Intelligent Automation And Soft Computing, 2020
Copyright © 2020, TSI® Press
Vol. 26, no. 2, 291–301
https://doi.org/10.31209/2020.100000164

Intelligent Speech Communication Using Double Humanoid Robots

Li-Hong Juang[1], Yi-Hua Zhao[2]

[1]School of Electrical Engineering and Automation, Xiamen University of Technology, No.600, Ligong Road, Jimei, Xiamen, 361024, P.R. China
[2]Engineering College, HuaQiao University, No.269, Chenghua North Road, Fengze District, Quanzhou, Fujian 362021, P.R. China

ABSTRACT

Speech recognition is one of the most convenient forms of human beings engaging in the exchanging of information. In this research, we want to make robots understand human language and communicate with each other through the human language, and to realize man–machine interactive and humanoid–robot interactive. Therefore, this research mainly studies NAO robots' speech recognition and humanoid communication between double-humanoid robots. This paper introduces the future direction and application prospect of speech recognition as well as its basic method and knowledge of speech recognition fields. This research also proposes the application of the most advanced method—establishment of the Hidden Markov Model (HMM) for the continuous word recognition in the speech recognition of NAO robots. In addition, this paper focuses on the establishment of a modelling algorithm and the extraction method of speech characteristics and the existing problems. Meanwhile, this research demonstrates the experiments of the NAO robot structured function and the voice interactive of double-humanoid robots. Through the use of an NAO-robot-controlled platform, Choregraphe software, and combined calling of Baidu speech recognition, the communication between double-humanoid robots has been achieved. Besides, a series of programming designs have realized the interactive functions such as NAO robots' daily dialogue and communication, arithmetic function (addition, subtraction, multiplication, and division), singing, making movement, and rhetorical-pattern dialogues. Finally, it shows the significance and contribution of research of interaction between double-humanoid robots.

KEYWORDS: Speech recognition, humanoid communication, voice interactive, Hidden Markov Model (HMM)

1 INTRODUCTION

A humanoid robot is a kind of intelligent robot that has the same shape as the human head, limbs, and trunk and can act and communicate like a human. The research of the humanoid robot (Yi et al., 2017) cannot exist without artificial intelligence (Matsuda, 2016), (Adawiah & Rahman, 2015) which is a branch of computer science. Research includes speech recognition (Chen, 2018), (Aldana-Murillo et al., 2018), (Chou & Nakajima, 2016), (Becerra, 2014), (Kim & Stern, 2016), image recognition, robot natural language processing, and so on. In the twenty-first century, with the increasing maturity and development of the artificial intelligence technology, which got a big breakthrough from deep learning seven years ago, we can take the deep learning as a super Excel form, put in a lot of data, and make a prediction, judgment, or classification. Speech is the most convenient interaction between human and machine (Sugiura & Zettsu, 2016), (Attawibulkul et al., 2017), (Zaman et al., 2017), (Ishi et al., 2014), and speech recognition contains great research value (Moubayed et al., 2013), (Cabibihan et al., 20113), (Meng et al., 2014), (Kumar et al., 2017).

At present, the simplest and most convenient way to exchange information is by way of voice for humans, but the voice application of intelligent robots is still relatively small. Through adding the speech recognition interface in the robot system (Miura et al., 2015), (Thoshith et al., 2018), (Achmad et al., 2016), (Miyanaga et al., 2013), (Moubayed et al., 2014), replacing the keyboard input with voice communication, then by the network interface, the

CONTACT Li-Hong Juang lipuu@qq.com

robot can be connected to the cloud to achieve human–computer interaction; therefore, the robot can not only understand a language but also give an answer. In recent years, the development of speech recognition technology has been very active. There also is a built-in voice recognition system in NAO robots (http://doc.aldebaran.com/2-1/family/robots/dimensions_robot.html), and by connecting to the cloud, the robot can achieve true intelligence. Any interaction between robots can be achieved by voice, and the development of the speech recognition technology promotes the development of artificial intelligence. In the meantime, the development of the intelligent robot also pushes a rapid development of speech recognition technology.

With the rapid development of science and technology and more and more usage of this technology in industry, education, business, medical, military, and other robotics fields, the speech recognition technology as one of the key research contents will have a greater impact. In addition, scientists not only do research on voice recognition but also on the more popular NAO robots—humanoid robots (https: // community Aldebaran. com) —that can add, subtract, multiply, divide, sing, act, and communicate on a given topic. The speech recognition technology, as a technology that translates the voice signal into a corresponding text or command, is also called automatic speech recognition. The three aspects of speech recognition technology mainly include: feature extraction, pattern matching, model training. The speech recognition tasks can be divided into three types in accordance with the different identifying objects: isolated word recognition, keywords recognition, continuous speech recognition. The speech recognition includes two stages, which are training and recognition. First, the recorded speech should be carried out via preprocessing and feature extraction, and second, be trained and recognized (Chou & Nakajima, 2016), (Becerra, 2014). Training refers to the work of various voice training, and extraction refers to the extraction of a characteristic parameter. Establishing a voice training library (Moubayed et al., 2014) is building a reference model through features, and outputting the characteristics of the reference model and the speech vector parameters to compare their similarity, then the output with the highest recognition is used as the recognition result, and the goal of speech recognition is finally realized. The NAO robots understand Human action and speeches by using visual and voice detection and recognition. Our research's major differences with the above researches are on the robot self and robot-robot recognizing each other speeches. The results can also apply for the medical therapy on human attention training.

