

# Multimodal Fusion Algorithm and Reinforcement Learning-Based Dialog System in Human-Machine Interaction

Hanif Fakhurroja<sup>1</sup>, Carmadi Machbub<sup>2\*</sup>, Ary Setijadi Prihatmanto<sup>3</sup> and Ayu Purwarianti<sup>4</sup>

Institut Teknologi Bandung, School of Electrical Engineering and Informatics, Indonesia

<sup>1</sup>hani002@lipi.go.id, <sup>2</sup>carmadi@lskk.ee.itb.ac.id, <sup>3</sup>asetijadi@lskk.ee.itb.ac.id,

<sup>4</sup>ayu@stei.itb.ac.id

\*Corresponding Author: carmadi@lskk.ee.itb.ac.id

**Abstract:** Studies on human-machine interaction system show positive results on system development accuracy. However, there are problems, especially using certain input modalities such as speech, gesture, face detection, and skeleton tracking. These problems include how to design an interface system for a machine to contextualize the existing conversations. Other problems include activating the system using various modalities, right multimodal fusion methods, machine understanding of human intentions, and methods for developing knowledge. This study developed a method of human-machine interaction system. It involved several stages, including a multimodal activation system, methods for recognizing speech modalities, gestures, face detection and skeleton tracking, multimodal fusion strategies, understanding human intent and Indonesian dialogue systems, as well as machine knowledge development methods and the right response. The research contributes to an easier and more natural human-machine interaction system using multimodal fusion-based systems. The average accuracy rate of multimodal activation, testing dialogue system using Indonesian, gesture recognition interaction, and multimodal fusion is 87.42%, 92.11%, 93.54% and 93%, respectively. The level of user satisfaction towards the multimodal recognition-based human-machine interaction system developed was 95%. According to 76.2% of users, this interaction system was natural, while 79.4% agreed that the machine responded well to their wishes.

**Keywords:** multimodal fusion; Indonesian dialogue system; reinforcement learning; natural language understanding; human-machine interaction

## 1. Introduction

Humans have the desire to improve the quality of technology. For this reason, they often make machines to interact and help them in various tasks. Technology creates a machine that interprets information derived from speech and human gestures, act according to the information, and communicate [1]. The most common form of communication is the use of human language and gestures to convey messages. Recently, studies have focused on human-machine interaction, including how systems understand the detection and recognition of gestures automatically under natural environmental conditions [2].

Human-machine interaction has gradually changed from being originally computer-centered to human-centered. Since speech and gesture are natural, intuitive and precise methods of everyday human communication, they are the mainstream of human-machine interaction, especially in control systems [3], virtual reality [4], and medical diagnosis [5]. The gesture recognition is concerned with the interpretation of the human gestures involving the hands, arms, face, head, and body [6]. Speech recognition is the process of converting signals into word sequences using computer algorithms/programs [7].

Since the first appearance of the graphical user interface (GUI), studies have focused on the increasingly natural ways to interact with developed systems. Several studies examined human-machine interaction through speech [8] [9] or gesture recognition systems [10] [11]. Others focused on the multimodal aspect, including integration between speech and gesture recognition [12] [13] [14].

A natural user interface system is among the potential technologies that may change the outlook of computer science and industry in 2022. However, the system faces challenges, including multisensory input; predictive, anticipatory and adaptive; and contextual awareness, such as the ability of the system to capture multimodal input which is unlimited to speech, touch, and gesture; responding to users in the most appropriate way; and, seeing the conversation context [15].

Most existing literature studies on human-machine interaction systems still use unimodal fusion, such as only speech or gestures. Therefore, this study develops a multimodal fusion-based human-machine interaction system with four modality inputs: skeleton tracking, face detection, speech recognition, and gestures. The system is equipped with a dialogue system and machine knowledge.

This study has the following contributions:

- Algorithms for understanding conversation context. Activation of human-machine interaction system developed with four modality inputs in the form of skeleton tracking, face detection, speech recognition, and gesture in humans to understand the context of conversation around them. Therefore, machines distinguish whether humans are talking to other machines or with fellow humans.
- Multimodal fusion algorithm. Integration of four input modalities using multimodal fusion from the results of face detection, skeleton tracking, and speech and recognition. Therefore, the interaction system between humans and machines can be carried out naturally.
- Dialogue system algorithm and machine knowledge development. The developed dialogue system understands human intent. Machines can interact with humans and increase their knowledge when the system does not understand their intent.

## 2. Proposed system

The following are the input modalities in the human-machine interaction system developed in this study: (1) Face detection by recognizing the position when facing Kinect, eyeball position when looking at Kinect, and open mouth; (2) Skeleton tracking to count the number of people captured by Kinect; (3) Speech recognition using the Google Cloud Speech API which converts human speech to text; (4) Gesture recognition using support vector machine (SVM).

The four multimodal inputs are then processed for the following three conditions of human-machine interaction: (1) When Kinect captures one user, the input modalities used are face detection, skeleton tracking, speech recognition, and gesture recognition; (2) When not caught by Kinect, then the input modalities used are skeleton tracking and speech recognition; (3) When Kinect caught more than one person, the input modality used are face detection, skeleton tracking, and speech recognition; (4) When in a noisy room, the input modalities used are skeleton tracking and gesture recognition.

Figure 1 shows the rationale for this research. Figure 2 shows the system architecture built based on the rationale in Figure 1.

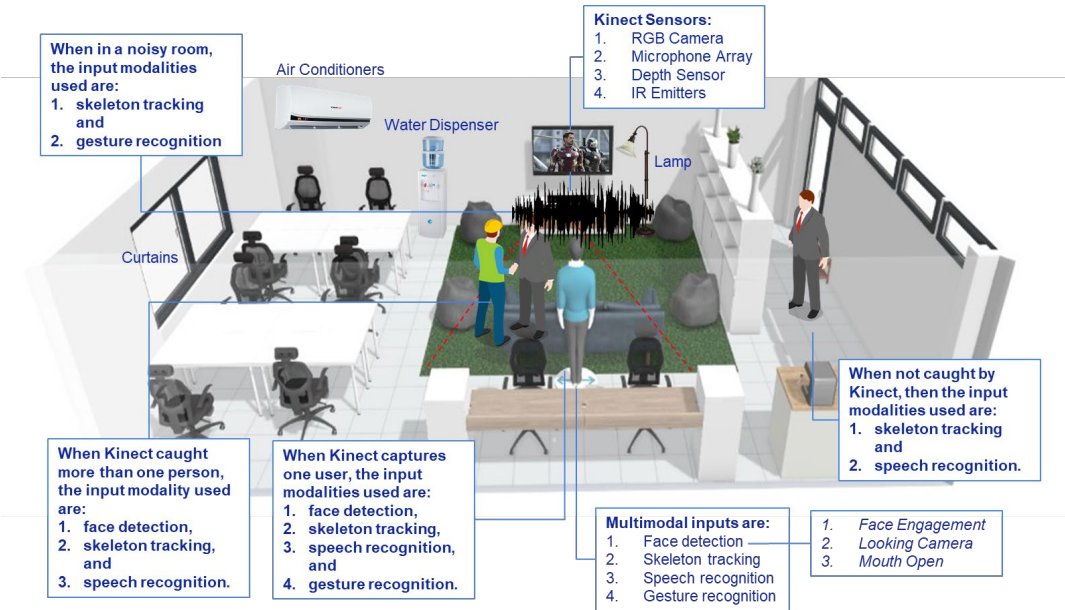


Figure 1. The rationale for the developed human-machine interaction system.

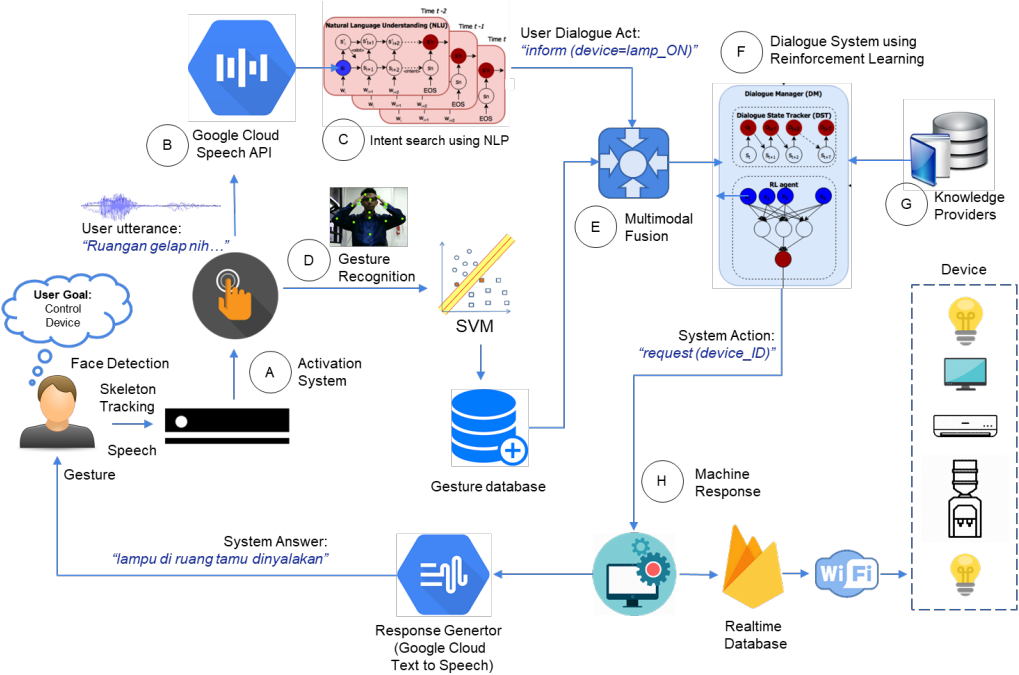


Figure 2. Design of multimodal fusion and dialog systems in human-machine interaction.

The multimodal recognition-based human-machine interaction system consists of eight modules, including *multimodal activation*, *Indonesian speech recognition*, *gesture recognition*, *intent classification*, *multimodal fusion*, *dialogue system*, *knowledge provider*, and *machine response*.

### A. Multimodal activation module

The multimodal activation feature allows humans to quickly activate the function of the human-machine interaction system using face detection, skeleton tracking, speech recognition and gesture. Figure 3 shows the activity diagram of the multimodal activation system.

