Online Robot Teaching With Natural Human–Robot Interaction

Guanglong Du, Mingxuan Chen, Caibing Liu, Bo Zhang, and Ping Zhang

***Abstract*—With the development of Industry 4.0, robots tend to be intelligent and collaborative. For one, robots can interact naturally with humans. For another, robots can work collaboratively with humans in a common area. The traditional teaching method is no longer suitable for the production mode with human–robot collaboration. Since the traditional teaching processes are complicated, they need highly skilled staffs. This paper focuses on the natural way of online teaching, which can be applied to the tasks such as welding, painting, and stamping. This paper presents an online teaching method with the fusion of speech and gesture. A depth camera (Kinect) and an inertial measurement unit are used to capture the speech and gesture of the human. Interval Kalman filter and improved particle filter are employed to estimate the gesture. To integrate speech and gesture information more deeply, a novel method of audio-visual fusion based on text is proposed, which can extract the most useful information from the speech and gestures by transforming them into text. Finally, a maximum entropy algorithm is employed to deal with the fusion text into the corresponding robot instructions. The practicality and effectiveness of the proposed approach were validated by five subjects without robot teaching skills. The results indicate that the online robot teaching system can successfully teach robot manipulators.**

***Index Terms*—Gesture, human–robot interaction, natural speech understanding, natural, online robot teaching.**

I. Introduction

With the development of Industry 4.0, robots in the future tend to be intelligent, multifunctionalized, and collaborative. Unlike a traditional robot, the collaboration robot is no longer fixed in a fixed place for a fixed task. Instead, the collaboration robot can work together with humans in a common space to accomplish common tasks. Therefore, it is important to realize the communication between humans and robots. With the arrival of the era of intelligence, human beings began to look forward to the higher level of demand for intelligent technology [1], which drives robots to develop toward the direction of intelligence and diversification. Human robot cooperation [2] is a growing field in the research of robot technology. The premise and foundation of the human computer cooperation technology are the robot teaching [3] and playback technology [4], which means that the humans teach robots knowledge in some way. The robot has the abilities of learning, remembering, and perceiving, which integrated the development results of many disciplines.

Throughout the development of robot teaching–playback technology, there are three main directions: using a joystick [5] [6], teaching based on speech recognition [7], [8], and using a teach pendant. The principles, advantages, and disadvantages of the three directions are different. The principle of joystick teaching–playback methods is to perceive the spatial location information of the joystick, then reproduce the changing process of these positions. In previous studies, numerous researchers have adopted joysticks and haptic devices to teach and control the robots. In [9], a method of robot teaching and rendering by wearing a metal skeleton is introduced. This playback mode is particularly suitable for a humanoid robot, and the utility model has the advantages of long distance control and no requirement of working environment. Besides, because of the flexibility of the humanoid structure, this mode can complete some more complex work. In [10] and [11], the omega haptic devices are used to control the robot arm, which allows the operator to feel feedback from the robot arm. Such a system allows the user to control the robotic arm precisely and perform sophisticated tasks in small workspaces such as surgery. In this system, however, the tools (i.e., the joystick and the haptic device) limit the workspace. The high accuracy of teaching pendants can meet the requirements of many tasks. Nevertheless, it needs complicated system parameter settings, and the operator should be skilled, which makes it unable to satisfy the naturalness and efficiency.

Natural human–computer interaction is an important interface to realize friendly collaboration of intelligent robot and human. Most of the communication between humans is done through speech and gesture, and the interaction between speech and gesture is natural and intuitive. Robot teaching by means of speech recognition [12] is a new way for teaching and playback, which uses the natural perception channels of human beings. This paper focuses on a teaching method based on the natural human–computer interaction. For some tasks that do not require high precision, such as welding, sorting, spray paint,

Manuscript received September 12, 2017; revised January 9, 2018; accepted March 17, 2018. Date of publication April 5, 2018; date of current version July 30, 2018. The work was supported in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2017A030306015, in part by the Pearl River S&T Nova Program of Guangzhou under Grant 201710010059, in part by the Guangdong Special Projects under Grant 2016TQ03X824, in part by the Fundamental Research Funds for the Central Universities under Grant 2017JQ009, and in part by the National Natural Science Foundation of China under Grant 61602182. *(Corresponding author: Guanglong Du.)*

The authors are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: csgldu@scut.edu.cn; 317460580@qq.com; 1044083971@qq.com; 550510024@qq.com; pzhang@scut.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2018.2823667

stacking, stamping, die casting forging, grinding polishing, assembly, product testing, etc., we can make use of the human speech and natural three-dimensional (3-D) gesture to achieve fast and natural teaching, Therefore, it solves the problem that traditional teaching needs well-trained experts. Especially in the unstructured environment of a diversified production line, there are always tasks that have never been encountered before. It is necessary for robots to perform new teaching tasks at any time and any place, so as to ensure the efficient production. The traditional teaching method has the disadvantages of slow speed, high time cost, and the program is too complicated. These disadvantages make it difficult to be competent in the future production mode. Besides the time cost, sensor cost is also a factor restricting the development of robots.

Compared with previously developed teach methods, motion-sensor technology can be used in a teaching–playback robot manipulator system. In [13], the Wii remote controller recognizes human motion and controls the 3-D movement and rotation of the robot arm. In [14], a new motion path teaching system is proposed. This new system is composed of a teaching pen and a motion capture system. When the operator moves the teaching pen to a desired position, it is easy to plan and record the moving path. Therefore, it can complete the entire movement plan of the robot arm. Both Du *et al*. [15] and Maric *et al*. [16] concerned the Kinect motion sensor as a way of online programming robots. Maric *et al*. [16] presented an approach based on the Kinect-based motion sensing system for control of robot motion by demonstrating human arm motion. The above-mentioned teaching–playback technologies have effect in the specific environment. However, some motion sensors are expensive. It is not flexible and fast to deploy big-scale motion sensors. Using motion-sensor technology in the teaching oriented to human–machine collaboration is limited.

At present, most natural human–robot interaction methods use one or more natural ways to interact with robots. Chao and Thomaz [17] used robot Simon to coordinate with humans to build a block house. In the process of the task execution, robots can communicate with humans in interactive ways such as speech, eyes, and gestures. Tang *et al.* [18] designed a small partner robot for the aging problem. Users can use gestures and voice to talk to a robot. Simonetti *et al*. [19] used noninvasive interactive interfaces (EEG, EMG, eye tracking) to identify the intention of user's movement for the interaction with robots. The center for artificial intelligence research at Stanford University proposed a multichannel fusion method based on weight. Voice and gestures can be used as the input streams, and output the fusion information [20]. Cherubini *et al.* [21] proposed a unified multimodal control framework. The framework mixes posture, vision, and force sense, using weighted combinations of different sensors to control the robot. Jiang *et al*. [22] proposed a wheelchair mechanical arm system based on multiple 3-D vision sensors to assist in the daily life of the objects. The system uses gesture, speech, and gesture speech hybrid interactive interface. However, the fusion way of the above-mentioned methods is rough. They just combine the interrelated information from different data structures according to relevant rules. Users need to follow certain rules during the interaction process.

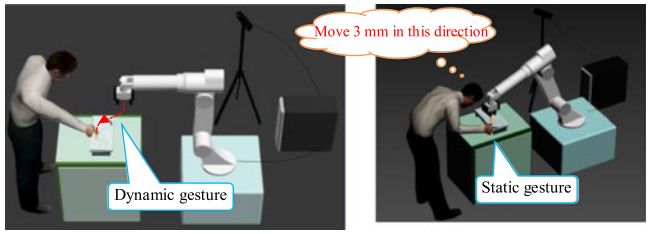

Fig. 1. Process of the online robot teaching system. (a) Rough teaching using dynamic gesture and basic speech commands. (b) Fine teaching using static gesture and teaching instructions of speech.

It is difficult to achieve the effect of natural interaction. Multimodal interaction is often used to combine multiple modes of interaction and uses them separately by switching. This combination is difficult to achieve the complementary effect between modes. The essence of fusion should be complementary information and complementary advantages. Data conversion and fusion inference are two key problems of information fusion.

