# Multimodal Interaction System for Home Appliances Control

https://doi.org/10.3991/ijim.v14i15.13563

Hanif Fakhrurroja (✉), Carmadi Machbub, Ary Setijadi Prihatmanto,
Ayu Purwarianti
Bandung Institute of Technology, Bandung, Indonesia
hani002@lipi.go.id

**Abstract**—This paper proposes a way to control home appliances using a multimodal interaction system such as speech, gestures, and smartphone applications. The Kinect sensor used to capture Indonesian speech and gestures from users. Dialogue system, speech and gesture recognition process with finite state machine, Google Cloud Speech and K-Means Clustering, respectively. Users can also use the smartphone application to remotely control home appliances through mobile devices that are connected directly to the real-time database. There are two output responses from this system, namely the audio response generator to provide feedback to the user through the sound of the computer speaker and also provide an action to control home appliances use Esp8266. The average level of accuracy testing of interaction using dialogue systems and gesture are 92.5% and 79,25%. Interaction using dialogue systems is better than gesture. Smartphone applications can control home appliances properly.

**Keywords**—Multimodal interaction, speech recognition, gesture recognition, smartphone application, home appliances.

## 1 Introduction

Humans communicate with each other do not only depend on speech but they use different modes or ways, such as gestures, hand expressions (sign language), facial expressions (gaze/eye movements), touch screen, keyboard or pointing device [1]. The rapid development of technology has a role to increase the comfort of home dwellers. Almost all home appliances use electricity so that it can be controlled automatically using the internet of things technology [2].

Today's home automation system widely used and popular. Home automation, also known as domotics, is building automation for a home [3]. Home automation systems will control entertainment systems, climate, lighting, and other electric home appliances [4]. Smart home technology offers a new opportunity to improve the comfort of people with computing technology that provides enhanced communication through a variety from multimodal inputs specifically, speech, gesture and mobile application. This communication translates into actions that help the smart home system to complete the tasks. To design a multimodal interface, users must be given two things; a way to instruct the smart home system uses a dialogue system and feedback about what is action to home appliances [5].

Dialogue is a conversation between two or more people through oral or written, which aims to exchange and share information or to resolve an argument. Dialogue between humans and systems is called a dialogue system, where the system can interact with humans in natural language [6]. Dialogue management important for a dialogue system to set the flow of communication between users and systems with natural language. Challenges in natural language processing (NLP) is the relationship between speech and the intent of the user towards user desires. Building a dialogue management system is a challenging work in speech recognition technology to understand speech from users. This research uses dialogue in Indonesian language. System for understanding words in the flow of the dialogue refers to artificial intelligence (AI) with a machine learning approach that depends on data or information. Dialogue classification is the decision to understand the intent is very important in understanding the user's speech in dialogue [7].

A multimodal interaction system for home appliances control has successfully developed. Multimodal inputs used are speech, gestures, and smartphone applications, so make it easy for users to control all their home appliances. This smart home system is also equipped with a dialogue system so humans can interact to communicate their intents, such as to control room temperature and lights. The novelty of this research is the multimodal interaction algorithm using a smartphone application, speech, and gesture recognition that is equipped with a dialogue system so that machines can interact with humans.

## 2 Related Work

Interaction system development for home automation is developing very fast. Existing home automation technology allows people to control home appliances using computers connected to the local network. However, the real challenge is to control home appliances naturally and comfort, allowing users to use greater freedom and flexibility.

Home automation interaction systems can use smartphone applications to remotely control home appliances through mobile devices such as tablets or smartphones [4] [8]. Home automation can also use speech control systems where speech is converted to text using automatic speech recognition such as the Microsoft Speech API or Google Cloud Speech API, then send to the server using Wi-Fi. The server system converts incoming text data into a form that can be used to handle home appliances[9] [10]. Another interaction system uses hand gesture recognition [11] or combines speech and gesture to control home appliances [12]. This research proposes a new way to control home appliances uses multimodal interaction such as speech recognition, hand gesture recognition, and smartphone application equipped with a dialogue system so users can interact with smart home systems. The comparison of related work of interaction system for home appliances control describes in Table 1.

**Table 1.** Comparison Analysis of interaction system for home appliances control.

| References | Speech Recognition | Gesture Recognition | Smart-phone Application | Dialogue System | Positive and Negative Points of Each System |
|---|---|---|---|---|---|
| Meja, et al [4] | No | No | Yes | No | (+) Inexpensive controller system.<br>(-) unnatural human-machine interaction. |
| Pavithra and Balakrishnan [8] | No | No | Yes | No | |
| Baig, et al [9] | Yes | No | Yes | No | (+) Home appliances controlled by spoken command using hanheld devices.<br>(-) The system is dependent on human speech. The machine can't interact with users. |
| Kamarudin, et al [10] | Yes | No | Yes | No | |
| Jadhav, et al [11] | No | Yes | No | No | (+) Simple and easy method of controlling the home appliances.<br>(-) The system is dependent on human gestures. The machine can't interact with users. |
| Anbarasan and Lee [12] | Yes | Yes | No | No | (+) Easy control of home appliances through combination of speech and gesture interactions.<br>(-) The machine can't interact with users. |
| This Work | Yes | Yes | Yes | Yes | (+) Human–Machine interaction more natural; Easy control of home appliances through combination of smartphone application, speech and gesture recognition; The machine can interact with users.<br>(-) Speech recognition use of third-party applications (Google Cloud Speech) |

## 3 Purpose Method

Figure 1 describes the overall architecture system proposed for the multimodal interaction to control home appliances. Microsoft Kinect 2.0 used as a sensor to capture the speech and gesture from the user. Smartphone application used to remotely control home appliances using mobile devices such as a tablet or smartphone.

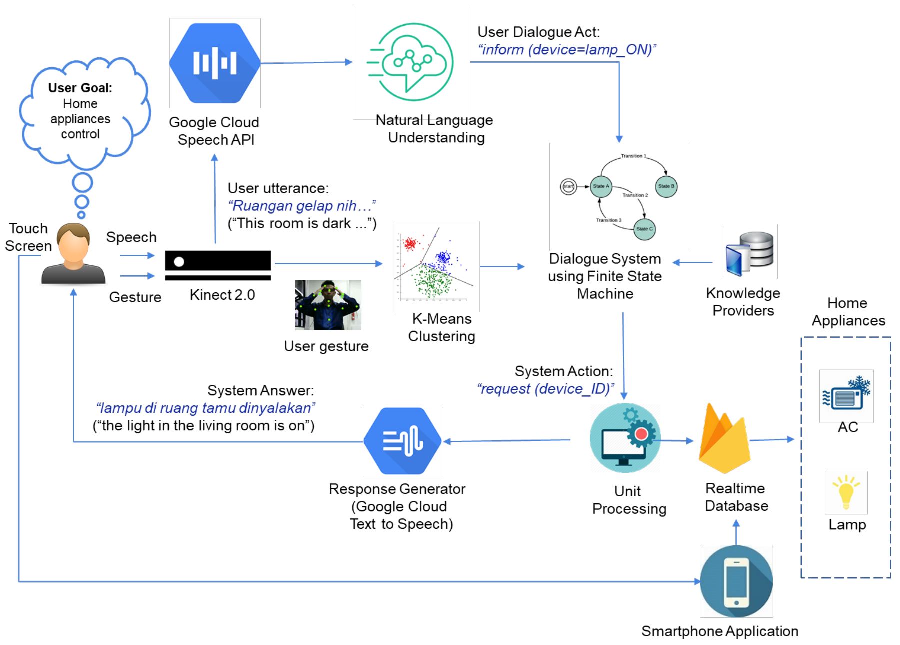


**Fig. 1.** Multimodal Interaction System Design

Speech recognition process with the Google Cloud Speech API integrated into Microsoft Visual Studio 2015. Gesture recognition process with the K-Means Clustering method used C# code. Natural language understanding and dialogue system with finite state machine method processed use phyton code. Users can also use the smartphone application to remotely control home appliances through mobile devices such as tablets or smartphones that are connected directly to the real-time database. There are two output responses from this system, namely the audio response generator to provide feedback to the user through the sound of the computer speaker and also provide an action to control home appliances use Esp8266 (NodeMCU).