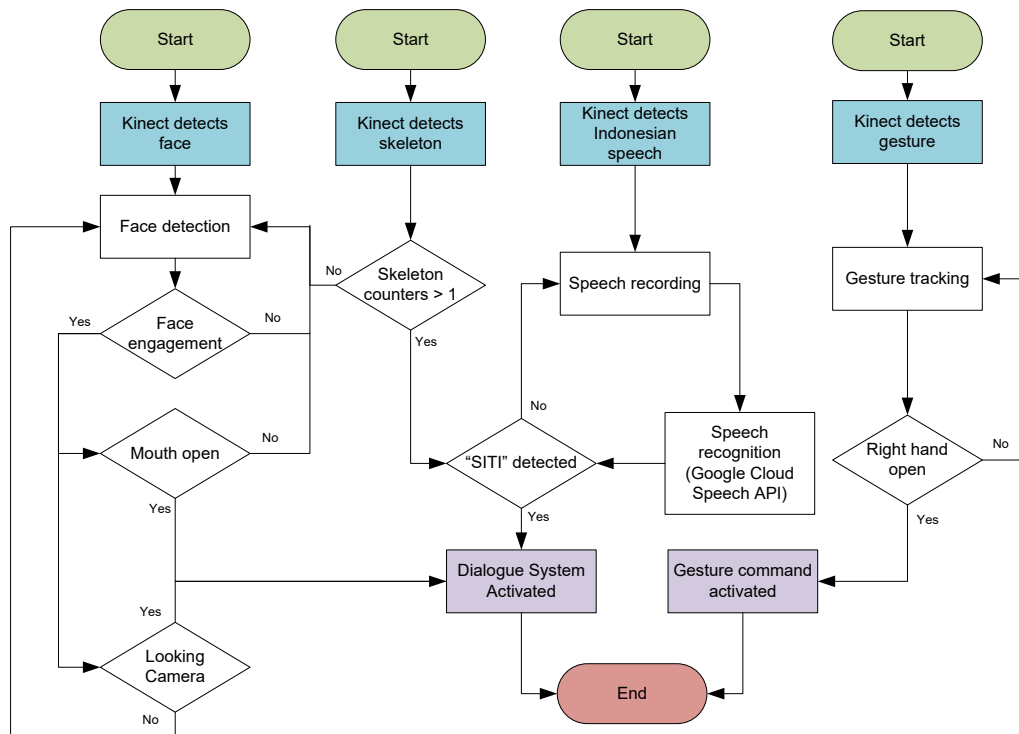


Figure 3. Flowchart of multimodal activation system.

The multimodal activation system can be carried out in the following.

#### 1). Activation Through Face Detection

In case only one person has detected a Kinect camera, face detection can be used. The voice-based interaction system is activated when the user's face is towards the Kinect camera, and the mouth is open. Also, the user's face can be fronting the Kinect camera while the eye looks at the camera.

Kinect is an active sensor for face detection and gesture tracking applications. This is because the Kinect camera has an integrated infrared sensor and captures streaming colour images with accurate data. The Kinect sensor receives three-dimensional data using colour camera components, infrared transmitters and receivers. The sensor is supported by a development kit of face tracking software [16].

Face detection on Kinect is carried out by analyzing the input image to calculate the head position and finding 121 face points, as shown in Figure 4(a). The Kinect SDK also measures the distance between the face point and the camera. It creates values used in the application at the same processing time.

The Face Tracking SDK uses an active appearance model for the two-dimensional tracker and data from the Kinect sensor and extended to three dimensions using depth value information [17]. The face detection is then performed on a three-dimensional Kinect coordinate system. Importantly, the values of the depth and skeleton space coordinates are expressed in meters. The x and y axes represent the skeleton space coordinates, while the z-axis represents depth, as shown in Figure 4(b). In this study, the Kinect sensor and Kinect Face

Tracking SDK are used to determine the face position when fronting the camera, the detection of the mouth condition when open, and the eye detection when looking at the camera.

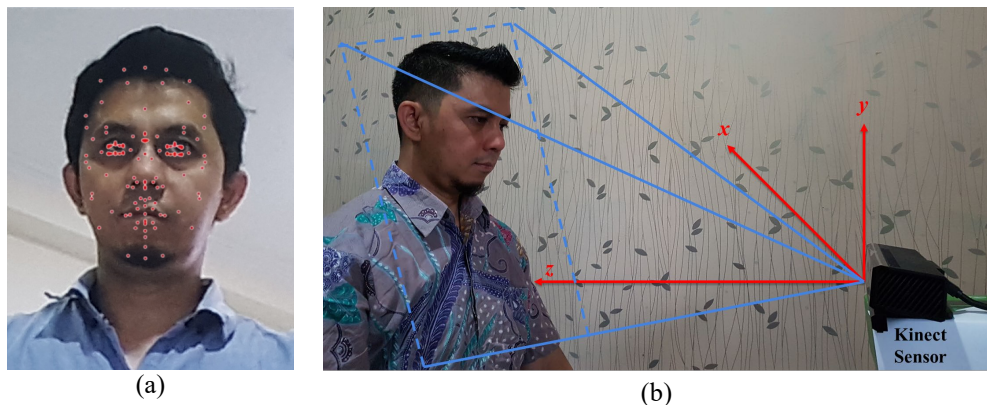


Figure 4. 121 Face points detected by the Kinect sensor.

### 2). Activation Through Speech

If humans are not captured by the Kinect camera or in a different room, then the system is active and called "SITI", an acronym for Intelligence Interaction System (*Sistem InTeraksi Intelijen*).

Voice activation provides speech input through a predetermined key phrase or an activation phrase. The term keyword detection describes the detection of activation phrases by hardware or software. It only occurs when the phrase "SITI" is pronounced, where the human-machine interaction system sounds "beep" to indicate it has entered listening mode (recording).

### 3). Activation Through Hand gestures

In case the environmental conditions around the Kinect have large noise, human speech cannot be detected. In this case, the interaction system can be activated by opening the right hand and pointing it at the Kinect.

Kinect supports hand-tracking to represent the commands to be defined based on palm gestures. There are several hand-states provided by Kinect, including open, close, lasso, unknown, and Not Tracked. In this activation system, the open gestures represent the system activation command.

The initial process of hand tracking is the capturing the gesture of the object in front of the Kinect v2 sensor. The gesture captured becomes the input for the system to track body parts of an object and obtain all joint positions. Afterwards, the system tracks the parts of the hands to detect the joint positions in each hand. Finally, it clarifies the joint to get a central area in each hand [18]. Figure 5 shows a hand-tracking process for the "Open" state.



Figure 5. Hand tracking for "Open" state.

To calculate the hand area, the center of the hand has to be captured, including the moment and its coordinates. The middle part of the hand is defined in Equation (1) below:

$$m_{0,0} = \iint x^p y^p f(x,y) dx dy \quad (1)$$

where  $f(x,y)$  is a gray value-function of an object. Integration is calculated in the object area. Generally, each pixel-based feature can be used to calculate the moment of an object instead of the gray value. In case it uses a binary image, the gray value function  $f(x,y)$  becomes:

$$f(x,y) = b(x,y) = \begin{cases} 1 \\ 0 \end{cases} \quad (2)$$

where 1 and 0 are object and background, respectively.

Therefore, the detected hand area in the 0<sup>th</sup> moment can be stated, as shown in the equation below:

$$m_{0,0} = \iint f(x,y) dx dy \quad (3)$$

The center of weight, first-order moments,  $m_{1,0}$  and  $m_{0,1}$  can be obtained from equations (4) and (5) as shown below.

$$m_{1,0} = \iint x dx dy f(x,y) \quad (4)$$

$$m_{0,1} = \iint y dx dy f(x,y) \quad (5)$$

The coordinate of  $x_c, y_c$  can be written as:

$$x_c = \frac{m_{1,0}}{m_{0,0}} \quad (6)$$

$$y_c = \frac{m_{0,1}}{m_{0,0}} \quad (7)$$

#### 4). Activation by Viewing Dialog Context

The developed human-machine interaction system understands the conversation between humans and machines by tracking the number of skeletons. The machine counts the number of people captured by the Kinect sensor and distinguishes whether humans are talking to machines or between themselves, as shown in Figure 6.

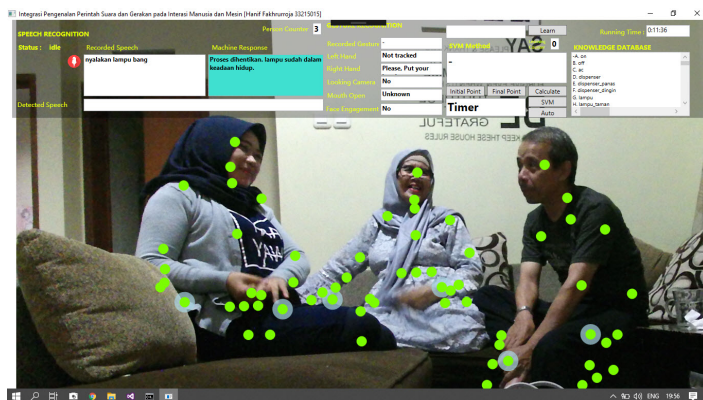


Figure 6. Activation system by looking at the dialogue context.

In case there is more than one person detected by a Kinect camera, the interaction system can be activated when the face is fronting the camera with an open mouth or eyes looking at the Kinect camera while saying the word "SITI".

### B. Indonesian Speech Recognition Module

The second module consists of the Indonesian speech recognition method that can be carried out in real-time and overcomes the noise problem in closed room conditions. Corpus voice recognition used at this stage is the Google Cloud Speech API.

Google Cloud Speech API can be integrated into an application. Cloud API determines an application software and interacts with cloud computing through the internet network. It offers applications that request information from the platform [19]. Development of cloud API has been increasing over time. For example, Google Cloud Speech API currently provides 120 languages, including Indonesian, and applies a deep-learning neural network algorithm with good accuracy.

Figure 7 shows the process of speech recognition with the Google Cloud Speech API. Human speech is captured using Kinect 2.0 as a speech sensor. The recorded speech data is sent through the cloud, then processed through the Cloud Speech API in the Google Cloud Platform. In case the audio is encrypted, the cloud speech API responds and sends it back to the user.

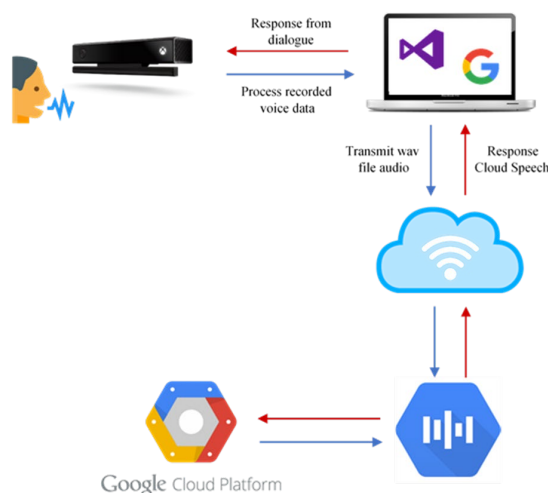


Figure 7. Indonesian speech recognition with the Google Cloud Speech API

### C. Intent Classification for Dialogue Module (Intent Search)

This module involves machine understanding of human speech to fulfil intent. The method used in this module is a simple natural language understanding. Essentially, three processes are carried out at this stage, including changing each sentence into a basic word (stemming), labeling the position/class of words, slot filling, and understanding the intent based on the rule. The stemming algorithm is based on Indonesian morphological rules, which are collected into one group and encapsulated on allowed and disallowed affixes. This algorithm uses an essential word dictionary and supports recoding, which is rearranging the words with excessive stemming [20].

After the stemming process, each word is labelled based on the essential words in the Indonesian dictionary corpus. The number of basic words used in this research corpus is 28,526 words. The Indonesian language has seven-word classes, specifically nouns, verbs, adjectives, pronouns, adverbs, numbers, and assignments.