Based on the feature of robot control instructions, the information in different structures is unified as text. Then, the robot control instruction can be extracted by using a text understanding method. In this paper, we proposed a method of audio-visual fusion based on text (AVFT). The advantage of this method is converting the speech information and gesture information into texts to achieve complementary advantages. The natural interactive teaching methods use a cheap sensor to realize online teaching for the tasks. Finally, the evaluation of the method is made by comparing with the existing methods [9], [13].

The rest of this paper is organized as follows. The overview is introduced in Section II. Section III details the recognition of the hand motion. The text understanding is introduced in Section IV. The realization of robot teaching is described in Section V. To verify the proposed method, experiments are designed in Section VI. Finally, discussion and conclusions are made in Sections VII and VIII, respectively.

II. Overview

Fig. 1 shows the process of the online robot teaching system, in which speech and gesture are used to teach the robots. This paper presents a method of AVFT for interacting with robots. The great advantage of this method is that by converting the speech data and the gesture data into a description text, the data of different structures can be unified to the same level. At the same time, this paper presents a fusion inference method to describe the text understanding of the robot instruction structure, thus extracting the interactive instruction of the robot. Speech information and gesture information are integrated into the text and solve the difficult problem of heterogeneous data fusion. In Fig. 1(b), the operator said, "Move 3 mm in this direction." Meanwhile, the operator's finger points at a certain direction (e.g., x, y, z), and the translated description text is, "Move 3 mm in the direction (x, y, z)," then the robot moves 3 mm in the direction (x, y, z). Fig. 1 shows the structure of the proposed system. What can be seen from Fig. 1(a) is that the Kinect [23] is

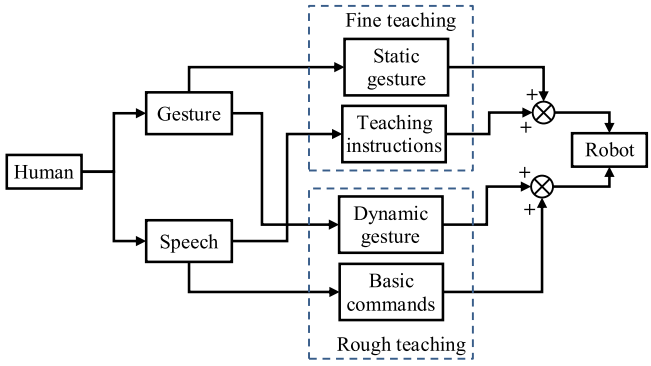

Fig. 2. Online robot teaching system process.

fixed beside the desk and the operator moves his or her hand in front of the Kinect. Kinect and inertial measurement unit (IMU) will collect the position and orientation of hand, respectively. The operator can pay more attention to how to complete the task using his or her hand because the robot will imitate the movement of the hand. As is seen from Fig. 1(b), the operator will give speech instructions to the robot, and the robot will move according to the given instructions. There are microphone arrays inside the Kinect, which can serve as an input device of phonetic information. Therefore, it will be more convenient for the operator to teach the robot.

According to the characteristics of the interactive mode, the two modes of rough teaching and fine teaching are defined (see Fig. 2), so as to improve the efficiency while ensuring the requirement of accuracy. Dynamic gestures and basic speech commands are suitable for robot teaching covering a certain distance. This kind of task requires less accuracy, such as approaching targets and avoidance programs. Moreover, dynamic gestures and basic speech commands are also suitable for the tasks with irregular paths or trajectories, which are difficult to describe by voice. Conversely, the tasks with a high precision require a fine-tuning of the position or orientation near the target position. The combination of speech and static gestures is needed to control. The specific direction is difficult to express in speech, so the static gesture in this paper is used to indicate the direction. The static gesture can be obtained from the finger skeleton point.

In addition, the maximum entropy algorithm is used to build the maximum entropy model, which can map the text into the robot control instructions. As for hand motion, generally, various filter tools such as Kalman filter (KF) [24] and particle filter (PF) [25] are used in the estimation process to get more accurate and reliable data. Since KF and PF can estimate the states of linear Gaussian state-space models and deal with highly nonlinear models, respectively, it would be a good method to integrate them to reduce the computational complexity [26]. However, the standard KF cannot deal with such situations with statistical parameters and inaccurate dynamics. Moreover, the standard PF may lead to severe particle degeneracy and dramatic loss of diversity in particles. Admittedly, $\mathrm{H}-\infty$ [27], [28] filter has been considered. There is an analysis [29] where an interval KF (IKF) is more suitable for this occasion to get the body's gesture data. In this paper, an IKF and an improved PF (IPF) are used to estimate the position and the orientation, respectively. And a formula from the factored quaternion algorithm (FQA), which can transform Euler angles into a quaternion, is employed in the orientation estimation process before using IPF to reduce the computation complexity. Therefore, the operator can easily teach the robot with the gesture and speech.

III. Recognition of Hand Motion

A. Hand Coordinate System

The proposed system uses a depth camera (Kinect) to locate the hand of the human operator and an IMU to measure the orientation of the hand. The position and orientation of the palm of the hand correspond to the end-effector of the robot in our method. The hand position is collected by the Kinect and the orientation is measured by an IMU. A sensor module, which is an IMU, is attached to the human hand, whose orientation (roll, pitch, and yaw) is to be determined. The IMU sensor consists of a three-axis accelerometer, two two-axis gyroscopes, and a three-axis magnetometer. We define the body frame as $X_bY_bZ_b$, sensor frame as $X_sY_sZ_s$, and Earth-fixed frame as $X_eY_eZ_e$. The sensor frame $X_sY_sZ_s$ corresponds to the axes of three orthogonally mounted accelerometers and magnetometers. The body frame $X_bY_bZ_b$ is assumed to coincide with the sensor frame $X_sY_sZ_s$.

B. Position Estimation Using the IKF

KF is used to estimate the position state because the position noise of Kinect is based on Gaussian distribution and the state-space model is linear. The IKF estimates a process in the form of feedback control. In other words, the filter estimates the process at some time and obtains feedback in the form of noise measurements. The updating process of the IKF was presented in method [30].

In the proposed method, the data of hand, such as velocity, position, and acceleration measured by Kinect, often contain noises. To minimize the errors, we define $P(p_x, p_y, p_z)$ as the coordinate of the center of the palm in the world frame. The IKF is used to estimate the state P of the position from a set of noise measurements [24]. The Kinect provides six measured parameters, which include three rotation angle components and three position components in the hand frame. Therefore, we define M_{H2W} as the direction cosine matrix from the hand frame to the world frame, which is as follows [31]:

$$M_{\mathrm{H2W}} = \begin{bmatrix} m_{X_x} & m_{Y_x} & m_{Z_x} \\ m_{X_y} & m_{Y_y} & m_{Z_y} \\ m_{X_z} & m_{Y_z} & m_{Z_z} \end{bmatrix} \tag{1}$$

where $m_{i_j} = \cos(\theta_{i_j})$ and $\theta_{i_j}\,(i, j \in (X, Y, Z))$ is the angle between the i-axis in the hand frame and the j-axis in the world frame. The acceleration of the hand in the world frame is calculated as follows [24]:

$$\dot{V}_x = m_{X_x} \cdot A_x + m_{Y_x} \cdot A_y + m_{Z_x} \cdot A_z$$
$$\dot{V}_y = m_{X_y} \cdot A_x + m_{Y_y} \cdot A_y + m_{Z_y} \cdot A_z$$
$$\dot{V}_z = m_{X_z} \cdot A_x + m_{Y_z} \cdot A_y + m_{Z_z} \cdot A_z - |g_l| \quad (2)$$

where $|g_l|$ represents the magnitude of the local gravity vector and (A_x, A_y, A_z) is the acceleration measurement component in each axis in the hand frame. Define the velocity component (V_x, V_y, V_z) in each axis in the local frame as

$$V_x = \dot{p}_x \quad V_y = \dot{p}_y \quad V_z = \dot{p}_z. \quad (3)$$

From (2) and (3), we can express the state x'_k of the position estimates by the IKF as follows:

$$x'_k = [p_{x,k}, V_{x,k}, A_{x,k}, p_{y,k}, V_{y,k}, A_{y,k}, p_{z,k}, V_{z,k}, A_{z,k}]. \quad (4)$$

x'_k is related to the values of variables at the time k. Therefore, the state-transition matrix Φ_k is defined as (5), shown at the bottom of this page.

