electronics

MDPI

Article

An Intelligent Hybrid–Integrated System Using Speech Recognition and a 3D Display for Early Childhood Education

Kun Xia, Xinghao Xie *, Hongliang Fan and Haiyang Liu

Department of Electrical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; xiakun@usst.edu.cn (K.X.); 182540395@st.usst.edu.cn (H.F.); 202421584@st.usst.edu.cn (H.L.)
* Correspondence: 192531442@st.usst.edu.cn

Abstract: In the past few years, people's attitudes toward early childhood education (PAUD) have undergone a complete transformation. Personalized and intelligent communication methods are highly praised, which also promotes the further focus on timely and effective human–computer interaction. Since traditional English learning that relies on parents consumes more time and energy and is prone to errors and omissions, this paper proposes a system based on a convolution neural network (CNN) and automatic speech recognition (ASR) to achieve an integrated process of object recognition, intelligent speech interaction, and synchronization of learning records in children's education. Compared with platforms described in the literature, not only does it shoot objects in the real-life environment to obtain English words, their pronunciation, and example sentences corresponding to them, but also it combines the technique of a three-dimensional display to help children learn abstract words. At the same time, the cloud database summarizes and tracks the learning progress by a horizontal comparison, which makes it convenient for parents to figure out the situation. The performance evaluation of image and speech recognition demonstrates that the overall accuracy remains above 96%. Through comprehensive experiments in different scenarios, we prove that the platform is suitable for children as an auxiliary method and cultivates their interest in learning English.

Keywords: intelligent speech interaction; three-dimensional display; early education; childhood care; internet of things (IoT)

check for updates

Citation: Xia, K.; Xie, X.; Fan, H.; Liu, H. An Intelligent Hybrid–Integrated System Using Speech Recognition and a 3D Display for Early Childhood Education. *Electronics* **2021**, *10*, 1862. https://doi.org/10.3390/electronics10151862

Academic Editors: Daniele Spoladore, Alberto Trombetta and Liming Luke Chen

Received: 30 June 2021
Accepted: 29 July 2021
Published: 3 August 2021

1. Introduction

Speech plays an uncomplicated but important role in human interaction because it has the ability to deliver useful information quickly and precisely [1]. Today, people can use different smart devices and obtain the medium of speech that they are committed to by extracting features from daily communication, which makes it a basis for the booming industry of voice assistants [2]. Language was originally a unique tool for communication and the exchange of information among people [3]. With the continuous improvement in science and technology, human beings can communicate with machines through voice commands and convey their instructions, which make the corresponding tasks possible [4]. Siri and Google assistant try to focus on the relation between human voice information and automatic operations. The distribution of methods based on real-time intelligent interaction and the content of the dialogue ranges from research to industry, education, and entertainment [5,6]. Automatic speech recognition (ASR) has been widely used in the process of human–computer interaction, which holds a pivotal position in people's daily lives [7]. ASR is an attempt to aim at a higher level of semantic understanding, which has an effect on the transformation of acoustic signals into textual information [8].

The focus of early childhood education (PAUD) is to bring about fundamental changes in curriculums in the form of a set of plans that provide benefits to suitable infants, toddlers, and preschoolers [9]. Basically, to attract attention and learn efficiently, games that are composed of education resources should convey happiness to children and arouse their interest. Child-centered education emphasizes that the motivation for growth originates from children themselves and believes that children are individuals with an independent spirit and that education should stimulate and extend students' independent spirit [10]. For example, the teaching of fairy tales in elementary school from a child-centered perspective emphasizes a form of classroom in which teachers and students grow together. The autonomy of students' learning is reflected in their willingness to learn spontaneously, and they become the master in order to take the initiative to solve problems in the face of difficulties [11].

As a carrier of specific features, the aim of augmented reality (AR) is to make people have a better understanding of the visual resources of the world [12]. AR is readily available and advances have been made in real-world applications, which are a credit to its smart and accessible insight. The application of AR in education is also proliferating. As a tool to deal with the real physical world, which makes it an enhanced version of the world, it is appropriate that AR help developers build a variety of virtual-reality teaching scenarios to improve various educational activities [13]. The coexistence of virtual objects and the real-life environment allows students to observe complex three-dimensional relationships and abstract scientific concepts, experience phenomena that cannot be felt in the real world, be immersed in a scenario that integrates two- or three-dimensional virtual information and real objects, and participate in training that cannot be done practically [14]. The characteristics of AR help users carry out learning tasks and the information that AR displays will not bore children.