### 3.1 Kinect v2 sensor

Many researchers and practitioners of robotics, electronic engineering, and computer science use Kinect to evolve new ways to interact with computers or machines. Kinect Software Development Kit (SDK) has the potential to change human-machine interactions in various industries, such as education, health, retail, transportation, and so on [13]. The second generation, Microsoft Kinect v2 released in 2013. Kinect v2 utilizes the principle of time of flight (ToF) and offers a field of view and larger resolution than Kinect v1 [14].

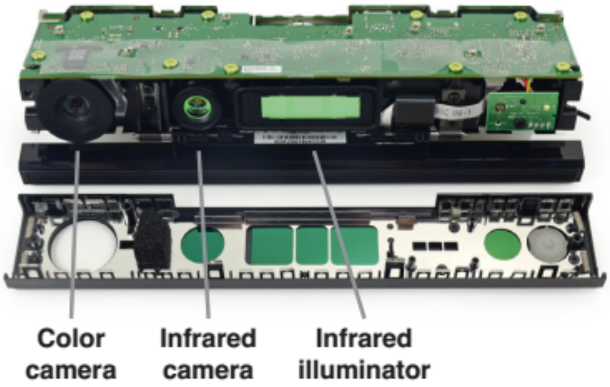


**Fig. 2.** The Kinect v2 Sensor [15]

Figure 2 shows the Kinect v2 sensor. Kinect v2 has three main components that work together for the Gesture Recognition process. RGB Camera functions to capture the image in front of it. In Kinect version 2, the captured image has a resolution greater than version 1, which is 1920 x 1080 pixels. The IR Emitters function to transmit infrared particles on the object in front of them and the Depth Sensor to read the distance of each particle that is on the object's surface. From the results of this reading, it can also be obtained the coordinates of objects in 3D space, including detecting each point of a person's skeleton. The detected skeleton can then be mapped to the captured image to fit the actual position. In Kinect version 2 the data processing is faster, and the accuracy is even higher than version 1 [16].

### 3.2 Speech recognition using google cloud speech

Speech recognition is the process of converting human voice signals into written language (text). Speech recognition is the most important part used in this dialogue system because reliable speech recognition can make a dialogue that is following the scenario. Cloud API development has been growing, one of which is the Google Cloud Speech API, which currently has 120 languages including Indonesian languages. We considering choosing Google because Google is the popular largest search engine and freely accessible API for the cloud-based speech recognizer [17]. Google Cloud Speech API (Application Programming Interface) is one of the speeches to text services provided by the largest search engine, Google, which can be integrated by the developer into the application. Cloud API can determine which application software to use can interact with cloud-based platforms [18]. Google Cloud Speech services help to recognize voice transmitted in requests and unify it to voice storage on Google Cloud Storage. This Google Cloud speech applies algorithms of a neural network to recognize user voice by generating high accuracy. Google's speech accuracy itself increases with time when Google companies improve speech recognition technology and internal software. The method used by Google Cloud Speech API is synchronous recognition by sending audio data in the form of a wav file to the Speech-to-Text API, and introducing the data and returning the results after all audio is processed [19].

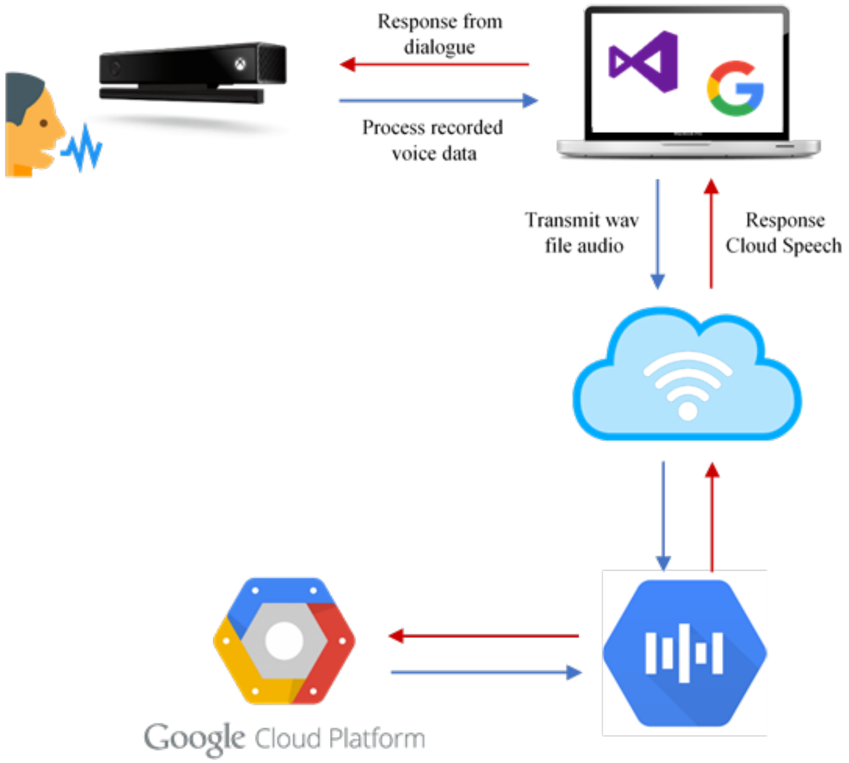


**Fig. 3.** Architecture Speech Recognition Using Google Cloud Speech

Figure 3 shows the speech recognition architecture using the Google Cloud Speech API, then processed recorded voice data through software on a personal computer (PC). Sound files are recorded in real time, then transmitted to Google Cloud Server, after the Google Cloud Speech Platform recognizes sound after receiving a sound package then sends the converted text back to the user [20].

### 3.3 Natural Language Understanding (NLU)

The purpose of NLU systems is to enable communication between man (users) and machines. The NLU system identifies the user's intention of natural language by extracting words that contain information and issuing queries to the back-end database to fulfill user requests [21].

The processes conducted in the NLU system are stemming, slot filling, and understanding intent by using rule-based. Stemming is the processing of a sentence to get the root word by separating each word from the base word and prefix and suffix. For example, the words “closes”, “closed”, and “closer” are clustered with the stem “close.” Indonesian stemming is a more complex effort than English stemming, so order of stemming rules requires careful consideration [22].

After the stemming process is done, then each word labeled a part of speech based on root words in the Indonesian dictionary corpus. The number of root words used in this research corpus is 28,526 words. The part of speech in Indonesian has seven types, namely nouns, verbs, adjectives, pronouns, adverbs, number words, and task words. After labeling the part of speech, the next is slot filling. The main task of language understanding is to automatically classify domains of user requests along with specific intent domains and fill in a set of slots to form sentence meanings (semantics). The popular IOB (in-out-begin) format is used to represent sentence slot tags [23] as shown in Figure 4. After slot filling is done, then understand the user's intent by using rule-based.

| | | | | |
|---|---|---|---|---|
| **W** (word) | *Ruangan*<br>(the room)<br>↓ | *Gelap nih,*<br>(is dark)<br>↓ | *Nyalakan*<br>(turn on)<br>↓ | *Lampu*<br>(the lamp)<br>↓ |
| **S** (IOB format) | O | B | I | O |
| **I** (Intent) | ***nyalakan_lampu* (light_on)** | | | |

**Fig. 4.** Examples of utterance with annotations semantic slots in IOB format (S) and intents (I)

### 3.4 Dialogue system using Finite State Machine (FSM)

FSM is used to model control and sequence processes in a system with a finite number of states. In particular, the actions of the system depending on the state do not depend only on the input to the system but also on what happened earlier in the system. State machines are very important for determining systems with behaviors that depend on significant circumstances [24].

Figure 5 shows the structure of the dialog scenario. Dialog system uses a finite state machine based on user requests. The system will continue the current state and will respond based on the previous transition and circumstances [25].