To be more refined, in this research, we aims to achieve the effect where: Two robots' daily dialogue and communication, arithmetic function (addition, subtraction, multiplication, and division), singing, making movement, and rhetorical-pattern dialogues In this paper, the process by which we convert the collected audio into text is called speech recognition. This process includes the collection of speech signals, the preprocessing of speech signals and the feature extraction of speech signals, and then the extraction of the features as our training template into the implementation of a set of voice library for matching comparison. Comparing similarity, if the similarity is higher, then it can be identified and output the comparison results. The similarity low will not be unrecognized.

This paper is organized as follows: for the first section, the basic math operation which is shown in II. Methods. In the experiment of robot speech identification and communication, we tries to achieve the voice mainly by filtering the noise with the following various algorithms to get the recognized voice strings to realize the feather representation accordingly.

For the second section, in the experiment of robot speech identification and communication, we realized the judgment on voice through searching the optimal algorithms in the methods library to realize identification and communication through two robots voice interaction. The logic detection experiment then uses the arithmetic function to complete the deduction between the different digital values. Finally, we made some conclusions and analysis.

2 METHODS

THERE are the three main steps of speech recognition: (1) the signal acquisition, processing, and feature extraction; (2) the model training and matching; and (3) the correction and evaluation through the existing knowledge. The relative formulas of speech recognition process which will form as the existing speech recognition Application Program Interface (API) in this research are shown in the following statement:

2.1 Hidden Markov Model

The Hidden Markov Model is called HMM for short, the application and promotion of which is the most important achievement in the field of speech recognition since the 1980s. HMM uses the connotative state to deal with the various acoustic pronunciation units. At the same time, it also introduces some probabilistic model statistics and utilizes the probability density to calculate voice parameters and determine the output probability by searching for the best status switch instead of using the dynamic time alignment method. HMM takes the examined maximum probability as a benchmark to look for the experimental results.

HMM is a statistical model with n states, S1, S2..., Sn output symbol sequence, changes from one state into another state by one cycle, and a corresponding

symbol output by it with another state, which is determined by the transition probability. We can determine its transfer status only by observing the output symbols, but not directly observe its transfer status, so the Markov model is hidden.

1). Forward—Backward algorithm

Forward algorithm

Define the forward variable as:

$$a_1(\mathrm{i}) = \mathrm{P}(o_1, o_2, \cdots o_t; q_t = s_i \mid \lambda), 1 \le \mathrm{t} \le T \tag{1}$$

The process of the forward algorithm is as follows:

Initialization:

$$\boldsymbol{a_1(i) = \pi_i b_i(o_1), 1 \le i \le N} \tag{2}$$

Recursive

$$a_{t+1}(\mathrm{j}) = \left[\sum_{i=1}^{N} a_t(\mathrm{i}) a_{ij}\right] b_{ij}(o_{t+1}), 1 \le t \le T-1, 1 \le j \le N \tag{3}$$

End:
$$\mathrm{p}(\mathrm{O} \mid \lambda) = \sum_{i=1}^{N} a_T(\mathrm{i}) \tag{4}$$

Backward algorithm

Define the backward variable as:

$$\beta_t(\mathrm{i}) = \mathrm{P}\left(o_{t+1,} o_{t+2,\cdots,} o_{t5} \mathrm{I} q_t = S_{\mathrm{i}}, \lambda\right), 1 \le t \le T-1 \tag{5}$$

The algorithm process is:

Initialize:
$$\beta_r(\mathrm{i}) = 1, 1 \le \mathrm{i} \le \mathrm{N} \tag{6}$$

$$\beta_t(\mathrm{i}) + \sum_{j=1}^{N} a_{ij} b_{ij}(\mathrm{o}_{t+1}) \beta_{t+1}(\mathrm{j})$$

Recursive: $\mathrm{t} = \mathrm{T}-1, \mathrm{T}-2, \cdots, 1; 1 \le i \le N$ (7)

End:
$$P(O \mid \lambda) = \sum_{i=1}^{N} \beta_1(\mathrm{i}) \tag{8}$$

Probability can also be calculated before and after variables

$$P(O \mid \lambda) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_t(\mathrm{i}) a_{ij} b_{ij}(\mathrm{o}_{t+1}) \beta_{t+1}(\mathrm{j}) \tag{9}$$

$$= \sum_{i=1}^{N} a_t(\mathrm{i}) \beta_t(\mathrm{i}) \tag{10}$$

2). Viterbi algorithm

The calculation procedure for solving the best condition is as follows:

Initialize:
$$\delta_1(i) = \pi_i b_i(\mathrm{o}_i), 1 \le i \le N \tag{11}$$

$$\Psi_1(i) = 0, 1 \le i \le N$$

Recursive:

$$\delta_{\mathrm{t}}(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right] b_{ij}(o_t), 2 \le t \le T, 1 \le j \le N \tag{12}$$

$$\Psi_t(i) = \arg\max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right], 2 \le t \le T, 1 \le j \le N \tag{13}$$

End:
$$P^* = \max_{1 \le i \le N} \left[\delta_T(i)\right] \tag{14}$$

$$q^*_T = \arg\max_{1 \le i \le N} \left[\delta_T(i)\right] \tag{15}$$

Path backtracking (determination of optimal state chain):

$$q^*_t = \Psi_{t+1}(q^*_{t+1}), t = T-1, T-2, \cdots, 1 \tag{16}$$

3). Baum–Welch algorithm

First defines two variables, define:

$$\gamma_t(i) = P(q_t = S_i \mid O, \lambda) = \alpha_t(i) \beta_{\mathrm{t}}(i) / \sum_{i=1}^{N} \alpha_t(i) \beta_t(i) \tag{17}$$

For the given model λ and the observation sequence O. r_t (i) is a probability measure that is bound to be satisfied

$$\sum_{i=1}^{N} \gamma_t(i) = 1 \tag{18}$$

Redefine

$$\xi_t(i, j) = P(qt = S_i, q_{t+1} = S_j \mid \mathrm{O}, \lambda) \tag{19}$$

The probabilities of being in the state Sj at time t +1 with forward and backward variables have:

$$\xi_t(i, j) = \alpha_t(i) \mathrm{a}_{ij} b_{ij}(O_{t+1}) \beta_{t+1}(j) / P(O \mid \lambda)$$
$$= \alpha_t(i) \mathrm{a}_{ij} b_{ij}(O_{t+1}) \beta_{t+1}(j) / \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) \mathrm{a}_{ij} b_{ij}(O_{t+1}) \beta_{t+1}(j) \tag{20}$$

According to the above definition, the following relationship can be seen:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i, j) \tag{21}$$

For left-to-right HMM, the Viterbi algorithm not only gets the score of the test tone, but also divides the class. Finally, the Viterbi algorithm can divide and rank the test tone in accordance with the left-to-right results.

The following is a left-to-right continuous HMM training process of the N state and M substate:

1. Selecting the initial parameter of the model.

2. The signature sequence of the Viterbi algorithm includes each training sound of the model. According to the classification of chronological order, the characteristic parameter of the conclusion is generated from the HMM, through the Viterbi algorithm the optimal solution can be obtained, and then the characteristic parameter can be assigned to the corresponding state. After that, the segmentation of the HMM phonetic feature sequence can be completed.

3. The characteristic vector belongs to each state Sj can be divided into the M class, thus the Gauss distribution center μ_{ijm} and variance $\sum_{ijm}$ of each state Sj can be achieved; the new HMM parameters can be figured out in accordance with Baum–Welch algorithm, and steps (2) (3) cannot stop until convergence of the HMM parameters.

2.2 Mel-Scale Frequency Cepstral Coefficient

In speech recognition and speaker recognition, the most commonly used speech feature is the Mel-scale frequency cepstral coefficient (MFCC). The reason this parameter can be suitable for the speech recognition is because it takes into account that human ears have different feelings of different frequencies. The following is a brief introduction for the process of solving the MFCC.

1). Pre-emphasis

The speech signal will pass through a high-frequency filter.

$$H(z) = 1 - a * z - 1 \tag{22}$$

Among the coefficients, a is between 0.9 and 1. If the focused signal can be expressed S2 (n) with a time-domain working equation, it will be:

$$s2(n) = s(n) - a * s(n-1) \tag{23}$$

This is the compensation of the high-frequency part of depression to achieve the vocal signal, which can be used to eliminate the effects of vocal cords and lips. It can also be the resonance peak that highlights the high-frequency part.

2). Frame blocking

We determine an observation unit to collect N sampling points, which is called "frame." Usually the N value is 256 or 512. To avoid the sudden and large change between them, we set the overlap area between two adjacent frames, in which there are M sampling points. Generally speaking, their value is 1/3 or half of that of N. The sampling frequency used for the speech recognition is generally 8,000 Hz or 16,000 Hz.

3). Hamming window

To increase the right-and-left continuity of the frame, each frame can be multiplied by a Hamming window. Assuming that the frame signal is S (n), N = 0..., N−1. Then, multiply the Hamming window:

$$S'(n) = S(n) * W(n) \tag{24}$$

The W (n) form as follows:

$$W(n, a) = (1 - a) - a\cos(2pn/(N-1)), 0 \leqq n \leqq N-1 \tag{25}$$

The different a values will get different Hamming windows. Usually we take $a = 0.46$.