After labelling the word class, slot filling is conducted. The main objective 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 semantics. The popular IOB (in-out-begin) format represents sentence slot tags [21], as shown in Figure 8.

|                       |                  |      |       |         |          |
|-----------------------|------------------|------|-------|---------|----------|
| <b>Word (W)</b>       | Room             | dark | here, | Turn on | The lamp |
|                       | ↓                | ↓    | ↓     | ↓       | ↓        |
| <b>IOB Format (S)</b> | O                | B    | O     | I       | O        |
| <b>Intent (I)</b>     | Turn on the lamp |      |       |         |          |

Figure 8. Examples of utterances with semantic slot annotations in IOB (S) and intent (I) format, B and I indicate the desired slot to turn on the lamp.

In this study, every utterance is converted into text. Furthermore, using stemming, each sentence is converted into basic word forms and given a position/class of words to find verbs and nouns. Intent classification is used to understand the user intent by searching the meaning of the relationship between adjectives, verbs and nouns.

#### D. Gesture Command Recognition Module

The features on the Kinect camera and proximity sensor are used to determine the x-y-z coordinate axis. Furthermore, the initial processing is carried out to determine the characteristics possessed by each gesture using statistical data [22].

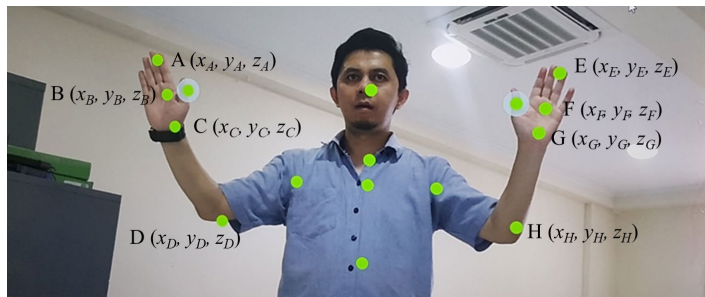


Figure 9. Eight skeleton coordinates (A to H)

The kinect sensor generates value at each joint of the human hand, including coordinate values of x, y, and z. This study uses four joints both on the right and left hand, as shown in Figure. 9. Equations (1), (2), and (3) produce pre-processing data, which is the value of the distance between each joint coordinate. It is calculated using statistical data such as average, variant, number, and median.

$$d(y_n, x_n) = x_n - y_n \quad (8)$$

$$d(z_n, x_n) = x_n - z_n \quad (9)$$

$$d(z_n, y_n) = y_n - z_n \quad (10)$$

where

$$n = A, B, C, \dots, H \quad (11)$$

Based on equations (8), (9), (10) and the calculation of the average value, variance, number, and median of each point, a 1x12 matrix is produced. The 12 lines represent feature values in the form of statistical data for each gesture, as shown in Figure 10.

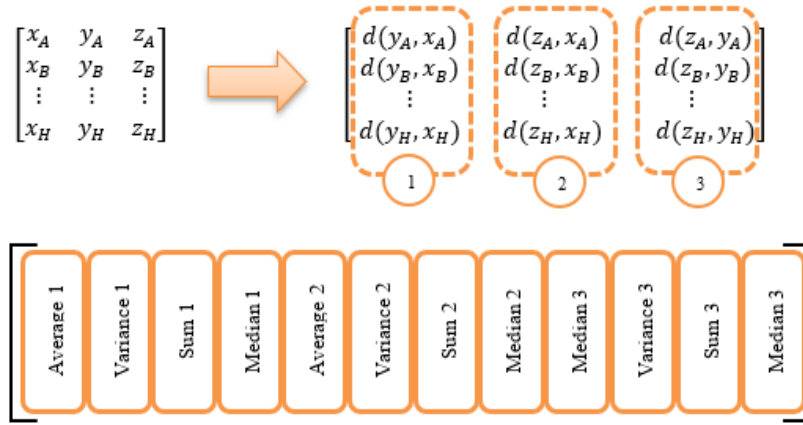


Figure 10. Matrix representation of skeleton coordinates.

After getting the dimensionless matrix value  $[1 \times 12]$ , the classification is performed to categorize the feature extraction results on each gesture command using a support vector machine (SVM). The SVM was chosen because of accurate results and fast computing time. It is a technique that determine the hyperplane with the most possibility of separating the two classes. This is carried out by measuring the hyperplane's margin and determining its maximum point. A margin refers to the distance between the corresponding hyperplane and the closest pattern of each class [23][24]. The best separator has the maximum margin and passes between the two classes. Margin is the minimum distance between the separator and the training sample. The sample closest to the separator is known as the support vector.

A sample data set of  $x_i$  pre-processing training results has the length  $L$ ,  $j = 1, \dots, L$ , where  $L$  is the point space in the  $L$  vector dimension. Linear classification is the point product between two vectors as expressed in Equation (12) below.

$$\langle w, x \rangle = \sum_{j=1}^L w_j x_j \quad (12)$$

The linear classification equation is:

$$f(x) = \langle w, x \rangle + b \quad (13)$$

$f(x)$  is a score for the  $x$ ,  $w$  input as a weight vector and  $b$  is a bias from the hyperplane. Multiclass SVM is an extension of binary classification. Multiclass proposes a DDAG strategy for classification using a binary classification model of  $\frac{k(k-1)}{2}$  where  $k$  is the number of

classes. The classification model is training with 2 class data, and a search solution for constraint optimization is as shown below.

$$\min_{w^{ij}, b^{ij}, \xi^{ij}} \frac{1}{2} (w^{ij})^T w^{ij} + C \sum_{\tau} \xi_{\tau}^{ij} \quad (14)$$

Where  $w^{ij}$  is the normal hyperplane for binary class,  $b^{ij}$  is bias,  $C$  is *penalty factor* and  $\xi^{ij}$  is *slack variable*.

DDAG is obtained by training each binary member -1 and +1. The points are evaluated to the decision node based on the first and last elements. Figure. 11 shows that in case a node prefers one of the two classes, the other one is removed from the list, and DDAG tests the first and last elements of the new member. However, DDAG stops in case only one class is left on the list. If there is a problem that has  $N$  class, the decision node should be evaluated to obtain the results. The selection of the class order listed is random for DAG-SVM [25].

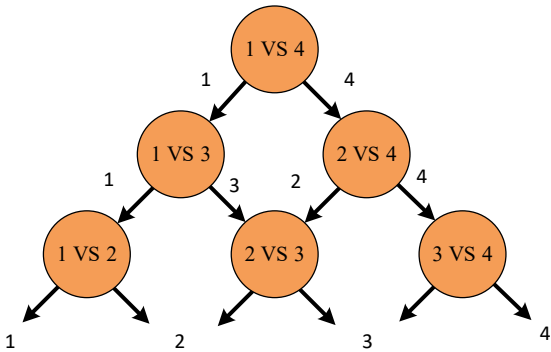


Figure 11. Directed Acyclic Graph (DAG)

E. Multimodal Fusion Module

The process of multimodal fusion influences the calculation model. It refers to where and when the fusion occurs. The fusion process can perform at the data or signal level [26], the feature level [27] [28] , and the decision or conceptual level [29]. The method of fusion calculation can divide into rule-based and statistical fusion (machine learning). Machine learning methods uses in multimodal fusion, such as Bayesian networks, neural networks, and graph-based fusion [30]. However, this study uses the rule-based fusion process using the logic gates algorithm at the signal level. Table 1 shows the difference between those methods.

Table 1. The differences between the fusion Methods

| Fusion Methods                     | Key issues  | Advantages   | Disadvantages   | Well Adopted in   |
|------------------------------------|---|--|---|---|
| Logic Gates Algorithm (this paper) | Data or signal fusion [29]<br>Fuse complementary semantics [31]                   | Fuse several modalities (more than two) [31]<br>Low computational cost [14]  | Cannot retrieve some missing signal from historical experience [32].            | Multi-modal biometric system [32], Multi-modal interaction system [14]  |
| Bayesian Decision                  | Data or signal fusion [29]<br>Maximize the joint distribution [31]                | Generate the missing modality [31]: retrieve some missing signal from historical experience and obtain the global optimal evaluation of the whole multimodal signal fusion [30]. | High computational cost [31]  | Face tracking, user behavior perception, robot pose estimation and obstacle avoidance, emotional understanding, and multi sensor information alignment and observation data analysis [30] |
| Neural Network                     | Feature fusion [29]<br>Preserve inter-modality and intra-modality similarity [31] | Measure the cross-modal similarity [31]<br>Good performance in nonlinear   | Hard to coordinate modalities more than two[31]<br>Designed for general purpose | Speech recognition, man-machine dialogue, machine translation, semantic   |

| Fusion Methods     | Key issues  | Advantages   | Disadvantages                         | Well Adopted in  |
|--------------------|---|--|---------------------------------------|--|
|                    |   | function fitting [30]  |                                       | understanding, object recognition, gesture detection and tracking, human body detection and tracking [30]. |
| Graph Based Fusion | Decision fusion [29]<br>Narrow the distribution difference [31] | Generate high quality novel samples [31]<br>Better tool for calculating uncertainty [30] | Suffer from training instability [31] | Scene segmentation, video content analysis, text semantic understanding [30]                               |

Figure 12 shows a schematic diagram that provides an overview of the fusion layers on different layers using rule-based. The layer at the bottom shows the sensor channel which functions as a recognition component in the input section. The output of the recognition component is visualized using a gray arrow pointing to the application. Hence, a set of rules has been applied in the knowledge-based fusion layer [33].

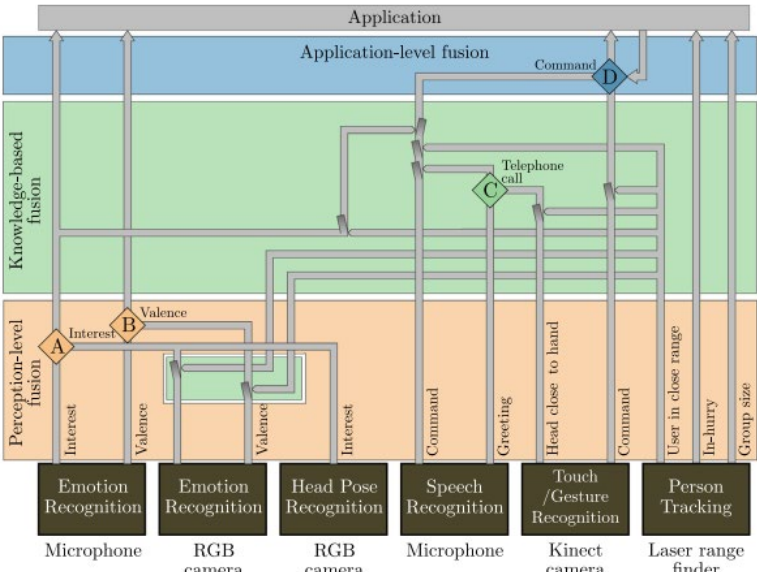


Figure 12. A schematic diagram that provides an overview of the fusion layers on different layers using rule-based [33].

