IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 19, 2019, accepted December 22, 2019, date of publication January 1, 2020, date of current version January 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963560

Dietary Composition Perception Algorithm Using Social Robot Audition for Mandarin Chinese

ZHIDONG SU, YANG LI, AND GUANCI YANG
Key Laboratory of Advanced Manufacturing Technology of Ministry of Education, Guizhou University, Guiyang 550025, China

Corresponding author: Guanci Yang (gcyang@gzu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61863005 and Grant 91746116, in part by the Science and Technology Foundation of Guizhou Province under Grant PTRC[2018]5702, Grant QKHZC[2019]2814, and Grant [2018]5781, and in part by the Fund for Hundred High-level Innovative Scholars of Guizhou Province.

ABSTRACT As the problem of an aging population becomes more and more serious, social robots have an increasingly significant influence on human life. By employing regular question-and-answer conversations or wearable devices, some social robotics products can establish personal health archives. But those robots are unable to collect diet information automatically through robot vision or audition. A healthy diet can reduce a person's risk of developing cancer, diabetes, heart disease, and other age-related diseases. In order to automatically perceive the dietary composition of the elderly by listening to people's chatting, this paper proposed a chat-based automatic dietary composition perception algorithm (DCPA). DCPA uses social robot audition to understand the semantic information and percept dietary composition for Mandarin Chinese. Firstly, based on the Mel-frequency cepstrum coefficient and convolutional neural network, a speaker recognition method is designed to identify speech data. Based on speech segmentation and speaker recognition algorithm, an audio segment classification method is proposed to distinguish different speakers, store their identity information and the sequence of expression in a speech conversation. Secondly, a dietetic lexicon is established, and two kinds of dietary composition semantic understanding algorithms are proposed to understand the eating semantics and sensor dietary composition information. To evaluate the performance of the proposed DCPA algorithm, we implemented the proposed DCPA in our social robot platform. Then we established two categories of test datasets relating to a one-person and a multi-person chat. The test results show that DCPA is capable of understanding users' dietary compositions, with an F1 score of 0.9505, 0.8940 and 0.8768 for one-person talking, a two-person chat and a three-person chat, respectively. DCPA has good robustness for obtaining dietary information.

INDEX TERMS Dietary composition perception, semantic understanding, robot audition, social robot, text information extraction.

I. INTRODUCTION

With the progress of artificial intelligence technology, social robot technology is developing at an unprecedented speed [1]. In 2016, it was reported that about 59.71 million social robots poured into people's homes. What is more, in 2016, the International Federation of Robotics predicted that more than 420,000,000 social robots will become a member of the family before 2019 [2], which means that social robots will have an increasingly significant influence on human life.

As we know, the problem associated with the aging of the population is becoming more and more serious. Namely, the population size of the elderly will soar, and more and more elderly people need to be offset by younger ones to take care of them [3]. Meanwhile, people generally hope that robots can take up some human work. Thus, a large number of scientists focus on the development of social robots to obtain more intelligent machines. In effect, it is reported that modifications of the diet reduced the study participants' risk of developing cancer, diabetes, heart disease and other age-related diseases [4]. Eating food scientifically and rationally is not only beneficial to health, but has a positive effect on the treatment of diseases [5]. With the improvement of people's health awareness, a healthy diet has attracted people's attention.

While current robot products provide the function of chatting, household appliance control, taking photographs, collecting health data, detecting falls, and so on, they have a

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Anwar Hossain.

shortage of intelligence and do not have a strong ability in cognition. Some works provide a method for judging if people are eating while speaking and what they are eating by monitoring eating sounds [6], [7], but in the home environment, most eating sounds are difficult to capture using robots, and it is also hard to distinguish between similar sounds, such as drinking water, drinking milk or drinking beer. To the best of our knowledge, no social robot is equipped with the function of figuring out what food is eaten by the elderly. Some products can observe the mental state of the elderly through regular question-and-answer conversation and can then establish health archives [8], but they do not use the sensors to collect diet information automatically. The achievement of the latter relies on the question-and-answer solution. However, most social robots are equipped with a camera or audio equipment, sometimes even a deep camera, which means that the robots have the potential to know the diet of the elderly through robot vision or audition. Robots can perceive most of the information of the physical world through vision or hearing, for robot vision, it consumes a lot of computing resources to realize real-time perception, but embedded devices have limited computing resources. While for robot audition, it is easy to realize real-time speech recognition and by monitoring people's conversion, it is a direct way to obtain eaten information through their talk.

Focusing on automatically understanding the dietary composition semantics of people by listening to people talking, this paper proposed a chat-based automatic dietary composition perception algorithm (DCPA) using social robot audition for Mandarin Chinese. If a doctor knows detailed dietary information of the patients, it will be beneficial in understanding the reasons for the occurrence of diseases and help doctors provide an efficient treatment solution. Otherwise, people can also adjust their daily diet schedule according to the collected dietary composition. The main contributions of this paper are summarized as follows:

- DCPA is proposed to offer an automatic method for acquiring users' dietary composition from their conversation, which provides a solution for mastering a person's dietary information in order to gain insight into the relationships between the occurrence of diseases and diets or to make an appropriate treatment prescription.
- According to the expression patterns of Mandarin Chinese, two kinds of automatic dietary composition semantics understanding algorithms are proposed to sensor food information from the conversation text.
- Two kinds of dialog corpuses are established to test the proposed algorithm. Otherwise, we implemented the proposed method in robot platform, and show that it is feasible to automatically perceive user dietary composition information.
- As we know, there is no report about the robot to automatically sense the dietary information. Although this research is language -oriented, this research is the first try to let the robot have the function of knowing user's dietary information.

The rest of this paper is organized as follows. Section II introduces the related works. Section III details the audio segment classification method. Section IV proposes the chat-based automatic dietary composition perception algorithm. Section V presents the hardware of the social robot platform and the software architecture of the proposed system. The training and test dataset, experimental results and analysis are presented in Section VI. Section VII summarizes this research and offers suggestions for future work.

II. RELATED WORKS

In recent years, many studies focus on social robots and their intelligent technology in relation to managing or assisting people in their daily lives. Some social robots have been put to commercial use, as Jibo [9], Aibo [10], Qrobot [11] and BaoBaoLong [12] explain. At the same time, the academic community have also done a lot of research on social robots. Yang *et al.* [13] proposed an improved neural network based on the YOLO model to detect embarrassing situations in a home environment and turn over the camera of the social robot to protect people's privacy. Do *et al.* [14] proposed the RiSH service robot platform for elderly fall detection and rescue, which integrated social robots, smart homes and body sensors. Zhou *et al.* [15] proposed a remote health-care system based on social robots, and this system employed a quick connection with families and doctors, automatic health data collection and object detection algorithms to achieve remote care. In order to enhance the physical fitness of the elderly and reduce the risk of falling, Foukarakis *et al.* [16] designed a robot vision system to identify and track users' behavior, and it is able to provide relevant exercises and feedback information to stimulate users. Wang *et al.* [17] proposed a home-auxiliary robot platform to help elderly and disabled people. This platform is able to control the social robot's movement, inside and outside of the field of vision, according to the obtained motion characteristics and eye movement features. Fernandes *et al.* [8] constructed a service robot platform for the cognitive assessment of the elderly. By analyzing the answers of the user, their cognitive status can be understood, which will provide a basis for doctors to diagnose the health of the user. To help the elderly live independently at home, Mast *et al.* [18] proposed a semi-autonomy strategy to solve the reliably-operated problem of social robots in the home environment. This strategy is able to connect different user groups, such as users, users' relatives, and professional telecom operators. Li *et al.* [19] proposed a human behavioral footprint-based approach to offer personalized services. This method employs an inverse reinforcement learning algorithm to learn human actions, and the social robot controls the indoor temperature automatically. A selective attention-directed active semantic cognitive algorithm in intelligent space is proposed to implement an efficient service [20], which enables robots to imitate human beings, perceive users and environments in unstructured home environments, and discover service tasks according to their cognition. In order to recognize the surrounding conditions cor-

rectly, Pyo *et al.* [21] proposed a social robot system with an informationally-structured environment, which enables the integration of various data from distributed sensors, to plan the robot's motion using real-time information about the surrounding environment. Those robots have various functions to assist the elderly in their daily life. However, they do not take into account the important impact of diet on the health of the elderly, and they cannot automatically acquire users' dietary data through the perception of sound and establish dietary information for users.

Text information extraction is to extract structured data from unstructured text [22], which can be useful in obtaining specific information from a large amount of text automatically. Recently, text information extraction has been thoroughly researched and used in different domains. In order to solve the problem that it is time-consuming and difficult to extract specially appointed information manually, Islam [23] developed a temporal information extraction system to find the temporal expressions from the textual data automatically, which used Long Short-Term Memory (LSTM), a recurrent neural network (RNN), along with word embedding and the existing annotated corpus QA-TempEva to train the model. When it comes to a specific area, there is no existing large-scale annotated corpus, and it takes time to annotate the corpus. Based on the electronic medical record (EMR) narrative text data from a tertiary hospital in China, Bao *et al.* [24] developed a corpus annotation platform to annotate the randomly selected electronic medical records and proposed regular expression extraction methods based on the rewritten extraction templates, which were summarized and induced based on the annotated corpora clinical text data. This method also requires a large number of tagged corpuses to summarize and generalize the extraction templates, which is time-consuming. Jonnalagadda *et al.* [25] also described a rule-based information extraction approach that automatically converts unstructured text into structured data to automatically extract the results of clinical trials. To deal with the mass of manual text and improve the knowledge acquisition speed, Wang *et al.* [26] proposed a method for extracting information from Chinese technical manual text based on rule-matching, and the extraction methods are based on domain ontology and syntactic analysis. Qizhi *et al.* [27] proposed a style and terminology-based method to extract emergency information, which extended the domain expert lexicon and used the lexicon to classify events and stylistic features to extract time information, and it combined lexical and stylistic features to extract location and casualty information. By using noise data filtering, named entity extraction, named entity disambiguation, and feedback loops, Habib *et al.* [28] established a mechanism for extracting information from Twitter texts containing informal grammar, short text, large amounts of irrelevant information, and uncertain content. Every subtask gives feedback to the preceding subtask, which allows for the possibility of iterations of refinement. When it comes to text information extraction, a rule-based information extraction approach is often proposed, based on the specific domain information, to extract structured data from unstructured text, but there is no suitable approach to extracting dietary information from conversation content.