This state-transition matrix is calculated by (1). The acceleration measurement can be achieved by gravity and the Z-axis is parallel to the gravity vector without control input. The input matrix of the system is written as

$$\Gamma_k \cdot u'_{k-1} = [0, 0, 0, 0, 0, 0, -|g_l| \cdot t^2/2, -|g_l| \cdot t, 0]^T \quad (6)$$

where Γ_k is the control-input model, which is applied to the input vector. So we get the process noise vector as

$$w'_k = [0, 0, w'_x, 0, 0, w'_y, 0, 0, w_z]^T \quad (7)$$

where (w'_x, w'_y, w_z) is the process noise of the hand acceleration.

The observation matrix for the estimated position is defined as

$$H_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (8)$$

We define $P(p_{x,k}, p_{y,k}, p_{z,k})$ as the optimal value of the position of the human hand at time k.

With the IKF, the covariance of the model error and observation error can be defined as

$$\begin{cases} Q_k^I = [Q_k - \Delta Q_k, Q_k + \Delta Q_k] \\ R_k^I = [R_k - \Delta R_k, R_k + \Delta R_k] \end{cases} \quad (9)$$

where ΔQ_k and ΔR_k are bounded and non-negative constant perturbation matrices. Because these disturbances come from the measured White Gaussian Noise, it can be considered as constant when environment is unchanged.

C. Orientation Estimation Using IPF

The FQA [32] is used to assist in estimating the orientation of the rigid body. The Euler angles are obtained by the accelerometer and magnetometer data and then transformed into quaternion in the traditional FQA. In this paper, the IMU measures the Euler angles (roll, pitch, and yaw) of the hand directly.

In order to reduce the error when integrating with the FQA, IPF [33] is adopted for optimizing data fusion. At time t_k, the approximation of the posterior density can be defined as

$$p(x_k | z_{1:k}, u_{0,k-1}) \approx \sum_{i=1}^{N} \omega_{i,k} \delta(x_k - x_{i,k}) \quad (10)$$

where x_k^i is the ith state particle at time t_k, N is the number of samples, ω_k^i is the normalized weight of the ith particle at time t_k, and $\delta(x)$ is the Dirac delta function.

An ensemble KF is used to approximate the probability density function of state variables $\{x_i\}_{i=1}^N$. Given the initial ensembles $\{x_{i,0}\}_{i=1}^N$, the forecast ensembles $\{x_{i,0}^f\}_{i=1}^N$ can be calculated as

$$x_{i,k}^f = f(x_{i,k-1}) + w_{i,k-1}, \quad w_{i,k-1} \sim N(0, Q_{k-1}) \quad (11)$$

where w_k is the model error. Q_{k-1} represents the covariance of the model error. The Kalman gain is obtained as (12), shown bottom of the next page.

where h represents the observation operator. Then, the analysis particles can be calculated as

$$x_{i,k}^a = x_{i,k}^f + K_k \left[z_k - h\left(x_{i,k}^f\right) + v_{i,k} \right], v_{i,k} \sim N(0, R_k)$$
$$\overline{x}_{i,k}^a = 1/N \sum_{i=1}^{N} x_{i,k}^a \quad (13)$$

where v_k is the observation error. Q_{k-1} represents the covariance of the observation error.

Finally, the weight updating formula is calculated as

$$\omega_{i,k} = \omega_{i,k-1} \frac{p(z_k | x_{i,k}^a) p(x_{i,k}^a | x_{i,k-1}^a)}{q(x_{i,k}^a | x_{i,k-1}^a, z_k)}. \quad (14)$$

A Markov chain Monte Carlo (MCMC) method is used to improve the diversity of particles after resampling. The Markov

$$\Phi_k = \begin{bmatrix} 1 & t & m_{X_x} \cdot t^2/2 & 0 & 0 & m_{Y_x} \cdot t^2/2 & 0 & 0 & m_{Z_x} \cdot t^2/2 \\ 0 & 1 & m_{X_x} \cdot t & 0 & 0 & m_{Y_x} \cdot t & 0 & 0 & m_{Z_x} \cdot t \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & m_{X_y} \cdot t^2/2 & 1 & t & m_{Y_y} \cdot t^2/2 & 0 & 0 & m_{Z_y} \cdot t^2/2 \\ 0 & 0 & m_{X_y} \cdot t & 0 & 1 & m_{Y_y} \cdot t & 0 & 0 & m_{Z_y} \cdot t \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & m_{X_z} \cdot t^2/2 & 0 & t & m_{Y_z} \cdot t^2/2 & 1 & t & m_{Z_z} \cdot t^2/2 \\ 0 & 0 & m_{X_z} \cdot t & 0 & 0 & m_{Y_z} \cdot t & 0 & 1 & m_{Z_z} \cdot t \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \quad (5)$$

transition kernel is defined as

$$\int p(x_k|z_{1:k})k(x_k^*|x_k) = p(x_k^*|z_{1:k}). \tag{15}$$

In the Metropolis Hasting algorithm, the resampled particles move to the proposed particle only if $u \leq a$, which are defined as

$$\begin{cases} a = \min\left(1, \frac{p(z_k|x_k^*)}{p(z_k|x_k)}\right) \\ u \sim U[0,1] \end{cases}. \tag{16}$$

After the MCMC step, the new particles will have more diversity since they have a distribution closer to the posterior probability density function. There are four states ($x_{\mathrm{PF},k}^i = [\,q0_k^i\ q1_k^i\ q2_k^i\ q3_k^i\,]$) in each particle orientation, which is represented by a unit quaternion and satisfies the following condition:

$$q0_k^{i^2} + q1_k^{i^2} + q2_k^{i^2} + q3_k^{i^2} = 1 \tag{17}$$

where $q0_k^i$, $q1_k^i$, $q2_k^i$, and $q3_k^i$ represent the four unit quaternion components. And we define the quaternion components of each particle at time t_{k+1} as

$$\begin{bmatrix} q0_{k+1}^i \\ q1_{k+1}^i \\ q2_{k+1}^i \\ q3_{k+1}^i \end{bmatrix} = \frac{1}{2} \times \begin{bmatrix} 2 & -\omega_{x,k}\cdot t & -\omega_{y,k}\cdot t & -\omega_{z,k}\cdot t \\ \omega_{x,k}\cdot t & 2 & \omega_{z,k}\cdot t & -\omega_{y,k}\cdot t \\ \omega_{y,k}\cdot t & -\omega_{z,k}\cdot t & 2 & \omega_{x,k}\cdot t \\ \omega_{z,k}\cdot t & \omega_{y,k}\cdot t & -\omega_{x,k}\cdot t & 2 \end{bmatrix} \cdot \begin{bmatrix} q0_k^i \\ q1_k^i \\ q2_k^i \\ q3_k^i \end{bmatrix} \tag{18}$$

where $\omega_{\mathrm{axis},k}$ is the angular velocity component and t is the sampling time. The rotation matrix ${}_b^f C_k^i$ of the *i*th particle from the body frame to the local frame at time t_k is demonstrated in [33].