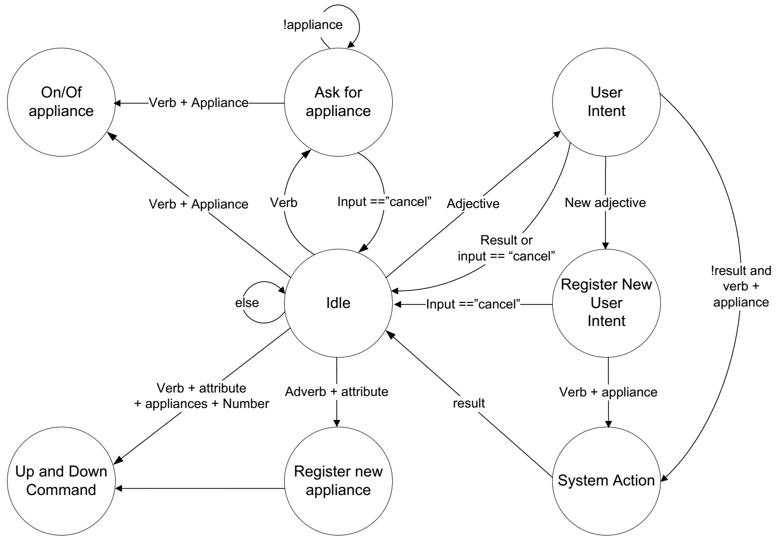


**Fig. 5.** Scenario of Dialog Using Finite State Machine

This dialog system has text processing to understand the intent of the user. This system filters certain words to recognize as slots. The dialogue system requires input slots in the form of verbs, appliances, adverbs, attributes, or adjectives to be able to move from one state to another. An idle state is an initiation state where a new program is run or command successfully execute. After the slots filled, FSM can be used to handle direct or indirect commands, such as:

1. Direct command:

   a) "*Tolong hidupkan lampu* (Please turn on the light)." Verb slot: *hidupkan* (ON); appliance slot: *lampu* (lamp).
   b) "*Naikkan suhu AC 2 derajat* (Increase AC temperature up to 2 degrees)". Verb slot: *naikkan* (increase); appliance slot: AC; attribute: *suhu* (temperature); Number: 2.

2. Indirect command:

   a) "*Duh, gelap banget* (It's so dark)." Adjective slot: *gelap* (dark) ➔ The system will turn on the lights.
   b) "*Panas banget nih* (It's so hot)." Adjective slot: *panas* (hot) ➔ The system will turn on the air conditioner (AC).

### 3.5 Gesture recognition using k-means clustering

The Kinect sensor captures skeletal joints on the right hand and left hand. This skeletal joint featured into three parts, namely thumb, elbow, and wrist, each having a position value for each axis (x, y, z). The data of each joint is then statistically calculated to get the value of average, variance, sum, and median. K-means algorithm is computed Euclidean Distance so each joint can be recognized in the form of clustering results.

Clustering is a data analyzing method to group data with the same characteristics in the same area and the distinct characteristics in the distinct area [26]. Figure 6 shows a proposed gesture recognition system using K-Means Clustering.

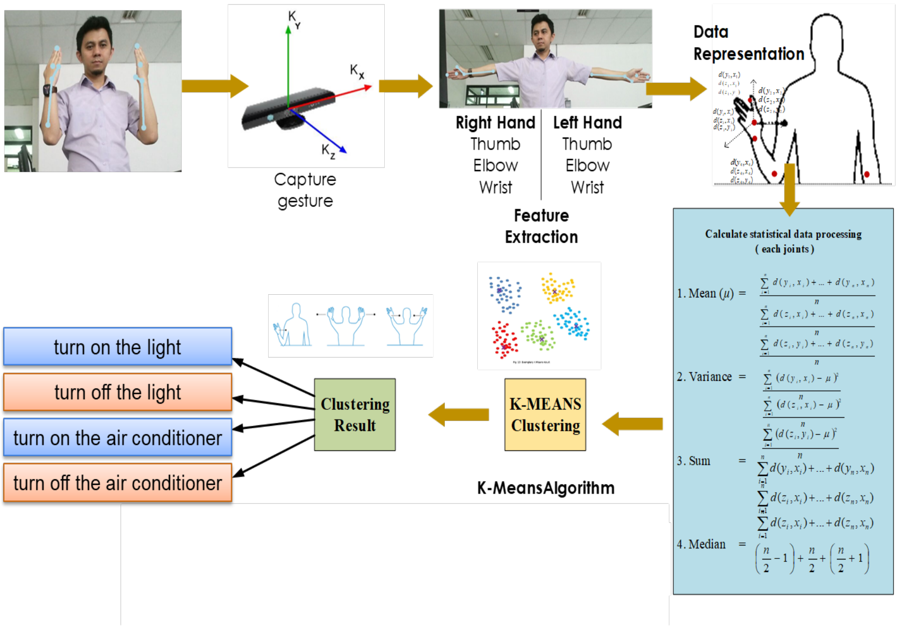

$K_Y$
$K_X$
$K_Z$
Capture gesture
Right Hand
Thumb
Elbow
Wrist
Left Hand
Thumb
Elbow
Wrist
Feature Extraction
Data Representation
Calculate statistical data processing ( each joints )
1. Mean ($\mu$) = $\frac{\sum_{i=1}^{n} d(y_i,x_i)+...+d(y_n,x_n)}{n}$ $\frac{\sum_{i=1}^{n} d(z_i,x_i)+...+d(z_n,x_n)}{n}$ $\frac{\sum_{i=1}^{n} d(z_i,y_i)+...+d(z_n,y_n)}{n}$
2. Variance = $\frac{\sum_{i=1}^{n} (d(y_i,x_i)-\mu)^2}{n}$ $\frac{\sum_{i=1}^{n} (d(z_i,x_i)-\mu)^2}{n}$ $\frac{\sum_{i=1}^{n} (d(z_i,y_i)-\mu)^2}{n}$
3. Sum = $\sum_{i=1}^{n} d(y_i,x_i)+...+d(y_n,x_n)$ $\sum_{i=1}^{n} d(z_i,x_i)+...+d(z_n,x_n)$ $\sum_{i=1}^{n} d(z_i,x_i)+...+d(z_n,x_n)$
4. Median = $\left(\frac{n}{2}-1\right)+\frac{n}{2}+\left(\frac{n}{2}+1\right)$
K-MEANS Clustering
K-MeansAlgorithm
Clustering Result
turn on the light
turn off the light
turn on the air conditioner
turn off the air conditioner

**Fig. 6.** Proposed gesture recognition system using K-Means Clustering

A cluster considers both single point clusters and centroids or means. The centroid will represent all points in the cluster if the data points cluster around the centroid. The spread size standard of points group average is the variance or the sum of distance squares between each point and average. If the data point closes to the average, the variance will be small. Variance generalization, where the centroid is replaced by a reference point that may or may not be a centroid, is used in cluster analysis to show the overall partition quality [27].

Skeletal joint position values for each axis (x,y,z) result of Microsoft SDK which comes with hand utility, named Coordinate-Mapper. Coordinate-Mapper is used to identify whether a point forms the 3D space. As input data, we find the difference of each axis i.e.:

$$d(y,x) = x - y \tag{1}$$

$$d(z,x) = x - z \tag{2}$$

$$d(z,y) = y - x \tag{3}$$

The formulas (1), (2), (3) are used to find the value of the distance between the axis so that no value is equal. Then, the value is calculated to statistical data, namely the average value, variance, sum, and median. This applies to any skeletal joint of the right hand and left hand, like thumb, wrist, and elbow as shown in Figure 7.

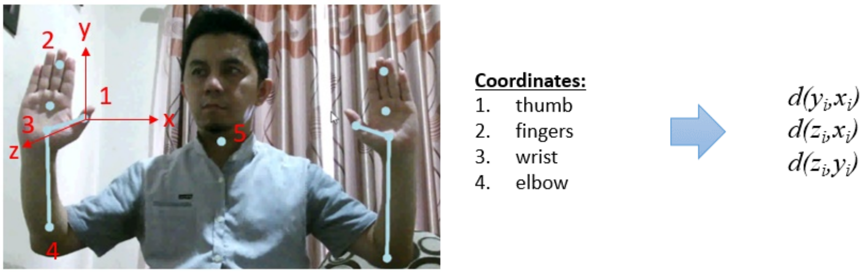


**Fig. 7.** Skeletal joint coordinates: thumb, fingers, wrist, and elbow

After gets the distance value of each axis, then calculated the value of mean, variance, sum, and median from the statistical data obtained [28].