Semantic parsing is the task of converting a natural language utterance to a logical form, which is a machine-understandable representation of its meaning. Semantic understanding is used to understand the intention or task-specific information of a natural language utterance. For semantic understanding, Hua *et al.* [29] claimed that short texts do not always obey the syntax of written language and do not have enough statistical signals to support many state-of-art methods for text processing, they used lexical semantic knowledge collected from YourDictionary and Probase [30], and proposed a text segmentation, type detection, and concept labeling framework to understand short texts. To induce and fill the semantic slot in an unsupervised way, Chen *et al.* [31] used a frame-semantic parser to parse the unlabeled ASR-transcribed data and employed a spectral clustering-based slot ranking model to align the output of the parser to the target semantic space. Most studies explored the sequence-to-sequence deep learning method for single domain, like slot filling or domain classification, Hakkani-Tur *et al.* [32] proposed an RNN-LSTM joint modeling approach to implement slot filling, intent determination, and domain classification for all user utterances. Different from many former semantic parsing methods which are based on the syntactic analysis of text, Grefenstette *et al.* [33] proposed a novel deep learning architecture for semantic parsing, it combines two neural models of language semantics. Without the need for parsing text, this architecture allows for the generation of ontology-specific queries from natural language statements and questions, which makes it especially suitable to grammatically malformed or syntactically atypical text.

While previous studies have enabled social robots to deal with speech interaction, semantic understanding and object detection functions to achieve the health management of the elderly and assist their lives, the important impact of their diet on their health has not been considered, and there is a lack of a method to record dietary information automatically. In addition, in previous studies, text information extraction was used to extract structured data from unstructured electronic medical records, user manuals, social media and other data to obtain interesting information based on the defined perception methods. In view of above problems, this paper uses the existing audition system of social robots to acquire the chat speech data of users and obtains text data from users' communication, then designs the semantic understanding algorithm to get dietary information from the converted text and obtains diet information concerning each user. This method will acquire dietary compositions automatically from chat speech data based on the audition of a social robot.

III. AUDIO SEGMENT CLASSIFICATION METHOD

As for dietary composition perception, the identity of different people should be used to preserve their dietary information and obtain the correct audio segment flow. Thus,

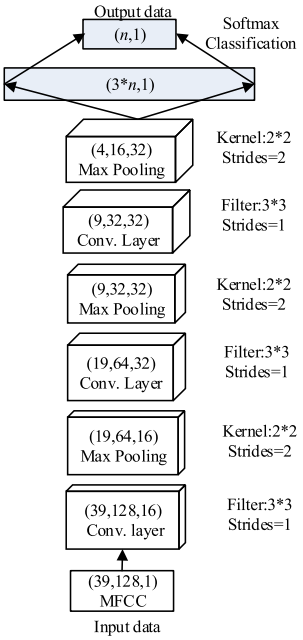

FIGURE 1. CNN structure of speaker recognition.

when it comes to the acquisition of speech information, especially multi-person dialogue speech data, the speech data are first segmented, then the identity of the speech segment is recognized, and the speech segments are grouped according to their identity. Finally, speech recognition is used to obtain the text information of the identified speech segments, preparing it for the subsequent dietary information extraction. In the following part, a speech segment classification method, combining speech segmentation and speaker recognition, is designed.

A. SPEAKER RECOGNITION METHOD BASED ON THE MEL-FREQUENCY CEPSTRUM COEFFICIENT AND CONVOLUTIONAL NEURAL NETWORK

Speaker recognition is employed to identify the identity of a person from the characteristics of speech data [34]. The identity is used to cluster audio segments and store identified food entities in this system. We designed a convolutional neural network structure to conduct speaker recognition. The 39-dimensional Mel-frequency cepstrum coefficient (MFCC) feature [35] is used as the input vector of the CNN network (MFCC-CNN). We tried different numbers of convolutional layers and the dimensions those layers and finally found out the architecture which has less parameters and good performance, and the architecture of the network is described as follows Figure 1, in where n is the number of people included in the training audio data. This neural network uses three two-dimensional convolutional layers and two fully connected layers, the dimension of the three convolution layers is 16, 32 and 32, each convolutional layer is followed by the Relu activation function, a max pooling layer and a batch normalization layer. The max padding length is set to 128, and the first fully connected layer, where L2 regularization is used, is followed by a batch normalization layer and a dropout layer. The last fully connected layer is followed by a softmax function to output the recognition result.

Algorithm 1 Audio Segment Classification Method Based on Audio Segmentation and the Speaker Recognition Algorithm

Input: conversation audio data and the speaker recognition model, which is obtained from the training audio data, including m people;

Output: connected audio segments set $C_a = \{p_1, p_2, \ldots, p_i, \ldots, p_n\}$, and the corresponding identity set $I_a = \{id_1, id_2, \ldots, id_i, \ldots, id_n\}$, where p_i is the connected audio segment set, according to the speaker, id_i is the corresponding identity of p_i, and n is the number of connected segments.

1) read the conversation audio data, which is in a wav format, and load the speaker recognition model;
2) obtain the audio segment set $S_a = \{S_{a1}, S_{a2}, \ldots, S_{aj}, \ldots, S_{aq}\}$ by applying the audio segmentation algorithm to the input conversation data, where S_{aj} is the audio segment, and q is the number of segments;
3) initialize set II_a as null, which is the set of identities of all the segments;
4) **For** each audio segment S_{aj}:
   employ the speaker recognition model presented in Section III-A to get the identified speaker id $k (k = 1, \ldots, m)$, assign $idd_j = k$, and $II_a = II_a \cup idd_j$;
5) initialize $p_j \in C_a$, $id_j \in I_a$ is null, set the temp variable $ID = idd_0$, $P = \emptyset$, and $j = 1$;
6) **For** each identity idd_i:
   a) **If** $ID == idd_i$ **Then** $P = P + S_{ai}$;
   b) **OTHERWISE** $p_j = P$, $P = \emptyset$, $id_i = ID$, $ID = idd_i$, and $j = j+1$;
7) output $C_a = \{p_1, p_2, \ldots, p_n\}$, and $I_a = \{id_1, id_2, \ldots, id_i, \ldots, id_n\}$.

B. AUDIO SEGMENT CLASSIFICATION METHOD BASED ON AUDIO SEGMENTATION AND THE SPEAKER RECOGNITION ALGORITHM

In this section, we designed an audio segment classification method based on audio segmentation and the speaker recognition algorithm (ASC-ASSR), which is used to distinguish different speakers in a speech conversation and store their identity information and the sequence of expression in the conversation. We firstly used the audio segmentation method based on Bayesian Information Criteria (BIC) [36] to segment the speech data, then utilized the MFCC-CNN method to identify the identities of the audio segments, and finally based on the identities and the sequence in the speech data, we concatenated the contiguous segments with the same identities. The method is detailed in Algorithm 1.

IV. CHAT-BASED AUTOMATIC DIETARY COMPOSITION PERCEPTION ALGORITHM

A. THE FRAMEWORK OF THE PROPOSED DCPA

To monitor and understand the dietary semantic information of family members, we propose a chat-based automatic dietary composition perception algorithm (DCPA) using social robot audition. Considering that the dietary text obtained from user's expression has many food names, such as fruit, vegetables, finished food, milk, meat and so on, it is difficult to segment the word correctly using only the general lexicon of the existent word segmentation system because of the variety of food types and food expression patterns. The incorrectness of word segmentation [37] will cause the incorrectness of part of speech tagging [38], and all those errors will eventually lead to an incorrect dietary information extraction. Thus, we built a custom dietetic lexicon, including 1563 commonly seen food names in daily life.

The sources of the diet lexicon include the following two aspects:

- Refer to the food classification system, provided by ''National Standard GB 2760-2014 of the people's Republic of China'' [39]. We grouped food names into about 15 categories, of foods that are commonly seen in daily life, including 826 food names;
- By using web crawler tools, we collected food names from food websites and the Midea ''CHINESEFOOD-NET'' dataset [40], including 737 food names that have no overlap with the former one.

A six-mic microphone array is used to capture speech data and suppress ambient noise, which still works within five meters, and a speech recognition system, provided by IFLYTEK Corp [41], is used to obtain text information from the audio segments, acquired from **Algorithm 1**. The speech recognition system is said to have an accuracy rate of 98% in its official website and can identify 17 Chinese dialects, the input audio format is WAV, the sound channel is mono, and the sampling rate is 16,000 Hz.

In order to better identify dietary information from the text information, we combine the LTP [42] general lexicon and the self-collected dietetic lexicon to form a dietary word dictionary, which is used in the Chinese word segmentation of our research.

Thus, to intelligently understand the semantics of dietary composition of a person through robot audition, we proposed the chat-based dietary composition perception algorithm (DCPA) using social robot audition, which is detailed in Algorithm 2.

B. DIETARY COMPOSITION SEMANTIC UNDERSTANDING ALGORITHM ABOUT DIETARY COMPOSITION

We consider the following factors to design the semantic understanding algorithms:

- Expression patterns of Chinese;
- The number of people involved in the conversation. When a single person talks to the robot, there is no reference to the words/content/sentences of other speakers' conversations; when it comes to a multi-person conversation, in one case, the speakers directly speak out what they have eaten, and there is no reference to the content of other speakers' conversations either; in another case, there are some dietary references among speakers' sentences.
- Part of speech information, part of speech and the position information of contextual vocabulary and the syntactic structure of linguistic units.