Velocity and position are estimated with the IKF for each particle orientation. The acceleration of the object is written as

$${}^f\dot{V} = {}_b^f C \cdot {}^b A + {}^f g \tag{19}$$

where ${}^f\dot{V}$ represents the velocity in the local frame, and ${}_b^f C$ is defined in [33]. ${}^b A$ represents the acceleration measurement in the body frame. ${}^f g$ is the local gravity vector. The acceleration (${}^f\dot{V}$) in the fixed frame is calculated with a large position error if a large error is presented in the rotation matrix. Therefore, the weights of each particle can be assigned by the accumulated differences between the estimated and calculated positions of the *i*thparticle via IKF. The position differences are defined as follows:

$$\mathrm{APD}_s^i = \sum_{k=(s-1)\cdot M_s+1}^{M_s\cdot s} \left\{ \left(P_{px,k}^i - P_{kx,k}^i\right)^2 + \left(P_{py,k}^i - P_{ky,k}^i\right)^2 + \left(P_{pz,k}^i - P_{kz,k}^i\right)^2 \right\} \tag{20}$$

where APD_s^i is the accumulated position difference of the *i*th particle at the *s*th orientation iteration, $M_s = \Delta T_s/t$, and $P_{p-\mathrm{axis},k}^i$ is the position state for the *i*th orientation particle at time t_k, which is calculated with the acceleration calculation from (19) without using the position measurement information. $P_{k-\mathrm{axis},k}^i$ is the position of the *i*th particle in each axis estimated by the IKF at time *k*.

IV. Text Understanding

This section mainly introduces how the robots understand complex text, which is fusion text including speech text and gesture text. First, we need to establish a thorough text control command corpus. Second, we designed a set of effective robot control commands. Third, how to transform the intentions of the user's natural language text into the corresponding robot control commands was considered. Based on the three problems mentioned above, we proposed a framework of an online robot teaching system.

The framework consists of three serial layers: input layer, interaction layer, and output layer. The task of the robot in the input layer is to obtain the user's information, mainly text information, and the information will be given to the interaction layer to process. The interaction layer is mainly composed of two modules: text fusion and intentions understanding. The text fusion module transforms the user's speech and gesture into text. The intentions understanding module will deal with the received fusion text information and transform the fusion text into the corresponding control commands. Finally, the output layer will transform the robot control commands into the corresponding robot motions.

Through the research and analysis, we found that when we wanted to get a robot motion, the intentions of the operator could be summarized as follows: an operation of the target. In this paper, four attributes ($C_{\mathrm{opt}}, C_{\mathrm{dir}}, C_{\mathrm{val}}, C_{\mathrm{unit}}$) are introduced to design the robot control commands. At the same time, the function of the corpus used in natural language understanding is not only giving control instructions but also including gesture instructions. According to the textual meaning of corpus, the fusion text corresponding robot control instructions is marked.

The framework of natural language instruction is divided into two parts: the training process and the testing process. The

$$\begin{cases} K_k = P_k^f H^T (H P_k^f H^T + R_k)^{-1} \\ \overline{x}_k^f = 1/N \sum_{i=1}^N x_{i,k}^f \\ P_k^f H^T = 1/(N-1)\sum_{i=1}^N \left(x_{i,k}^f - \overline{x}_k^f\right)\left[h\left(x_{i,k}^f\right) - h\left(\overline{x}_k^f\right)\right]^T \\ H P_k^f H^T = 1/(N-1)\sum_{i=1}^N \left[h\left(x_{i,k}^f\right) - h\left(\overline{x}_k^f\right)\right]\left[h(x_{i,k}^f) - h\left(\overline{x}_k^f\right)\right]^T \end{cases} \tag{12}$$

control command corpus is also divided into two parts: the training corpus and the test corpus. During the training process, first, in the training corpus, the text features are extracted from training text, then the term frequency-inverse document frequency (TF-IDF) was used to attach weight to text feature before classification. Finally, we express the training text as a text feature vector. These are detailed in [34]. The TF value is a local variable, and IDF is a global variable as follows:

$$\mathrm{TF}_{i,j} = n_{i,j} / \sum_k n_{k,j} \tag{21}$$

$$\mathrm{IDF}_i = \log(|D|/|\{j : t_i \in d_j\}|) \tag{22}$$

$$\mathrm{TFIDF}_{I,J} = \mathrm{TF}_{i,j} * \mathrm{IDF}_i \tag{23}$$

where $n_{i,j}$ in (21) means the appearance time of the word in the corpus text. $\sum_k n_{k,j}$ denotes the number of all words that the corpus contains. $|D|$ in (22) denotes the number of all texts in training corpus. $\{j : t_i \in d_j\}$ denotes the number of the corpus text that contains this word.

Then the text feature vector and the corresponding class labels are used to train so as to obtain the maximum entropy classification model. The core idea of the maximum entropy model [35] is satisfying the known conditions while predicting the probability distribution of a random variable. At that time, the information entropy of the probability distribution is the maximum, which preserves all kinds of possibilities and minimizes the risk of prediction.

Assume that x is a text feature vector, and y is the corresponding intent output tag. The maximum entropy algorithm is to model the conditional probability $p(y|x)$ so as to obtain the most uniform distribution model. Introduce conditional entropy $H(p)$ to measure the uniformity of the distribution of conditional probability $p(y|x)$. According to the definition of entropy made by Shannon, computational formula $H(p)$ is shown as follows:

$$H(p) \equiv -\sum_{x,y} {}^{\sim}p(x)p(y|x)\log p(y|x) \tag{24}$$

where ${}^{\sim}p(x)$ means the empirical distribution of the text feature vector x in the training set, and $p(y|x)$ is the conditional probability distribution in the model requiring solution. Then, the formula to solve the maximum entropy model is shown as follows:

$$p^* = \arg\max H(p). \tag{25}$$

The maximum entropy problem can be summed up as the following optimization problem:

$$\max H(p) \qquad \text{s.t.}{}^{\sim}p(f_i) = p(f_i), \quad i = 1, 2, \ldots, n. \tag{26}$$

According to the Lagrange multiplier algorithm, we can obtain the probability distribution p^*. In the Kullback–Leibler distance, p^* is the closest to the empirical probability distribution $\sim p(x, y)$:

$$p^* = \frac{1}{Z(x)} \exp\left[\sum_{i=1}^{n} \lambda_i f_i(x, y)\right] \tag{27}$$

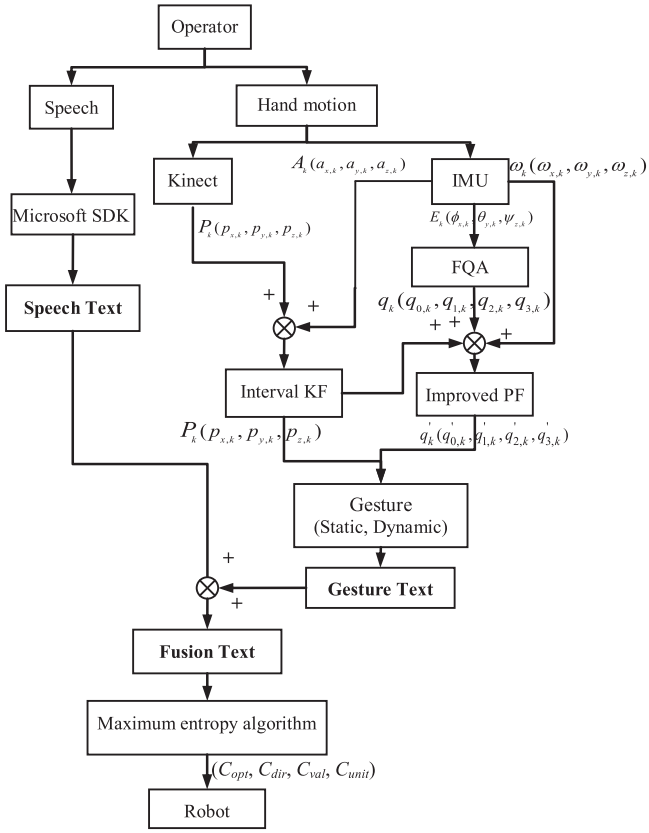

Fig. 3. Processing of online robot teaching.

where p^* is the maximum entropy probability distribution, $f_i(x, y)$ is the ith characteristic function, λ_i is the weight of $f_i(x, y)$, n is the number of the characteristic function, and $Z(x)$ is the normalized factor. Through the study of the training set, we can get the value of the parameter λ_i so as to get the probability distribution p^*, which needs to be solved. In order to obtain the parameters λ_i, the GIS algorithm [36] is used in this paper. So far, the establishment of the maximum entropy model has been completed.