Traditional intelligent speech interaction systems place servers that store and process very large amounts of data in the runtime environment, and some are directly embedded in the gateway [15]. Due to hardware performance limitations, the current arrangements are inadequate to meet the needs of data collection and data mining. On the other hand, the data are stored in a single device, which inevitably causes an isolated data island to form. It is not conducive to the analysis of user data by enterprises. The rapid development of the Internet of Things (IoT) is considered to provide auxiliary support for the prosperity of intelligent speech interaction systems [16]. In some scenarios, the terminal must face a complex electromagnetic environment, extreme temperatures, and insufferable humidity. It is very difficult to effectively extract data under various disturbances, which is a catalyst for deep learning. Machine learning needs detailed differences in features between each primitive in advance. The effectiveness of machine learning is dependent on the accuracy of feature extraction, and it is not suitable in efficient, low-cost, and networked terminals [17]. While the characteristics of data are not a necessity for deep learning algorithms, such as the typical convolutional neural network (AR), the only thing you need to do is feed the data directly into the CNN network to train it or obtain results. CNNs have been widely used in image recognition, speech recognition, and many other fields [18].

A typical "child-centric" content-sharing education platform starts with looking at the needs of students and understanding what they are trying to do, which taps into the intrinsic passion of learning within them. However, in most cases, it comes with a heavy dose of memorization, which puts an emphasis on rote-based assessments of words and formulas rather than developing the mind to influence the ways that children see the world. It used to be a natural way to make more effective use of children's time since it provides the quickest path and makes it easier for students to get started. However, with the gradual entry of ASR, 3D displays, CNNs, and other new technologies into the area of education, it is time to wake up to the fact that a hybrid–integrated system that combines the above three basic features will help us seize the immediate market opportunities, give children early exposure to embedded systems, cultivate children's personal interests, and enable children to pursue cooperative learning on their own initiative [10]. The proposed platform is primarily oriented towards young children and, due to the individual instruction and intangibles of education services, the customer satisfaction has some characteristics that are different from those in other application areas. Hence, the message to the user and the emotional experience that the product provides have become important aspects to take into account. In summary, the focus of this paper is as follows.

The present paper contributes to the integration of smart voice commands and AR in education, which is implemented through a new type of early education platform embedded in the storage box. It is suitable for homes and kindergartens where preschool education takes place. This platform focuses on exploring a new way of learning English, can serve children as an auxiliary, and can cultivate children's interest in imagining virtual objects, which makes it possible to establish an association between their mathematical skills and cognitive abilities [19]. To obtain a pleasant and meaningful education from which children gain useful knowledge is the final goal.

The rest of this paper is organized as follows. Section 2 presents a literature review, describing studies that apply image and speech recognition to education. The history and applications of AR in education are also briefly reviewed. Section 3 presents the main methodologies of CNNs and intelligent speech interaction. Section 4 introduces the design of the proposed early education platform, including the hardware and methods of programming. Section 5 shows the performance of the experimental platform and discusses how it could be improved compared with previous works. A summary and possible future improvements are given in the final section.

2. Background and Objects

Deep learning, an emerging discipline in machine learning, aims to automatically extract and represent multi-layer features from data samples using a series of nonlinear transformations in a data-driven manner to derive characteristics from the original resources in order from low-level to high-level, from concrete to abstract, and from general to specific [20]. Deep learning has already made breakthroughs in educational applications. From the literature, it can be concluded that traditional image recognition mainly includes four parts (image acquisition, image preprocessing, feature extraction, and pattern matching) and most of the work is concentrated in the image preprocessing and feature extraction stages, which involve more complicated processes [17]. In most recognition experiments, images with a single background are used and the data collection is completed in the same experimental environment, which circumvents the problems of light intensity changes, the depth of shadows, and occlusion during the acquisition and eliminates the influence of external factors [21]. In addition, the algorithm removes the complex background of images, which improves the speed and accuracy of the process.