In the case of this study, the focus combines multiple multimodal, such as speech, gesture, face detection, and skeleton tracking with low computation. Based on the results of the survey paper, as shown in Table 1, the suitable method for multimodal fusion involving several modalities (4 modalities) with a low computation should use the logic gates algorithm.

Input modalities of the human-machine interaction system developed in this study includes: Skeleton tracking  $s(t)$ , Face detection by observing the state of looking camera  $l(t)$ , face engagement  $e(t)$ , and mouth open  $m(t)$ ; speech  $U(t)$ ; and gestures  $G(t)$ .

Data from various input modalities are captured by the Kinect camera continuously and in real-time. Input data obtained are then extracted, recognized and combined to provide semantic representation and sent to the dialogue system. Figure 13 shows the multimodal fusion framework developed in this study. The proposed method is shown in Figure 14.

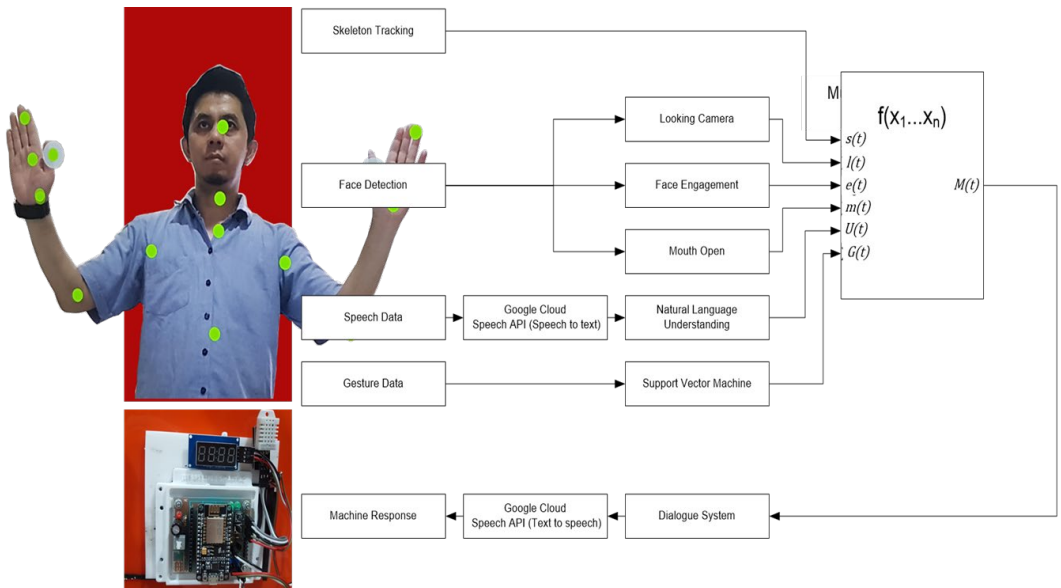


Figure 13. Multimodal fusion framework.

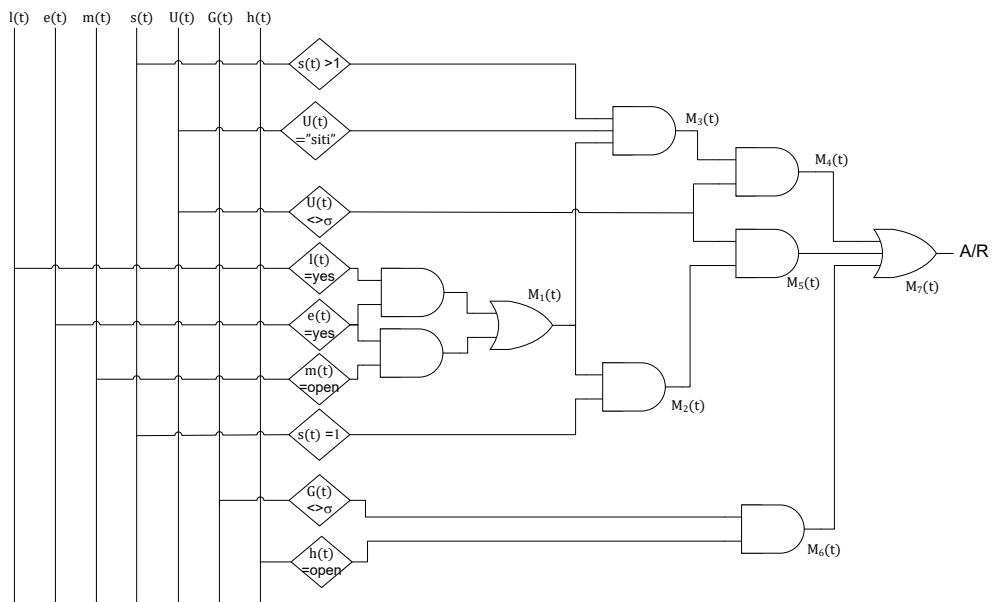


Figure 14. The developed multimodal fusion method.

The multimodal fusion method used for face detection,  $M_1(t)$  is obtained when the user's face is facing the Kinect camera  $e(t)$  in a "yes" state, and the user's mouth  $m(t)$  is in an "open" state. Still, the user's face might be facing the Kinect camera  $e(t)$  in a "yes" state, and the eyes looking at the Kinect camera  $l(t)$  also in a "yes" state.

$$M_1(t) = \{l(t) \times s(t)\} + \{s(t) \times m(t)\} \quad (15)$$

The developed human-machine interaction system sees the context of the conversation. Machines distinguish when humans talk to them, and therefore speech needs to be recognized and acted upon as a command. To perform this function, another input modality is added: skeleton tracking  $s(t)$ . In case only one person is caught,  $s(t)$  has a value of 1. This multimodal fusion method,  $M_2(t)$ , is described in equation (16).

$$M_2(t) = M_1(t) \times s(t) \mid s(t) = 1 \quad (16)$$

The machine also distinguishes when humans talk between them, and therefore all their speech is ignored because the context of the conversation is not to the machine. The maximum value of a human skeleton that can be identified based on the Kinect v2 specification is 6. In case the number of human skeletons is more than 1,  $s(t) > 1$ , the system becomes active when another input modality is added, specifically the detection of the word "SITI" or the value of  $U(t) = \text{"SITI"}$ . This indicates that humans are talking to machines. To activate the human-machine investment system, the following multimodal fusion method  $M_3(t)$  is used:

$$M_3(t) = M_1(t) \times s(t) \times U(t) \mid s(t) > 1, U(t) = \text{"siti"} \quad (17)$$

After the human-machine interaction system is active (the  $M_3(t)$  value is 1), the machine captures and recognizes all human speech. The speech is interpreted as a form of dialogue or command that must be answered or carried out. The speech recognition  $U(t)$  should fulfill the threshold value, specifically,  $U(t) \geq \sigma$ , to be included in the multimodal fusion equation. The threshold value is obtained in case human speech can be recognized using the Google Cloud Speech API.

The multimodal fusion method when the Kinect camera captured more than one human being ( $M_4(t)$ ) is described in equation 18 below.

$$M_4(t) = M_3(t) \times U(t) \quad (18)$$

In case only one person is captured, the multimodal fusion method ( $M_5(t)$ ) is described in equation 19.

$$M_5(t) = M_2(t) \times U(t) \quad (19)$$

In case there is an input modality from humans in the form of gesture  $G(t)$ , the gesture recognition value appears when there are two conditions. When an open right hand facing the Kinect camera,  $h(t) = \text{open}$ , there is a gesture recognition value  $G(t)$  that fulfills the threshold value of  $G(t) \geq \sigma$ . The threshold value is obtained from the changes classification result in the skeleton coordinates using the support vector machine. Multimodal fusion equations for gesture recognition are described in equation (20).

$$M_6(t) = G(t) \times h(t) \quad (20)$$

The final results of the multimodal fusion system developed by the human-machine interaction are described in equation (21). This means that human desires conveyed to the machine can be in the form of speech, gestures, or a combination of the two. Multimodal fusion results can be accepted (accept, A) or rejected (Reject, R) by the machine. Decisions of rejected and accepted are determined by the relationship between human intention and machine response available in the knowledge database.

$$M_7(t) = M_4(t) + M_5(t) + M_6(t) \quad (21)$$

#### F. Dialog System Module

The dialogue system shows that the machine provides answers to human needs. In this case, machines do not understand human desires or run electrical equipment in smart homes based on human needs. The dialogue system can be applied based on text, speech or pictures. The whole system requires a module called a dialogue system to regulate the conversations carried out by the system against humans. The dialogue system developed in this study uses the reinforcement learning (Q-learning) method.

Reinforcement learning is a method of mapping each state towards the selected actions to maximize the reward received [34]. Each state and action is given a value and represented as a

table. Learners are not told the action to choose. They have to determine the action that produces the greatest reward through trials. In some cases, actions affect both the delayed and directly received rewards. Also, reinforcement learning considers the problem and aims at the goal when interacting with an uncertain environment. The agent receives the state and selects the action. Furthermore, the agent receives a reward value from the selected action and maximizes it regularly [35].

Q-learning is one of the most important breakthroughs in reinforcement learning. It updates the action-value function, as shown in the following equation.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)) \quad (22)$$

where:

$\alpha$  = Learning rate,  $0 < \alpha \leq 1$ , determine the size of the rate, where the new value will replace the old value.

$\gamma$  = Discount rate,  $0 \leq \gamma \leq 1$ , determine the value of future rewards. With a smaller value of  $\gamma$ , the agent prioritizes the close rewards.

Q-learning updates the value function based on the largest action-value in the next state.

The state of the adaptive dialogue system developed in this study is the user's intent on the status condition of an electronic device in a smart home, while the action is the response of an electronic device. The relationship between state and action is represented in the form of Table Q, as shown in Figure 15.

|                    |         | Action (Device) |      |           |       |
|--------------------|---------|-----------------|------|-----------|-------|
|                    |         | AC              | Lamp | Dispenser | ....? |
| State/<br>(intent) | Hot     |                 |      |           |       |
|                    | Cold    |                 |      |           |       |
|                    | Dark    |                 |      |           |       |
|                    | Bright  |                 |      |           |       |
|                    | Bored   |                 |      |           |       |
|                    | Thirsty |                 |      |           |       |
|                    | Sultry  |                 |      |           |       |
|                    | ... ?   |                 |      |           |       |




Figure 15. The relationship between states and actions represented in the form of a Q Table.

### G. Knowledge provider module

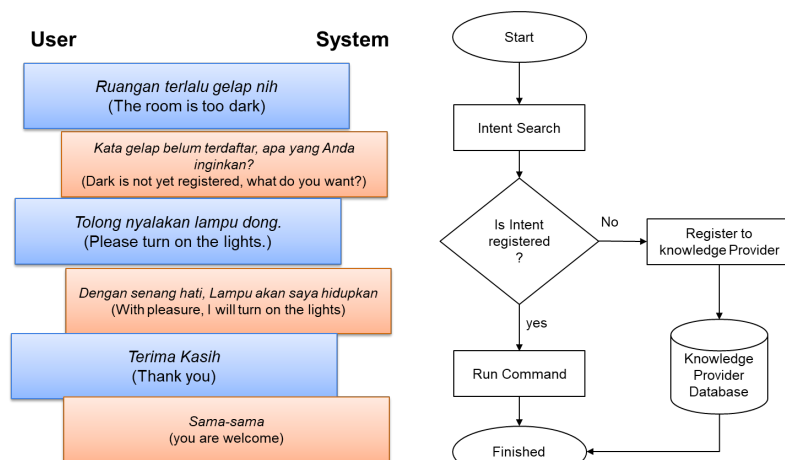


Figure 16. Knowledge provider algorithm for adding new state (intent)

The knowledge provider module is an algorithm in the smart home system. It is based on knowledge gained from a dialogue system that maps the relationship between human intents and the expected machine responses. The stored knowledge continues to develop when there is a new relationship between the intent and the machine response. The knowledge provider algorithm in the form of a flow chart is shown in Figure 16 for adding new states (intents).