V. Realization of Robot Teaching Using AVFT

Fig. 3 shows the process of online robot teaching based on AVFT. In this paper, a Kinect and an IMU are used to capture the speech and the hand movements of the operator. The speech can be converted into text by the Microsoft Azure SDK. Hand motion can be captured by the Kinect and the IMU. The position $P_k(p_{x,k}, p_{y,k}, p_{z,k})$ of the operator hand can be obtained by Kinect. The IMU can measure the acceleration $A_k(a_{x,k}, a_{y,k}, a_{z,k})$, the angular velocity $\omega_k(\omega_{x,k}, \omega_{y,k}, \omega_{z,k})$, and the orientation $E_k(\phi_{x,k}, \theta_{y,k}, \psi_{z,k})$ of the operator hand. The IKF is used to estimate the position from the measured position and the measured acceleration. The IPF uses the measured angular velocity, the measured orientation, and the estimated position to estimate the orientation of the hand. Since the static gesture is the movement used to indicate the direction, the static gesture can be obtained from the finger skeleton point. The

TABLE I
EXAMPLE OF MAPPING OF NATURAL INTERACTION INSTRUCTIONS AND ROBOT MOVEMENT COMMANDS

Type	Speech	Gesture	Instructions			
			C_{opt}	C_{dir}	C_{val}	C_{unit}
Base command	"Stop"	--	--	--	0	--
Speech teaching	"Move 1 mm in the right (X)"	--	Move	X	1	mm
Speech and Gesture (static)	"Move 3 mm in this direction"	Direction: $\vec{P}$	Move	$\vec{P}$	3	mm
Speech and Gesture (dynamic)	"Follow my hand"	$P_0, P_1, \ldots, P_n$	Move	$\overrightarrow{P_0 - P_{current}}$, $\overrightarrow{P_1 - P_0}$,… $\overrightarrow{P_n - P_{n-1}}$	$\lvert P_0\text{-}P_{current}\rvert$, $\lvert P_1\text{-}P_0\rvert$, …$\lvert P_n\text{-}P_{n-1}\rvert$	m

dynamic gestures are mainly used for rough trajectory planning, so the dynamic gestures can be obtained from tracking the index finger. Therefore, the static gesture can be converted into the text of a direction description, and the dynamic gesture can be transform into the text of the point sequence. Hence, the fusion text, consisting of the speech text and the gesture text, is processed by the maximum entropy algorithm to extract robot control commands. By analyzing a large number of robot instructions, four attributes $(C_{\text{opt}}, C_{\text{dir}}, C_{\text{val}}, C_{\text{unit}})$ are introduced to design the robot control commands.

To ensure the efficiency and accuracy of the teaching, rough teaching and fine teaching are combined. In order to ensure the naturalness of teaching interaction, the operator does not need to do extra action to switch between rough teaching and fine teaching. He just needs to say what he wants and act accordingly, then the system can identify the instructions and move accordingly. Table I shows the classic example of mapping natural interaction instructions and robot commands. For the basic commands, such as "STOP" and "PAUSE," the operator just says the corresponding word and does not need to make any gestures. The system just sets C_{val} as 0, and then the robot will stop. For the fine teaching, the operator can adjust the robot end-effector by speech or the combination of speech and gesture. For example, the operator says "Move 1 mm in the right." Suppose the right side of the operator is the X-direction of the robot, then the robot instruction will be identified as $(C_{\text{opt}} = \text{MOVE}, C_{\text{dir}} = X, C_{\text{val}} = 1, C_{\text{unit}} = \text{mm})$. When the operator says "Move 3 mm in this direction," meanwhile he points at one direction with his finger. The system captures his gesture and transforms the gesture as text "Direction: $\vec{P}$" or "Direction: $[x, y, z]$." Then the fusion text, combined with speech text and gesture text, becomes "Move 3 mm in this direction: $\vec{P}$ (or $[x, y, z]$)." Finally, the system analyzes the fusion text and extract the robot instruction as $(C_{\text{opt}} = \text{MOVE}, C_{\text{dir}} = \vec{P}$ (or $[x, y, z]), C_{\text{val}} = 3, C_{\text{unit}} = \text{mm})$. For the rough teaching, dynamic gestures are used for rapid trajectory planning, and the speech is used to instruct the robot to follow the hand gesture. In fact, the dynamic gesture is a series of points. When the operator says "Follow my hand" and draws a track at the same time, the fusion text can be formed as "MOVE TO $P_0, P_1, \ldots, P_n$." Then, a series of robot instructions are constructed out as $(C_{\text{opt}} = \text{MOVE}, C_{\text{dir}} = \overrightarrow{P_0 - P_{\text{current}}}, C_{\text{val}} = |P_0 - P_{\text{current}}|, C_{\text{unit}} = m)$, $(C_{\text{opt}} = \text{MOVE}, C_{\text{dir}} = \overrightarrow{P_1 - P_0}, C_{\text{val}} = |P_1 - P_0|, C_{\text{unit}} = m), \ldots,$ $(C_{\text{opt}} = \text{MOVE}, C_{\text{dir}} = \overrightarrow{P_n - P_{n-1}}, C_{\text{val}} = |P_n - P_{n-1}|, C_{\text{unit}} = m)$.

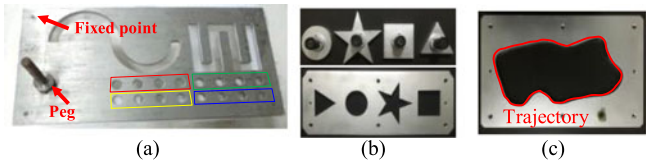

Fig. 4. Experimental plates and metal objects. (a) Steel plate. (b) Wire-cut shaped objects and corresponding holes. (c) Irregular trajectory.

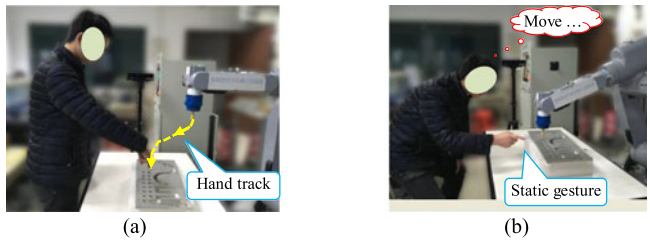

Fig. 5. Process of the peg-into-hole experiment. (a) Rough teaching. (b) Fine teaching.

VI. EXPERIMENT

A. Experiments Environment

The proposed online robot interface system allows the operator to teach the robot by using the combination of gesture and speech. The Kinect (version II) and IMU (LPMS-B2, Bluetooth transmission) were used to measure the hand's position and orientation. The robot EE can repeat the motion of the operator in real time, to accomplish the function of robot rough teaching. Meanwhile, speech recognition technology was adopted to implement robot fine teaching. In order to verify the proposed method, which embodies the naturality of human–robot interaction, a series of experiments were performed to assess the effectiveness, and the results of the three experiments were compared with two methods described in [9] and [13]. A GOOGOL GRB3016 robot and KUKA KR 6 R700 sixx were used for the experiments.

Five subjects who lack teaching skills participated in this study. A series of peg-into-hole experiments were carried out to assess the accuracy and stability of the proposed method. A steel plate, which contains 16 holes, was placed on the experiment platform [see Fig. 4(a)]. The length and width of the steel plate are 50 and 24.5 cm, respectively. The diameters of the four holes in the red rectangle frame, the yellow rectangle frame, the green rectangle frame, and the blue rectangle frame are 20, 19, 18, and 17 mm, respectively. The diameter of the peg is 15 mm. The

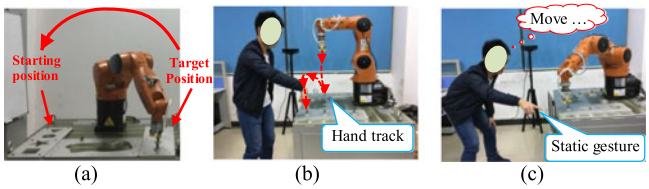

Fig. 6. Process of the placing workpieces experiment, which uses the wire-cutting shapes. (a) Starting position and the target position. (b) Rough teaching. (c) Fine teaching.