1. Mean (μ)

$$\frac{\sum_{i=1}^{n} d(y_i, x_i) + \dots + d(y_n, x_n)}{n} \tag{4}$$

$$\frac{\sum_{i=1}^{n} d(z_i, x_i) + \dots + d(z_n, x_n)}{n} \tag{5}$$

$$\frac{\sum_{i=1}^{n} d(z_i, y_i) + \dots + d(z_n, y_n)}{n} \tag{6}$$

2. Variance

$$\frac{\sum_{i=1}^{n} \left(d(y_i, x_i) - \mu\right)^2}{n} \tag{7}$$

$$\frac{\sum_{i=1}^{n} \left(d(z_i, x_i) - \mu\right)^2}{n} \tag{8}$$

$$\frac{\sum_{i=1}^{n} \left(d\left(z_i, y_i\right) - \mu\right)^2}{n} \tag{9}$$

3. Sum

$$\sum_{i=1}^{n} d(y_i, x_i) + ... + d(y_n, x_n) \tag{10}$$

$$\sum_{i=1}^{n} d(z_i, x_i) + ... + d(z_n, x_n) \tag{11}$$

$$\sum_{i=1}^{n} d(z_i, y_i) + \cdots + d(z_n, y_n) \tag{12}$$

4. Median

$$\left(\frac{n}{2} - 1\right) + \frac{n}{2} + \left(\frac{n}{2} + 1\right) \tag{13}$$

K-means unsupervised learning method uses to recognize hand gestures. It needs four motion data for four gesture commands. Each movement has four joint axes (x-y-z) for the right hand and four joint axes (x-y-z) for the left hand. Each joint has four features (mean, variant, sum, and median). Then get the matrix of 12 columns and 40 rows which will be input for the clustering process.

The process of sorting is by entering statistical data on each axis for the first movement until the fourth movement. Each movement sampled ten times to get raw data clustering. K-Means clustering applied to the hand gesture recognition into four clusters based on the axes (x-y-z) obtained from the Kinect hand joint. K-Means clustering minimizes the within-cluster sum of squares [29].

$$\underset{c}{\arg min} \sum_{i=1}^{k} \sum_{p_j \epsilon c_i} \| \; p_j - \mu_i \|^2 \tag{14}$$

This gesture recognition system is running in real-time which is using Microsoft Kinect v2 and K-Means clustering is applied whenever there is an input data source, and the value of K is 4. The centroid point initialization is 0 and the iteration is 1000. The selection of centroid values has an effect on the outcome of clustering. If initial centroid initialization values are different, it will give different results. The next step is to compute the range of each data to the nearest centroid. The distance of each data to the nearest centroid in two dimensions defined as [30]:

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_1 - p_1)^2} \quad (15)$$

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (16)$$

Where $q$ is the input of the source data and $p$ is the centroid point. The range between the input source data and the centroid value of each cluster is computed and the result is assigned to the nearest cluster to obtain new cluster means. Then calculated and changed again the value of each cluster centroid by means of the cluster member until the members of each cluster do not change or convergent, then the step stopped and obtained clustering results. Figure 8 shows flowcharts of gesture recognition using K-Means clustering for home appliances control use Kinect v2.

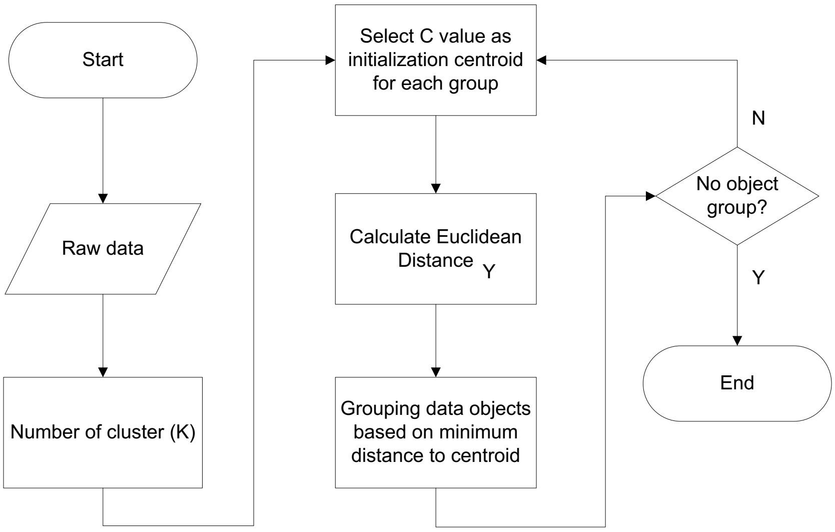

Start
Raw data
Number of cluster (K)
Select C value as initialization centroid for each group
Calculate Euclidean Distance
Y
Grouping data objects based on minimum distance to centroid
N
No object group?
Y
End

**Fig. 8.** Algorithm of K-Means clustering

# 4 Results and Discussion

## 4.1 Overview the system

Figure 9 shows the structure of integrating software and hardware. The design of a hardware system to control home appliances use a personal computer (PC) as a server and Arduino esp8266 (NodeMCU) as microcontroller. Commands from speech and gesture recognition results are sent from PC to Firebase use Wi-Fi. Action result from Firebase send to Arduino esp8266 use Wi-wf a78ijnnnnnnnFi to control home appliances.

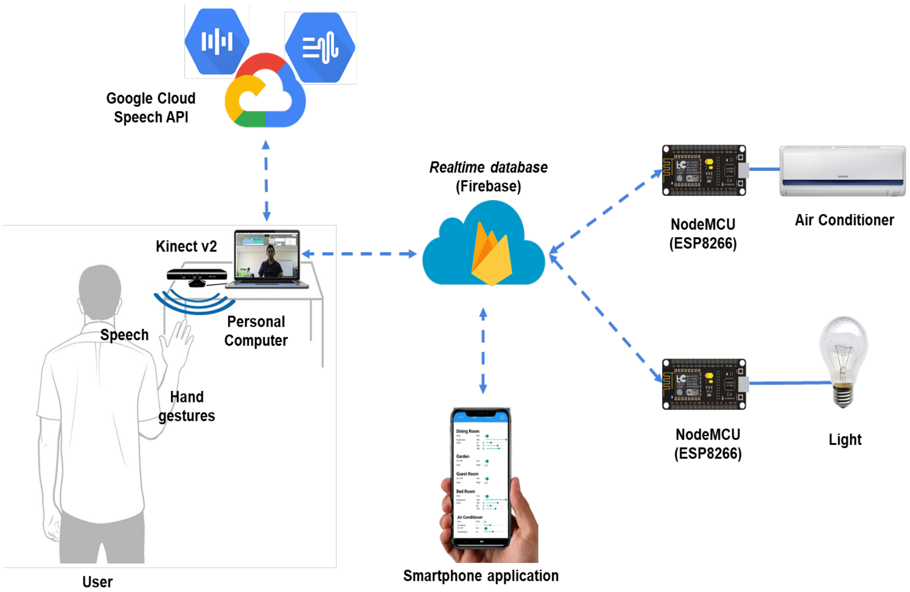


**Fig. 9.** Multimodal Interaction System Design

Figures 10, 11(a), and 11(b) show the Kinect v2.0 connects to the personal computer (unit processing), the Esp8266 connects to the servo motor to move the light switch to the ON or OFF position, and the Esp8266 connects to an infrared sensor to control the air conditioner, respectively.