#### H. Response Machine Module

The machine responds to the results of the dialogue system through speech and actions to control the electrical equipment based on users' needs. To determine answers based on user desires, several alternatives are provided in the response generator system to be chosen randomly. The answer text is converted into speech in Indonesian using the Google Cloud Text to Speech API. Figure 17 shows the generator response and text to speech.

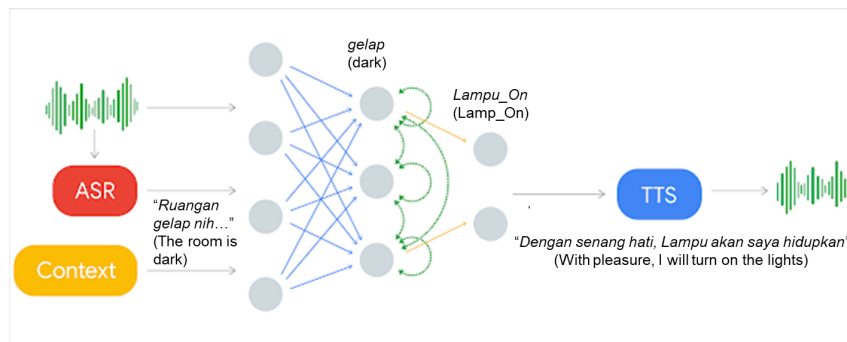


Figure 17. The diagram of the generator response and text to speech.

### 3. Result and Discussion

#### A. Overview of the human-machine interaction system

Figure 18 shows a multimodal fusion in human-machine interaction system.

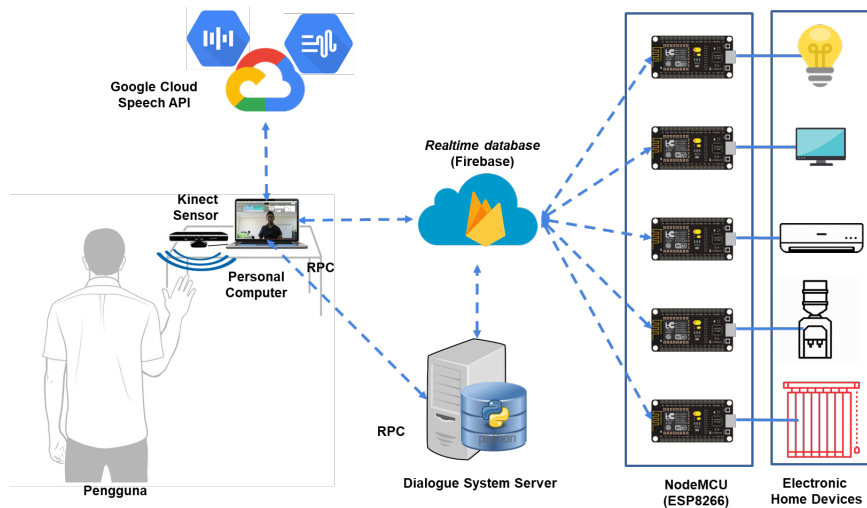


Figure 18. Multimodal fusion in human-machine interaction system.

The modality input in the form of speech, gestures, face detection, and skeleton tracking of the user is captured in real-time by Kinect v2. Furthermore, the speech recorded by Kinect v2 is converted to text via the Google Cloud Speech API. The recorded gestures from the user are processed by the SVM method to be converted into text. Data from various input modalities are

extracted, recognized and combined to provide semantic representation and sent to the dialogue system. Subsequently, the text results received by the system are processed using NLU, and therefore the system knows the user's intent. The intention is then sent to the dialogue system server using Remote Procedure Call (RPC). The system provides answers after being processed using the reinforcement learning method (Q-learning). RPC allows access to a process contained in another computer/platform. It uses a procedural programming paradigm with a socket to communicate with other processes [36]. In case the system updates the status in the real-time database, the equipment connected to the smart room can be controlled to suit the user's wishes. Communication between the dialogue system and the electrical equipment in the smart room is sent via Wi-Fi and received by NodeMcu installed on every electrical equipment, as shown in Figure 19.

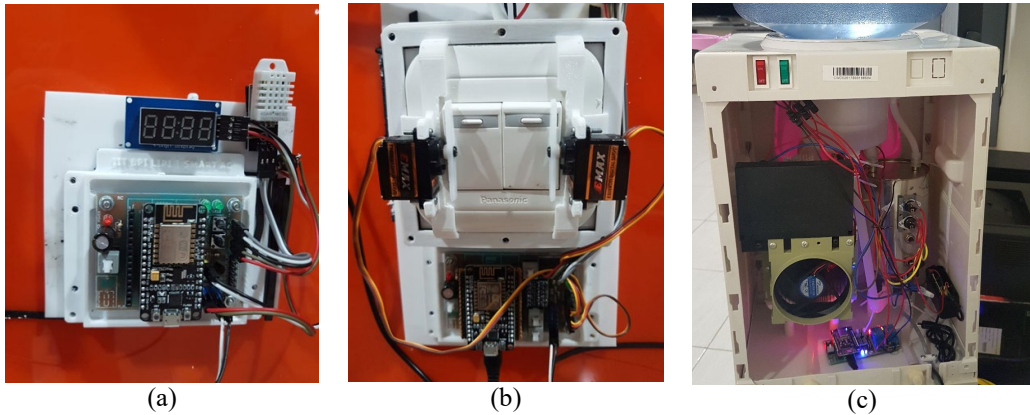


Figure 19. NodeMCU to control electrical equipment, such as (a) air conditioners, (b) room lamp, and (c) dispensers.

### B. Multimodal Fusion Based Human-machine Interaction System Interfaces

Figure 20 shows display interface systems equipped with Kinect sensors to capture input modalities from humans.



Figure 20. Multimodal Recognition Based Human-machine Interaction System Interfaces.

The human-machine interaction system interface is equipped with the following features.

- 1) Indonesian speech recognition feature displays the results of changes in human speech signals in the form of text (speech to text).

- 2) Recording of the human speech intent when the system has been successfully activated. This feature changes color from white to light blue when the system is successfully activated through multimodal recognition, such as face detection, skeleton tracking, speech and gestures recognition.
- 3) The response feature displays the results of dialogue response processing from the engine in the form of text, then it will be automatically changed to speech (text to speech).
- 4) The skeleton tracking feature displays the number of people captured by the Kinect camera sensor.
- 5) Face detection feature that displays the condition of the human eye, mouth, and human face.
- 6) Gesture recognition feature that displays the results of changes processing in skeleton coordinates using SVM.
- 7) Features to train and enter new gesture commands into the database automatically.
- 8) A machine knowledge database feature that displays Q tables as a result of dialogue system processing using reinforcement learning. This shows the relationship between human intents and machine responses.

### C. Testing and Implementation

The testing of human-machine interaction system developed was carried out in three stages, including multimodal activation, dialogue system, and multimodal fusion testing. Testing is carried out in a closed room with the least noise disturbance. The intensity of light in the room during the day is between 300 to 400 lux. The Kinect sensor is placed on a static plane with a distance to a fixed user, which is about 150 cm. This test was carried out on 40 people with various genders and age ranges. The respondent profile includes 65% of men and 35% women. The age group between 10-20, 21-30, 31-40, 41-50, and above 50 years is 7%, 27%, 35%, 13%, and 18%, respectively. The process is shown in Figure 21.



Figure 21. The respondent's profile in human and machine interaction systems.

### 1). The Activation Testing of Human-Machine Interaction Systems Using Face Detection, Speech and Gesture Modality.

Each test sample performs system activation through detection of faces, speech, and gestures 10 times. Therefore, the total testing was carried out 120 times (3 modalities x 10 times x 40 people).

The accuracy level graph in system activation testing with multimodal fusion is shown in Figure 22. The accuracy level of system activation using faces, gestures and speech modalities are 81.25%, 93.25%, and 87.75%, with an average of 87.42%. The system easily understands activation by the gesture because it only reads the position and state of the user's palm. The lowest level of activation accuracy uses face detection because the system must read three conditions at a time, such as the user's eyes, mouth, and face or face engagement. The lowest accuracy level for face detection occurs in the age group above 50 years.

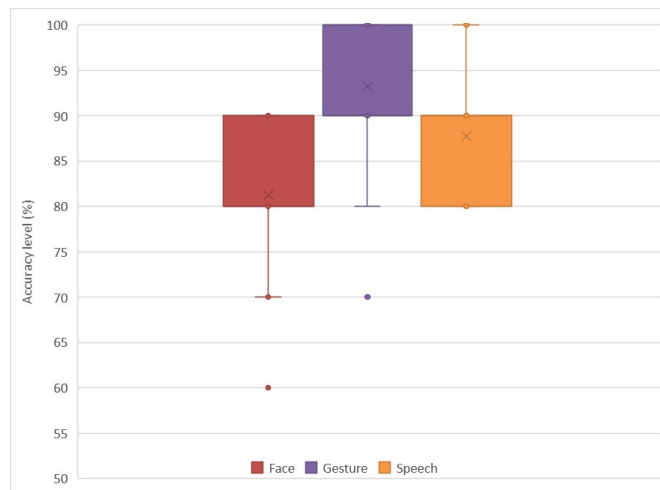


Figure 22. The results of activation testing in the human-machine interaction system.

### 2). Testing of the Dialog System for human-machine interaction

The dialogue system in the developed human-machine interaction needs to be responsive to understand every human utterance. The dialogue system testing uses Indonesian scenario, as shown in Table 2. Each test sample spoke 10 times for 14 dialogues determined. The total dialogue system testing is 5,600 times (14 dialogues x 10 times x 40 people). Figure 23 shows the testing results for the accuracy level of the developed dialogue system.