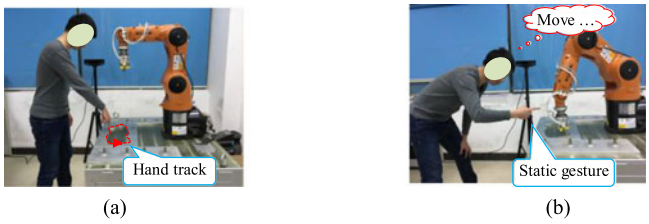

Fig. 7. Process of the irregular trajectory tracking experiment. (a) Rough teaching. (b) Fine teaching.

operator needed to guide the robot near the hole and put the peg into the hole (see Fig. 5). Since the guidance process did not require high accuracy, dynamic gesture was used to complete the guidance process. However, the insertion process requires high precision, and then fine teaching came in handy.

Experiment 2 was a placing workpieces task of wire-cutting shapes. As shown in Fig. 4(b), there are four metal objects with the shapes of triangle, circle, star, and square, and a steel plate housing four holes of corresponding shapes was also used in this experiment. The operators were required to manipulate the robot to place the four wire cutting workpieces into their corresponding holes from the starting position to the target position [see Fig. 6(a)]. Different from the peg-into-hole task of simple circle shapes mentioned in experiment 1, in order to implement the placing workpieces task of wire cutting shapes, the operators were supposed to make both the centers and the shapes into alignment. To meet this requirement, we imported speech control to rotate the objects in the plane with the rotation axis set to the Z-axis. In placing workpieces experiment, the data of the hand motion collected by the sensor were conducted by robot EE [see Fig. 6(b)], and then the control instructions of speech and static gesture, which were made by the operator to achieve fine teaching, were given to move the end-effector [see Fig. 6(c)]. In the end, the results were compared with methods [9] and [13] to indicate the proposed method is more accurate.

An irregular trajectory tracking task was designed to verify the stability and accuracy of the proposed method. As shown in Fig. 4(c), a steel plate that contains an irregular trajectory whose external rectangle was 450 × 210 mm was used in this experiment. In this experiment, the operator moved along the trajectory with hand motion [see Fig. 7(a)], and then the robot EE repeats the same movement according to the motion data collected by the sensor [see Fig. 7(b)]. The hand motion of the operator conducted the robot rough teaching. In addition, the operator adjusted the robot end-effector by speech and static

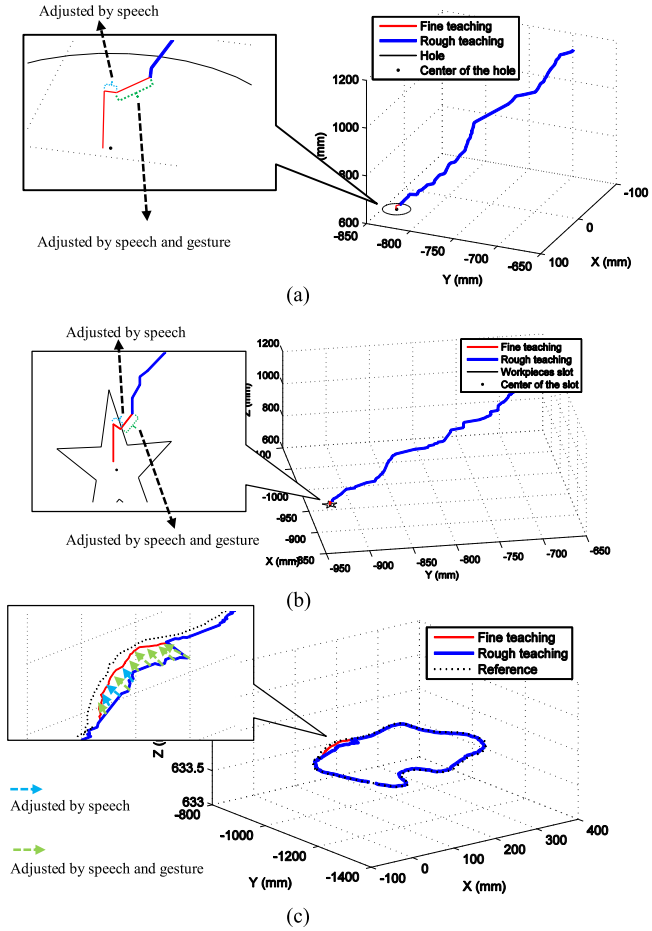

Fig. 8. Robot teaching by using gestures and adjustment after using speech recognition. (a) Peg-into-hole experiment. (b) Placing workpieces experiment. (c) Irregular trajectory tracking experiment.

gesture when the robot end-effector deviates from the reference trajectory.

B. Result Analysis

Fig. 8 shows the trajectory of the robot end-effector during robot teaching. The blue bold lines represent robot teaching by using three-dimensional gestures, while the red thin lines represent adjustment after using fine teaching. Fig. 8(a) shows the result of the peg-into-hole experiment. When moving closer to the target hole, the use of gestures can still complete the task. But with the adjustment of fine teaching, the robot end-effector can be accurately moved to the center of the hole. Fig. 8(b) shows the result of placing workpieces experiment. Similar to the peg-into-hole experiment, using fine teaching to adjust the robot EE can help it precisely move to the center of the slot.

However, different from the peg-into-hole experiment, the robot would fail to complete the task without adjustment by using fine teaching. Because only when the workpiece and the corresponding slot match exactly, the workpiece can be placed into the slot. Fig. 8(c) shows the result of irregular trajectory tracking. By employing dynamic gestures for robot teaching, the actual trajectory of the robot EE had a big deviation

TABLE II
OPERATION TIME OF 16 HOLES FOR THREE METHODS (S)

Holes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Our method	28.2	27.2	27.5	26.7	27.3	25.3	27.4	27.6	25.6	26.7	26.5	27.4	19.6	20.2	19.5	19.8
Method [9]	44.6	42.3	40.8	42.3	39.8	38.9	39.7	40.6	41.2	44.3	42.1	43.2	33.2	30.5	31.8	31.7
Method [13]	--	--	--	--	47.9	--	--	--	48.8	--	--	--	41.4	46.8	42.3	39.8

T1: Test 1 and T2: Test 2.

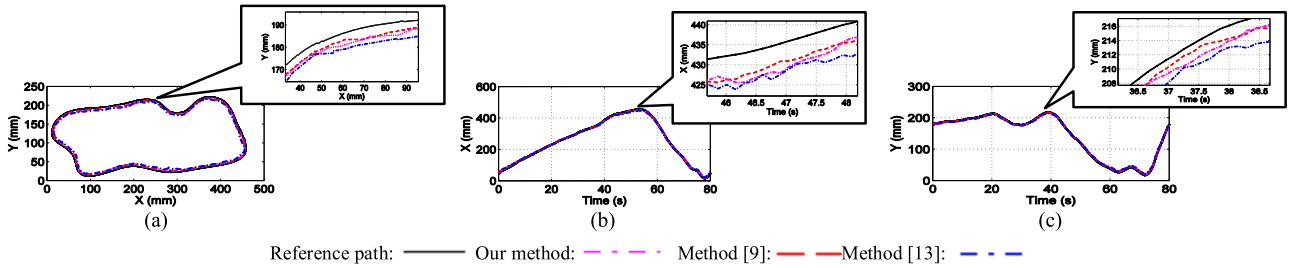

Fig. 9. Tracking results of experiment 3. (a) Trajectory of the robot's movement. (b) Trajectory in the *x*-direction. (c) Trajectory in the *y*-direction.

from the reference trajectory. After adjusting by using fine teaching, the actual trajectory of the robot fits closer to the reference trajectory. The results indicated that high-precision tasks can be performed well after fine teaching using speech and static gesture.

