for communicating with the smart home appliances

Voice command detection refers to detecting the spoken commands generated by a human, automatically and converting them into text or codes which are understandable for the machine. The block diagram of the Persian speaker-independent voice command detection system for communicating with smart home appliances is depicted in Fig. 1.

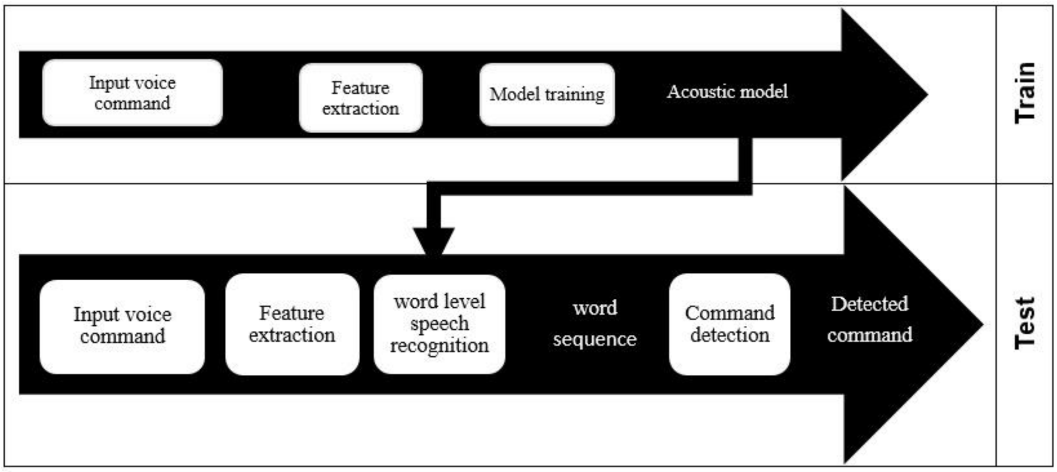


Fig. 1 The block diagram of the proposed Persian speaker-independent voice command detection system for communicating with the smart home appliances

As it is clear from Fig. 1, the proposed voice command detection system includes train and test phases. In the training phase, first, the suitable features are extracted from the input in the form of spoken commands. After that, the Cepstral Mean-Variance Normalization (CMVN) method (Bocklet and Marek 2020) is used in order to normalize the extracted features and remove the channel noise effects. The second step is to train the word-level speech recognizer which plays the main role in the accuracy of the voice command detection system (Ittichaichareon et al. 2012). In the current research, we trained the word-level speech recognizer based on HMM-based and long short-term memory (LSTM)-based methods. In the test phase, the features are extracted from the input voice commands and are normalized using the CMVN method. Then, the trained HMM or LSTM-based word-level speech recognizer recognizes the input voice commands based on their keywords. All the commands are made of at least two (appliance name and appliance task) and at most three (appliance name, a quantity value (low, medium, high), and appliance task) parts. The first and the last parts of each Persian command are the appliance name and the appliance task, respectively. At the command detection part, a simple rule-based language model or a more complex structure such as ontology has been exploited to detect commands based on the recognized list of words. The command detection part has been depicted in Fig. 2.