Table 2. Dialogue scenarios for testing human-machine interaction systems

| No  | User  | Machine Response  |
|-----|---|---|
| D01 | <i>Helo SITI</i><br>(Hello SITI)  | <i>Halo! Apa kabar? Ada yang bisa saya bantu?</i><br>(Hello! How are you? Is there anything I can help?)        |
| D02 | <i>Aduh, cuaca hari ini terasa GERAH banget nih</i><br>(Gosh, the weather today is so SULTRY) | <i>Rasa gerah belum terdaftar. Apa yang anda inginkan?</i><br>(Sultry is not yet registered. What do you want?) |
| D03 | <i>Tolong nyalakan AC dong</i><br>(Turn on the AC)  | <i>Dengan senang hati, AC akan saya dihidupkan</i><br>(With pleasure, I will turn on the air conditioner)       |

| No  | User   | Machine Response  |
|-----|--|---|
| D04 | <i>Saya masih kepanasan nih</i><br>(I'm still overheated)                                | <i>Apakah Anda ingin suhu AC diturunkan?</i><br>(Do you want the temperature of the air conditioner to be lowered?)   |
| D05 | <i>Turunkan Suhu AC dua derajat</i><br>(Lower the AC temperature by two degrees)         | <i>Baik, suhu AC akan segera diturunkan dua derajat</i><br>(Okay, the temperature of the air conditioner will soon be reduced by two degrees)   |
| D06 | <i>Nyalakan kulkas dong</i><br>(Turn on the fridge)                                      | <i>Maaf device kulkas belum terdaftar, apakah Anda ingin mendaftarkannya?</i><br>(Sorry, the fridge device is not registered yet, do you want to register it?)  |
| D07 | <i>Daftarkan</i><br>(Register)   | <i>Baik, device kulkas didaftarkan ke dalam sistem</i><br>(Okay, the fridge device is registered to the system)   |
| D08 | <i>Nyalakan kulkas lagi dong</i><br>(Turn on the fridge again)                           | <i>Maaf, device kulkas belum terinstalasi secara fisik</i><br>(Sorry, the fridge device has not been physically installed)  |
| D09 | <i>Ruangan kok gelap banget ya</i><br>(Why is the room so dark?)                         | <i>Apakah Anda ingin lampu dihidupkan</i><br>(Do you want the lights on)  |
| D10 | <i>Iya, nyalakan lampu</i><br>(Yes, turn on the lights)                                  | <i>Sepertinya lampu belum disebutkan pada perintah Anda, lampu apa yang ingin dinyalakan?</i><br>(It looks like the light has not been mentioned in your command, what kind of light do you want to turn on?) |
| D11 | <i>Semua lampu, tolong nyalakan</i><br>(Turn on all the lights)                          | <i>Siap, semua lampu akan dinyalakan sesuai dengan keinginan Anda</i><br>(Okay, all the lights will be turned on according to your wishes)  |
| D12 | <i>Saya haus nih</i><br>(I'm thirsty)  | <i>Rasa haus belum terdaftar. Apa yang anda inginkan?</i><br>(Thirsty is not yet registered. What do you want?)   |
| D13 | <i>Nyalakan power dingin pada dispenser</i><br>(Turn on the cold power on the dispenser) | <i>Dengan senang hati, power dingin pada dispenser akan saya dihidupkan</i><br>With pleasure, I will turn on the cold power on the dispenser  |
| D14 | <i>Terima kasih</i><br>(Thank you)   | <i>Asyiap... sama-sama</i><br>(Okay ... you are welcome)  |

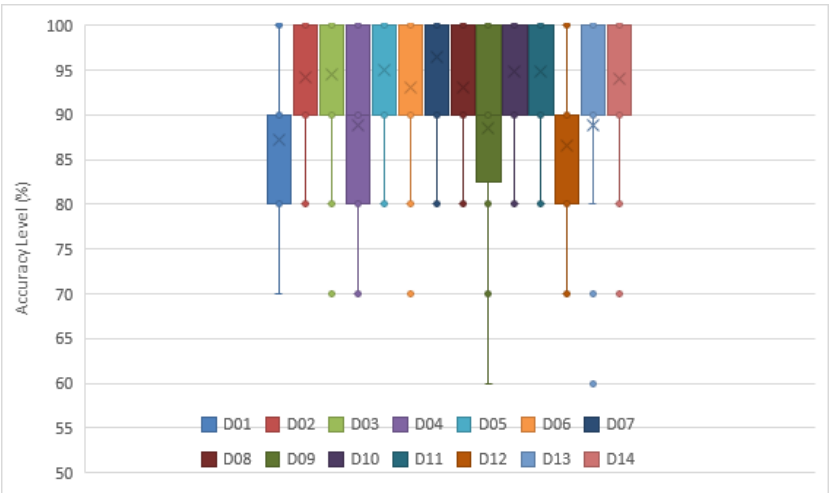


Figure 23. Graph of dialogue system testing.

The average accuracy level of the dialogue system testing is 92.11%. The highest accuracy level occurs in the word "register (D07)" with 96.5%. Almost all test samples are clear enough to say "register", and therefore, they can be easily translated by speech recognition systems. The lowest accuracy level occurs in the dialogue "I am thirsty (D12)" with 86.5%. The pronunciation of the word "haus (thirsty)" has been translated several times by the speech recognition system with the word "hapus (delete)". The second-lowest accuracy level occurs in the dialogue "Hello SITI (D01)" with 87.25%. The pronunciation of the word "SITI" has also been translated several times by the speech recognition system with the word "Kitty".

The gender of the test sample did not significantly influence the accuracy of the dialogue system. However, the pronunciation of the test sample with an age group of more than 50 years has a significant effect on the accuracy of the dialogue system. From the test sample of more than 50 years, a total of 7 people have an average accuracy level of 79.39%, or 12.72% lower than the average accuracy level of the overall test sample, which is 92.11.

3). Testing of Gesture Recognition Systems

Each test sample performs gesture commands to control electronic devices. This includes the commands to turn the lights, TV, and AC on and off, and raise or lower the temperature, open or close the curtains, turn the dispenser on or off, and turn off the system. Table 3 shows examples of gesture commands.

Table 3. Examples of gesture commands

| No.  | Gesture Position      |          | No.  | Gesture Position |          |
|------|-----------------------|----------|------|------------------|----------|
|      | Starting Point        | Endpoint |      | Starting Point   | Endpoint |
| G01. |                       |          | G07. |                  |          |
|      | Turn on the dispenser |          |      | Open the curtain |          |
| G02. |                       |          | G08. |                  |          |
|      |                       |          |      |                  |          |



The average accuracy of the gesture recognition system is 93.46%. The lowest accuracy level occurs in recognition of the command "Turn on the lights" of 88.25%. The highest accuracy level occurs in recognition of the command "Turn on the AC" of 98.25%, while the lowest is in the age group of 10 to 20 years at 80.27%. This is because during the gesture recognition training process using the SVM method, the sample is the adult age group. The graph for the testing results of the gesture recognition system is shown in Figure 24.

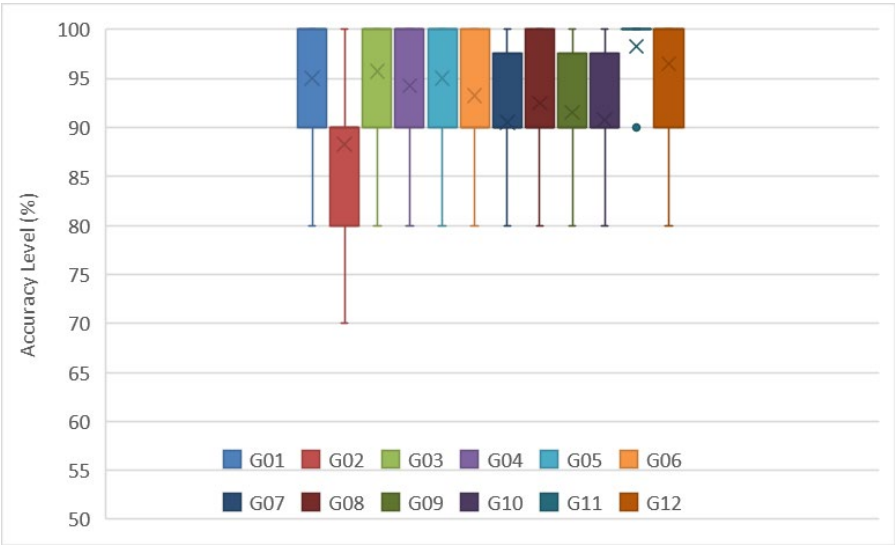


Figure 24. Graph for the testing results of the gesture recognition system.

4). Multimodal Fusion Testing for Human-machine Interaction Systems

Testing for the accuracy of the human-machine interaction system involves multimodal recognition, which is the integration of speech and gesture recognition with multimodal fusion algorithms.

Each test sample (40 people) simultaneously said: "Turn on the light" and moved the hand command to turn the light on 10 times. This means the total test was carried out 400 times (40 test samples x 10 times). Figure 25 shows the graph of the multimodal fusion test results. The average accuracy level is 93%. The lowest level occurs in the age group above 50 years. This is because the pronunciation "turn on the lights" is less clear.

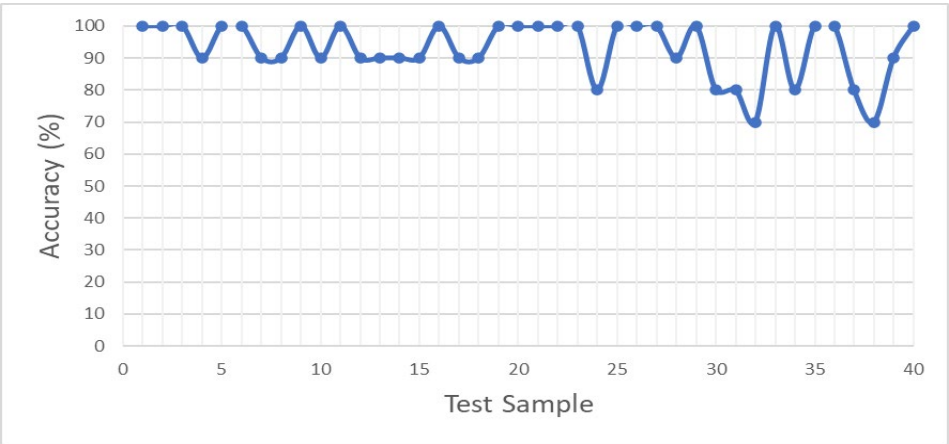


Figure 25. Graph for accuracy level of multimodal fusion testing (speech and gesture).

5). Testing of Machine Knowledge Development

The machine knowledge development based on knowledge gained from a dialogue system that maps the relationship between human intents and the expected machine responses. The stored knowledge continues to develop when there is a new relationship between the intent and the machine response. Figure 26 shows the initial environment initiation for the dialogue system using reinforcement learning (Q-learning)

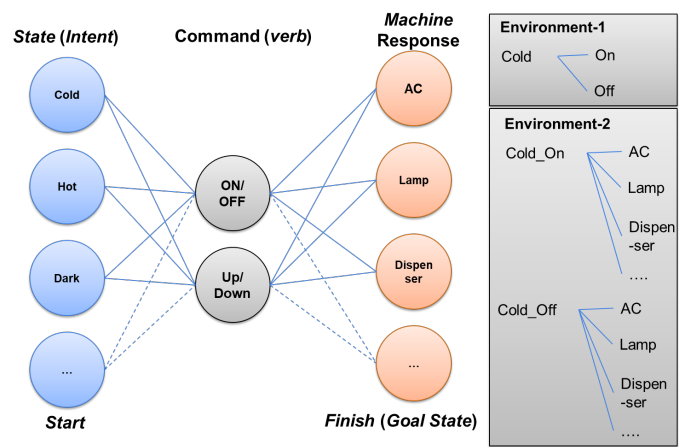


Figure 26. Environment for a dialogue system

The reinforcement learning-based dialogue system training process to build a database of machine knowledge. The steps of the training process are as follows: (1) Create a training data

set of 676 intents from the observations of verbs and adjectives in the Indonesian Dictionary; (2) Train each data set using the reinforcement learning method; (3) from 676 intents datasets that trained, the relationship between the intent and the power device obtained was 156 data; and (5) the training results are stored in a database.