Algorithm 2 Chat-Based Dietary Composition Perception Algorithm Using Social Robot Audition

Input: The connected audio segment set $C_a = \{p_1, p_2, \ldots, p_n\}$, corresponding identity set $I_a = \{id_1, id_2, \ldots, id_i, \ldots, id_n\}$ acquired from **Algorithm 1**, and the dietetic lexicon.

Output: $Food_a = \{Food_1, Food_2, .., Food_m\}$, where $Food_i$ is the dietary composition set of person i.

1) load the identity set $I_a = \{id_1, id_2, \ldots, id_i, \ldots, id_n\}$ and set the length of I_a to np;
2) translate the audio segment set C_a into text set $T_a = \{t_1, t_2, \ldots, t_n\}$ using the speech recognition function;
3) based on the constructed dietary word dictionary, use the algorithm, provided by LTP, to carry out the word segmentation, part of the speech tagging and the dependency syntax analysis [43] of text set T_a;
4) **If** $np == 1$, **Then** use **Algorithm 3** to obtain diet information, which will be detailed in Section IV-B-1), to assign the $Food_a$;
5) **If** $np > 1$, **Then** use **Algorithm 4** to obtain diet information, which will be detailed in in Section IV-B-2), to assign the $Food_a$;
6) output diet information $Food_a = \{Food_1, Food_2, ..., Food_m\}$.

Based on these considerations, we designed the following semantic understanding algorithm 3 for single-person talking and semantic understanding algorithm 4 for a multi-person conversation.

1) DIETARY COMPOSITION SEMANTIC UNDERSTANDING ALGORITHM FOR SINGLE-PERSON SESSION SPEECH DATA

It is a critical factor to confirm whether the dietary action actually occurs in order to determine the dietary composition. In daily situation, there are some sentences, such as ''wo zai chaoshi maile yixie mianbao''(English: I bought some bread in the supermarket), ''wo xihuan chi mianbao''(English: I like to eat bread), ''wo xianzai xiang chi mianbao''(English: I want to eat bread now), ''wo kewang chi mianbao''(English: I am eager to eat bread), although the fact that there is a verb, ''chi''(English: eat), the speaker did not necessarily eat the bread. Thus, when a conversation includes such a session, it is necessary to figure out whether the ''eat'' action actually occurs. If it occurs, then we can get the right information

regarding dietary composition by extracting foods around that verb.

a: DEFINITIONS OF THE LANGUAGE UNITS FOR SEMANTIC UNDERSTANDING

1)Set $sentence = \{\text{word}_1, \text{word}_2, \ldots, \text{word}_i, \ldots, \text{word}_N\}$ presents the language unit set of text T_{text} using the word segmentation method, where N is the total number of the language units.

2)Set V_{eat} = {"chi", "he", "xiangyong", "pinchang", "chang", "pin", "lai", "dian", "shi"} (English: {"eat", "drink", "enjoy", "taste", "taste", "taste", "order", "order", "try"}) denote the action feature word unit set, whose elements are able to express the meaning of eat.

3)Set V_{negative} = {"xiang", "xihuan", "ai", "xi'ai", "meiyou", "mei", "bu", "taoyan", "yanwu", "fangan" } (English: {"want", "like", "love", "adore", "not", "no", "no", "hate", "detest", "disgust"}), whose elements show the meaning of like, will, plan, and denial. If a sentence includes words/phrases belonging to V_{negative}, the "eat" action does not happen.

4)Set V_{enhance} = {"lei", "de"} (English: they are auxiliary words, meaning accomplishment) is the language unit set of the words/phrases expressing that the actions are complete

5)Set V_{cuisine} = {"chao", "baochao", "zheng", "qingzheng", "zhu", "zha", "youzha","dun", "jian", "youjian", "lu", "kao", "shao", "liangban", "ban", "men"} (English:{"saute", "stir-fry", "steam", "steamed in clear soup", "boil", "fry", "deep-fry", "stew", "decoct", "panfry", "stew", "roast", "burn", "cold and dressed with sauce", "mix", "stew"}) is the language unit set of the verb words/phrases expressing the cooking action.

6)Set W_{comment} = {"weidao", "haochi", "nanchi", "youwei"} (English: {"taste", "tasty", "tasty", "unpalatable" }) is the language unit set of words/phrases expressing the implicit diet action.

7)Set W_{jump} = {"bimengeng", "kui", "yabakui"} (English: {"cold-shoulder treatment", "loss", "suffering that cannot be told to others"}) is the spurious language unit set of the noun to express "eat" meanings.

b: SEMANTIC UNDERSTANDING PROCESS DESCRIBED BY FIRST-ORDER PREDICATE LOGIC THEORY

Suppose that $T_{\text{word}-1}$, T_{word} and $T_{\text{word}+1} \in sentence$ are adjacent language units. Based on the Chinese expression patterns, T_{text} may have the meaning of having eaten something, when $T_{\text{word}} \in V_{\text{eat}}$. Thus, by judging the relationship between $T_{\text{word}-1}$ and V_{negative}, $T_{\text{word}+1}$ and V_{enhance}, we can judge whether the "eat" action really happens, so as to determine the users' true dietary compositions. If the language unit $T_{\text{word}} \in sentence$ and $T_{\text{word}-1} \in V_{\text{negative}}$, we can judge that the user did not eat the foods that have a relation with T_{word}. If the language unit $T_{\text{word}} \in sentence$ and $T_{\text{word}-1} \cap V_{\text{negative}} = \emptyset \&\& T_{\text{word}+1} \in V_{\text{enhance}}$, we can judge that the user ate the foods that have a relation with T_{word}.

According to the above definitions and analysis, using the first-order predicate logic theory [44], the predicate logic reasoning formula (1) can be used to determine whether the eating action occurs or not.

$$Contains\left(V_{\text{eat}}, T_{\text{word}}\right), \neg Contains\left(V_{\text{negative}}, T_{\text{word}-1}\right), Contains\left(V_{\text{enhance}}, T_{\text{word}+1}\right) \Rightarrow RealEat\left(T_{\text{word}}\right) \quad (1)$$

where $Contains(A, a)$ judges if set A contains element a, $RealEat(T_{\text{word}})$ is the logical consequence of predicate logic reasoning, and the True value indicates that the user has eaten the foods that have a relation with T_{word}.

When the value of $RealEat(T_{\text{word}})$ is TRUE, the contextual words of T_{word} are analyzed to extract the dietary information.

1)For the words before T_{word}, if there are no food words between T_{word} and the next verb $T_{\text{word}+m}$ or the end word T_{end} of the *sentence*, and there is a word $T_{\text{word}-n}$ in W_{comment} before T_{word}, then we can judge that there are target food names before T_{word}. The predicate logic reasoning formula (2) can be used to determine whether there are target food names before T_{word}.

$$NoFoodNames\left(T_{\text{word}}, T_{\text{word}+m}\ or\ T_{\text{end}}\right), Contains\left(W_{\text{comment}}, T_{\text{word}-n}\right) \Rightarrow ExistBefore(T_{\text{word}}) \quad (2)$$

where $T_{\text{word}+m}$ is a verb after T_{word}, $T_{\text{word}+m} \cap V_{\text{cuisine}} = \emptyset$, $T_{\text{word}-n}$ is a word before T_{word}, $NoFoodNames(a,b)$ judges whether there are food names between a and b, and $ExistBefore(T_{\text{word}})$ judges whether there are target food names before T_{word}.

2)When $ExistBefore(T_{\text{word}})$ is TRUE, if $T_{\text{word}-n-i}$ is a verb before $T_{\text{word}-n}$ and does not belong toV_{cuisine}, then stop the traverse; If $T_{\text{word}-n-i}$ is a noun before $T_{\text{word}-n}$ in dietetic lexicon, then $T_{\text{word}-n-i}$ is the target food information. Based on this reasoning, the predicate logical formulas (3) and (4) are established:

$$Verb\left(T_{\text{word}-n-i}\right), \neg Contains\left(V_{\text{cuisine}}, T_{\text{word}-n-i}\right) \Rightarrow Stop(T_{\text{word}-n-i}) \quad (3)$$

$$Contains\left(Lexicon, T_{\text{word}-n-i}\right) \Rightarrow EatFood(T_{\text{word}-n-i}) \quad (4)$$

where $Verb(T_{\text{word}-n-i})$ judges if $T_{\text{word}-n-i}$ is a verb, $Stop(T_{\text{word}-n-i})$ judges if traversing is to be stopped or not, $Lexicon$ is the dietetic lexicon, $EatFood(B)$ is the logical result of predicate logic reasoning, and its TRUE value indicates that the user has eaten the food B.

3)For the words after T_{word}, if the next noun after T_{word} in the dietetic lexicon or the next verb after T_{word} belongs to V_{cuisine}, traverse backward from T_{word}; if not, conduct syntactic dependency analysis for the linguistic units after T_{word}. The predicate logic reasoning formula (5) can be used to determine which pattern should be employed to obtain target food names from the linguistic units after T_{word}.

$$Contains\left(Lexicon, nextnoun\left(T_{\text{word}}\right)\right) orContains\left(V_{\text{cuisine}}, nextverb\left(T_{\text{word}}\right)\right) \Rightarrow Traverse(T_{\text{word}}) \quad (5)$$

where $nextnoun(T_{\text{word}})$ is the next noun after T_{word}, $nextverb(T_{\text{word}})$ is the next verb after T_{word}, and the TRUE

value of $Traverse(T_{\text{word}})$ indicates that the extraction pattern is traverse.

4)If $Traverse(T_{\text{word}})$ is TRUE, use the formulas (3) and (4), which have already been proposed, to traverse the linguistic units after T_{word}.

5)If $Traverse(T_{\text{word}})$ is FALSE, use the following considerations to extract the food names.