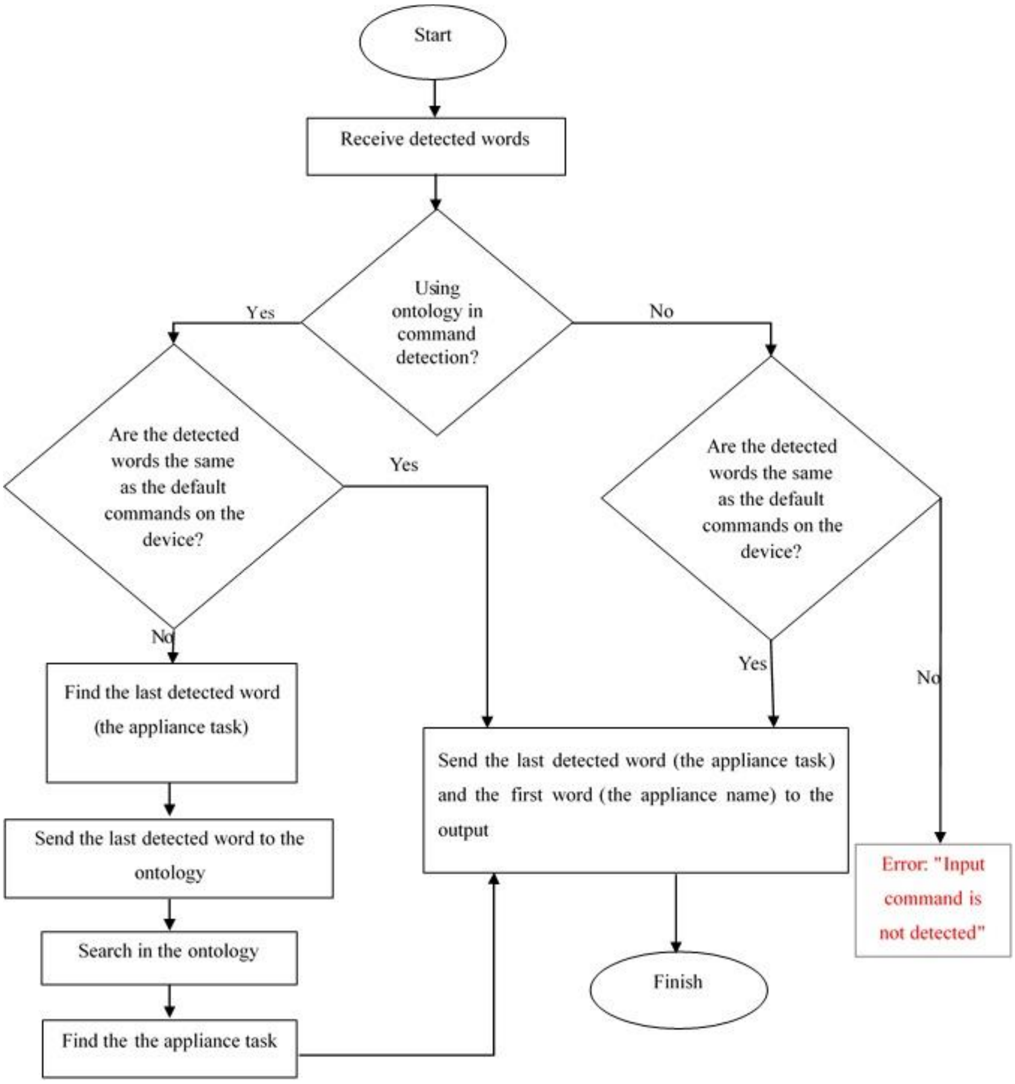

Start
Receive detected words
Using ontology in command detection?
Yes
No
Are the detected words the same as the default commands on the device?
Yes
No
Are the detected words the same as the default commands on the device?
Yes
No
Find the last detected word (the appliance task)
Send the last detected word to the ontology
Search in the ontology
Find the the appliance task
Send the last detected word (the appliance task) and the first word (the appliance name) to the output
Finish
Error: "Input command is not detected"

Fig. 2 Command detection part of the proposed smart home voice command detection system

As it is shown in Fig. 2, the input spoken command is detected with or without ontology. The first detected word in the input command refers to the appliance name, and the last one refers to the appliance task. In the “without ontology” case, if the detected task matches the pre-defined appliance phrase, the first and the last detected words (for two-part commands) will form the final output. In the case of a three-part command, the first, the quantity value (low, medium, high) and the last detected words will form the final output. If the detected task does not match the predefined appliance phrase, an error of expressing not-defined command will be reported. In the "with ontology" case, if the detected task does not match the predefined appliance phrase, but has the same meaning as the predefined appliance phrase, the ontology will infer it. Then, the inferred task, the appliance name, and the quantity value (for a three-part command) will form the final detected command. For example, in an indirect command such as "the room is dark", since the device must recognize the requested action according to the meaning of the user's sentence, there is a need to use ontology. Moreover, in a direct command such as "Connect Lamp!", since the user's command is different from the predefined command of the lamp ("Turn on the Lamp!"), the ontology searches the input command in the knowledge base. It identifies the user's desired operation. In this paper, we exploited ontology as KB to create a semantic dependence between the various command keywords. In addition, ontology helps the appliances understand the detailed structure of the voice commands. To prepare the control commands of smart home appliances, a group of different users has been chosen. We asked them to use their own commands for controlling smart home appliances. A large set of different commands for controlling each of the appliances has been generated. We exploited this set to create the proposed smart appliances ontology as shown in (Fig. 3).