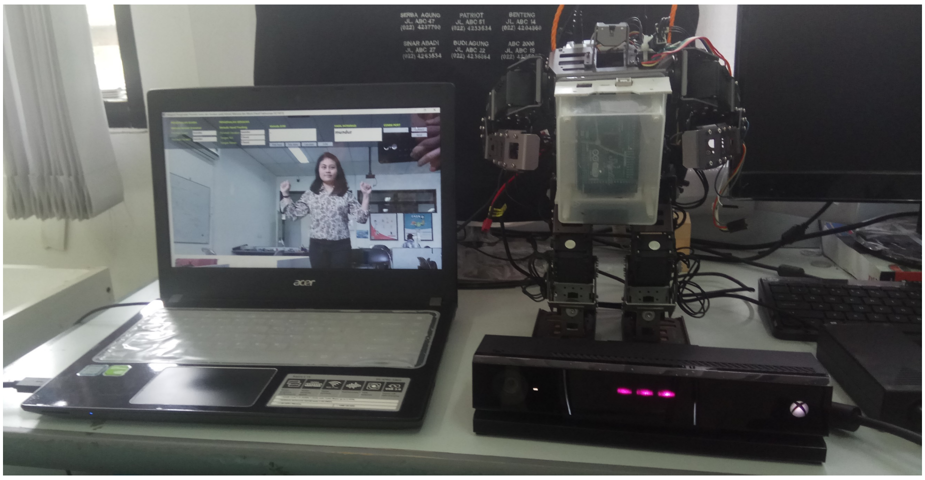
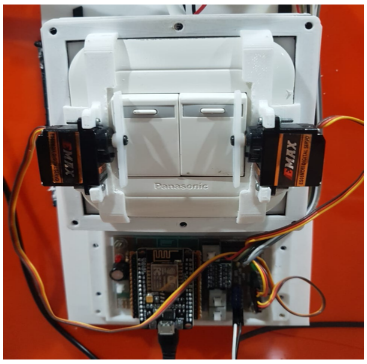
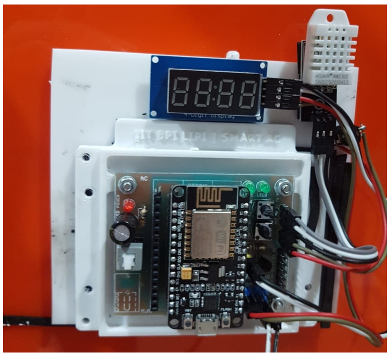


**Fig. 10.** The Kinect v2.0 connects to the personal computer (unit processing)





a) The Esp8266 connects to the servo motor to move the light switch

b) The Esp8266 connects to an infrared sensor to control the air conditioner(b)

**Fig. 11.**

### 4.2 Smartphone application

In this research, users can also use the smartphone application to remotely control home appliances through mobile devices such as tablets or smartphones. Figure 12 shows the smartphone application flowchart.

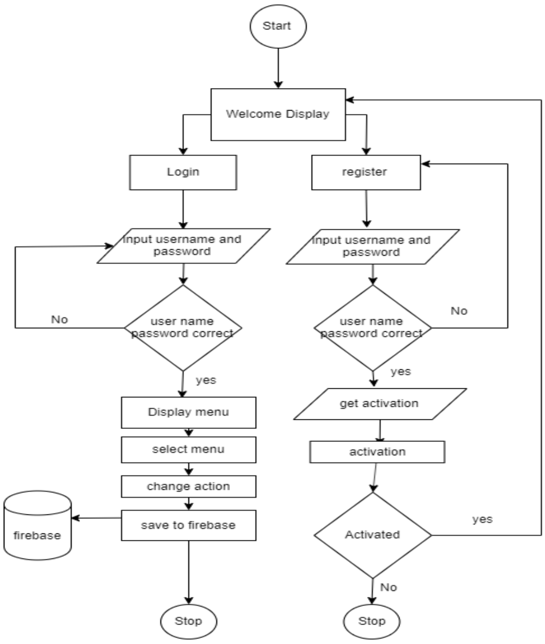
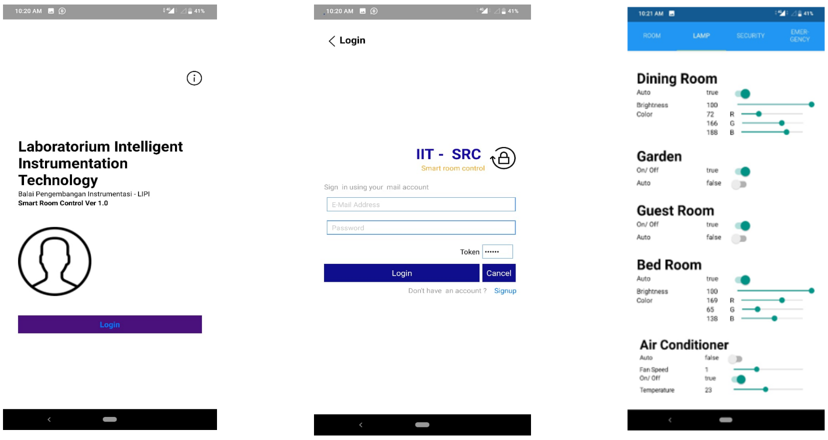


**Fig. 12.** The smartphone application flowchart

The smartphone application develops to control home appliances that have been connected to the NodeMCU (Esp8266) as a microcontroller. The application connects directly to the real-time database (firebase) to change the status of home appliances such as on/off or up/down. Figure 13 displays the features of the application.



Dining Room
Auto true
Brightness 100
Color 72 R
166 G
188 B

Garden
On/ Off true
Auto false

Guest Room
On/ Off true
Auto false

Bed Room
Auto true
Brightness 100
Color 169 R
65 G
138 B

Air Conditioner
Auto false
Fan Speed 1
On/ Off true
Temperature 23

(a) (b) (c)

**Fig. 13.** Smartphone application for controlling home appliances.

### 4.3 Testing and implementation

The test carried out in three stages. Firstly, testing the speech recognition system and dialogue system, secondly testing the gesture recognition system, and thirdly testing the smartphone Application. The testing environment is in the room which allows for the smallest possible noise disturbances. The intensity of the light in the room matches the general conditions of the room during the day, which is between 300 to 400 lux. The location of the Kinect sensor in the static field with user distance remains about 150 centimeters. Multimodal interaction system tested by twenty people from different gender, ages ranges, and dialects because in Indonesia it consists of many cultural tribes with different dialects. The profile of the respondent shows in Figure 14-16.

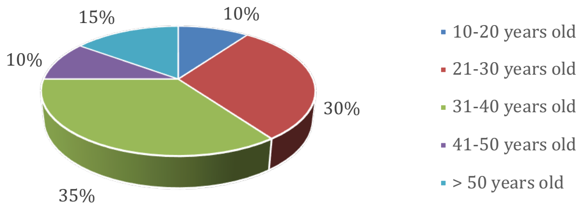


**Fig. 14.**The respondent's profile of ages ranges

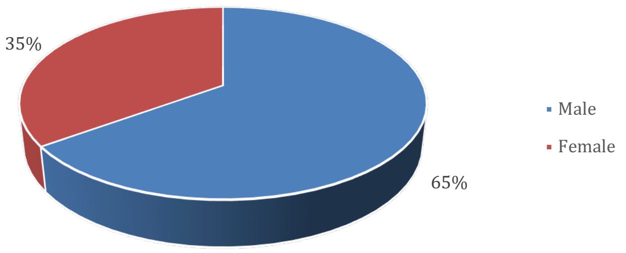


**Fig. 15.** The respondent's profile of gender

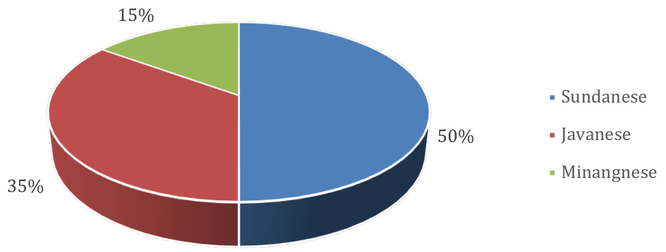


**Fig. 16.** The respondent's profile of dialect

The dialogue system test for 12 dialogue scenarios in Indonesian. Each dialogue test 10 times for 12 dialogues sentences. Everyone who tests says 120 dialogues. Table 2 shows the testing scenarios between user and SITI. SITI is the name of the smart home system.