The results of the peg-into-hole experiment are shown in Table II. In this experiment, in order to verify the naturality of the proposed interactive method, we considered the task failed if the operator did not successfully place the pegs into their corresponding holes within 50s. The shorter the time required for the selected volunteers to complete the corresponding experiment through the proposed interactive method, the interaction method is considered more natural. We are based on the idea that the higher the naturalness of the human–robot interaction method, the less time it takes for a new operator to control or use the robot. The referee started timing when the operator gave speech command to let the robot begin its motion. And the referee stopped timing when the peg was placed into the hole covering the thickness of the steel plate. As Table II shows, there were only six tasks completed by using method [13], and most of which were successfully performed when the diameter of the hole is larger. All the tasks were completed by the proposed method and method [9]. With the help of combination of rough teaching and fine teaching, the operation time of our method is much shorter than that of method [9], which signified our method is more efficient than methods [9] and [13].

In placing workpieces experiment, five volunteers (They are junior and senior students in the school and they have not been exposed to the control of the robot before the relevant knowledge) were invited to perform the peg-into-hole tasks. Table III shows the success times and the mean of operation time for each volunteer. The tasks were barely implemented by using methods [9] and [13] for the lack of speech control, which made it hard for operators to perform the tasks requiring high precision. Although there were two volunteers successfully performing the task once using method [9], they spent too much time to adjust the position of the circular object, which caused the mean operation time of the task relatively long. It can be indicated from the results that our method performs better under the condition that requires high precision. Our method is more accurate than [9] and [13] when fine teaching (speech and static gesture) is used for placing workpieces task. Another irregular trajectory tracking experiment was performed for the comparison of the three methods. In this experiment, we did not import fine teaching because the speech control might take a long time during the moving process of the robot EE. The tracking result is shown in Fig. 9. Fig. 9(b) and (c) shows the trajectories unfolding in *X*-axis and *Y*-axis, respectively. The small windows show the parts of the amplified tracking trajectories. The black solid line represents the reference path, and the sapphire dash-dot line represents our method. The red dashed line and the fuchsia dash-dot line represent the trajectories of methods [9] and [13], respectively. Because the speech control instructions were not introduced, the accuracy of our method is relatively lower than that of method [9]. It can be seen from these figures that the trajectories of method [9] are the closest to the reference path, which signifies that method [9] is the most accurate one.

TABLE III
RESULT OF PLACING WORKPIECES TASK

	V1		V2		V3		V4		V5	
	N	Tm	N	Tm	N	Tm	N	Tm	N	Tm
Our method	4	64	4	66	4	62	4	64	4	64
Method [9]	1	82	--	--	--	--	--	--	1	86

V: Volunteer, *N*: The successful times, and *Tm*: The mean of operation time (s).

The results of the tracking task based on hand motion are shown in Table IV. The mean errors of the two tests using our method and methods [9] and [13] were 4.41 and 4.62 mm, 3.62 and 3.67 mm, and 5.73 and 5.96 mm, respectively, which means that method [9] is of the highest precision.

VII. DISCUSSION

This paper focuses on the robot teaching application based on the natural human–robot interaction. This paper presents an

TABLE IV
TRACKING ERRORS OF THE THREE METHODS. (T1: TEST 1, T2: TEST 2)

	Minimum (mm)		Mean (mm)		Maximum (mm)		Time (s)	
	T1	T2	T1	T2	T1	T2	T1	T2
Our method	1.31	1.23	4.41	4.62	12.34	13.15	32	33
Method [9]	1.21	1.36	3.62	3.67	9.68	10.12	145	139
Method [13]	1.33	1.25	5.73	5.96	18.23	17.98	167	175

online robot teaching method that fuses speech and gesture. The application conditions of this method are to fix a depth camera and a microphone near the robot to capture the user's gestures and voice. At the same time, the user's hand wears an IMU, which is used to accurately measure the hand gestures of the user. This method allows the staffs who lack teaching skills to perform the task of robot teaching quickly and effectively. However, because the method is still in the embryonic stage, the above-mentioned constraint conditions are harsh. In future research, our team hopes to use the augmented reality technology and the movement tracking technology. The worn sensor will be used to measure the user's gestures at anytime and anywhere. The user's movement will be no longer restricted by the region. At the same time, the designed virtual robot can verify the teaching task in the process of teaching. There is no need for real robots to debug repeatedly so that the safety of the user and teaching efficiency can be ensured. Since the hand is moving continuously, in future research, the teaching system will sample the continuous gestures by a novel Lyapunov–Krasovskii functional [37], [38], which takes full advantage of the available information about the actual sampling pattern, so that the conservativeness of the condition can be further reduced. Moreover, the advantage of sliding mode control [39]–[41] is that it can overcome the uncertainty of the system and has strong robustness and fast response to the disturbed and dynamics systems, especially for the nonlinear systems. In the future research work, the sliding mode control can be used to make the robot respond quickly according to the user gesture and plan a smooth trajectory as much as possible, which improves the teaching stability and efficiency. Further investigation of force feedback would be considered to improve the online teaching method.

VIII. Conclusion

This paper presented an online robot teaching system that utilized a Kinect sensor and an IMU. The IKF was developed to estimate the position of the human hand more stably. FQA and IPF were used to estimate the orientation of the hand. Microsoft speech SDK was used in the system to collect the speech of the operator. AVFT was proposed to fuse the speech and the hand gesture. The maximum entropy algorithm was employed to transform the text into the instructions. The feasibility of manipulation for high-precision tasks was validated in the experiment.

The proposed interface could be extended to the actual industrial scene. By using gesture and speech, operators can control the robot without complex operations. Besides, the system can teach two collaborative manipulators using two hands easily.

References

[1] S. Wang, F. Xu, L. Chen, F. Zou, and B. Li, "Industrial robot components assembly based on machine vision technology," *Modular Mach. Tool Automat. Manuf. Techn.*, vol. 8, pp. 107–110, 2015.

[2] P. D. Labrecque, J. M. Hache, M. Abdallah, and C. Gosselin, "Low-impedance physical human-robot interaction using an active-passive dynamics decoupling," *IEEE Robot. Autom. Lett.*, vol. 1, no. 2, pp. 938–945, Jul. 2016.

[3] H. Lee and J. Kim, "A survey on robot teaching: Categorization and brief review," *Appl. Mech. Mater.*, vol. 330, pp. 648–656, 2013.

[4] Y. Mohammad and T. Nishida, "Learning interaction protocols by mimicking understanding and reproducing human interactive behavior," *Pattern Recog. Lett.*, vol. 66, pp. 62–70, 2015.

[5] K. B. Cho and B. H. Lee, "Intelligent lead: A novel HRI sensor for guide robots," *Sensors*, vol. 12, no. 6, pp. 8301–8318, 2012.

[6] J. Du, C. Mouser, and W. Sheng, "Design and evaluation of a teleoperated robotic 3-D mapping system using an RGB-D sensor," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 46, no. 5, pp. 718–724, May 2016.

[7] L. Yuan, "Improved hidden Markov model for speech recognition and POS tagging," *J. Central South Univ.*, vol. 19, no. 2, pp. 511–516, 2012.

[8] K. Livescu *et al.*, "Speech production in speech technologies: Introduction to the CSL special issue," *Comput. Speech Lang.*, vol. 36, pp. 165–172, 2016.

[9] H. Lee, J. Kim, and T. Kim, "A robot teaching framework for a redundant dual arm manipulator with teleoperation from exoskeleton motion data," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2014, pp. 1057–1062.

[10] L. Meli, C. Pacchierotti, and D. Prattichizzo, "Sensory subtraction in robot-assisted surgery: Fingertip skin deformation feedback to ensure safety and improve transparency in bimanual haptic interaction," *IEEE Trans. Bio-Med. Eng.*, vol. 61, no. 4, pp. 1318–1327, Apr. 2014.

[11] J. Buzzi, C. Gatti, G. Ferrigno, and E. D. Momi, "Analysis of joint and hand impedance during teleoperation and free-hand task execution," *IEEE Robot. Autom. Lett.*, vol. 2, no. 3, pp. 1733–1739, Jul. 2017.

[12] V. Gonzalez-Pacheco, M. Malfaz, F. Fernandez, and M. A. Salichs, "Teaching human poses interactively to a social robot," *Sensors*, vol. 13, no. 9, pp. 12406–12430, 2013.