In recent years, many studies, conducted to look for different IoT smart applications, would like to make human–computer interaction more natural and comparable with a simple mouse, keyboard, or Tablet PC. A number of new devices and technologies that satisfy certain basic requirements have been proposed for recognizing voice commands and giving appropriate feedback [22]. The construction of traditional speech interaction systems is focused on the stages of speech recognition and speech synthesis [23]. In most recognition tasks, the audio used in the experiments is normally sampled in a strictly limited environment in order to get rid of the disturbances that make results unreliable [24]. In contrast, the actual audio will have problems, such as noise interference, a low dialect recognition rate, and poor tolerance for errors of semantic understanding, which affect the accuracy of speech interaction systems to varying degrees [25,26].

AR has the ability to build an immersive virtual environment, new interactive methods, and contextual navigation scenarios, and these advantages help to raise awareness of engagement in order to gain better outcomes [27]. In addition, AR also facilitates the understanding of auxiliary knowledge and the construction of collaborative learning methods by vividly displaying the inaccessible real world in the form of three-dimensional virtual models in front of learners, making it easier to implement concrete observations of the real world and thus explore the nature and laws of things [28]. AR integrates multiple information resources and provide learners with knowledge access interfaces that reduce the cognitive load, expanding the application of teaching assistant systems in three dimensions: physical perception, cognition, and context [29]. An education platform based on AR requires less human intelligence, which significantly increases the digestion and absorption efficiency of knowledge [30]. On the other hand, the combination of fictitious and realistic approaches conveys emotions through a scenario description, allowing students to develop their own innovative works, which exercises students' ability with the help of AR to express complex concepts and solve problems [31]. Due to AR, solutions for designing instructional programs and reducing costs emerge. In scientific demonstrations and experiments, full use of many high-quality resources is able to be made [32].

Scientific practices should not have ethical attributes, but the development and application of technology are done by human beings, which inevitably leads to research misconduct [33]. The new technique plays an indelible role in transforming the world for human beings, which is also double-sided like other things, and we should examine it from a moral point of view. With the corresponding ethical principles, people are constantly faced with ethical choices in scientific and technological practice. Every choice and decision may bring about both positive and negative effects and the accumulation of such possibilities and uncertainties forms the ethical risks we have to face. Ethical risk refers to "uncertain events or conditions that may arise in the ethical relationship between people, people and society, people and nature, and people and themselves due to positive or negative effects, especially the uncertain negative ethical effects, such as dysfunctional ethical relationships, social disorder, uncontrolled mechanisms, people's misbehavior, and psychological imbalances". Focusing on ethical risks is important to the current rapid development of society and continuous technological progress. Thus, in this paper, as the relationship between science and society has changed from "science in society" to "science with society", which requires prototypes to respond to social values, moral expectations and security demands, the proposed platform mainly concerns economic, privacy, and environmental health and safety issues. Power savings, personal information protection, and completely recyclable materials are factors to be taken into account in the proposed platform's design.

3. Related Methodology

3.1. Convolution Neural Network (CNN)

The CNN is currently the most widely used deep learning model for image recognition. The CNN can directly take two-dimensional images as the input, avoiding complex preprocessing procedures. The connection between each layer of the network is reduced by means of sparse connections and weight sharing, which results in a decrease in parameters for training and improves the generalization performance.

The CNN is fundamentally viewed as a model trained with known patterns so that the network learns a large number of mapping relationships between inputs and outputs. Figure 1 shows its structure, which consists of an input layer, a convolutional layer, a pooling layer, and a fully connected layer.

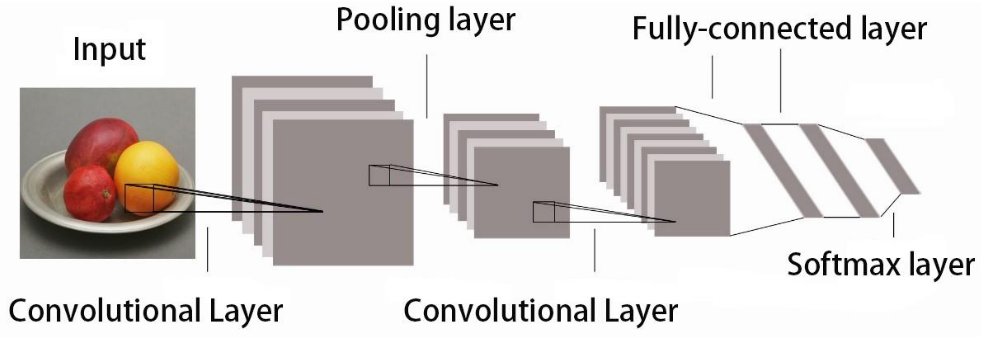


Figure 1. Frame diagram of the CNN model.