The steps of the reinforcement learning-based dialogue system testing process are as follows: (1) Create test data sets for each class; (2) The class based on the relationship between the intent and machine response; (3) The number of data set test for each class is 40 data. The 20 data for turn on the device (device\_on) and 20 data for turn off the device (device\_off). The test data get from the results of interviews with respondents regarding what they want to control electrical equipment, such as televisions, air conditioners (AC), and lights; (4) Testing the data for each class with reinforcement learning. In this testing process, the training process deactivates so that the dialogue system can make predictions from the results of the previous training process to build a database of machine knowledge; (5) Displays data confusion matrix; and (6) Each class calculated using the classification report method, namely the values of accuracy (CA), precision (PR), recall (RE), F1-measure (F1), and support. Figure 27 shows the formula of classification report.

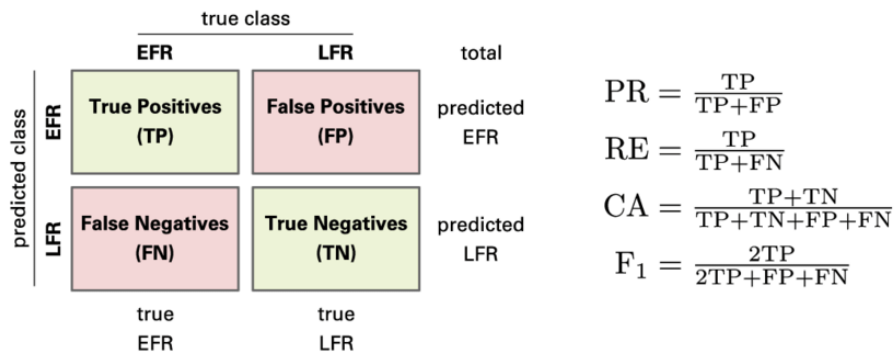


Figure 27. Confusion matrix with the formulas of precision (PR), recall (RE), accuracy (CA), and F1-measure (F1)[37].

The machine knowledge development testing uses an Indonesian dialogue scenario. Table 4 shows examples of dialogue scenarios for developing knowledge. Each human intent (in blue) tested it to get a machine response (in red). Figure 28 shows the results of machine knowledge development using reinforcement learning (Q-learning).

Table 4. Dialogue scenario for testing Machine Knowledge Development

| Relation         | User  | Machine Response  |
|------------------|---|---|
| Human Intent     | <i>Aduh, hari ini sangat mem-BOSAN-kan.</i><br>(Gosh, today is so <i>BORING</i> ) | <i>Rasa BOSAN belum terdaftar. Apa yang anda inginkan?</i><br>( <i>BORING</i> is not yet registered. What do you want?) |
| Machine Response | <i>Tolong NYALAKAN TV dong</i><br>( <i>TURN ON THE TV</i> )                       | <i>Dengan senang hati, TV AKAN DIHIDUPKAN</i><br>(With pleasure, I will <i>TURN ON THE TV</i> )                         |



Table 8. The values of accuracy (CA), precision (PR), recall (RE), F1-measure (F1)

| Devices | Accuracy | Precision | Recall | F1-Score | Support |
|---------|----------|-----------|--------|----------|---------|
| AC      | 85%      | 95%       | 85%    | 89%      | 40      |
| TV      | 80%      | 92%       | 80%    | 83%      | 40      |
| Lights  | 80%      | 88%       | 57%    | 66%      | 40      |

6). Testing of Dialogue Contexts on Human-machine Interaction Systems

Dialogue context testing is carried out to determine whether the developed human-machine interaction system distinguishes when the user is talking to a machine or a fellow human. Therefore, the system ignores what the user is saying because the context is talking to a fellow human rather than a machine.

This dialogue context testing was carried out in two groups of three and four people, as shown in Figure 29. The dialogue scenario was divided into three parts, including counting the people caught on the Kinect camera, start a conversation by and without looking at the Kinect camera, as shown in Table 9. When there is more than one person caught by the camera, the human-machine interaction system is active in case the first person detected by Kinect looks at the camera and calls the system "SITI". Figure 30 shows the graph of the testing results.



Figure 29. A sample of dialogue context testing in a human-machine interaction system: (a) A group of three people and (b) A group of four people

Table 9. Scenarios for the dialogue context testing on multimodal recognition-based human-machine interaction

| No  | User 1 / User 2   | User 3 / User 4  | System Response   |
|---|---|--|---|
| Person Counter  |   |  |   |
| PC  | Count the number of people detected by Kinect camera (Person Counter)                         | -  | Displays the number of people detected by the Kinect camera   |
| Start the conversation without looking at the Kinect camera |   |  |   |
| DO1   | <i>Bagaimana kabar "SITI" teman kita?</i><br>(How is our friend "SITI"?)                      | <i>Kabar "SITI" baik dia sekarang tinggal di Jakarta</i><br>("SITI" is fine, she now lives in Jakarta) | No response   |
| DO2   | <i>Apakah "SITI" sudah menikah?</i><br>(Is "SITI" married?)                                   | <i>"SITI" belum menikah</i><br>("SITI" is not married yet.)  |   |
| Start a conversation by looking at the Kinect camera        |   |  |   |
| DK1   | SITI  | -  | <i>Apakah Anda memanggil saya?</i><br>(Are you calling me?)   |
| DK2   | <i>Iya</i><br>(Yes)   | -  | <i>Silahkan bicara</i><br>(Please speak)  |
| DK3   | <i>Saya kegerahan nih</i><br>(I feel hot)   | -  | <i>AC sudah hidup, apakah Anda ingin menurunkan suhu?</i><br>(The air conditioner is on, do you want to lower the temperature?) |
| DK4   | <i>Turunkan suhu tiga derajat lagi dong</i><br>(Lower the temperature by three degrees again) | -  | <i>Siap suhu AC diturunkan tiga derajat</i><br>(Okay, the temperature of the air conditioner is reduced by three degrees)       |
| DK5   | <i>Terima kasih</i><br>(Thank you)  |  | <i>Asyiap... sama-sama</i><br>(Okay ... you are welcome)  |

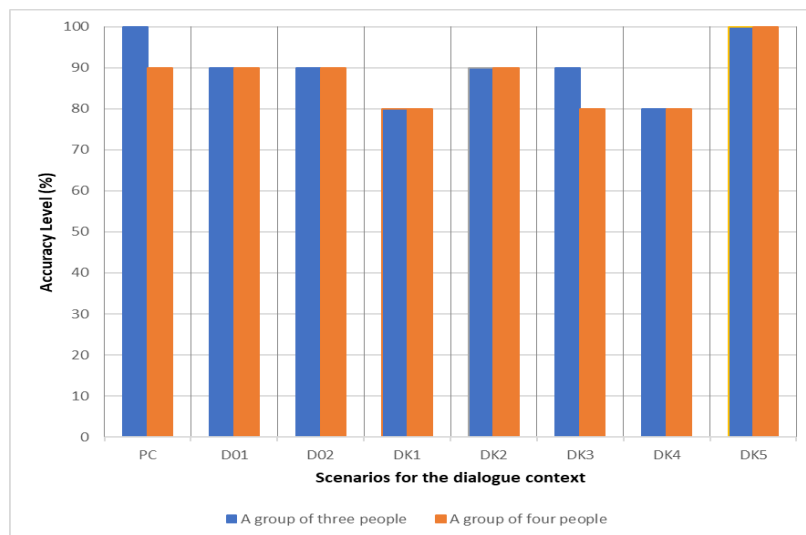


Figure 30. Graph for the results of dialogue context testing on human-machine interaction systems.

The average accuracy level of dialogue context in multimodal recognition-based human-machine interaction systems is 88.75%. The system calculates the exact number of people caught on the Kinect camera through a developed skeleton tracking algorithm with an accuracy level of 95%. The accuracy level of the dialogue context with groups of four and three people was 87.5% and 90%, respectively. A group with more people has lower accuracy compared to a group with fewer people. This is because the system detects more skeletons, and therefore requires a longer process.

#### 7). User Acceptance Test (UAT)

UAT is used to measure the extent to which the multimodal recognition-based human-machine interaction system developed can be accepted by the user. A total of respondents participated in the UAT survey. Furthermore, 40 people directly tried the system, while 23 people only watched when the system testing process was in progress.

The UAT questionnaire distributed consisted of six parts as the following:

- 1) Introduction to the questionnaire containing an explanation of the multimodal recognition-based human-machine interaction system developed.
- 2) Respondent's identity for profile data.
- 3) Six preliminary questions on the respondents' knowledge and opinions about the existing human-machine interaction system.
- 4) Twenty-five core questions on the UAT survey, including questions about the multimodal activation process, recognition and fusion, intent search and dialogue systems, and machine responses for smart home case studies.
- 5) Conclusions or opinions of respondents regarding the final assessment of the multimodal recognition-based human-machine interaction system developed.
- 6) Acknowledgements and suggestions for researchers to improve the multimodal recognition-based human-machine interaction system developed.

Respondents with and without knowledge the human-machine interaction system were 85% and 14.3%, respectively. Respondents that used a dialogue system, such as the applications of CORTANA (Microsoft); OK, GOOGLE (Google); SIRI (Apple); or ALEXA (Amazon) were 84.1%, while those that never used it were 15.9%. A total of 63 respondents (100%) agreed to

develop a human-machine interaction system with additional modalities, such as face detection and gesture recognition.

The level of user satisfaction with multimodal recognition-based human-machine interaction systems shows that 60% are very satisfied, 35% satisfied, and 5% are quite satisfied. The delay during the multimodal fusion process and the response of the machine, which takes about 3 to 5 seconds, dissatisfy users. This is because the dialogue system algorithm and machine response need to wait for the results of the multimodal fusion algorithm. The fusion algorithm also waits for the queue process from the results of face detection and the process of speech and gesture recognition that runs simultaneously. A total of 76.2% of users agreed that the interaction system developed was natural, while 23.8% were still doubtful. A total of 79.4% of users agreed that the machine responded to the user's desires well while 20.6% were still doubtful.

#### **4. Further Research**

Contributions to the multimodal recognition-based human-machine interaction system method have opened up opportunities for further research, including the development of speech recognition systems integrated with human-machine interaction systems. It reduces the waiting time when converting speech signals into text. Furthermore, the accuracy of multimodal fusion systems can be improved by adding machine learning methods, such as deep reinforcement learning.

#### **5. Conclusion**

The activation process of a human-machine interaction system using face detection, skeleton tracking, and speech and gesture recognition, has been successfully developed and responds well. The average accuracy level of multimodal activation is 87.42%, which is generated with the gesture recognition of 93.25%. The lowest accuracy level is generated from the activation results with face detection of 81.25%.