**Table 2.** Scenario dialog of testing

| **Dialogue Number.** | **User (*Indonesian*/English)** | **System (*Indonesian*/English)** |
|---|---|---|
| D01 | *Halo* SITI. | *Halo. Ada yang dapat saya bantu?* |
| | Hello SITI. | Hello. What do you want me to do? |
| D02 | *Aduh, cuaca hari ini terasa GERAH banget nih* | *Rasa gerah belum terdaftar. Apa yang anda inginkan?* |
| | Oops, the weather is hot and sultry day | Sultry is not yet registered. What do you want? |
| D03 | *Tolong nyalakan AC dong* | *Dengan senang hati, AC akan saya dihidupkan* |
| | Please turn on the air conditioner. | With pleasure, the air conditioner will turn on |
| D04 | *Saya masih kepanasan nih* | *Apakah Anda ingin suhu AC diturunkan?* |
| | I'm still hot | Do you want the temperature of the air conditioner to be lowered? |

| | | |
|---|---|---|
| D05 | *Turunkan Suhu AC dua derajat* | *Baik, suhu AC akan segera diturunkan dua derajat* |
| | Lower the AC temperature by two degrees | Ok, the temperature of the air conditioner will be reduced by two degrees soon |
| D06 | *Nyalakan kulkas dong* | *Maaf device kulkas belum terdaftar, apakah Anda ingin mendaftarkannya?* |
| | Turn on the fridge, please | Sorry the fridge device is not registered yet; do you want to register it? |
| D07 | *Daftarkan* | *Baik, device kulkas didaftarkan ke dalam system* |
| | Register it. | Okay, the fridge is registered in the system. |
| D08 | *Nyalakan kulkas lagi dong* | *Maaf, device kulkas belum terinstalasi secara fisik* |
| | Turn on the fridge again. | Sorry, the fridge has not been physically installed |
| D09 | *Ruangan kok gelap banget ya* | *Apakah Anda ingin lampu dihidupkan* |
| | The room is dark. | Do you want the lights on? |
| D10 | *iya, nyalakan lampu* | *Lampu di ruang mana yang ingin dinyalakan?* |
| | yes, turn on the lights | Which lamp do you want to light on? |
| D11 | *Semua lampu, tolong nyalakan* | *Siap, semua lampu akan dinyalakan.* |
| | All lights please turn on | Ok, all lights will be turned on |
| D12 | *Terima kasih* | *Asyiap sama-sama* |
| | Thank you. | Ok your welcome |

Figure 17 shows the accuracy level from testing result. The average level of accuracy testing of human and machine interaction dialogue systems from a total of 2,400 times of testing (20 respondents x 12 dialogues x 10 times) is 92.5%.  The highest level of accuracy occurs in the word's "*daftarkan*" (D07) which is 96.5%. All test samples are clear enough to say "*daftarkan*" so it can be easily translated by speech recognition systems. The lowest level of accuracy occurs in the dialogue "Halo SITI" which is 87.5%. The pronunciation of "SITI" is often translated by the speech recognition system with the word "Kitty".

The gender and dialect of the respondents did not affect the accuracy of the dialogue system significantly. However, the pronunciation of respondents more than 50 years old affects the accuracy of the dialogue system significantly. From respondents more than 50 years old, as many as 3 people, have an average accuracy rate of 81.7% or 10.83% lower than the average level of accuracy of the entire test sample, which is 92.5%.

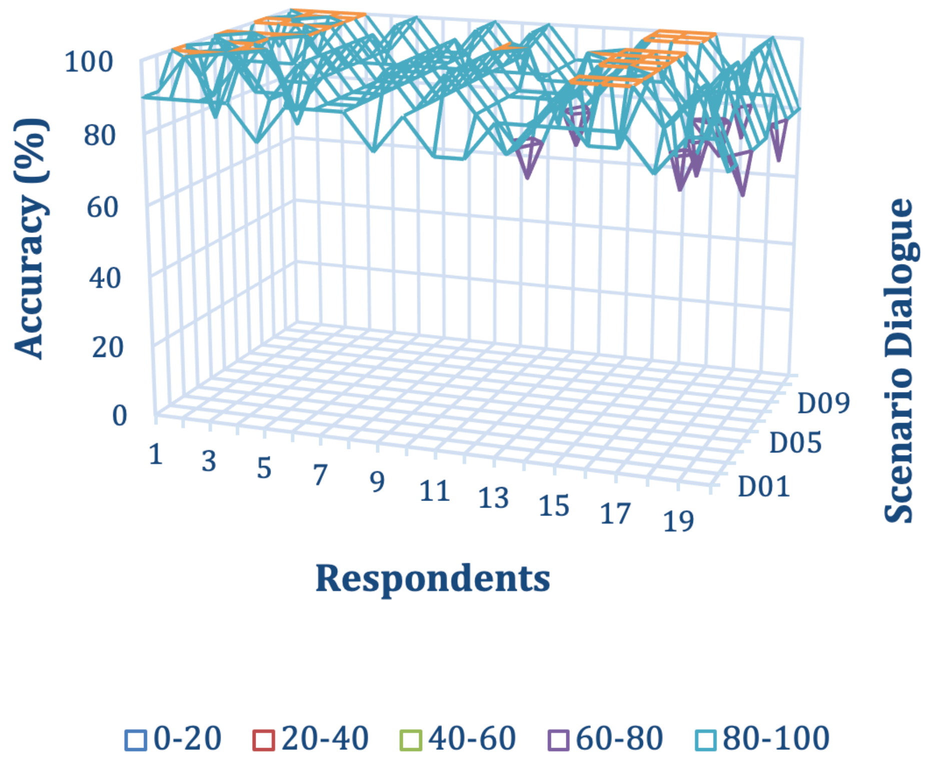
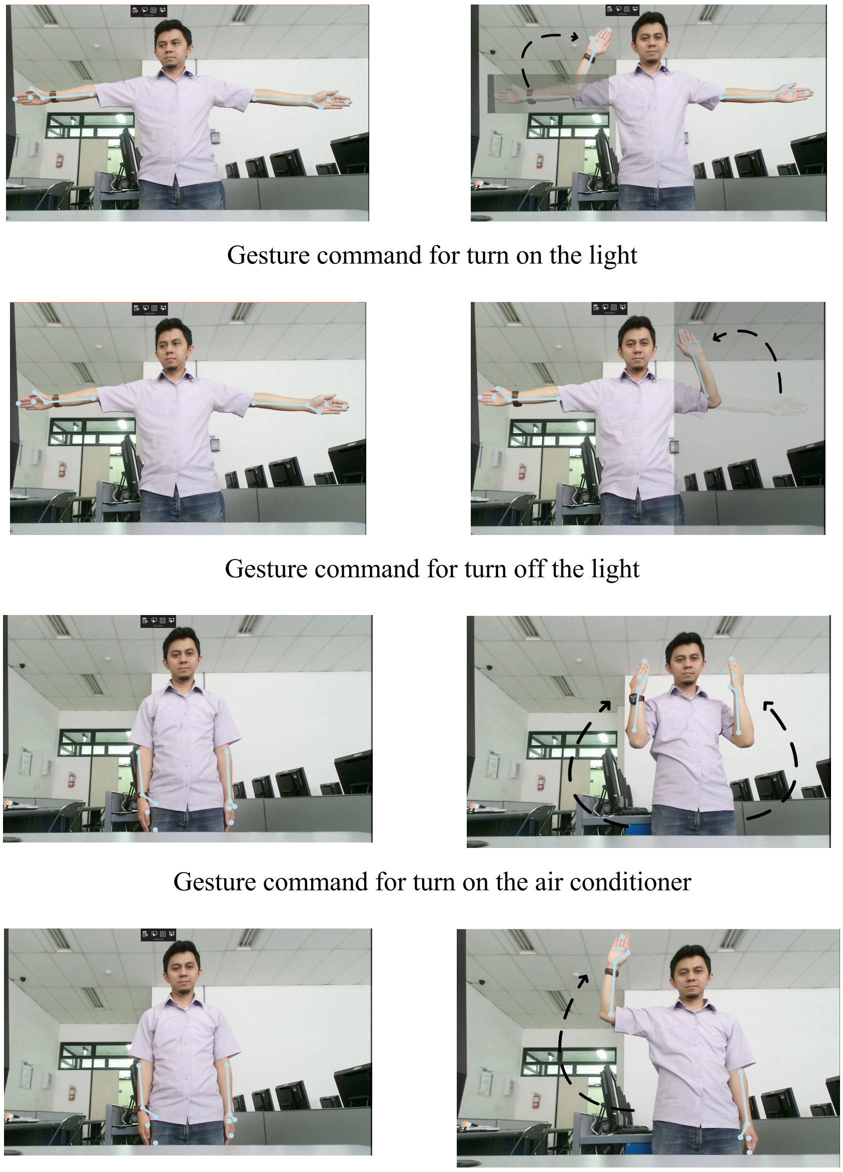


**Fig. 17.** The accuracy level from dialogue system testing result

Figure 18 shows a gesture recognition dataset using K-Means Clustering for command: turn on the light, turn off the light, turn on the air conditioner, and turn off the air conditioner.