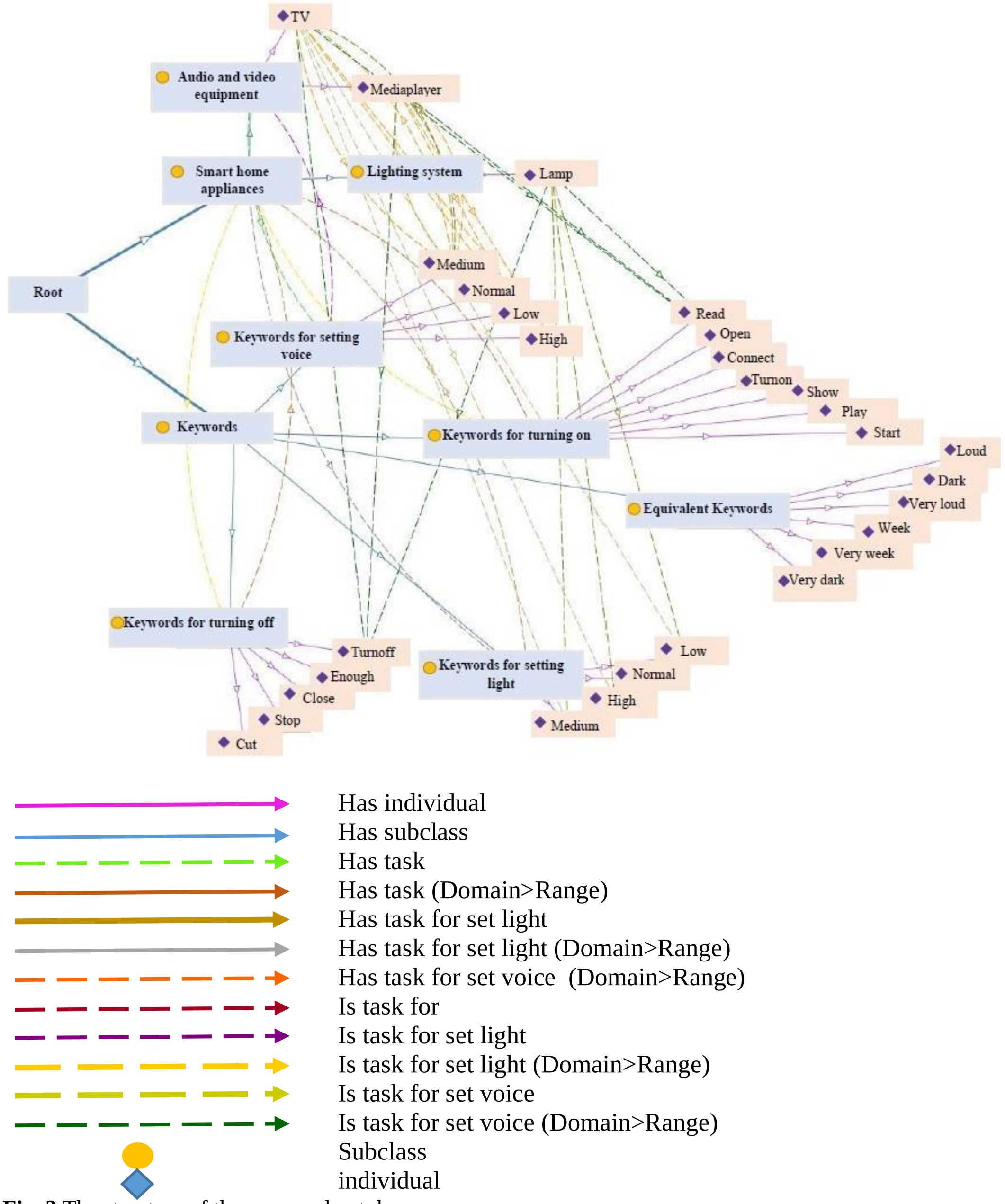


Fig. 3 The structure of the proposed ontology

As it is clear from Fig. 3, the proposed ontology includes two super-classes and seven sub-classes. The "smart home appliances" super-class includes audio and video equipment (media player, TV) and the lighting system (lamps) sub-classes. The "Keywords" super-class contains 'voice setting' keywords, 'light setting' keywords, 'device turning off', and 'on' keywords and 'equivalent keywords' sub-classes. The sub-class 'equivalent keywords' include those keywords that are used in indirect commands. For example, "very dark" contains an indirect concept of setting the light as 'medium'. Thus, "very dark" is equal to medium or normal light. Super-classes and sub-classes communicate with each other through the logical relationships that exist between them. The input in the form of detected words will be inferred or decoded into the corresponding command using these dependencies. Thus, it is claimed that, by using t ontology, the appliances will “understand” the concept of the input command words. Table 3 includes the user input commands and the ontology inferred commands for commands presented in Table 1.

Table 3 The user commands and the ontology inferred commands for commands presented in Table 1

Spoken commands	Inferred commands
TV! Stop it (It’s enough/ Shut down/ Shut up)	TV! Shut down.
TV! Get started (Open up/ Show/ Play it/ Turn on)	TV! Turn on.
TV! The volume is very weak TV! Turn up the volume	TV! Turn up the volume.
TV! The volume is weak (loud) TV! Make the volume medium (normal)	TV! Make the volume normal.
TV! The volume is very loud. TV! Turn down the volume	TV! Turn down the volume.
TV! Make the screen light medium (normal)	TV! Make the screen light normal.
TV! Turn down the screen light	TV! Turn down the screen light.
TV! Increase the screen light	TV! Increase the screen light.

The proposed Persian speaker-independent voice command detection system for communicating with smart home appliances is evaluated in the following section.

5 Experimental results

The proposed voice command detection system has been evaluated on the PVC-SHA dataset. The train and test sets contain 1777 and 960 voice commands, respectively. The speakers of the two sets are independent of each other. In order to evaluate the experimental results, accuracy, correctness and word error rate (WER) evaluation measures have been used and computed according to the following equations.

$$\text{Accuracy}= (H-I)/N \tag{1}$$

$$\text{Correctness}=H/N \tag{2}$$

$$\text{WER}= (D+S+I)/N \tag{3}$$

where H indicates the number of true detections. True detection here means the number of classes, namely words or commands, correctly detected according to the target of classification which can be speech recognition or voice command detection. N indicates the total number of identifiable classes (words or commands). S, I and D indicate the substitution, insertion and deletion errors, respectively (Klakow and Peters 2002). The evaluation results have been presented in the following sections.

5.1 Evaluation of the PVC-SHA dataset

The experimental results of implementing the HMM-based and LSTM-based methods on the PVC-SHA dataset have been analyzed in the next sub-sections.

5.1.1 The experimental results of evaluating the PVC-SHA dataset using HMM

The number of HMMs is 32 for 32 keywords, one silence model, and one filler model for the non-keyword parts of the voice commands. For each model, a different number of states (8, 10, 12, 14, 16, and 18) have been considered. Additionally, the number of Gaussian mixture functions in each state is a power of two and varies from 4 to 64. The optimum model configuration has been chosen after analyzing various evaluation results. The feature vector extracted from each speech segment contains 12 Mel Cepstrum coefficients plus energy coefficient and their first-order, second-order, and third-order derivatives. Again the optimum number of features is obtained after performing different evaluations. Table 4 presents the results of evaluating the PVC-SHA using HMM.