(1) if $T_{\text{word+1}}$ ="Le", the start word $T_{\text{start}} = T_{\text{word}}$;

(2) if $T_{\text{word+1}}$ ="De", the start word $T_{\text{start}} = T_{\text{word+}j}$, where $T_{\text{word+}j}$ has subject-verb relation (SBV) with $T_{\text{word+1}}$;

(3) $word_k$, which has a verb-object relation (VOB) with T_{start}, and not with W_{jump}, and $word_m$, which has a coordinate (COO) relation with $word_k$, and not with W_{jump}, are target food names. The predicate logical formulas (6) and (7) are established based on this reasoning:

$$RelationVOB\left(T_{\text{word}}, word_k\right), \neg Contains\left(W_{\text{jump}}, word_k\right) \Rightarrow EatFood(word_k) \quad (6)$$

$$RelationCOO\left(word_m, word_k\right), \neg Contains\left(W_{\text{jump}}, word_m\right) \Rightarrow EatFood(word_m) \quad (7)$$

where $RelationVOB(A, B)$ and $RelationVOB(A, B)$ are used to judge the VOB and COO relation between A and B.

(4) set $word_n \in V_{\text{eat}}$, which has COO relation with T_{start}. Then, $word_{kn}$, which has a verb-object relation (VOB) with $word_n$, and not with W_{jump}, and $word_{mn}$, which has a coordinate (COO) relation with $word_{kn}$, and not with W_{jump}, are target food names. The predicate logical formulas (8) and (9) are established based on this reasoning:

$$Contains\left(V_{\text{eat}}, word_n\right), RelationCOO\left(T_{\text{word}}, word_n\right), RelationVOB\left(word_n, word_{kn}\right), \neg Contains\left(W_{\text{jump}}, word_{km}\right) \Rightarrow EatFood(word_{kn}) \quad (8)$$

$$RelationCOO\left(word_{kn}, word_{km}\right), \neg Contains\left(W_{\text{jump}}, word_{km}\right) \Rightarrow EatFood(word_{km}) \quad (9)$$

(5) when $T_{\text{after}} \in V_{\text{cuisine}}$ has a COO relation with T_{word}, output the former and the next noun word as food information.

$$Contains\left(V_{\text{cuisine}}, T_{\text{after}}\right), VCOO\left(T_{\text{after}}, T_{\text{word}}\right), Near\left(word_k, word_{\text{near}}\right) \Rightarrow EatFood(word_{\text{near}}) \quad (10)$$

where $Near(word_k, word_{\text{near}})$ judge whether $word_{\text{near}}$ is the former and next one is a noun word of $word_k$.

(6) when $VCOO(T_{\text{after}}, T_{\text{word}})$ =TRUE && $T_{\text{after}} \cap V_{\text{cuisine}} = \emptyset$ && $T_{\text{after}} \cap V_{\text{eat}} = \emptyset$, stop the food information extraction triggered by T_{word}.

$$\neg Contains\left(V_{\text{cuisine}}, T_{\text{after}}\right), \neg Contains\left(V_{\text{eat}}, T_{\text{after}}\right), VCOO\left(T_{\text{after}}, T_{\text{word}}\right), \Rightarrow \neg EatFood(word_{\text{after}}) \quad (11)$$

C: DIETARY COMPOSITION SEMANTIC UNDERSTANDING ALGORITHM FOR SINGLE-PERSON SESSION SPEECH DATA

Based on the definition and description above, we designed semantic understanding algorithm 3 for dietary composition perception for a single-person session.

Algorithm 3 Dietary composition Semantic Understanding Algorithm for Single-Person Session Speech Data

Input: $Sentence = \{word_1, word_2, ..., word_i, ..., word_N\}$
Output: the eaten food set $Food$;
1) initialize $Food = \emptyset$;
2) **For** each $word_i$ in $Sentence$:
(1) Using formulation(1), calculate the logical value of $RealEat(word_i)$;
(2) **If** $RealEat(word_i)$ ==TRUE, **Then**
a) using formulation (2), calculate the logical value of $ExistBefore(word_i)$;
b) **If** $ExistBefore(word_i)$ ==TRUE, **Then**
$Sub_{\text{Sentence}} = \{word_i, word_{i-1}, \ldots, word_1\}$;
For T_{word} in Sub_{Sentence}
using formulation (3), calculate the logical value of $Stop(T_{\text{word}-n-i})$;
using formulation (4), calculate the logical value of $EatFood(T_{\text{word}-n-i})$;
If $Stop(T_{\text{word}-n-i})$ ==TRUE, **Then Break**
If $EatFood(T_{\text{word}-\text{n}-\text{i}})$,
Then $Food = Food \cup T_{\text{word}-n-i}$;
c) using formulation (5), calculate the logical value of $Traverse(word_i)$;
d) **If** $Traverse(word_i)$ ==TRUE, **Then**
$Sub_{\text{Sentence}} = \{word_i, word_{i+1}, \ldots, word_N\}$;
For T_{word} in Sub_{Sentence}
using formulation (3), calculate the logical value of $Stop(T_{\text{word}-n-i})$;
using formulation (4), calculate the logical value of $EatFood(T_{\text{word}-n-i})$;
If $Stop(T_{\text{word}-n-i})$ ==TRUE, **Then Break**;
If $EatFood(T_{\text{word}-n-i})$,
Then $Food = Food \cup T_{\text{word}-n-i}$;
e) **OTHERWISE**
(a) $Sub_{\text{Sentence}} = \{word_i, word_{i+1}, \ldots, word_N\}$;
(b) *using* the algorithm, analyze the dependency syntax of Sub_{Sentence};
(c) **For** T_{word} in Sub_{Sentence}
① using formulation (6), calculate the logical value of $EatFood(word_k)$;
② **If** $EatFood(word_k)$ ==TRUE, **Then**
$Food = Food \cup word_k$;
using *formulation* (7), calculate the logical value of $EatFood(word_m)$;
If $EatFood(word_m)$ ==TRUE, **Then** $Food = Food \cup word_m$;
③ using formulation (8), calculate the logical value of $EatFood(word_{kn})$;
④ **If** $EatFood(\text{word}_{\text{kn}})$, **Then**
$Food = Food \cup word_{kn}$;
using *formulation* (9), calculate the logical value of $EatFood(word_{km})$;
If $EatFood(\text{word}_{\text{km}})$, **Then**
$Food = Food \cup word_{km}$;
⑤ using formulation (10), calculate the logical value of $EatFood(word_{\text{near}})$;

Algorithm 3 *(continued.)*

⑥ **If** $EatFood(word_{near})$, **Then** $Food = Food \cup word_{near}$;
⑦ using formulation (11), calculate the logical value of $EatFood(word_{after})$;
⑧ **If** $EatFood(word_{after})$, **Then Break**;
3) **Return** *Food*.

2) DIETARY COMPOSITION SEMANTIC UNDERSTANDING ALGORITHM FOR MULTI-PERSON CHAT SPEECH DATA

a: SEMANTIC UNDERSTANDING PROCESS DESCRIBED BY FIRST-ORDER PREDICATE LOGIC THEORY

When a chat involves more than one person, there are references to the expression content among different speakers. In order to acquire the dietary information precisely, first, judge if the former speaker $former_{speaker}$ asks about the dietary information of the current speaker, then judge whether the current speaker $current_{speaker}$ expresses the food information directly or not.

1)Referring to Chinese expression patterns, if $former_{speaker}$ asks about the dietary information directly, by saying, for instance, ''ni chi le shenme?'' or ''ni chi de sha?''(English: what do you eat?), judge that $former_{speaker}$ asks about the dietary information. The predicate logical formula (12) is established based on this reasoning:

$$Regex(former_{sentence}, \text{"(ni chi(English : youeat))(le|de)}\ \text{(shenme | sha) English : what)"}) \Rightarrow Asked(former_{speaker}) \tag{12}$$

where the TRUE value of $Regex(A, B)$ means A matches the regular expression B, the TRUE value of $Asked(former_{speaker})$ means the content $former_{sentence}$ of $former_{speaker}$ asks about the dietary information concerning the current speaker.

2)When $former_{speaker}$ expresses what he has eaten, then asks ''ni ne?''(English: what about you?), we can judge that $former_{speaker}$ asks about the dietary information concerning the current speaker. The predicate logical formula (13) is established based on this reasoning:

$$Regex\ (former_{sentence}, \text{(chi)(le|de))}\ , Regex\ (former_{sentence}, \text{(ni ne)}(English : what\ about\ you)\text{"}) \Rightarrow Asked(former_{speaker}) \tag{13}$$

3)When $Asked(former_{speaker})$ is TRUE, and the answer content $current_{sentence}$ of the current speaker $current_{speaker}$ includes expressions, such as ''wo ye chile ni shuo de nage''(English: I also ate what you said) and ''zhege weidao bucuo, woye chile''(English: It tastes good, I also ate it), judge that $current_{speaker}$ answers the dietary information question of $former_{speaker}$ indirectly. We considered four types of indirect expressions, formulas (14) to (17) are used to judge which type does the $current_{speaker}$ answer the dietary question indirectly:

$$Regex\ (current_{sentence}, \text{(wo ye chi le).} * \text{(na|zhe)(ge|xie))} \Rightarrow Indirect(type1(current_{speaker}) \tag{14}$$

$$Regex\ (current_{sentence}, \text{(na|zhe)(ge|xie).} * \text{(wo ye chi le))} \Rightarrow Indirect(type2(current_{speaker}) \tag{15}$$

$$Regex\ (current_{sentence}\text{". *(woyechile). *")} \Rightarrow Indirect(type3(current_{speaker}) \tag{16}$$

$$Regex\ (current_{sentence}, . * \text{chi.} * \text{yiyang(English : same).*)} \Rightarrow Indirect(type4(current_{speaker}) \tag{17}$$

where the true value of $Indirect(type1(current_{sentence}))$ means the $current_{sentence}$ matches the regular expression ''(na|zhe)(ge|xie). *(wo ye chi le)'', and the $current_{speaker}$ answers the dietary question indirectly. Formulas (15) to (17) have similar interpretations to that of formula (14).