[13] P. Neto, J. Pires, and A. Moreira, "High-level programming and control for industrial robotics using a hand-held accelerometer-based input device for gesture and posture recognition," *Ind. Robot*, vol. 37, no. 2, pp. 137–147, 2015.

[14] H. Lin and Y. H. Lin, "A novel teaching system for industrial robots," *Sensors*, vol. 14, no. 4, pp. 6012–6031, 2014.

[15] G. Du, P. Zhang, and D. Li, "Human-manipulator interface based on multisensory process via Kalman filters," *IEEE Trans. Ind. Electron.*, vol. 61, no. 10, pp. 5411–5418, Oct. 2014.

[16] F. Maric *et al.*, "Kinematics-based approach for robot programming via human arm motion," *J. Braz. Soc. Mech. Sci. Eng.*, vol. 39, no. 7, pp. 2659–2675, 2017.

[17] C. Chao and A. Thomaz, "Timed Petri nets for fluent turn-taking over multimodal interaction resources in human-robot collaboration," *Int. J. Robot. Res.*, vol. 35, no. 11, pp. 1330–1353, 2016.

[18] D. Tang *et al.*, "A novel multimodal communication framework using robot partner for aging population," *Expert Syst. Appl.*, vol. 42, no. 9, pp. 4540–4555, 2015.

[19] D. Simonetti *et al.*, "Multimodal adaptive interfaces for 3D robot-mediated upper limb neuro-rehabilitation," *Robot. Auton. Syst.*, vol. 85, pp. 62–72, 2016.

[20] M. Johnston and S. Bangalore, "Finite-state multimodal parsing and understanding," in *Proc. 18th Conf. Comput. Linguist.*, 2000, pp. 369–375.

[21] A. Cherubini *et al.*, "A unified multimodal control framework for human–robot interaction," *Robot. Auton. Syst.*, vol. 70, pp. 106–115, 2015.

[22] H. Jiang *et al.*, "Enhanced control of a wheelchair-mounted robotic manipulator using 3-D vision and multimodal interaction," *Comput. Vision Image Understand.*, vol. 149, pp. 21–31, 2016.

[23] K. Yang, D. Yong, and S. Lv, "Relative distance features for gait recognition with Kinect," *J. Vis. Commun. Image Represent.*, vol. 39, pp. 209–217, 2016.

[24] S. P. Won, W. W. Melek, and F. Golnaraghi, "A fastening tool tracking system using an IMU and a position sensor with Kalman filters and a fuzzy expert system," *IEEE Trans. Ind. Electron.*, vol. 56, no. 5, pp. 1782–1792, May 2009.

[25] T. Khalid, Z. Mourad, C. Jean-Bernard, and B. Mohammed, "Bayesian bootstrap filter for integrated GPS and dead reckoning positioning," in *Proc. IEEE Int. Symp. Ind. Electron.*, 2007, pp. 1520–1524.

[26] S. P. Won, W. W. Melek, and F. Golnaraghi, "A Kalman/particle filter-based position and orientation estimation method using a position sensor/inertial measurement unit hybrid system," *IEEE Trans. Ind. Electron.*, vol. 57, no. 5, pp. 1787–1798, May 2010.
[27] Y. Wei, J. Qiu, and H. R. Karimi, "Quantized ∞ filtering for continuous-time Markovian jump systems with deficient mode information," *Asian J. Control*, vol. 17, no. 5, pp. 1914–1923, 2015.
[28] Y. Wei *et al.*, "Filtering design for two-dimensional Markovian jump systems with state-delays and deficient mode information," *Inf. Sci.*, vol. 269, no. 4, pp. 316–331, 2014.
[29] Y. Wei, J. Qiu, and H. R. Karimi, "Fuzzy-affine-model-based memory filter design of nonlinear systems with time-varying delay," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 504–517, Apr. 2018.
[30] X. He, Y. Le, and W. Xiao, "MEMS IMU and two-antenna GPS integration navigation system using interval adaptive Kalman filter," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 28, no. 10, pp. 22–28, Oct. 2013.
[31] P. Zhang, B. Li, and G. Du, "A wearable-base and markerless human-manipulator interface with feedback mechanism and Kalman filters," *Ind. Robot*, vol. 42, no. 5, pp. 485–495, 2015.
[32] X. Yun, E. R. Bachmann, and R. B. McGhee, "A simplified quaternion based algorithm for orientation estimation from Earth gravity and magnetic field measurements," *IEEE Trans. Instrum. Meas.*, vol. 57, no. 3, pp. 638–650, Mar. 2008.
[33] G. Du, P. Zhang, and X. Liu, "Markerless human–manipulator interface using leap motion with interval Kalman filter and improved particle filter," *IEEE Trans. Ind. Informat.*, vol. 12, no. 2, pp. 694–704, Apr. 2016.
[34] W. Wang, Q. Zhao, and T. Zhu, "Research of natural language understanding in human-service robot interaction," *Microcomput. Appl.*, vol. 31, no. 3, pp. 45–49, 2015.
[35] X. Sun *et al.*, "Urban expressway traffic state forecasting based on multi-mode maximum entropy model," *Sci. China Technol. Sci.*, vol. 53, no. 10, pp. 2808–2816, 2010.
[36] F. J. S. Fernández-Caramés and V. Moreno, "A real-time indoor localization approach integrated with a geographic information system (GIS)," *Robot. Auton. Syst.*, vol. 75, pp. 475–489, 2016.
[37] Y. Wang, H. Shen, and D. Duan, "On stabilization of quantized sampled-data neural-network-based control systems, " *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3124–3135, Oct. 2017.
[38] Y. Wang, Y. Xia, and P. Zhou, "Fuzzy-model-based sampled-data control of chaotic systems: A fuzzy time-dependent Lyapunov–Krasovskii functional approach," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1672–1684, Dec. 2017.
[39] Y. Wang, Y. Xia, H. Shen, and P. Zhou, "SMC design for robust stabilization of nonlinear Markovian jump singular systems," *IEEE Trans. Automat. Control*, vol. 63, no. 1, pp. 219–224, Jan. 2018.
[40] Y. Wang, H. Shen, H. R. Karimi, and D. Duan, "Dissipativity-based fuzzy integral sliding mode control of continuous-time T-S fuzzy systems," *IEEE Trans. Fuzzy Syst.*, pp. 1–1, 2017, doi: 10.1109/TFUZZ.2017.2710952.
[41] Y. Wang, Y. Gao, H. R. Karimi, H. Shen, and Z. Fang, "Sliding model control of fuzzy singularly perturbed systems with application to electric circuit," *IEEE Trans. Syst., Man, Cybern., Syst.*, pp. 1–9, 2017, doi: 10.1109/TSMC.2017.2720968.

Guanglong Du received the Ph.D. degree in computer application technology from South China University of Technology, Guangzhou, China, in 2013.

He is currently an Associate Professor with the Computer Science and Engineering School, South China University of Technology. His research interests include intelligent robotics, human–computer interaction, artificial intelligence, and machine vision.

Mingxuan Chen received the B.S. degree in computer science from Hunan University, Changsha, China, in 2015. He is currently working toward the Ph.D. degree in computer science and technology at the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China.

His research interests include human–swarm interaction and swarm robotics.

Caibing Liu is currently working toward the Undergraduate degree in computer science and technology at the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China.

Her research interests include machine vision and human–computer interaction.

Bo Zhang received the B.S. degree in computer science and technology from the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China, in 2017. He is currently working toward the Postgraduate degree in computer science and technology at the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China.

His research interests include robot teaching, machine vision, and human–computer interaction.

Ping Zhang received the B.S. degree in mechanical engineering and the M.S. and Ph.D. degrees in robotics from Tianjin University, Tianjin, China, in 1985, 1988, and 1994, respectively.

He is currently a Professor with the Computer Science and Engineering School, South China University of Technology, Guangzhou, China. His research interests include intelligent networked robotics, intelligent networked manufacturing, human–computer interaction, and real-time embedded systems.