Table 4 The results of evaluating the PVC-SHA using HMM

Number of Gaussian mixtures	Number of states	feature vector size	Accuracy	Correctness	WER
16	8	39	**%**91,23	**%**92,54	**%**10,13
16	10	39	**%**89,75	**%**91,25	**%**10,24
16	12	39	**%**90,88	**%**92,07	**%**9,12
16	14	39	**%**90,24	**%**91,23	**%**9,75
16	16	39	**%**89,97	**%**90,84	**%**10,12
16	18	39	**%**89,03	**%**90,00	**%**10,96
4	8	39	**%**89,69	**%**91,66	**%**10,00
8	8	39	**%**91,41	**%**92,83	**%**8,58
32	8	39	**%**87,70	**%**89,22	**%**12,29
64	8	39	**%**81,66	**%**83.44	**%**18,40
8	8	52	**%**81,27	**%**83,62	**%**18,73
16	**8**	**52**	%**91,64**	%**93,67**	%**8,42**
64	8	52	**%**82,29	**%**83,42	**%**17,70

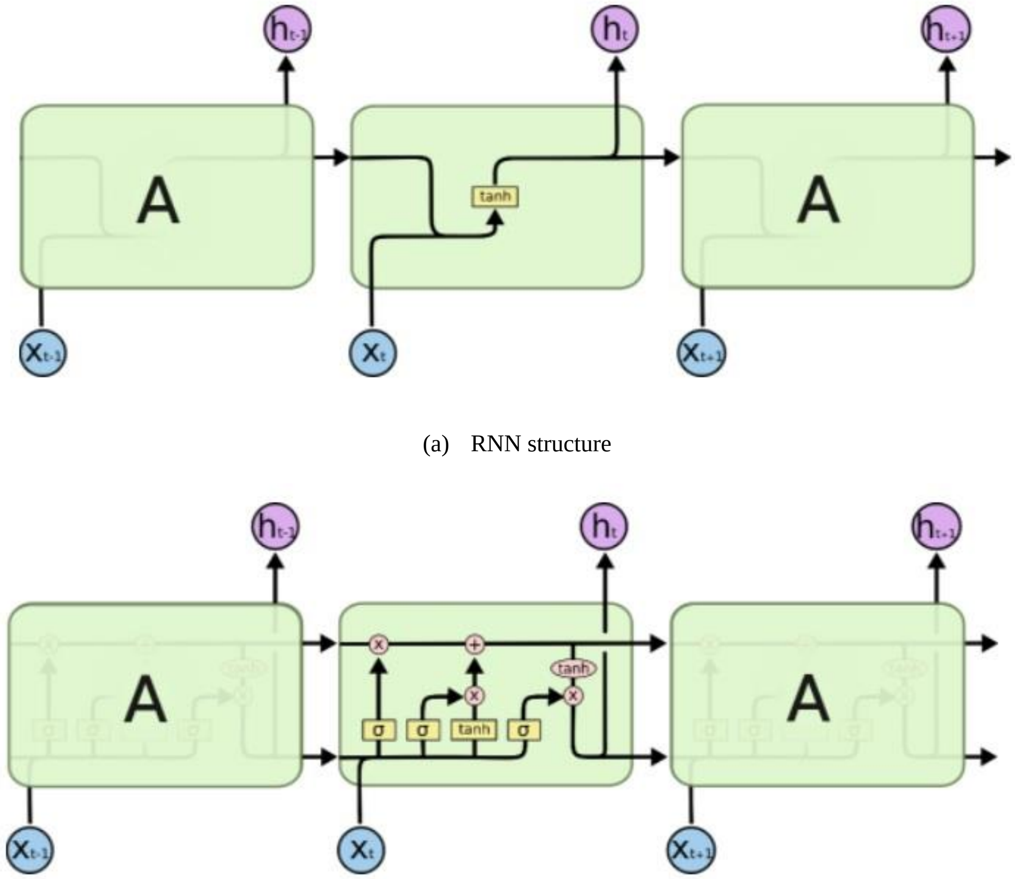
The reported results in Table 4 are based on the number of samples (keywords, silence, or non-keywords) which have been correctly detected. As it is obvious from Table 4, the evaluation results of modeling the PVC-SHA keywords using HMM, have been reported based on different evaluation measures such as accuracy, correctness, and WER. Different configurations based on the number of HMM states, the number of Gaussian mixture functions in each state, and the feature vector size have been chosen to train HMM-based word models. The number of HMM states is usually estimated based on the average number of phonemes that make the words modeled using the corresponding HMM. The Gaussian mixture functions in each state, models the observation probability matrix in each state. Increasing the number of Gaussian mixture functions may lead to the model accuracy and also model complexity increment.