4). Fast Fourier transform

Because it is difficult to see the characteristics of change of signal in the time domain, we usually convert it into the energy of the frequency domain to observe. Thus, after the windowing processing, which must be carried out, the Fourier transform strengthens the right-and-left and continuous windowing. Due to the Fourier transform and the observation of different energy distributions, the different phonetic characteristics can be observed. If there is no periodic signal, two Fourier transforms will produce some non-existent energy to conform to the right-and-left continuity, which results in the mistakes in our observation. Of course, if we can make the signal in frame contain the integer multiples of a basic cycle when taking frames, then the left and right terminal of the frame will be continuous, and we don't need the Hamming window. But in fact, our basic cycle usually does not realize the integer multiple, so we still need to achieve continuity by windowing and have a high reliability in the observed results.

5). Triangular band-pass filters

The relationship between the Mel frequency and the general frequency is:

$$mel(f) = 2595 * log10(1 + f/700) \tag{26}$$

or

$$mel(f) = 1125 * ln(1 + f/700) \tag{27}$$

The Mel frequency is the feeling of ordinary human ears to frequency. Equations (26) and (27) are designed as the voice range which robot can recognize. In the low-frequency part, the feeling is more sensitive; in the high-frequency part, the feeling is rougher.

The functions of the triangular band-pass filter are the following:

- Smooth the spectrum
- Reduce the data volume

6). Discrete cosine transform (or DCT)

Take the above logarithmic energy, Ek, in the formula to figure out the L rank MFCC. The DCT formulation is:

$$Cm = Sk = 1N\cos\left[m * (k - 0.5) * p/N\right] * Ek,\ m = 1, 2, \ldots, L \tag{28}$$

Ek is the inner product value calculated by the former step; the number of triangle filters is N. The

reason for taking the DCT conversion is because we want to return the similar time domain, which is also called the frequency domain, actually, cepstrum. Also because Mel-frequency is used to convert to the Mel frequency, we call it Mel-scale cepstrum.

7). Logical energy

An important feature of voice is the volume of a frame (i.e., energy). And it is very easily calculated.

8). Difference cepstrum

In the practical application of MFCC in the speech recognition, we usually take the measure of adding delta cepstrum to show the change of delta cepstrum in terms of time. For the time slope, delta cepstrum represents the dynamic change in time. The formulation is as follows:

$$Cm(t)=[St=-MMCm(t+t)/t] /[St=-MMt2] \quad (29)$$

Here, the value of M is usually 2 or 3.

2.3 Feature Template Training Method

Through the feature template matching for voice recognition, we need to go through a large number of users to read aloud to training, and then generate a template library. The more times the same vocabulary is trained, the richer the feature template is and the better the recognition effect. Currently, there are three more training methods:

1). Accidental training method

This is a speech recognition system suitable for a specific person or a relatively small number of vocabularies. We can read the vocabulary many times and then generate a feature template based on the results of each reading. And then through the Dynamic Time Warping (DTW) algorithm you can determine the degree of distortion with each template and finally get the smallest distortion of the template, so as to identify it. This method is relatively simple, but the voice feature parameters are very large. The training time may produce errors, so this method also has a great limitation.

2). Robust training method

This training method takes into account the consistency of training. It is mainly a number of training for each word, and then hangs the best training sequence. Finally, the DTW algorithm is used to obtain a better performance template.

2). Non-specific identification—clustering method

When we want to get a higher recognition rate for non-specific people to identify, we need to conduct multiple sets of training and talk about training data classification.

3 EXPERIMENTS

IN this research, Two Nao robots were used to realize the smart communication between them. The humanoid robot NAO [5-6], designed and developed by Aldebaran Robotics company, has the open source framework and is very suitable for research and education. Nowadays, the NAO robot has been applied in more than 500 universities in the world, and there are many universities in China that purchased NAO robots for studying and developing. The NAO robot is also applied at some technology companies because it is the world's most intelligent open source robot. The robot is very suitable for a variety of applications on the market, because it has a good and broad interface. The company also provides a comprehensive programming environment, and the robot can be programmed through Python and C++. The company has also launched the visual programming software Choregraphe, which is simple and intuitive and suitable for most people. For people who have basic phonetics, a good open-source framework can be developed further.

NAO robots use not only wired networks but also wireless networks. In addition, robots can communicate with each other through infrared, wireless networks, microphones, and so on.

3.1 General Characteristics and Configuration of NAO Robots

The NAO robot's weight is 4.3 KG, height is 57.3 cm, and weight is 27.3 cm. The NAO robot is made by high-tech special plastic, and there is a built-in 21.6 V battery, which can be used for about 1.5 hours with full electricity. Most of the time, the NAO robot can be used when in charge, and there are many sensors, such as an ultrasonic sensor, gravity sensor, and light sensor, and multimedia, such as a microphone and camera, installed in its body.

The system used by the NAO robot is the Gentoo Linux operating system. All operations of the NAO robot can be programmed with naoqi architecture. The interaction management of a system user can deliver information through a Choregraphe, Monitor, Motion Module, or Audio module. Implementing naoqi delivers information and command with Broker.