The convolutional layer is used to extract features from the inputs. It contains multiple convolution kernels, each element corresponding to a weight factor W and a deviation amount b. The convolutional layer contains an excitation function f to assist in expressing complex features. The relationship between the input x^{l-1} and the output x^l of the current layer is specified as follows:

$$x^l = f\left(W^l x^{l-1} + b^l\right) \tag{1}$$

where l denotes the number of channels, x^{l-1} and x^l represent the inputs and outputs, respectively, f is an activation function, W is the weight matrix of the convolution kernel, and b is the bias value.

The pooling layer is used to perform feature selection and information filtering on the data, and the general pooling model is represented as:

$$A_k^l(i,j) = \left[\sum_{x=1}^{f}\sum_{y=1}^{f} A_k^l(s_0 i + x, s_0 j + y)^p\right]^{\frac{1}{p}} \tag{2}$$

where s_0 represents the step size and p is the pre-specified parameter.

Forward propagation stage. In this stage of the CNN, the samples in the dataset are put into the network, which outputs results after the transformation of each layer. In the network, the output of the previous layer is used as the input of the current layer.

Back propagation stage. After forward propagation, a deviation needs to be defined to characterize the state of the network after this propagation. The back propagation process passes the deviation forward through the reverse transfer method layer by layer, so that neurons in the upper layer update their own weights according to the deviation.

The error of samples is calculated by the following formula:

$$E^{\mathrm{n}} = \frac{1}{2}\sum_{k=1}^{c}\left(t_k^n - y_k^n\right)^2 = \frac{1}{2}\|t^n - y^n\|_2^2 \tag{3}$$

where c is the number of classes of samples in the dataset, t_k^n is the target value corresponding to the kth dimension of the nth sample, and y_k^n is the kth dimension of the output corresponding to the nth sample.

The formula that presents the output of the current layer is as follows:

$$x^l = f\left(u^l\right) \tag{4}$$

where $u^l = W^l x^{l-1} + b^l$ is the variable of the activation function.

With respect to biases and weights in the network, the calculation of the partial derivative of the error is shown as follows:

$$\frac{\partial E}{\partial b} = \frac{\partial E}{\partial u}\frac{\partial u}{\partial b} = \delta \tag{5}$$

where $\partial u/\partial b = 1$ results in $\partial E/\partial b = \partial E/\partial u = \delta$.

The error signal and the partial derivative of the error with respect to all inputs of a cell are equal. Then, the sensitivity of the implied layer can be derived:

$$\delta^l = \left(w^{l+1}\right)^T \delta^{l+1} \times f\left(u^l\right) \tag{6}$$

Moreover, the sensitivity of the output layer is defined as:

$$\delta^l = f\left(u^l\right) \times (y^n - t^n) \tag{7}$$

The partial derivative of the error with respect to each weight in the layer is equal to the cross product of the input of the layer and the sensitivity. Then, the computed partial derivative is multiplied by a negative learning rate to obtain an update of the neuron weights in the layer.

$$\frac{\partial E}{\partial w^l} = x^{l-1}\left(\delta^l\right)^T \tag{8}$$

$$\Delta w^l = -\eta \frac{\partial E}{\partial w^l} \tag{9}$$

The process we followed to train the image recognition algorithm is as follows:

1. Randomly initialize all filters as well as other parameters and weight values.
2. Input the images, perform forward propagation to obtain the image features extracted by the CNN, and finally pass them through to the output layer, which outputs a vector containing the probability values of each class prediction.
3. Calculate the error, apply back propagation to calculate the gradient of the error corresponding to each weight in the network, and update the weight value of each filter to reduce the output error.
4. Repeat steps 2 to 4 until the training session reaches a set value.