4)When *Indirect*() is FALSE, $current_{speaker}$ answers the dietary question of $former_{speaker}$ directly, and there are two kinds of expression patterns:

(1) $current_{speaker}$ says the food names directly; or

(2) $current_{speaker}$ answers the question in a single-person session;

5)Set $V_{jump} = V_{cuisine} \cup$ {''haiyou'', ''yiji''} (English: {''and'', ''as well as''}), if the part of speech of the first language unit of $current_{sentence}$ is a noun, the $current_{sentence}$ is traversed to extract the noun as a piece of dietary information until there is a verb that does not belong to V_{jump}. Then, **Algorithm 3** is used to obtain the rest of the $current_{sentence}$. If the part of speech of the first language unit of the $current_{sentence}$ is not a noun, **Algorithm 3** is used to acquire the $current_{sentence}$ directly.

b: DIETARY COMPOSITION SEMANTIC UNDERSTANDING ALGORITHM FOR MULTI-PERSON CONVERSATION SPEECH DATA

Based on the definition and description above, we proposed algorithm 4 for dietary composition semantic understanding in a multi-person conversation.

V. SOCIAL ROBOT PLATFORM AND SYSTEM FRAMEWORK

A. SOCIAL ROBOT PLATFORM

Our MAT social robot [45] is used as an experimental platform for conducting research, it is shown in Figure 2. This platform consists of a mobile chassis, a host computer, data acquisition equipment, and a mechanical frame. The mobile chassis is EAI DashGO B1, which is capable of remote mobile control and is provided with 5V∼24V independent power for the host. The auditory system is based on the six-mic microphone array module, which can recognize speech and locate the position of a sound in a noisy environment. The vision system uses a depth camera, Kinect 2.0, with a maximum resolution of 1920∗1080 and a transmission frame rate of 30fps. The MAT social robot processing system is the Intel NUC (Next Unit of

Algorithm 4 Dietary Composition Semantic Understanding Algorithm for Multi-Person Conversation Speech Data

Input: $I_a = \{id_1, id_2, \ldots, id_i, \ldots, id_n\}$, audio segment text set $T_a = \{t_1, t_2, .., t_i, .., t_n\}$, $sentence(id_i) = \{word_1, word_2, ..., word_i, ..., word_N\}$, and the set of language units, obtained by conducting word segmentation on t_i;
Output: $Food_a = \{Food_{id1}, Food_{id2}, \ldots, Food_{idi}, \ldots, Food_{idn}\}$, and $Food_{idi}$ is the dietary composition set of id_i.

1) initialize $Food_a = \emptyset$;
2) **For** each t_i in set T_a:
(1)$Food_{idi} = \emptyset$, and $sentence(id_i) = \{word_1, word_2, ..., word_i, ..., word_N\}$;
(2) Using (12) and (13), obtain the logical value of $Asked(id_{i-1})$;
(3) **If** $Asked(id_{i-1})$ ==TRUE
a)**If** $Indirect(type1(sentence(id_i)))$ == TRUE || $Indirect(type2(sentence(id_i))$ ==TRUE, **Then**
For $word_i$ in $sentence(id_i)$:
If $RealEat(word_i)$ ==TRUE, **Then**
(a)$Sub_{Sentence}(id_i) = word_i, word_{i-1}, \ldots, word_1\}$;
(b)Define an empty variation *Temp* to save the food names;
(c)**For** $word_j$ in $Sub_{Sentence}(id_i)$:
If $word_j$ in $Food_{idi-1}$, **Then**$Temp = Temp \cup word_j$;
(d)$Food_{idi} = Food_{idi} \cup Temp$;
(e) **If** $Temp$ == null, **Then**
$Food_{idi} = Food_{idi} \cup Food_{idi-1}$;
$Sub_{Sentence}(id_i) = \{word_j, word_{j+1}, .., word_N\}$;
$Food_{idi} = Food_{idi} \cup$**Algorithm 3** $(Sub_{Sentence}(id_i))$;
OTHERWISE If $word_j$ is a verb, and $word_j$ is not in $V_{cuisine}$, **Then Break**;
b) **OTHERWISE**
(a) $Sub_{Sentence}(id_i) = \{word_j, word_{j+1}, \ldots, word_N\}$;
(b) **If** $word_1 \in Sub_{Sentence}(id_i)$ is a noun, **Then**
For $word_j$(j>0) in $Sub_{Sentence}(id_i)$:
If $word_j$ is a noun, **Then**$Food_{idi} = Food_{idi} \cup word_j$;

Algorithm 4 (*continued*.)

OTHERWISE If $word_j$ is a verb, and $word_j$ is not in V_{jump}, **ThenBreak**;
$TSub_{Sentence}(id_i) = \{word_j, word_{j+1}, \ldots, word_N\}$;
$Food_{idi} = Food_{idi} \cup$ **Algorithm 3** $(TSub_{Sentence}(id_i))$;
(c)**OTHERWISE** $Food_{idi} = Food_{idi} \cup$ **Algorithm 3**$(Sub_{Sentence}(id_i))$;
c) **If**$Indirect(type3(sentence(id_i)))$ == TRUE, **Then**
For $word_i$ in $sentence(id_i)$:
(a)**If** $RealEat(word_i)$, **Then**
$Sub_{Sentence}(id_i) = \{word_i, word_{i-1}, \ldots, word_1\}$;
For $word_j$ in $Sub_{Sentence}(id_i)$:
If $word_j$ in $Food_{i-1}$, **Then**
$Food_{idi} = Food_{idi} \cup wordj$;
OTHERWISE If $word_j$ is a verb, and $word_j$ is not in $V_{cuisine}$, **Then Break**;
(b) $Sub_{Sentence}(id_i) = \{word_j, word_{j+1}, \ldots, word_N\}$;
(c) $Food_{idi} = Food_{idi} \cup$ **Algorithm 3** $(Sub_{Sentence}(id_i))$;
d)**If** $Indirect(type4(sentence(id_i)))$ == TRUE, **Then**
$Food_{idi} = Food_{idi} \cup Food_{idi-1}$;
For $word_i$ in $sentence(id_i)$:
If $RealEat(word_i)$, **Then**
$Sub_{Sentence}(id_i) = \{word_i, word_{i+1}, \ldots, word_N\}$;
$Food_{idi} = Food_{idi} \cup$ and **Algorithm 3** $(Sub_{Sentence}(id_i))$;
(4) **OTHERWISE** $Food_{idi} = Food_{idi} \cup$ and **Algorithm 3**$(Sub_{Sentence}(id_i))$;
3) **Return** $Food_a$.

Computing), which has a Core i7-6770HQ processor, with a clock at 2.6-3.5GHz, and the graphics processor is Intel IRIS Pro. Besides, we installed a 16GB memory, with DDR4-2133 frequency, in the host computer. The operating system is Ubuntu 16.04, with the Robot Operation System (ROS).

B. ARCHITECTURE OF THE DIETARY COMPOSITION PERCEPTION SYSTEM

Figure 3 is the architecture of the dietary composition perception system. The input data of the system is speech data,

FIGURE 2. Social robot platform of MAT.

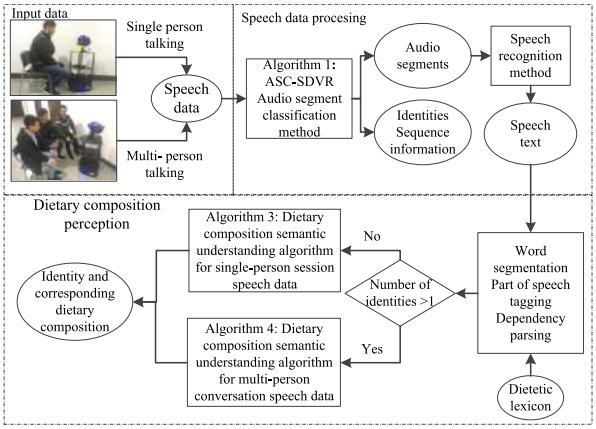

FIGURE 3. The architecture of the chat-based dietary composition perception system.

it includes single-person talking speech and multi-person conversion speech. The system firstly processes the speech data utilizing the proposed ASC-ASSR algorithm to obtain identities, speech segments and sequence information. Next, the speech recognition method is used to obtain the speech text of the segmentations. The general lexicon of the word segmentation system is expanded by collecting the dietetic lexicon, then the speech text is processed using word segmentation, part of speech is tagged and dependency parsing obtains the language features. The number of people participating in the speech conversation is acquired by Algorithm 1. If the number of people is one, Algorithm 3 is used to extract food names, if there are two or more people in the conversation, Algorithm 4 is used to extract food names. Finally, match the food names to the users' identities to obtain the results.

VI. PERFORMANCE OF THE TEST RESULTS AND ANALYSIS OF THE PROPOSED SYSTEM

A. TRAINING DATASET FOR SPEAKER RECOGNITION

The MAT robot, equipped with a six-mic microphone array, is used to collect the speech data. To test this system, we established a training dataset, including 20 persons' speech data, to register the speaker recognition system. From reading the given text, 30 different speech data are obtained for each person. The duration of each speech is 10 seconds, so we have a total of 5 minutes of speech data for each person.

When the full training data are ready, the method introduced in Section III-A will be employed to obtain the speaker recognition model.

B. TEST DATASET FOR DIETARY COMPOSITION PERCEPTION

It is very difficult to compare and check the performance of the system if the test uses real-time speech data. Thus, to check the performance of the proposed algorithms, we established a test dataset, which includes the speech data of a single person talking, a two-person conversation and a three-person conversation.