There is an embedded system in the head of an NAO robot that can control the robot. And there is a microcontroller in its chest to control the power source and the engine. The embedded system in its head uses the embedded Linux (32 bits x 86 ELF), and its H/W is composed of x86 AMD GEODE 500MHz CPU, 256MB SDRAM, and flash. The NAO robot also supports the network connections of ethernet (cable) and Wi-Fi (wireless, IEEE 802.11g).

3.2 Audio System

The NAO robot microphone configuration is shown in Figure 1. There are usually four microphones in the NAO robot, including one in forehead, one in the occiput, and two in both sides of ears. Through the speaker it can play music and read out the written text. The microphone can capture audio and carry out the positioning of the sound source.

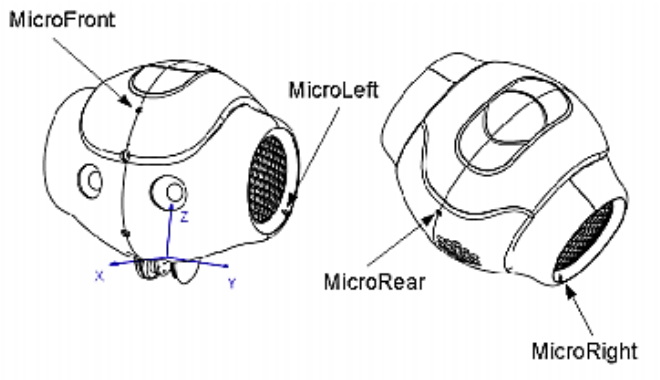

Figure 1. Nao structure system figure.

Through the description of the audio structure system of the NAO robot, the WAV file is captured through the four channels in the front, left, and right. If the files of the four audio tracks are handled, it will greatly increase the processing workload and consume a longer time. In fact, there is always one closest sound source in the four audio files recorded by the four microphones of the NAO robot. The best selection of this track channel is significant for subsequent identification.

3.3 Choregraphe Software

Choregraphe software is specially developed by Aldebaran Robotics company for the NAO robot and Pepper robot development as shown in Figure 2. In the Choregraphe software, we can write a python program; connect NAO real robot and virtual robot.

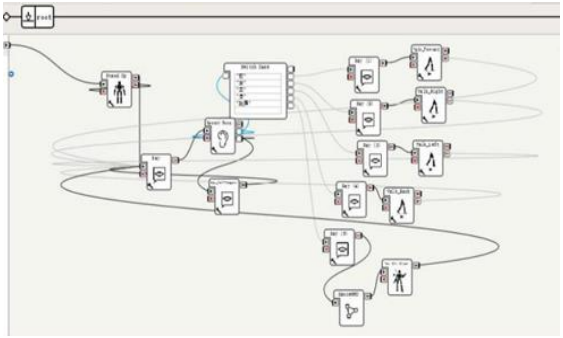

Figure 2. Choregraphe main interface.

It needs to note when using Choregraphe for NAO programming:

1. To prevent the motor from overheating, we should unload the motor stiffness when the robot is idle.
2. Punctuation must be in English.
3. Develop a habit of keeping the programs at any time.
4. If the installed programs are showed in the lower right corner of the software, we can delete or import operation in the graphical interface.
5. When the Choregraphe software is inputting a program to the robot, its version must be the same with naoqi.

3.4 The Control and implementation of Speech Recognition in the NAO Robot

On the basis of the research of the speech recognition system of network transmission, the existing speech recognition API has been used. Through the Secure File Transfer Protocol (SFTP) protocol model, the collected voice signal by the robot can be transmitted to the server to recognize and process, then feedback to the robot, which can be played by the robot microphone to realize voice communication.

The establishment and configuration procedures for calling recognition of voice recognition, the NAO microphone to record the sound and the function for the signal processing are shown in the following pseudo code implementation.

```
If head touched:
  Start listen
Get Record From Nao (IP, PORT):
Upload record file to server
A series of functions that directs NAO to act
If voice recognized:
  Searching the optimal algorithms in the methods library
Analysis the result and find the best answer in the corpus
From text to speech, voice synthesis
  Speak the answer
```

It will need to deal with the corresponding voice signal function in the compiled program as shown in the following command. For example, to identify the results of an addition:

```
If result[i] == u 'plus':
Num1 = cn2dig (result[0: i])
Num2 = cn2dig (result[i + 1:])
Result = num1 + num2
Print "% s +% s =% s"% (num1, num2, result)
```

3.5 Experimental Results and Analysis

1. In the interaction of double-humanoid robots, the communicated content can be imported into the Red robot, then the program that was debugged in advance can be written into the Blue robot to complete the dialogue. The contents of the program includes the voice signal collected by the robot, which can be delivered to the server to recognize and process by SFTP, then the result can be fed back to the robot and played by the microphone as shown in Figure 3.

2. The dialogue of the calculation of add, subtract, multiply, and divide is also similar with the above. The calculated content can be imported in the Red robot, then the program that was debugged in advance can be written into the Blue robot to complete the dialogue. The character string of the program is processed by the Baidu voice and API speech recognition, then the result can be fed back to the robot and played by the microphone as shown in Figure 4.