Each respondent performs every gesture command 10 times to control home appliances such as turn on the lights, turn off the lights, turn on the air conditioner, and turn off the air conditioner. The total test data is 800 times (20 test samples x 4 gesture commands x 10 times).

Figure 19 shows the accuracy level from the testing result for interaction using gestures. The average level of testing accuracy is 79,25%. The highest is 90% in the gesture command for "Turn on the air conditioner". The lowest is 70% in the gesture command for "Turn off the air conditioner". Respondents between 10 and 20 years old who were shorter than the other respondents, as many as two people, had an average accuracy rate of 65% or 14.25% lower than the average level of accuracy of the entire test sample, which is 79.25%. This is due to the training data performed on adults with an age range between 31-40 years old.



Gesture command for turn on the light

Gesture command for turn off the light

Gesture command for turn on the air conditioner

Gesture command for turn off the air conditioner

**Fig. 18.** Gesture recognition dataset using K-Means.

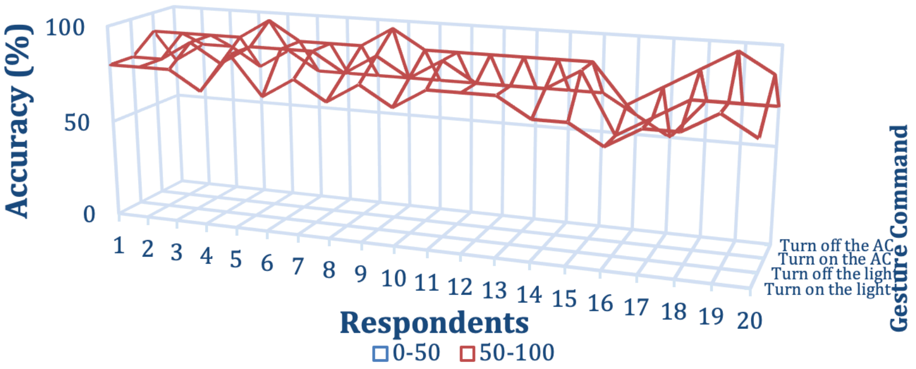


**Fig. 19.** The accuracy level from gesture recognition testing result

Black box testing uses to test smartphone applications to control home appliances properly [31]. The result shows that every button runs appropriately as displayed in Table 3.

**Table 3.** Black box testing results

| No. | Page | Testing Button | Status |
|---|---|---|---|
| 1. | Welcome display | | ( √ ) Succeed ( ) Failed |
| 2. | Register for user | | ( √ ) Succeed ( ) Failed |
| 3. | Login | | ( √ ) Succeed ( ) Failed |
| 4. | Input username and password | | ( √ ) Succeed ( ) Failed |
| 5. | Display menu | | ( √ ) Succeed ( ) Failed |
| 6. | Select menu | | ( √ ) Succeed ( ) Failed |
| 7. | Change action | | |
| | Lamp in dining Room | Auto | ( √ ) Succeed ( ) Failed |
| | | Brightness | ( √ ) Succeed ( ) Failed |
| | | Color | ( √ ) Succeed ( ) Failed |
| | Lamp in garden | On/Off | ( √ ) Succeed ( ) Failed |
| | | Auto | ( √ ) Succeed ( ) Failed |
| | Lamp in guest Room | On/Off | ( √ ) Succeed ( ) Failed |
| | | Auto | ( √ ) Succeed ( ) Failed |
| | Lamp in bed | Auto | ( √ ) Succeed ( ) Failed |
| | | Brightness | ( √ ) Succeed ( ) Failed |
| | | Color | ( √ ) Succeed ( ) Failed |
| | Air Conditioner | Auto | ( √ ) Succeed ( ) Failed |
| | | Fan Speed | ( √ ) Succeed ( ) Failed |
| | | On/Off | ( √ ) Succeed ( ) Failed |
| | | Temperature | ( √ ) Succeed ( ) Failed |
| 8. | Save to firebase | | ( √ ) Succeed ( ) Failed |
| 10. | Quit | | ( √ ) Succeed ( ) Failed |

### 4.4 Comparison with previous work

Table 4 shows the comparison between the previous works of the interaction system for home appliances control and this experiment.

**Table 4.** Comparison with previous work.

| *References* | *Comparison* | | | |
|---|---|---|---|---|
| | *Modalities (Smartphone application/ Speech/Gesture)* | *Interaction with machine (Yes/No)* | *Number of Sensor* | *Natural Interaction System (Poor/Fair/ Good/Excellent)* |
| Meja, et al [4] | 1 | No | 1 | Poor |
| Pavithra and Balakrishnan [8] | 1 | No | 4 | Poor |
| Baig, et al [9] | 2 | No | 1 | Fair |
| Kamarudin, et al [10] | 2 | No | 1 | Fair |
| Jadhav, et al [11] | 1 | No | 1 | Poor |
| Anbarasan and Lee [12] | 2 | No | 2 | Fair |
| This Work | 3 | Yes | 1 | Good |

## 5 Conclusion

The designs and steps that have been developed, expected to be a new proposal, to control home appliances using multimodal interaction such as speech, gesture, and smartphone application. The average level of accuracy testing of interaction using dialogue systems and gesture are 92.5% and 79,25%. Interaction using dialogue systems is better than gesture. Smartphone applications can control home appliances properly. In the future, face detection and skeleton tracking can be added to multimodal interaction so human and machine interactions can run more naturally.

## 6 Acknowledgement

The first author acknowledges support from the Lembaga Pengelola Dana Pendidikan (Indonesia Endowment Fund for Education) scholarship, Ministry of Finance, The Republic of Indonesia.

## 7 References

[1] H. Fakhrurroja, Riyanto, A. Purwarianti, A. S. Prihatmanto, and C. Machbub, ‘Integration of Indonesian Speech and Hand Gesture Recognition for Controlling Humanoid Robot’, in 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV), Singapore, Nov. 2018, pp. 1590–1595, https://doi.org/10.1109/icarcv.2018.8581071.

[2] N. A. Prasetyo, A. G. Prabawati, and S. Suyoto, ‘Smart Home: Power Electric Monitoring and Control in Indonesia’, Int. J. Interact. Mob. Technol., vol. 13, no. 03, p. 143, Mar. 2019, https://doi.org/10.3991/ijim.v13i03.10070.

[3] H. Fakhrurroja, A. Abdillah, U. Nadiya, and M. Arifin, 'Hand State Combination as Gesture Recognition using Kinect v2 Sensor for Smart Home Control Systems', in 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), 2019, pp. 74–78. https://doi.org/10.1109/iotais47347.2019.8980390.

[4] R. Miramontes Meza, L. V. Escamilla del Río, and R. T. Aquino Santos, 'Mobile Remote Control for Home Automation', Int. J. Interact. Mob. Technol., vol. 7, no. 4, p. 21, Oct. 2013, https://doi.org/10.3991/ijim.v7i4.3178.

[5] N. T. Wei, A. S. Baharudin, L. A. Hussein, and M. F. Hilmi, 'Factors Affecting User's Intention to Adopt Smart Home in Malaysia', Int. J. Interact. Mob. Technol., vol. 13, no. 12, p. 39, Dec. 2019, https://doi.org/10.3991/ijim.v13i12.11083.

[6] S. Mallios and N. Bourbakis, 'A survey on human machine dialogue systems', in 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA), Chalkidiki, Greece, Jul. 2016, pp. 1–7, https://doi.org/10.1109/iisa.2016.7785371.