In order to determine the best HMMs configuration, at the first stage, the number of Gaussian mixture functions and feature vector size were set to 16 and 39, respectively. Then, the number of states varied from 8 to 18 with step 2. As Table 4 shows, the best HMMs accuracy, correctness, and WER have been obtained with the number of states equal to 8 (you can see the first six rows of Table 4). At the second stage, with the fixed number of states equal to 8 and 39-dimension feature size, the number of Gaussian mixture functions varied from 4 to 64. It is better to select a number in the form of powers of two for this parameter. As Table 4 shows, the best HMMs accuracy, correctness, and WER have been obtained when the number of Gaussian mixture functions is equal to 16 as can be seen on the sixth to the tenth rows of Table 4. Finally, we tried another number for feature vector size (52 MFCC) which means 12 Mel Cepstrum coefficients, one energy coefficient, and their first-order, second-order, and third-order derivatives. The Final HMM was trained with the number of states equal to 8, the number of Gaussian mixture functions equal to 16, and the feature vector size equal to 52. This configuration has an accuracy of about 92%, the correctness of about 94%, and the WER of about 8% on the test set and is the best-obtained result as can be seen on the 12th row of Table 4.

5.1.2 The experimental results of evaluating the PVC-SHA dataset using LSTM

LSTM is an artificial neural network used in artificial intelligence and deep learning-based approaches (Eramo et al. 2021). LSTMs were developed to deal with the vanishing gradient problem encountered when training traditional RNNs. Unlike standard feedforward neural networks, LSTM has feedback connections. Such a recurrent neural network can process single data points and entire sequences of data (such as speech or video). For example, LSTM has been applied to tasks such as continuous handwriting recognition, speech recognition, machine translation, robot control, video games, and healthcare. LSTM prediction layer performs the time series prediction by providing the storage of the internal states (Eramo et al. 2020). Thus, it is also well-suited for classifying, processing, and making predictions based on time series data.

As mentioned above, the purpose of designing the LSTM was to solve the problem of remembering long-term dependencies in RNNs. RNNs have a very similar structure to multilayer perceptron networks, except that the latent layer neurons, in addition to the anterior edges, have a recurrent edge with a delay time. Such a structure ensures short-term dependency recall. However, it is not possible to learn the long-term dependencies. To solve this problem, hidden neurons were replaced with a more complex memory block, leading to the emergence of networks based on long short-term memory, or LSTM. Fig. 5 shows the structural differences between RNN and LSTM architectures.



(a) RNN structure

(b) LSTM structure

Fig. 4 The structural differences between RNN (a) and LSTM architectures (b) (Olah 2015)

As Fig. 4 shows, in RNN architecture, iterative modules have a very simple structure consisting of only one layer of the hyperbolic tangent. The input of each network includes the output of the previous module and the new input. The output of each module is the result of applying the hyperbolic tangent function (or any other activity function) to the weighted combination of the two inputs. In the LSTM architecture, the iterative module structure is more complex than that of the RNN, in order to remember long-term dependencies. The input of each module includes a new input and two outputs from the previous module. Each module has also two outputs. To produce each of these two outputs, sigmoid activity and hyperbolic tangent functions (or any other activity function) are applied to the outputs of the previous module and the new input and their weighted combinations. Details of this memory block, its components, and the related relationships are provided in (Olah 2015).

LSTM training algorithm has two stages; forward training and backward training (Hochreiter and Schmidhuber 1997). One period refers to a reciprocal movement in the network. That is, a period equal to one reciprocal movement in the entire network. The number of periods used to train the network and the number of hidden layers in the network can be changed. Although more periods and hidden layers lead to a more accurate network, it increases network training time. To get the best results on the PVC-SHA dataset using the LSTM-based method, we chose different values for the number of hidden units, the number of epochs, and the number of mini-batch sizes. The evaluation results have been shown in Table 5.

Table 5 The evaluation results of the PVC-SHA dataset using the LSTM-based method

Mini-batch Size	Max epochs	Number of units in each hidden layer	Accuracy
27	300	300	96.41%
30	1000	1000	97.18%
20	1500	1500	97.29%
30	800	500	97.94%
5	1200	2000	98.05%

10	700	500	98.16%
5	**800**	**500**	**98.20%**

As it is clear from Table 5, increasing the number of units in each hidden layer and number of maximum epochs usually increases the accuracy. According to Table 5, the best result is obtained for 500 hidden units in each hidden layer, 800 maximum epochs, and a minimum batch size equal to 5. These settings have an accuracy of about 98.2% on the test set. As it is clear from the evaluation results, the LSTM-based method performance is about 6% higher than the HMM-based one. It can be due to the fact that deep learning-based methods are more powerful to model patterns and so spoken utterances.

It has to be noted that all the experiments in this sub-section (5.1) have been done to evaluate the accuracy of the HMM-based or LSTM-based methods in recognizing the words constructing the smart home voice commands. Thus, the obtained evaluation results in this sub-section are the speech recognition accuracy at the word level. In the following sub-section, we have to evaluate the results of the voice command detection system. Thus, the system performance has been evaluated based on correctly detecting the whole voice commands that are highly likely to be smaller than the word-level speech recognition accuracy. This is due to the fact that, if the system does not detect correctly even one of the keywords of the input voice command, the whole voice command will not be detected by the system. In this case, if the input voice command is composed of for example four keywords, the speech recognition accuracy at the word level is equal to (3/4 or 75%) for that command. However, the accuracy in the command (sentence) level is equal to zero, since the input voice command could not be detected correctly. However, due to high word-level accuracy in both approaches, the mentioned situation is not highly likely to occur. Thus, the command (sentence)-level accuracies are expected to be a little smaller than that of the word-level ones.