3. The interaction of body movements, the movement made by Red, can be written into the Blue robot and set on a delay program. Then a code that was used to implement the movement can be written into the Red robot, which can make the corresponding movements when it hears the voice signal form Blue as shown in Figure 5.

4. This is the interaction of instructing the other to sing and giving evaluation. This experiment can be achieved through the visual program of the instruction box in Choregraphe software as shown in Figure 6.

5. The interaction of rhetorical form is also by the SFTP protocol model. The voice signal collected by the robot can be delivered to the server through Baidu speech to recognize and process. Then semantic identification can be carried out in the Turing machine and output to the robot, then played by microphone to achieve voice communication function as shown in Figure 7.

This experiment mainly completed the control of the interaction and communication of the double-humanoid robot voice. The naoqi and Choregraphe software that controlled the environment of the NAO robot were allocated in the computer. For the realized model, programming was implemented, and some Python header files were installed in the process of debugging code. In the recognition module, the Baidu speech recognition API was the main tool. After the register of Baidu speech recognition, App ID, API Key, and Secret Key can be obtained, which will be used in debugging. Through these open-source speech recognition platforms, the process of training templates and building a voice library by ourselves can be eliminated. Their recognition effect and the interface function with the Turing machine are very mature and, through understanding the semantics, can achieve real intelligence.

We took the Choregraphe instruction box to complete some models, which can carry out the visual programming to an NAO robot. In the official document, we find some Python codes used to control the robot's behavior, which can be placed in the command box and let the robot make the corresponding behavior through debugging and running. Finally, the code programming and visual software programming achieves five interactive models, including daily dialogue; question-and-answer mode; dialogue of add, subtract, multiply, and divide; rhetorical questions; and control and evaluation of the robot's movements and singing.

After preliminary statistics, a comparison chart of the five pattern recognition rates was obtained as shown in Table 1.

The speech recognition technology is a key part of researching the artificial intelligence robot, and the artificial intelligence plays a promoting role in the industrialization of the service robot. In the future, workshop production workers will be replaced by intelligent robots, and humans only need to manage and control the robots in the background, so the communication and interaction with robots is essential to achieve the production. While the interactive behavior with any robot should be achieved through voice, the development of speech recognition technology promotes the development of artificial intelligence. Meanwhile, the development of intelligent robots pushes the rapid development of speech recognition technology. If the service robot wants to be promoted, it must have the ability to replace humans and also have a high intelligence. The industrial service robot is still in development in accordance with the current application maturity and market space. The success rate was calculated by the success times divided by total experiment times. Each core function has its different parameters, the parameter value also affects the cognitive accuracy, and it can test multiple core functions to allow the better cognitive accuracy reaching the highest. The special point of the proposed work compared with the existing algorithm is quick response, time consumption is low and real time. To compare the performance of the NAO robot to other robots based on this topic, it can find that the Nao robot is easy to realize the actual human visual communication. In the future, the research of voice interaction and communication of double-humanoid robots will have a profound impact on the application in production.

Figure 3. The Video 1 for daily dialogue.

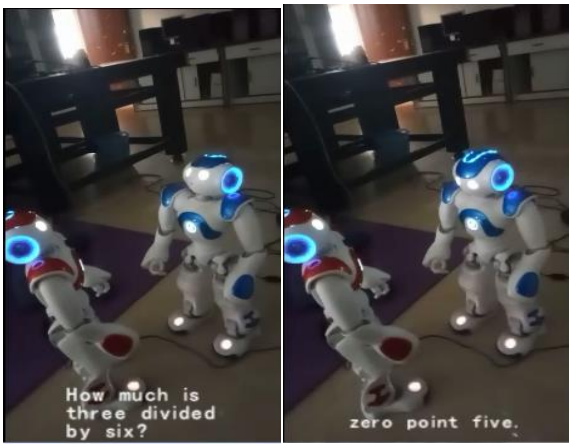

Figure 4. The Video 2 for calculation dialogue.

Figure 5. The Video 3 for body movements.

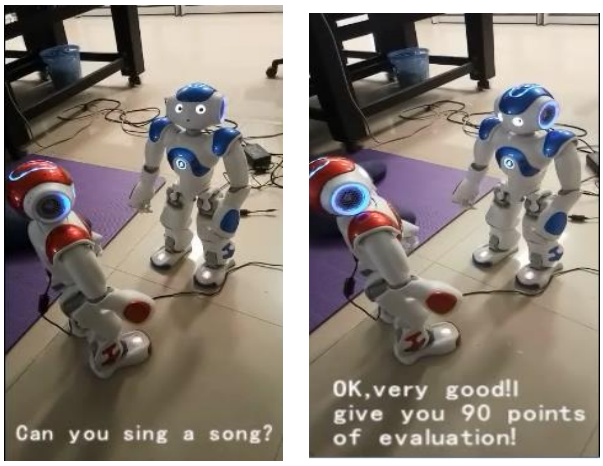

Figure 6. The Video 4 for singing and giving evaluation.