The accuracy level of the multimodal recognition-based human-machine interaction dialogue system was 92.11% from a total of 5,600 tests. The gender of the test sample did not significantly influence the accuracy level of the dialogue system. However, the pronunciation of the test sample with an age group of more than 50 years has a significant effect on the accuracy level of the dialogue system. For instance, a total of 7 peoples have an average accuracy level of 79.39%.

The average accuracy level for interactions using gesture recognition is 93.54%. The highest average accuracy level of gesture recognition is in the adult age group of 97.71% while the lowest level is in the children age group of 88.96%. This is because during the training process for gesture recognition, the sample is the adult age group.

The accuracy level for interactions using two multimodal at once, speech and gesture, with multimodal fusion algorithms on human-machine interaction systems is better compared to interactions through speech recognition alone. However, when compared with gesture recognition, interaction through multimodal recognition has a lower accuracy level of 0.54% than gesture alone. This is because interactions through gesture recognition do not need to wait for the results of speech to text processing by the Google Cloud Speech API, which is dependent on internet connection.

The level of user satisfaction with multimodal recognition-based human-machine interaction systems shows that 95% are satisfied. The delay during the multimodal fusion process and the response of the machine, which takes about 2 to 3 seconds, dissatisfy users. This is because the dialogue system algorithm and the machine response need to wait for the results of the multimodal fusion algorithm. Similarly, the multimodal fusion algorithm needs to wait for the queue process from the results of face detection and the process of speech and gesture recognition that runs simultaneously. A total of 76.2% of users agreed that this interaction system was natural. Also, 79.4% of users agreed that the machine responded well to their wishes.

## 6. Acknowledgements

The authors express gratitude to the Educational Fund Management Institution (LPDP), Ministry of Finance of the Republic of Indonesia for partly funding this research. The authors are also grateful to anonymous reviewers for their valuable comments.

## 7. References

- [1]. H. Fakhrrurroja, 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, doi: 10.1109/ICARCV.2018.8581071.
- [2]. S.-T. Cheng, C.-W. Hsu, and J.-P. Li, 'Combined Hand Gesture — Speech Model for Human Action Recognition', *Sensors*, vol. 13, no. 12, pp. 17098–17129, Dec. 2013, doi: 10.3390/s131217098.
- [3]. 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, doi: 10.1145/3264996.3265002.
- [4]. F. Lamberti, V. Gatteschi, A. Sanna, and A. Cannavò, 'A Multimodal Interface for Virtual Character Animation Based on Live Performance and Natural Language Processing', *International Journal of Human–Computer Interaction*, vol. 35, no. 18, pp. 1655–1671, Nov. 2019, doi: 10.1080/10447318.2018.1561068.
- [5]. F. Febriansyah, N. A. Suwastika, and H. Fakhrrurroja, 'Patient necessity notification system based on gesture recognition (Kinect V2) and internet of things using selection frame method', *J. Phys.: Conf. Ser.*, vol. 1192, p. 012051, Mar. 2019, doi: 10.1088/1742-6596/1192/1/012051.
- [6]. S. Mitra and T. Acharya, 'Gesture Recognition: A Survey', *IEEE Trans. Syst., Man, Cybern. C*, vol. 37, no. 3, pp. 311–324, May 2007, doi: 10.1109/TSMCC.2007.893280.
- [7]. M. H. Tambunan, Martin, H. Fakhrrurroja, Riyanto, and C. Machbub, 'Indonesian speech recognition grammar using Kinect 2.0 for controlling humanoid robot', in *2018 International Conference on Signals and Systems (ICSigSys)*, Bali, May 2018, pp. 59–63, doi: 10.1109/ICSIGSYS.2018.8373568.
- [8]. M. Katore and M. R. Bachute, 'Speech based human machine interaction system for home automation', in *2015 IEEE Bombay Section Symposium (IBSS)*, Mumbai, India, Sep. 2015, pp. 1–6, doi: 10.1109/IBSS.2015.7456634.
- [9]. K. Nakadai, T. Mizumoto, and K. Nakamura, 'Robot-Audition-based Human-Machine Interface for a Car', in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Hamburg, Germany, Sep. 2015, pp. 6129–6136, doi: 10.1109/IROS.2015.7354250.
- [10]. M.-S. Kim and C. H. Lee, 'Hand Gesture Recognition for Kinect v2 Sensor in the Near Distance Where Depth Data Are Not Provided', *International Journal of Software Engineering and Its Applications*, vol. 10, no. 12, pp. 407–418, 2016.
- [11]. P. Tu and C. Huang, 'Mechanical arm teleoperation control system by dynamic hand gesture recognition based on kinect device', *The Journal of Engineering*, vol. 2019, no. 23, pp. 9110–9113, Dec. 2019, doi: 10.1049/joe.2018.9196.
- [12]. Z. Lei, Z. H. Gan, M. Jiang, and K. Dong, 'Artificial robot navigation based on gesture and speech recognition', in *Proceedings 2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, Wuhan, China, Oct. 2014, pp. 323–327, doi: 10.1109/SPAC.2014.6982708.
- [13]. D. Yongda, L. Fang, and X. Huang, 'Research on multimodal human-robot interaction based on speech and gesture', *Computers & Electrical Engineering*, vol. 72, pp. 443–454, Nov. 2018, doi: 10.1016/j.compeleceng.2018.09.014.

- [14]. H. Fakhurroja, C. Machbub, A. S. Prihatmanto, and A. Purwarianti, 'Multimodal Interaction System for Home Appliances Control', *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 14, no. 15, Sep. 2020.
- [15]. H. Alkhatib *et al.*, 'Ieee cs 2022 report', *IEEE Computer Society*, pp. 25–27, 2014.
- [16]. A. Yargic and M. Dogan, 'A lip reading application on MS Kinect camera', in *2013 IEEE INISTA*, Albena, Bulgaria, Jun. 2013, pp. 1–5, doi: 10.1109/INISTA.2013.6577656.
- [17]. Q. Wang and X. Ren, 'Facial Feature Locating Using Active Appearance Models With Contour Constraints From Consumer Depth Cameras', *. Vol.*, vol. 45, p. 5, 2012.
- [18]. H. Fakhurroja, 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 (IoT&IS)*, 2019, pp. 74–78.
- [19]. 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, doi: 10.5220/0003388101660169.
- [20]. 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, doi: 10.1145/1316457.1316459.
- [21]. X. Li, Y.-N. Chen, L. Li, J. Gao, and A. Celikyilmaz, 'End-to-End Task-Completion Neural Dialogue Systems', *arXiv:1703.01008 [cs]*, Mar. 2017, Accessed: Aug. 08, 2019. [Online]. Available: <http://arxiv.org/abs/1703.01008>.
- [22]. D. A. Maharani, H. Fakhurroja, 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, doi: 10.1109/ISCAIE.2018.8405435.
- [23]. M. Satone and G. Kharate, 'Feature selection using genetic algorithm for face recognition based on PCA, wavelet and SVM', *International Journal on Electrical Engineering and Informatics*, vol. 6, no. 1, p. 39, 2014.
- [24]. A. Turnip, M. F. Amri, H. Fakurroja, A. I. Simbolon, M. A. Suhendra, and D. E. Kusumandari, 'Deception detection of EEG-P300 component classified by SVM method', in *Proceedings of the 6th International Conference on Software and Computer Applications - ICSCA '17*, Bangkok, Thailand, 2017, pp. 299–303, doi: 10.1145/3056662.3056709.
- [25]. P. Chen and S. Liu, 'An improved dag-svm for multi-class classification', in *2009 Fifth International Conference on Natural Computation*, 2009, vol. 1, pp. 460–462.
- [26]. R. Riyanto, C. Machbub, H. Hindersah, and W. Adiprawita, 'Slope Balancing Strategy for Bipedal Robot Walking Based on Inclination Estimation Using Sensors Fusion', *ijeei*, vol. 11, no. 3, pp. 527–547, Sep. 2019, doi: 10.15676/ijeei.2019.11.3.6.
- [27]. F. M. T. Retno Kinasih, C. F. Dommaris Saragih, C. Machbub, P. H. Rusmin, L. Yulianti, and D. Andriana, 'State Machine Implementation for Human Object Tracking using Combination of MobileNet, KCF Tracker, and HOG Features', *ijeei*, vol. 11, no. 4, pp. 697–712, Dec. 2019, doi: 10.15676/ijeei.2019.11.4.5.
- [28]. D. Andriana, A. S. Prihatmanto, E. M. Idris Hidayat, and C. Machbub, 'Combination of Face and Posture Features for Tracking of Moving Human Visual Characteristics', *ijeei*, vol. 9, no. 3, pp. 616–631, Sep. 2017, doi: 10.15676/ijeei.2017.9.3.14.
- [29]. M. Palcari and C. L. Lisetti, 'Toward multimodal fusion of affective cues', in *Proceedings of the 1st ACM international workshop on Human-centered multimedia - HCM '06*, Santa Barbara, California, USA, 2006, p. 99, doi: 10.1145/1178745.1178762.
- [30]. M.-H. Yang and J.-H. Tao, 'Data fusion methods in multimodal human computer dialog', *Virtual Reality & Intelligent Hardware*, vol. 1, no. 1, pp. 21–38, Feb. 2019, doi: 10.3724/SP.J.2096-5796.2018.0010.
- [31]. W. Guo, J. Wang, and S. Wang, 'Deep Multimodal Representation Learning: A Survey', *IEEE Access*, vol. 7, pp. 63373–63394, 2019, doi: 10.1109/ACCESS.2019.2916887.

- [32]. C. N. Manivannan N, ‘Multimodal Biometrics for Robust Fusion Systems using Logic Gates’, *J Biom Biostat*, vol. 06, no. 01, 2015, doi: 10.4172/2155-6180.1000218.
- [33]. M. Glodek *et al.*, ‘Fusion paradigms in cognitive technical systems for human–computer interaction’, *Neurocomputing*, vol. 161, pp. 17–37, Aug. 2015, doi: 10.1016/j.neucom.2015.01.076.
- [34]. S. C. Sari, K. Kuspriyanto, A. S. Prihatmanto, and W. Adiprawita, ‘Online State Elimination in Accelerated reinforcement Learning’, *ijeei*, vol. 6, no. 4, pp. 665–680, Dec. 2014, doi: 10.15676/ijeei.2014.6.4.3.
- [35]. R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [36]. M. A. Holliday, J. T. Houston, and E. M. Jones, ‘From sockets and RMI to web services’, in *Proceedings of the 39th SIGCSE technical symposium on Computer science education - SIGCSE '08*, Portland, OR, USA, 2008, p. 236, doi: 10.1145/1352135.1352221.
- [37]. S. Bittrich, M. Kaden, C. Leberecht, F. Kaiser, T. Villmann, and D. Labudde, ‘Application of an interpretable classification model on Early Folding Residues during protein folding’, *BioData Mining*, vol. 12, no. 1, p. 1, Dec. 2019, doi: 10.1186/s13040-018-0188-2.



**Hanif Fakhurroja** 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.



**Carmadi Machbub** 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 leads of the Control and Computer Systems Research Group, 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 was 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.