The content of the test speech data scripts includes the following kinds of conversation:

- Category a: Self-designed conversation, according to the Chinese language tradition. For the single person talking, the contents include the following classes: (a) The name of the eaten food is mentioned directly, (b) The name of the food that the person wants/plans/is eager to eat, c) A mixed sentence, including (a) and (b); and (d) A deceptive sentence. This kind of data includes 100 samples. While for the multi-person conversation, the contents refer to (e) No food name reference and (f) An existing food name reference, referred to by the speakers. The two-person conversation and three-person conversation include 40 samples.
- Category b: The conversations were collected from the students of our university by simulating real family scenes. In this research, 120 students took part, and we selected 180 completed text samples as the script of the speech data. Then, we employed three people to record the speech data based on the 180 samples. The sample sizes of a single person talking, a two-person conversation and a three-person conversation are 100, 40 and 40, respectively. This kind of data is characterized by randomness, which is used to check the robustness of the system.

We acted the collected scripts in the real family room, the volunteers sat in front of the social robot and recorded the speech data. They talked at a normal speed and volume, and the distance between robot and the interlocutor is between 0.5 meters and 1 meter. There is no noise in the room, but the noise of students' entertainment in the campus can be heard from the opened window. The total length of time for the test speech data is one hour, 51 minutes and 30 seconds, which includes 1,005 sentences and 2,546 food entities mentioned in those sentences. The training and test dataset are available on https://github.com/ZhidongSu/Dietary-Composition-Perception-Dataset.

There are some sample scripts of the test data.

Sample	English Meaning
今天去奶奶家做客，品尝了奶奶做的清蒸鱼，味道好极了。	Today, I went to Grandma's house and tasted steamed fish made by Grandma. It tasted great.
红烧肉口味的泡面是真的好吃，今天中午吃的真舒服。	The instant noodles with braised meat taste are really delicious. It's comfortable to eat at noon today.
甲：我今天晚上吃了一碗牛肉面，也喝了一瓶啤	A: I had a bowl of beef noodles and a bottle of beer

酒，吃的很饱。你吃了什么？ 乙：我吃的不多，我去超市买了一些香蕉和苹果，还有一些西瓜，但是我就吃了一碗米饭，喝了一碗粥。 甲：确实吃的不多，我也很想吃西瓜啊，但是这里没有卖的，所以就很伤心。 乙：哈哈，不要伤心，下次我请你吃西瓜。	this evening. I'm so full now. What did you eat? B: I didn't eat much. I went to the supermarket to buy some bananas and apples, and some watermelons, but I ate a bowl of rice and drank a bowl of gruel. A: You really didn't eat much. I'd like to eat watermelon, too. But there are no watermelons sold here, so I'm very sad. B: Ha-ha, don't be sad. Next time I'll invite you to eat watermelon.
甲：今天中午点了一份外卖，吃的是青椒炒鸡蛋和米饭，分量很足啊。 乙：我跟你就不一样了，我点的是米粉，那真是太少了分量，一点都吃不饱啊。 丙：你们下次还是不要吃外卖的好，因为外卖很不卫生的，我中午就在食堂吃的饭，吃了一份番茄炒鸡蛋、一份冬瓜炖排骨和一份米饭，干净又卫生。 甲：恩，好吧，那我们下次听你的，我们一起去食堂吃饭。 乙：对，我们再也不吃外卖了。	A: I ordered a take-out at noon today, eating green pepper scrambled eggs and rice, it is very good. B: I'm not the same as you, I just ordered rice noodles, the amount is so little that it is not enough at all for me. C: I suggest that you don't want to take a takeaway next time, because the takeaway is unsanitary. I ate at the cafeteria at noon, ate a tomato scrambled egg, a melon stewed pork ribs and a rice and the foods are clean and hygienic. A: Well, okay, we'll follow your suggestion to go to the cafeteria to eat next time. B: Yes, we won't eat takeout anymore.

C. METRICS

The eaten food is judged as correct when it matches the label and corresponding identity simultaneously. The modified recall value R, precision value P and F_1 [46] score of each conversation are used to check the recognition accuracy.

For each person in a conversation/talking, if N_t is the number of the tagged food manually, N_r is the dietary composition number reported by the system, and N_c is the number of the correctly eaten items of food, reported by the system by comparing them with the tagged food. Considering that the N_t values of some sentences, which have deceptive meanings or the meaning of wanting to eat food, are zeros, and the N_r values of some sentences, which are recognized falsely, are also zeros, we set the recall value R, precision value P and F_1 score as follows:

1) The precision P is the number of correctly eaten items of food, divided by the reported dietary composition number. Namely,

$$P = \begin{cases} \frac{N_c}{N_r}, & N_r \neq 0 \\ 0, & N_r = 0, N_t \neq 0 \\ 1, & N_r = 0, N_t = 0 \end{cases} \tag{18}$$

when N_r is zero, if N_t is not zero, it means that the system fails to recognize any of the food names in this sample, so we set P to 0; if N_t is zero, it means that this sample has deceptive meanings or the meaning of wanting to eat food, and the system is not cheated by this sample successfully, so we set P to 0.

2) The recall value R is the number of correctly eaten items of food, divided by the number of tagged foods. Namely,

$$R = \begin{cases} \frac{N_c}{N_t}, & N_t \neq 0 \\ 0, & N_t = 0, N_r \neq 0 \\ 1, & N_t = 0, N_r = 0 \end{cases} \tag{19}$$

when N_t is zero, if N_r is not zero, it means that the system is cheated by this sample which has deceptive meanings or the meaning of wanting to eat food, so we set R to 0; if N_r is zero, it means that the system is not cheated by this sample, so we set R to 1.

3) The F_1 score is a value that is related to precision and recall. Namely,

$$F_1 = \begin{cases} 2 \times \frac{P \times R}{P + R}, & R \neq 0 \; or \; P \neq 0 \\ 0, & R = P = 0 \end{cases} \tag{20}$$

D. THE TEST OF MFCC-CNN SPEAKER RECOGNITION MODEL

To evaluate the performance of the speaker recognition model proposed in Section III-A, the audio dataset Aishell [47] is employed, where 400 people with different accent from different accent areas in China were invited to participate in the recording. For each people, we chose 80 audios with the length of 6 seconds as the train set, which has totally 8 minutes, and 10 audios with the length of 6 seconds as the test set, which has totally 1 minutes. The total size of the training set and test set is 32,000 and 4,000.

The train epoch is set to 32, batch size is 64, and the dropout rate is 0.5. The test accuracy is 96.12%, which means that the proposed MFCC-CNN model has good performance for speaker recognition.

E. TEST RESULTS AND ANALYSIS

We implemented the proposed DCPA with Python and the C language and installed them on the MAT robot. We used the test dataset to check the performance of the system.

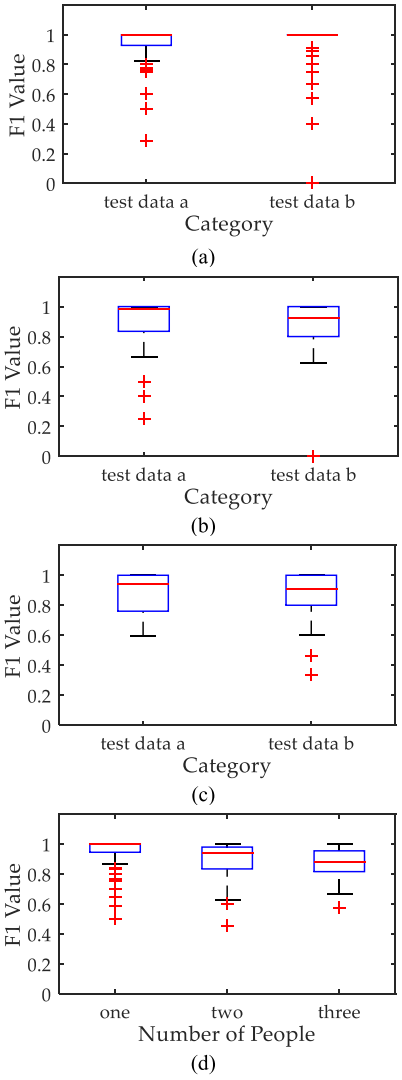

FIGURE 4. The Boxplot of F_1. (a) F1 value of one person talking; (b) F_1 value of a two-person conversation; (c) F_1 value of a three-person conversation; and (d) mean F_1 value of different people.

TABLE 1. Statistical Results of the Metric F_1, P, and R, with Different Test Data

Test Data	Number of people	F_1		P		R	
		Mean	Variance	Mean	Variance	Mean	Variance
a	One person	0.9505	0.0118	0.9753	0.0053	0.9361	0.0196
	Two-person	0.8940	0.0309	0.8820	0.0419	0.9306	0.0185
	Three-person	0.8768	0.0177	0.9283	0.0169	0.8609	0.0322
b	One person	0.9477	0.0333	0.9605	0.0431	0.9452	0.0337
	Two-person	0.8793	0.0334	0.8873	0.0412	0.8968	0.0404
	Three-person	0.8683	0.0254	0.8840	0.0296	0.8729	0.0336

1) TEST RESULTS OF THE SYSTEM

The statistic results of F_1 are shown in Table 1. Figure 4 is the boxplot of the F_1, and Figure 5 is the boxplot of the R and P. The results indicate the following:

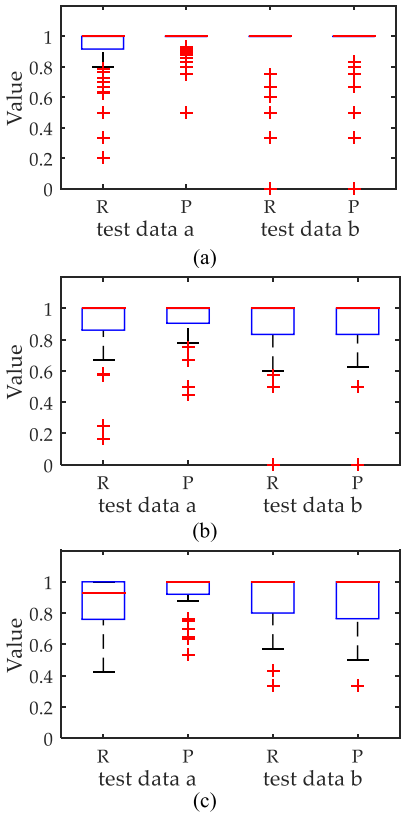

FIGURE 5. The Boxplot of R and P. (a) R and P in category a and b for one person talking; (b) R and P in category a and b for a two-person conversation; (c) R and P in category a and b for a three-person conversation.