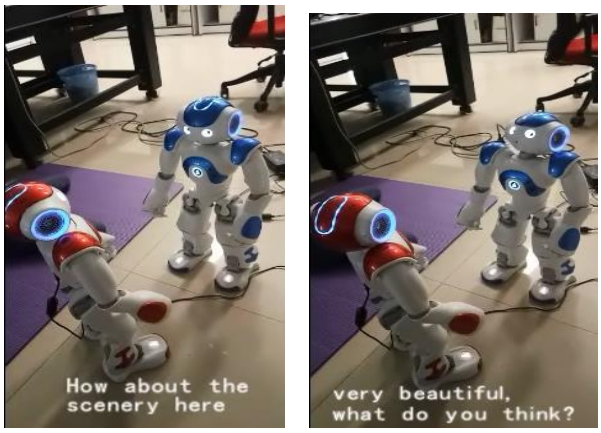

Figure 7. The Video 5 for rhetorical question.

Table 1. A comparison chart of the five pattern recognitions rates

Mode	Recognition rate (%)
Daily dialogue	92
Calculation dialogue	95
Body movements	97
Sing and giving evaluation	100
Rhetorical question	89

4 CONCLUSIONS

IN this research, for the realized model, the programming design is carried out, and the Baidu speech recognition API is also used. Through the open source speech recognition platform, the process of training templates and building a voice library by ourselves can be eliminated, and the semantics can be clarified by connecting with the Turing machine to achieve intelligent interactive communication. The Choregraphe software also can carry out the visual program of the robot's behavior and action. The experimental results are relatively good, to achieve a certain degree of interactivity. In the official document, the relevant codes can be found and placed in the command box, and the robot can make the corresponding behavior through debugging and running.

The development of speech recognition technology mainly lies in the improvement of algorithms. At present, the simplest and most convenient way to exchange information is via voice for humans. Through adding the speech recognition interface in the robot system, replacing the keyboard input with voice communication, then by the network interface, the robot can be connected to the cloud to achieve human–computer interaction; therefore, the robot not only can understand a language but also give an answer. This technology sounds very attractive, so the speech recognition technology is the key point of intelligent robots. We also hope to apply it to the medical therapy on human attention training in future.

5 ACKNOWLEDGEMENT

THE authors deeply acknowledge the financial support from Xiamen University of Technology, Fujian, P.R. China under the Xiamen University of Technology Scientific Research Foundation for Talents plan.

6 REFERENCES

Achmad, M.; Muttaqin, R.; Sumpeno, S. (2016), "Implementation of Face Detection and Recognition of Indonesian Language in Communication between Humans and Robots," International Conference on Information & Communication Technology and Systems (ICTS), pp. 53-57.

Adawiah, R.; Rahman, A. (2015), "Use of Humanoid Robot in Children with Cerebral Palsy: The Ups and Downs in Clinical Experience," IEEE International Symposium on Robotics and Intelligent Sensors, pp. 115 – 118.

Aldana-Murillo, N. G.; Hayet, J. B.; Becerra, H. (2018), "Comparison of Local Descriptors for Humanoid Robots Localization Using a Visual Bag of Words Approach," Intelligent Automation and Soft Computing, VOL. 24, NO. 3, JUNE, pp. 471-481.

Aldebaran Community Website [EB / OL]. (2018), Available at https: // community Aldebaran. com, Accessed 11 October 2018.

Attawibulkul, S.; Kaewkamnerdpong, B.; Miyanaga, Y. (2017), "Noisy Speech Training in MFCC-Based Speech Recognition with Noise Suppression toward Robot Assisted Autism therapy," Biomedical Engineering International Conference (BMEiCON) 10th, pp. 1-5, 2017

Becerra, H. (2014), "Fuzzy Visual Control for Memory-Based Navigation Using the Trifocal Tensor", Intelligent Automation and Soft Computing, VOL. 20, NO. 2, March, pp. 245-262.

Cabibihan, J. J.; Javed, H.; Jr, M. A.; Aljunied, S. M. (2013), "Why Robots? A Survey on the Roles and Benefits of Social Robots in the Therapy of Children with Autism," International Journal Social Robotics, VOL. 5, No. 4, pp. 593-618, 2013.

Chen, M. Y. (2018), "The SLAM Algorithm for Multiple Robots Based on Parameter Estimation," Intelligent Automation and Soft Computing, VOL. 24, NO. 3, JUNE, pp. 593-607.

Chou, Y. C. (2016); Nakajima, M., "Particle Filter Planar Target Tracking with a Monocular Camera for Mobile Robots," Intelligent Automation and Soft Computing, VOL. 24, NO. 3, JUNE, pp. 117-125.

Ishi, C.T.; Ishiguro, H.; Hagita, N. (2014), "Analysis of Relationship between Head motion Events and Speech in Dialogue Conversations," Speech Communication VOL. 57 pp. 233–243.