[7] L. Meng and M. Huang, 'Dialogue Intent Classification with Long Short-Term Memory Networks', in Natural Language Processing and Chinese Computing, vol. 10619, X. Huang, J. Jiang, D. Zhao, Y. Feng, and Y. Hong, Eds. Cham: Springer International Publishing, 2018, pp. 42–50. https://doi.org/10.1007/978-3-319-73618-1_4.

[8] D. Pavithra and R. Balakrishnan, 'IoT based monitoring and control system for home automation', in 2015 Global Conference on Communication Technologies (GCCT), Thuckalay, Kanya kumari district, India, Apr.2015, pp. 169–173, https://doi.org/10.1109/gcct.2015.7342646.

[9] F. Baig, S. Beg, and M. F. Khan, 'Zigbee Based Home Appliances Controlling Through Spoken Commands Using Handheld Devices', International Journal of Smart Home, vol. 7, no. 1, p. 8, 2013.

[10] R. Kamarudin and A. F. Yusof, 'Low Cost Smart Home Automation via Microsoft Speech Recognition', International Journal of Engineering, vol. 13, no. 03, pp. 6–11, 2013.

[11] J. Jadhav and P. Avhad, 'Hand Gesture Based Home Appliances Control System', International Research Journal of Engineering and Technology (IRJET), vol. 04, no. 05, pp. 2553–2555, May 2017.

[12] Anbarasan and J. S. A. Lee, 'Speech and Gestures for Smart-Home Control and Interaction for Older Adults', in Proceedings of the 3rd International Workshop on Multimedia for Personal Health and Health Care - HealthMedia'18, Seoul, Republic of Korea, 2018, pp. 49–57, https://doi.org/10.1145/3264996.3265002.

[13] Z. Zhang, 'Microsoft Kinect Sensor and Its Effect', IEEE MultiMedia, vol. 19, no. 2, pp. 4–10, Feb. 2012, https://doi.org/10.1109/mmul.2012.24.

[14] P. Fankhauser, M. Bloesch, D. Rodriguez, R. Kaestner, M. Hutter, and R. Siegwart, 'Kinect v2 for mobile robot navigation: Evaluation and modeling', in 2015 International Conference on Advanced Robotics (ICAR), Istanbul, Turkey, Jul. 2015, pp. 388–394, https://doi.org/10.1109/icar.2015.7251485

[15] Xbox One Kinect Teardown', iFixit, Nov. 21, 2013. https://www.ifixit.com/Teardown/Xbox+One+Kinect+Teardown/19725 (accessed Jan. 20, 2020). https://doi.org/10.1049/et.2011.0320

[16] M. Rahman, Beginning Microsoft Kinect for Windows SDK 2.0: Motion and Depth Sensing for Natural User Interfaces. Apress, 2017. https://doi.org/10.1007/978-1-4842-2316-1

[17] P. Lange and D. Suendermann-Oeft, 'Tuning Sphinx to outperform Google's speech recognition API', Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2014, pp. 32–41.

[18] D. Petcu and C. Ciprian, 'Towards a cross platform cloud API', in Proceedings of the 1st International Conference on Cloud Computing and Services Science, Noordwijkerhout, Netherlands, 2011, pp. 166–169. https://doi.org/10.5220/0003388101660169.

[19] 'Cloud Speech-to-Text - Speech Recognition Cloud Speech-to-Text Google Cloud'. https://cloud.google.com/speech-to-text/.https://doi.org/10.5626/ktcp.2019.25.3.191.

[20] M. Assefi, G. Liu, M. P. Wittie, and C. Izurieta, 'An Experimental Evaluation of Apple Siri and Google Speech Recognition', in Proccedings of the 2015 ISCA SEDE, pp. 1–6.

[21] R. Sarikaya, G. E. Hinton, and A. Deoras, 'Application of Deep Belief Networks for Natural Language Understanding', IEEE/ACM Trans. Audio Speech Lang. Process., vol. 22, no. 4, pp. 778–784, Apr. 2014, https://doi.org/10.1109/taslp.2014.2303296.

[22] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, 'Stemming Indonesian: A confix-stripping approach', TALIP, vol. 6, no. 4, pp. 1–33, Dec. 2007, https://doi.org/10.1145/1316457.1316459.

[23] X. Li, Y.-N. Chen, L. Li, J. Gao, and A. Celikyilmaz, 'End-to-end task-completion neural dialogue systems', arXiv preprint arXiv:1703.01008, 2017.

[24] H. Gomaa, Real-Time Software Design for Embedded Systems. Cambridge University Press, 2016.

[25] S. Yi and K. Jung, 'A chatbot by combining finite state machine, information retrieval, and bot-initiative strategy', Proc. Alexa Prize, 2017.

[26] H. Purnawansyah, 'K-Means clustering implementation in network traffic activities', presented at the 2016 International Conference on Computational Intelligence and Cybernetics, Makassar, Indonesia. IEEE, 2016.https://doi.org/10.1109/cyberneticscom.2016.7892566.

[27] V. Faber, 'Clustering and the continuous k-means algorithm', Los Alamos Science, vol. 22, no. 138144.21, p. 67, 1994.

[28] D. A. Maharani, H. Fakhrurroja, Riyanto, and C. Machbub, 'Hand gesture recognition using K-means clustering and Support Vector Machine', in 2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, Apr. 2018, pp. 1–6, https://doi.org/10.1109/iscaie.2018.8405435.

[29] A. Gaur and S. Yadav, 'Handwritten Hindi character recognition using k-means clustering and SVM', 2015, pp. 65–70.

[30] Y. Li, 'Hand gesture recognition using Kinect', 2012, pp. 196–199.

[31] S. Nidhra, 'Black Box and White Box Testing Techniques - A Literature Review', International Journal of Embedded Systems and Applications (IJESA), vol. 2, no. 2, pp. 29–50, Jun. 2012, https://doi.org/10.5121/ijesa.2012.2204.

## 8 Authors

**Hanif Fakhrurroja** received the bachelor's degree in Physics from the Universitas Padjadjaran (Unpad) in 2003 and the master's degree in Informatics from Institut Teknologi Bandung (ITB) in 2010. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Informatics, Institut Teknologi Bandung (ITB). Since 2006, he has been with the Indonesian Institute of Sciences as a Researcher. His research interests include Human-Machine Interaction and Intelligent Instrumentation Technology. Email: hani002@lipi.go.id

**Carmadi Machbub** is received the bachelor's degree in electrical engineering from the Institut Teknologi Bandung (ITB), in 1980, and the master's degree (DEA) in control engineering and industrial informatics and the Ph.D. degree in engineering sciences majoring in control engineering and industrial informatics from Ecole Centrale de Nantes, in 1988 and 1991, respectively. He is currently a Professor and the Head of the Control and Computer Systems Research Division, School of Electrical Engineering and Informatics, ITB. His current research interests include machine perception, optimization, and control.

**Ary Setijadi Prihatmanto** graduated with B.E. and M.S. in Electrical Engineering at Institut Teknologi Bandung in 1995 and 1998, and received his PhD in Applied Informatics from Johannes Kepler University of Linz, Austria in 2006. He is an associate professor & lecturer of School of Electrical Engineering & Informatics, Institut Teknologi Bandung since 1997. He is also the president of Indonesia Digital Media Forum since 2009. His main interests are Human-Content Interaction, Computer Graphics & Mixed-Reality Application, Machine Learning & Intelligent System, Intelligent Robotics, and Cyber-Physical System.

**Ayu Purwarianti** is a lecturer at Institut Teknologi Bandung, Indonesia since 2008. She graduated from Institut Teknologi Bandung for her undergraduate and master degree. She received his doctoral degree at Toyohashi University of Technology in December 2007 with research topics of cross language question answering. She has interest in computational linguistics, especially for Indonesian language. She has written several publications in conferences and journals related with computational linguistics for Indonesian language. She also provides Indonesian natural language processing tools to be used by other researchers.

Article submitted 2020-02-04. Resubmitted 2020-07-01. Final acceptance 2020-07-08. Final version published as submitted by the authors.