5.2 The evaluation results of the voice command detection system

The proposed voice command detection system for controlling smart home appliances has been evaluated on the PVC-SHA dataset using both HMM and LSTM methods. We evaluated the proposed system with and without ontology for decoding the voice commands. The evaluation results have been shown in Table 6.

Table 6 The evaluation results of the proposed voice command detection system for controlling smart home appliances

Method	With ontology	Without ontology
LSTM	**96.72%**	**56.89%**
HMM	95.34%	43.27%

As it is obvious from Table 6, the true detection rate of the LSTM-based method with and without ontology has increased by about 1.38 and 13 percent in comparison to the HMM-based method, respectively. Moreover, the true detection rate of both the LSTM-based and HMM-based methods increases by about 39.83 and 52.07 percent, respectively, when using ontology. In other words, the LSTM-based voice command detection system, in the “with ontology” scenario, detects 561 out of 580 commands, correctly and 19 commands incorrectly. In the “without ontology” scenario, only 329 of 580 commands are correctly detected. Moreover, the HMM-based method, voice command detection system, in the “with ontology” case, detects 553 out of 580 commands, correctly and 27 commands incorrectly. In the “without ontology” scenario, only 250 out of 580 commands are correctly detected. When a user expresses his/her needs indirectly, “TV! The volume is very loud.” as an example, the voice command detection system without the ontology could not infer the user’s indirect command. However, the user indirect command can be inferred using the ontology. In our example, the system realizes that the user needs to decrease the TV volume using ontology. Additionally, in the “without ontology” case, if users express their commands using words that do not match the predefined appliances’ words, the system does not detect them. However, using ontology results in decoding the direct commands by paying attention to their meaning instead of their exact structure composed of predefined device words and hence leads to the significant improvement of the voice command detection accuracy in comparison to the “without ontology” case. Thus, ontology improves the voice command detection accuracy for both direct and indirect commands. It should be noted that this accuracy is at the command level, not at the word level. In other words, the reported accuracy is equal to the percentage of commands detected and executed, correctly. In some cases, in the "without ontology" case, the words forming the command were recognized correctly, but the input command was not detected, because the command did not match the system’s predefined commands. In the "without ontology" case, although the word level detection rate is as high as that of the "with ontology" case, the command level accuracy decreases because of the misunderstanding of the indirect and the direct commands with non-defined device words. Thus, although the LSTM-based method performance, at the word level, is about 6 percent better than the HMM-based method performance, its performance, at the command level, is only about 1.5 percent better than HMM-based method performance. Various methods in the smart home appliances domain have been compared with our proposed method in Table 7.

Table 7 A comparison between our proposed method and similar methods in the field of smart home appliances

Reference	Goal	Dataset	Command detection approach	Ontology	Experimental results
Current paper	Increasing the accuracy of the voice command detection in the smart home appliances domain	PVC-SHA (Persian)	LSTM-based and HMM-based voice command detection methods	✓	97% command-level true detection rate (40% improvement using ontology with LSTM-based method)
(Rubio-Drosdov et al. 2017)	Reducing the natural language processing complexity in the IoT context	2170 grammatical sentences from the online form (English)	A system created from ready-made modules that use natural language as a common interface for simple and complex orders	×	92% command-level true detection rate
(Mittal et al. 2015)	Proposing a multi-functional Smart Home Automation System (SHAS) with the ability to control the home appliances, and gadgets using voice commands	Commands for controlling five groups of appliances: access, fan, light, utility, and safety (English)	The voice commands are recognized using dedicated hardware (Voice Recognition Module) and a microcontroller (Arduino-UNO).	×	The best-achieved distance for correct detection of a voice command is between 5 cm and 10 cm (command-level true detection rate not reported)
(Han et al. 2016)	Proposing a speech-based control system for home appliances exploiting the context information	Commands for controlling light, music player, TV, clock, air conditioner, and some kitchen appliances (English)	A system for controlling home appliances using human speech and context information composed of speech recognition (CMU-Sphinx 4) and command executor modules.	✓	A quite low voice command detection accuracy is reported.

(Milward and Beveridge 2003)	Examining the possibility of Replacing the handcrafted dialogue design with a combination of a generic dialogue components and ontological domain knowledge	WordNet includes more than 118.000 different word forms and more than 90.000 different word senses for both home control and cancer referral applications (English)	An approach based on Hypernyms and Hyponyms term recognition	✓	No formal evaluation has been done
(Alexakis et al. 2019)	Introducing an IoT Agent for monitoring and controlling a smart home, remotely	No specific dataset is used due to exploiting a ready-made speech recognition system	Web Speech API for enriching	×	70% reduction in system response time (command-level true detection rate not reported)
(Chenxuan 2021)	Providing a speech recognition system based on deep learning to translate speech audio into instructions according to speech features	data_thchs3 Chinese speech dataset (25000 Chinese famous sentences in novels and poems), Free ST Chinese Mandarin Corpus (more than 300000 daily conversation recordings) and H1228 (6000 instructions recorded at different distances)	Auto Speech Recognition Tool based on deep learning	×	80% command-level true detection rate
(Wang 2020)	Increasing the accuracy of speech recognition systems in order to use them in the design of smart homes	3600 speech samples of 10 commonly used short words include “turning off the light”, “turning on the light”, “closing the door”, “opening the door”, “opening the window”, and “closing the	An acoustic model based on speech feature recognition and adopted DNN-HMM	×	90.1% command-level true detection rate

window”