Kim, C.; Stern, R. M. (2016), "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," IEEE/ACM Transactions on Audio Speech and Language Processing, VOL. 24, NO. 7 JULY, pp. 1315-1329.

Kumar, P.; Nagapushpa, K.; Raj, V.; Kumar, V.; Kumaraswamy, R. (2017), "Multi-Functional Intelligent Humanoid Using Speech Processing," International Conference on Communication and Signal Processing (ICCSP), pp. 0244 – 0248.

Matsuda, Y. (2016), "Teaching Interface of Finger Braille Teaching System Using Smartphone," International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) 8th, pp. 115 – 118.

Meng, X.; Liu, C.; Zhang, Z.; Wang, D. (2014), "Noisy Training for Deep Neural Networks," IEEE China Summit & International Conference on Signal and Information Processing 4th, September, pp. 16-20.

Miura, Y.; Yihao, H.; Kuchii, S.; Sern, G. C. (2015), "Research and Development of the Social Robot Using Speech Recognition and Image Sensing Technology," International Conference on Information Technology and Electrical Engineering (ICITEE) 7th, pp. 66-69.

Miyanaga, Y.; Takahashi, W.; Xihao, S. (2013), "Robust Speech Communication and its Embedded Smart Robot System", International Conference on Systems, Signals and Image Processing (IWSSIP) 20th, pp. 151-154.

Moubayed, S. A.; Skantze, G.; Beskow, J. (2013), "The Furhat Back-Projected Humanoid Head - Lip reading Gaze and Multiparty Interaction," International Journal of Humanoid Robotics, VOL. 10, NO. 1, pp. 1793-6942.

Moubayed, S. A.; Beskow, J.; Bollepali, B.; Gustafson, J.; Hussen-Abdelaziz, A.; Johansson, M.; Koutsombogera, M.; Lopes, J.D.; Novikova, J.; Oertel, C.; Skantze, G.; Stefanov, K.; Varol, G. (2014), "Human-Robot Collaborative Tutoring using Multiparty Multimodal Spoken Dialogue", ACM/IEEE International Conference on Human-Robot Interaction (HRI) 9th, pp. 112-113.

Moubayed, S. A.; Beskow, J.; Skantze, G. (2014), "Spontaneous Spoken Dialogues with the Furhat Human-like Robot Head", ACM/IEEE International Conference on Human-Robot Interaction (HRI) 9th, pp. 326-326.

NAO-technical overview (2018), Available at http://doc.aldebaran.com/2-1/family/robots/dimensions_robot.html, Accessed 10 December 2018.

Sugiura, K.; Zettsu, K. (2016), "Analysis of Long-Term and Large-Scale Experiments on Robot Dialogues Using a Cloud Robotics Platform," ACM/IEEE International Conference on Human-Robot Interaction (HRI) 11th, pp. 525 – 526.

Thoshith, S.; Mulgund, S.; Sindgi, P.; Yogesh, N.; Kumaraswamy, R. (2018), "Multi-Modal Humanoid Robot," International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 1-4.

Yi, C. A.; Min, H.; Zheng, G. (2017), "Affordance Learning and Inference Based on Vision-Speech Association in Human-Robot Interactions," IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 743 – 749

Zaman, H. U.; Khisha, J.; Zerin, N.; Jamal, M. H. (2017), "Speech Responsive Mobile Robot for Transporting Objects of Different Weight Categories," International Conference on Advanced Robotics (ICAR) 18th, pp. 395-400.

7 DISCLOSURE STATEMENT

NO potential conflict of interest was reported by the authors.

8 NOTES ON CONTRIBUTORS

Li-Hong Juang received the B.S. degree in civil engineering from the National Chiao Tung University, Taiwan, in 1990 and the M. S. degree in applied mechanics from the National Taiwan University, Taiwan, in 1993, and Ph.D. degree in control and embedded system group from Department of Engineering at Leicester University, UK, in 2006. After his master's degree, he worked for machinery, electrical and computer industries for over ten years, then he went to the Department of Engineering, Leicester University, and began his PhD study for electrical engineering about motor design and control for three years. Finally, he published six SCI papers including two IEEE transaction journal papers. After finishing his PhD, he worked at several universities as an assistant professor to the chair professor for over another ten years. During that period, his research interest was focused on power systems, medical systems, system control, and AI robots. Now, he joins the School of Electrical Engineering and Automation, Xiamen University of Technology, and serves as a distinguished professor. He continues to make his contribution to the smart system application for engineering and science, especially for the computer vision, intelligent robot, medical system, and cloud computer platforms for the Internet of things. He has gotten sixteen project grants and published fifty-three SCI papers, twenty-five EI papers, six book chapters, one whole book, and twenty-four patents. Currently, he services as the Editor in Chief for Journal of Mechanical and Automation Engineering and the Associate Editor for IEEE Access Journal.

Yi-Hua Zhao received the B. S. degree in Department of IOT, National HuaQiao University, P. R. China in 2017. She research interests are in audio signal process.