- For the result of test data a, the average value F_1 of one person talking, two-person conversation and three-person conversation are 0.9505, 0.8940 and 0.8768, respectively, and the variances of F_1 are 0.0118, 0.0309 and 0.0177, respectively, which are small and have no great fluctuations, indicating that the system has a good dietary composition perception performance for the language expression considered. For test data b, the results are a little worse than test data a. The average recognition F_1 values are 0.9477, 0.8793 and 0.8683, respectively, and the F_1 value decreased by 0.0028, 0.0147 and 0.0085, respectively. The variances of F_1 are 0.0333, 0.0334 and 0.0254, respectively, which have larger fluctuations than in category a. Test data b was collected from 120 students using an online form, which has more uncertainty and more randomness than test data a because the number of people involved in collecting test data b is bigger than a, so the decrease in the F_1 value and fluctuations in variances are explainable within reasonable limits.
- For the F_1 value in test data a, one person talking has a better performance than a two- or three-person conversation, with a difference of 0.0565 and 0.0737, and a two-person conversation has a better performance than a three-person conversation, with a difference of 0.0172,

which is smaller than the difference between one person talking and a two-person conversation and that between one person talking and a three-person conversation. For test data b, the trend is similar. The differences between one person talking and a two- and a three-person conversation are 0.0684 and 0.0794, respectively, and a two-person conversation has a better performance than a three-person conversation, with a difference of 0.0110. This indicates that the number of people involved in the conversation will affect the recognition result, and the performance declines as the number of people increases.

From the Box plots in Figure 4, the following can be found:

- The median lines for test data a and b in Figure 4(a), 4(b) and 4(c) are all above 0.85, and the rectangular region of the boxplots is small, indicating that the system has a good recognition performance for test data a and b for one person talking, a two-person conversation and a three-person conversation.
- In Figure 4(a), although there are some outliers outside the rectangular region, the values of the two median lines are 1.0, and the area of the rectangular regions is the smallest of all the figures in Figure 4, indicating that the system can recognize most of the food names in the test data of one person talking, with a high F_1 value and the best performance among the different conversation types.
- The position of the median line in Figure 4(d) declines as the number of people involved in the conversation increases, indicating that the number of people involved in the conversation will worsen the recognition performance of the system.

From the Box plots in Figure 5, the following can be found:

- The median lines of R and P for test data a and b in Figure 5(a) are all 1.00, we can see that most of the R and P values are 1.00, and most of the other R and P values are above 0.60. Only three R values and one P value are in test data a, and six R values and three P values in test data b have a lower R and P values under 0.60. These results indicate that our system has a good performance in extracting food names, and the extracted food names fit the labels well.
- From Figure 5(b) and Figure 5(c), we can see that the area of the rectangular region is a little larger than that of one person talking, but most of the positions of the median lines are still 1.00. From those data, we can see that the number of lower R and P values increases slightly as the number of people involved in the conversation increases, i.e., the greater the number of people, the greater the complexity of the system, which has to employ the ASC-ASSR speech processing method and a more complex reasoning algorithm, based on one person talking, to deal with a multi-person conversation situation.

The above results are obtained from the test set with a total length of time of one hour, 51minutes and 30seconds, including 1,005 sentences and 2,546 food entities, which is collected from 120 people imitating the real-life scene. The number of people registered in the system is 20, and our MAT social robot is used to record the test speech data in home environment. The test results show that DCPA can obtain user's dietary composition with the F_1 score of 0.9505, 0.8940 and 0.8768 for one-person talking, a two-person conversation and a three-person conversation. Thus, we have enough data collected from real life scene, many people registered in the system and good statistical results to support us to draw the conclusion that our system have good performance and can be employed in real applications.

TABLE 2. Comparison Results of the Mean F_1, P, and R for the Text Corpus and Speech Data.

Test Data	Number of people	F_1		P		R	
		Text	Speech	Text	Speech	Text	Speech
a	One person	0.9704	0.9505	0.9922	0.9753	0.9642	0.9361
	Two-person	0.9451	0.8940	0.9571	0.9306	0.9478	0.8820
	Three-person	0.9436	0.8768	0.9358	0.9283	0.9599	0.8609
b	One person	0.9508	0.9477	0.9505	0.9605	0.9543	0.9452
	Two-person	0.9473	0.8793	0.9331	0.8873	0.9758	0.8968
	Three-person	0.9346	0.8683	0.9278	0.8840	0.9529	0.8729

2) ERROR ANALYSIS OF THE SYSTEM

While the system has a good performance in dietary composition perception, there are still some recognition errors. The cause of these errors are as follows.

1) The speech data processing affects the input of the text information extraction, which will eventually affect the performance of the system. In order to evaluate the extent to which the speech data processing affects the system, the collected conversation text corpus is used directly as the input of the text information extraction to obtain the recognition result in order to compare this with the result shown in Table 1, and the comparison results of the mean F_1, P, and R for the text corpus and speech data are shown in Table 2. From Table 2, we can see that the average F_1 values decreased by 0.0199, 0.0511 and 0.0668, from the text corpus to speech data in test data a, and decreased by 0.0031, 0.0680 and 0.0663 in test data b. The decline of the performance of two and three persons talking is greater than that of one person talking. This can be explained as follows:

- There will be some speech segments with the length of 1 ± 0.5s, while the speaker recognition system needs sufficient data to identify the identity of the speech, and short-term speech data will seriously affect the performance of the speaker recognition system. Incorrectly identified speech data will cause errors in the sequence of corresponding text data in the conversation text, while the reasoning algorithms are based on the sequence text information and contextual information of the identified speaker.

- When the speech segmentation method fails to detect the speaker identity transition point, two or more speakers will appear in a speech segment. Regardless of which speaker's identity the speech segment is identified with, it will result in a redundancy of the speaker's expression and a lack of another speaker's expression.
- The speaker's accent and dialect will affect the performance of the speech recognition system, especially when the keywords, such as eating action words and food names, have recognition errors, which will result in errors in food information extraction.

2) Language expressions are diverse and arbitrary, so the designed semantic understanding algorithms are not able to consider all situations. There are expressions that do not explicitly express dietary information but implicitly include dietary information and include deceptive and ambiguous meanings. Especially when it comes to a multi-person conversation, the context of the reference will be more difficult to identify.

VII. CONCLUSION AND FUTURE WORK

In conclusion, this paper proposed a chat-based automatic dietary composition perception algorithm (DCPA) using social robot audition for Mandarin Chinese, which can improve the function of social robots in healthy diet management. The audio segment classification method, based on audio segmentation and the speaker recognition algorithm (ASC-ASSR), is proposed to process conversation speech data to obtain the text content and identities of different speakers. At the same time, the semantic understanding algorithms for one-person speech and a multi-person conversation are designed to extract dietary information from the obtained text content and identities. We implemented a social robot, with the function of dietary composition perception, which employs the proposed dietary composition perception algorithm. We used the collected test data to conduct an experiment using our social robot, and the experiment results show that the proposed algorithm has a good recognition performance in dietary composition perception.

In our future work, we will optimize the speech data processing method, address the problem of environment robustness, such as noise and relative movement between the speaker and robot, update the semantic understanding algorithms of food text information perception, and improve the performance of multi-person conversation food information recognition. Otherwise, it would be a reliable method to use food name entity recognition to recognize food names, which will reduce the complexity of semantic understanding algorithms, but it is essential to have a lot of tagged diet-related corpuses. Besides, further understanding of how to combine wearable equipment with the proposed DCPA, which will be able to help us to establish the relationships between our diets and diseases and monitor people's health status in relation to their eating habits, is required.

REFERENCES

[1] P. Share and J. Pender, "Preparing for a robot future? Social professions, social robotics and the challenges ahead," *Irish J. Appl. Social Stud.*, vol. 18, pp. 45–62, Mar. 2018.

[2] International Federation of Robotics. (2017). *Executive Summary World Robotics 2017 Service Robots*. [Online]. Available: https://ifr.org/downloads/press/Executive_Summary_WR_Service_Robots_2017_.pdf

[3] J. Zhao and X. Li, "The status quo of and development strategies for healthcare towns against the background of aging population," *J. Landscape Res.*, vol. 10, pp. 41–44, Mar. 2018.

[4] M. Wei, S. Brandhorst, M. Shelehchi, H. Mirzaei, C. W. Cheng, J. Budniak, S. Groshen, W. J. Mack, E. Guen, S. Di Biase, P. Cohen, T. E. Morgan, T. Dorff, K. Hong, A. Michalsen, A. Laviano, and V. D. Longo, "Fasting-mimicking diet and markers/risk factors for aging, diabetes, cancer, and cardiovascular disease," *Sci. Transl. Med.*, vol. 9, no. 377, Feb. 2017, Art. no. eaai8700.

[5] N. Zmora, J. Suez, and E. Elinav, "You are what you eat: Diet, health and the gut microbiota," *Nature Rev. Gastroenterol. Hepatol.*, vol. 16, no. 1, pp. 35–56, Jan. 2019.

[6] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A. E.-D. Mousa, and B. Schuller, "I hear you eat and speak: Automatic recognition of eating condition and food type, use–cases, and impact on ASR performance," *PLoS ONE*, vol. 11, no. 5, May 2016, Art. no. e0154486.

[7] H. Kaya, "Fisher vectors with cascaded normalization for paralinguistic analysis," in *Proc. 16th Annu. Conf. Int. Speech Commun.*, Dresden, Germany, 2015, pp. 909–913.

[8] F. E. Fernandes, H. M. Do, K. Muniraju, W. Sheng, and A. J. Bishop, "Cognitive orientation assessment for older adults using social robots," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Macau, China, Dec. 2017, pp. 196–201.