As Table 7 shows, the research studies in the field of smart homes are divided into two categories. The first group has used existing ready-made speech recognition tools, APIs, and software for voice command detection. Thus, the main purpose of research studies is not focused on improving speech recognition rate but rather to find the best distance between the microphone and the device or how devices communicate with each other, and so on. The second category which also relates to our current work focuses on proposing an approach in order to achieve acceptable voice command detection performance. Thus, in these research studies, the command-level true detection rate is reported as a result. As it is clear from Table 7, the datasets in the smart home appliances domain were mostly based on the English language. However, our proposed system was trained in the Persian language. Additionally, the performance of our proposed voice command detection system is higher than that of the other research studies in the mentioned field (5% and 17% in comparison to the best and worst cases, respectively). Moreover, an ontology based on the Persian language to understand the meaning of Persian commands in the smart home appliances domain has been created in this paper for the first time. It leads to a 40% improvement in the performance of the voice command detection system.

Finally, in order to evaluate the users' preferences in using speech, touch, or mouse and keyboard for controlling the smart home appliances, we interviewed users of our proposed system. The subjective results have shown that about 80% of the users prefer the speech-based user interface due to its desirability, usability, memorability, learnability, and accessibility. This result is in line with other research studies such as (Lau et al. 2018), it seems that most users prefer to communicate with the smart home user interface based on speech in comparison to other modalities such as mouse, keyboard, and touch.

6 Conclusion

Nowadays, despite the technological improvement in the field of speech-based user interfaces, people are hesitant to use them. Some devices have a limited set of commands. Thus, they can only detect the same mentioned set of commands. It makes users dissatisfied since they are not free to use their own phrases or commands while interacting with those devices. In this paper, an ontology-based approach is used to resolve this problem by inferring the direct and indirect various user's requests. In the indirect commands the users' requirements are not mentioned, explicitly, while in the direct commands, the users instruct the device directly by expressing their request, explicitly. The good point about the proposed system is that we can even incorporate cipher commands, similar to (t indirect commands, selected by users to avoid controlling the smart home by intruders. Moreover, due to the unavailability of Persian voice commands for communicating with smart home appliances, we designed, assembled, and evaluated a dataset of Persian speaker-independent voice commands for communicating with smart home applications such as TV, voice media player, and lamp. The voice command detection system is trained using both the Hidden Markov Model (HMM)-based and Long Short Term Memory (LSTM)-based approaches. HMM-based and LSTM-based approaches have been selected according to their remarkable performance in modeling the time-varying nature of speech signals compared to other machine learning approaches.

As experimental results show, the accuracy of the LSTM-based system is about 1.5 percent higher than that of the HMM-based system. By using the ontology, the performance of the HMM-based and LSTM-based voice command detection systems have been improved by about 50 and 40 percent, respectively. The system's response time for commands with a time duration between two and four seconds is less than two seconds. Thus, the proposed voice command detection system is able to correctly identify user's commands in low time complexity. Therefore, increasing the commands volume mainly affects offline computational complexity and does not lead to higher response time complexity, because increasing the volume commands involves steps such as collecting the new commands and adding them to the Persian dataset, training the voice command detection system with new commands, and updating the ontology using new commands. Thus, the proposed system has acceptable scalability according to the command size. Additionally, the most critical work during the maintenance process of the system is monitoring the customer requirements for adding new possible commands for the existing or new devices and trained again the voice command system to consider changes in these requirements. Thus, the proposed system can be considered as a maintainable system.

In future studies of this research, we will focus on increasing the number of indirect commands ("Lamp! It is dark", "Lamp! It is the time to study", "Lamp! It is morning", "Lamp! It is night", "TV! It is a crowded room" and "Media player! It is a crowded room") to use ontological reasoning to better understanding and recognizing the commands and also reducing users' privacy concerns about controlling the smart home by intruders using the ciphered commands. In addition, we will exploit LSTM in the form of the sequence to sequence recognition to improve the voice command detection performance. Moreover, we will consider the microphone distance from the voice command recognition system. In addition to the above, another future direction of this research could be training the proposed voice command detection system, speaker-dependently, to support personalization and hence to increase the true detection rate of the commands in personal usage.

Statements and Declarations

Competing interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability: The datasets generated during the current study are not publicly available. However, further information about the data and conditions for access are available from the corresponding author on reasonable request.