[9] *Jibo*. Accessed: Jun. 19, 2019. [Online]. Available: https://www.jibo.com/

[10] *Aibo*. Accessed: Jun. 19, 2019. [Online]. Available: https://aibo.sony.jp/en/

[11] *Qrobot*. Accessed: Jun. 19, 2019. [Online]. Available: https://qrobot.qq.com/#/

[12] *Baobaolong*. Accessed: Jun. 19, 2019. [Online]. Available: http://ainirobot.com/baobaolong/

[13] G. Yang, J. Yang, W. Sheng, F. Junior, and S. Li, "Convolutional neural network–based embarrassing situation detection under camera for social robot in smart homes," *Sensors*, vol. 18, no. 5, p. 1530, May 2018.

[14] H. M. Do, M. Pham, W. Sheng, D. Yang, and M. Liu, "RiSH: A robot-integrated smart home for elderly care," *Robot. Auton. Syst.*, vol. 101, pp. 74–92, Mar. 2018, doi: 10.1016/j.robot.2017.12.008.

[15] B. Zhou, K. Wu, P. Lv, J. Wang, G. Chen, B. Ji, and S. Liu, "A new remote health–care system based on moving robot intended for the elderly at home," *J. Healthcare Eng.*, vol. 2018, pp. 1–11, Feb. 2018, doi: 10.1155/2018/4949863.

[16] M. Foukarakis, I. Adami, D. Ioannidi, A. Leonidis, D. Michel, A. Qammaz, K. Papoutsakis, M. Antona, and A. Argyros, "A robot-based application for physical exercise training," in *Proc. Int. Conf. Inf. Commun. Technol. Ageing Well e-Health*, 2016, pp. 45–52.

[17] F. Wang, X. Zhang, R. Fu, and G. Sun, "Study of the home–auxiliary robot based on BCI," *Sensors*, vol. 18, no. 6, p. 1779, Jun. 2018.

[18] M. Mast, M. Burmester, B. Graf, F. Weisshardt, G. Arbeiter, M. Španěl, Z. Materna, P. SmrŽ, and G. Kronreif, "Design of the human-robot interaction for a semi-autonomous service robot to assist elderly people," in *Ambient Assisted Living*, R. Wichert and H. Klausing, Eds. Cham, Switzerland: Springer, 2015, pp. 15–29.

[19] K. Li and M. Q.-H. Meng, "Personalizing a service robot by learning human habits from behavioral footprints," *Engineering*, vol. 1, no. 1, pp. 79–84, Mar. 2015.

[20] H.-Z. Chen, G.-H. Tian, and G.-L. Liu, "A selective attention guided initiative semantic cognition algorithm for service robot," *Int. J. Autom. Comput.*, vol. 15, no. 5, pp. 559–569, Oct. 2018.

[21] Y. Pyo, K. Nakashima, S. Kuwahata, R. Kurazume, T. Tsuji, K. Morooka, and T. Hasegawa, "Service robot system with an informationally structured environment," *Robot. Auto. Syst.*, vol. 74, pp. 148–165, Dec. 2015.

[22] D. E. Appelt, "Introduction to information extraction," *AI Commun.*, vol. 12, no. 3, pp. 161–172, 1999.

[23] S. M. M. Islam, "Temporal information extraction from textual data using long short term memory recurrent neural network," *J. Comput. Technol. Appl.*, vol. 9, no. 2, pp. 1–6, 2018.

[24] X. Y. Bao, W. J. Huang, K. Zhang, M. Lin, Y. Li, and C. Z. Niu, "A customized method for information extraction from unstructured text data in the electronic medical records," *J. Peking Univ. Health Sci.*, vol. 50, pp. 256–263, Apr. 2018.
[25] S. R. Jonnalagadda, A. K. Adupa, R. P. Garg, J. Corona-Cox, and S. J. Shah, "Text mining of the electronic health record: An information extraction approach for automated identification and subphenotyping of HFpEF patients for clinical trials," *J. Cardiovasc. Trans. Res.*, vol. 10, no. 3, pp. 313–321, Jun. 2017.
[26] L. Wang, Y. Qian, Y. Liu, Q. Meng, and T. Xu, "Information extraction method and its application in Chinese equipment technical manual based on rule-matching," in *Proc. 6th Int. Conf. Softw. Comput. Appl. (ICSCA)*, Bangkok, Thailand, 2017, pp. 92–97.
[27] Q. Qi, S. Zhou, C. Liu, and H. Chen, "Emergency information extraction based on style and terminology," *J. Chin. Inf. Process.*, vol. 32, pp. 56–65, Sep. 2018.
[28] M. Badieh Habib Morgan and M. Van Keulen, "Information extraction for social media," in *Proc. 3rd Workshop Semantic Web Inf. Extraction*, Dublin, Ireland, 2014, pp. 9–16.
[29] W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Short text understanding through lexical-semantic analysis," in *Proc. IEEE 31st Int. Conf. Data Eng.*, Apr. 2015, pp. 495–506.
[30] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: A probabilistic taxonomy for text understanding," in *Proc. Int. Conf. Manage. Data (SIGMOD)*, 2012, pp. 481–492.
[31] Y.-N. Chen, W. Y. Wang, and A. I. Rudnicky, "Unsupervised induction and filling of semantic slots for spoken dialogue systems using frame-semantic parsing," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 120–125.
[32] D. Hakkani-Tür, G. Tur, A. Celikyilmaz, Y.-N. Chen, J. Gao, L. Deng, and Y.-Y. Wang, "Multi–domain joint semantic frame parsing using bi–directional RNN–LSTM," in *Proc. Interspeech*, Aug. 2016, pp. 715–719.
[33] E. Grefenstette and P. Blunsom, "A deep architecture for semantic parsing," *Comput. Sci.*, vol. 30, pp. 1–15, Jun. 2014.
[34] Z. Saquib, N. Salam, R. Nair, and N. Pande, "Voiceprint recognition systems for remote authentication-a survey," *Int. J. Hybrid Inf. Technol.*, vol. 4, no. 2, pp. 79–98, 2011.
[35] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *J. Comput. Sci. Technol.*, vol. 16, no. 6, pp. 582–589, Nov. 2001.
[36] A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: Background, derivation, and applications," *WIREs Comput. Statist.*, vol. 4, no. 2, pp. 199–203, Mar. 2012.
[37] J. R. Saffran, E. L. Newport, and R. N. Aslin, "Word segmentation: The role of distributional cues," *J. Memory Lang.*, vol. 35, no. 4, pp. 606–621, Aug. 1996.
[38] R. W. Brown, "Linguistic determinism and the part of speech.," *J. Abnormal Social Psychol.*, vol. 55, no. 1, pp. 1–5, 1957.
[39] *National Standard GB 2760-2014 of the People's Republic of China*, Nat. Health Family Planning Commission People's Republic China, Beijing, China, 2015.
[40] X. Chen, Y. Zhu, H. Zhou, L. Diao, and D. Wang, "ChineseFoodNet: A large-scale image dataset for chinese food recognition," Oct. 2017, *arXiv:1705.02743v3*. [Online]. Available: https://arxiv.org/pdf/1705.02743.pdf
[41] W. Che, Z. Li, and T. Liu, "LTP: A Chinese language technology platform," in *Proc. 23rd Int. Conf. Comput. Linguistics*, Stroudsburg, PA, USA, 2010, pp. 13–16.
[42] *IFLYTEK*. Accessed: Jun. 19, 2019. [Online]. Available: https://www.xfyun.cn/
[43] C. Samuelsson, "A statistical theory of dependency syntax," in *Proc. 18th Conf. Comput. Linguistics*, 2003, pp. 684–690.
[44] S. Lucas and R. Gutiérrez, "Automatic synthesis of logical models for order-sorted first-order theories," *J. Automated Reasoning*, vol. 60, no. 4, pp. 465–501, Apr. 2018.
[45] G. Yang, Z. Chen, Y. Li, and Z. Su, "Rapid relocation method for mobile robot based on improved ORB-SLAM2 algorithm," *Remote Sens.*, vol. 11, no. 2, p. 149, Jan. 2019.
[46] M. Buckland and F. Gey, "The relationship between recall and precision," *J. Assoc. Inf. Sci. Technol.*, vol. 45, no. 1, pp. 12–19, Jan. 1994.
[47] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," Sep. 2017, *arXiv:1709.05522v1*. [Online]. Available: https://arxiv.org/pdf/1709.05522.pdf

ZHIDONG SU received the B.S. degree in mechanical engineering from the He'nan University of Technology, Zhengzhou, China, in 2016, and the M.S. degree in mechanical engineering from Guizhou University, Guiyang, China, in 2019. His research interests include natural language processing and intelligent autonomous social robots.

YANG LI received the B.S. degree in computer science and technology from Anyang Normal University, Anyang, China, in 2017. He is currently pursuing the Ph.D. degree with Guizhou University, Guiyang, China. His research interests include intelligent and autonomous robots, human action recognition.

GUANCI YANG was born in Jiahe, Hunan, China, in 1983. He received the B.S. degree in computer science and technology from the Hunan Institute of Science and Technology, Yueyang, China, in 2006, the M.S. degree in mechanical engineering from Guizhou University, Guiyang, China, in 2009, and the Ph.D. degree in computer software and theory from the University of Chinese Academy of Sciences, Chengdu, China, in 2012.

From 2013 to 2017, he was an Associate Professor with the Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, China. From 2015 to 2016, he worked as a Visiting Scholar at the Laboratory for Advanced Sensing, Computation and Control, Oklahoma State University, Stillwater, OK, USA. Since 2018, he has been a Professor with the Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, China. He is the author of two books, and more than 60 articles. His research interests include intelligent and autonomous robots, computational intelligence, smart home, and intelligent control systems.

Dr. Yang was a recipient of the Scientific Talents Foundation for the Outstanding Youth of Guizhou province, in 2015, and the Second Prize for the Progress in Science and Technology of Guizhou Province, in 2019.