References

Al-Osaimi R, Karim NA Ontology Powered Knowledge Modeling for a Smart Home. In: 9th IEEE-GCC Conference and Exhibition (GCCCE), 2017. pp 1-6

Alexakis G, Panagiotakis S, Fragkakis A, Markakis E, Vassilakis K (2019) Control of smart home operations using natural language processing, voice recognition and IoT technologies in a multi-tier architecture Designs 3:32-49

Bajpai S, Radha D Smart phone as a controlling device for smart home using speech recognition. In: International Conference on Communication and Signal Processing (ICCSP), 2019. pp 0701-0705

Bird S, Boguraev B, Kay M, McDonald D, Hindle D, Wilks Y (1997) Survey of the state of the art in human language technology vol 12. Cambridge university press

Bocklet T, Marek A (2020) Cepstral variance normalization for audio feature extraction. Google Patents

Chenxuan H Research on Speech Recognition Technology for Smart Home. In: IEEE 4th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), 2021. pp 504-507

Chuang LL, Glatz C, Krupenia S Using EEG to understand why behavior to auditory in-vehicle notifications differs across test environments. In: Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2017. pp 123-133

Elsayed EK, Fathy DR (2020) Sign Language Semantic Translation System using Ontology and Deep Learning Sign 11

Eramo V, Lavacca FG, Catena T, Di Giorgio F (2020) Reconfiguration of optical-NFV network architectures based on cloud resource allocation and QoS degradation cost-aware prediction techniques IEEE Access 8:200834-200850

Eramo V, Lavacca FG, Catena T, Salazar PJP (2021) Application of a Long Short Term Memory neural predictor with asymmetric loss function for the resource allocation in NFV network architectures Computer Networks 193:108104

Han Y, Hyun J, Jeong T, Yoo J-H, Hong JW-K A smart home control system based on context and human speech. In: 2016 18th International Conference on Advanced Communication Technology (ICACT), 2016. pp 165-169

Hochreiter S, Schmidhuber J (1997) Long short-term memory Neural computation 9:1735-1780

Huang C-C, Liu A, Zhou P-C Using ontology reasoning in building a simple and effective dialog system for a smart home system. In: IEEE International Conference on Systems, Man, and Cybernetics, 2015. pp 1508-1513

Huxohl T, Pohling M, Carlmeyer B, Wrede B, Hermann T Interaction Guidelines for Personal Voice Assistants in Smart Homes. In: International Conference on Speech Technology and Human-Computer Dialogue (SpeD), 2019. IEEE, pp 1-10

Ittichaichareon C, Suksri S, Yingthawornsuk T Speech recognition using MFCC. In: International Conference on Computer Graphics, Simulation and Modeling, 2012. pp 135-138

Kalkhoran LS, Tabibian S, Homayounvala E Improving the accuracy of Persian HMM-based Voice Command Detection System in Smart Homes Based on Ontology Method. In: 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), 2020. pp 1-5

Këpuska V (2011) Wake-up-word speech recognition Speech Technologies:237-262

Khan YI, Ndubuaku MU Ontology-based automation of security guidelines for smart homes. In: IEEE 4th World Forum on Internet of Things (WF-IoT), 2018. pp 35-40

Klakow D, Peters J (2002) Testing the correlation of word error rate and perplexity Speech Communication 38:19-28

Lau J, Zimmerman B, Schaub F (2018) Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers Proceedings of the ACM on Human-Computer Interaction 2:1-31

Mehrabani M, Bangalore S, Stern B Personalized speech recognition for Internet of Things. In: 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), 2015. pp 369-374

Milward D, Beveridge M Ontology-based dialogue systems. In: Proc. 3rd Workshop on Knowledge and reasoning in practical dialogue systems (IJCAI03), 2003. pp 9-18

Mittal Y, Toshniwal P, Sharma S, Singhal D, Gupta R, Mittal VK A voice-controlled multi-functional smart home automation system. In: Annual IEEE India Conference (INDICON), 2015. pp 1-6

Munir A, Ehsan SK, Raza SM, Mudassir M Face and Speech Recognition Based Smart Home. In: 2019 International Conference on Engineering and Emerging Technologies (ICEET), 2019. IEEE, pp 1-5

Olah C (2015) Understanding LSTM Networks.[(accessed on 10 September 2022)]. http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Preece J, Sharp H, Rogers Y (2015) Interaction design: beyond human-computer interaction. John Wiley & Sons

Rabiner L, Juang B (1986) An introduction to hidden Markov models ieee assp magazine 3:4-16

Reedoy AV, Dayal SB, Govender P, Fonou-Dombeu JV An Ontology for Smart Home Design. In: International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), 2021. pp 1-6

Rubio-Drosdov E, Díaz-Sánchez D, Almenárez F, Arias-Cabarcos P, Marín A (2017) Seamless human-device interaction in the internet of things IEEE Transactions on Consumer Electronics 63:490-498

Rubio-Drosdov E, Díaz-Sánchez D, Arias-Cabarcos P, Almenárez F, Marín A Towards a seamless human interaction in IoT. In: International Symposium on Consumer Electronics (ISCE), 2015. pp 1-2

Saba D, Degha HE, Berbaoui B, Maouedj R Development of an ontology based solution for energy saving through a smart home in the city of Adrar in Algeria. In: International Conference on Advanced Machine Learning Technologies and Applications, 2018. pp 531-541

Saba D, Sahli Y, Hadidi A (2021) An ontology based energy management for smart home Sustainable Computing: Informatics and Systems 31:100591

Wang P Research and design of smart home speech recognition system based on deep learning. In: International Conference on Computer Vision, Image and Deep Learning (CVIDL), 2020. pp 218-221

Weng F, Angkititrakul P, Shriberg EE, Heck L, Peters S, Hansen JH (2016) Conversational in-vehicle dialog systems: The past, present, and future IEEE Signal Processing Magazine 33:49-60

Zhang Y, Wei Z, Yang Y, Song C Ontology description of smart home appliance based on semantic web. In: International Conference on Computer Science and Service System, 2012. pp 695-698