3.2. Intelligent Speech Interaction

The main purpose of voice commands that apply in an IOT platform is to exchange information and share resources between terminals whose usages are different without manual control [34]. However, identifying what people say is probably a task that the machine has difficulties dealing with. In order to respond immediately, a Cloud Speech API is needed. First, the audio signal is recorded and sampled, which is a prerequisite of encoding. During natural language processing, the analog signal is transformed into a digital format, and with the help of the speech recognition engine the platform derives a translation so as to match the collected information stored in the database. As soon as the comparison between the input and the existing information is complete, the result obtained from a successful match will be converted into text messages and programs that activate or deactivate some function of the system [35]. Compared with natural language processing, speech synthesis is the conversion of corresponding texts or commands into sound outputs. The speech recognition database can be placed in either the cloud or the terminal. Since the algorithm that the speech recognition database currently uses is relatively complex and the data volume is huge, most researchers choose to place the database in the cloud. Thus, the online connection takes the place of traditional buttons and the cloud server replaces the local storage.

Normally, a speech recognition application includes network functions, audio playback, speech services (keyword extraction and data messaging), cloud services, etc. The process of speech interaction is shown in Figure 2.

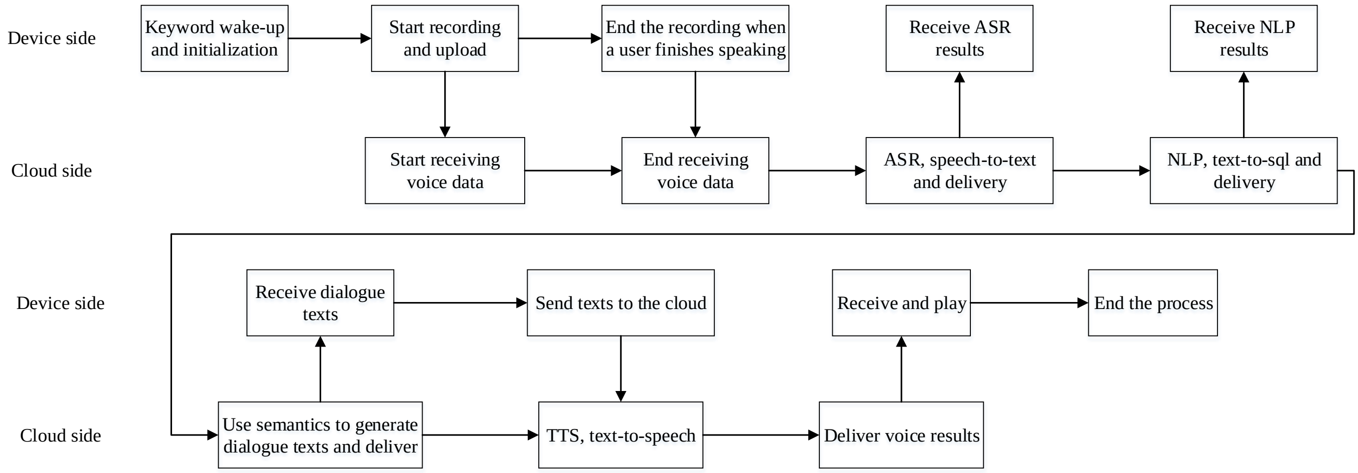

Device side
Keyword wake-up and initialization
Start recording and upload
End the recording when a user finishes speaking
Receive ASR results
Receive NLP results
Cloud side
Start receiving voice data
End receiving voice data
ASR, speech-to-text and delivery
NLP, text-to-sql and delivery
Device side
Receive dialogue texts
Send texts to the cloud
Receive and play
End the process
Cloud side
Use semantics to generate dialogue texts and deliver
TTS, text-to-speech
Deliver voice results

Figure 2. Speech recognition principle.

The signal input to the microphone is processed by echo cancellation, noise reduction, and keyword extraction modules, then pushed to the cloud for semantic recognition. Speech synthesis is performed on the dialogue text generated by a semantics module, which has more flexibility. Finally, a generation module receives new speech data and plays it.

3.3. Holographic Imaging

A holographic imaging system mainly consists of a cabinet, holographic film, video playback equipment, and facilities for holographic video source production, as shown in Figure 3. The cabinet is a five-sided hollow hexahedron bearing a spectroscope and light sources. The upper part is the device placement area, and the other four sides constitute the holographic imaging area where the best way to get messages to the particular audience lies. Generally, the holographic film is made of PVC that is optically transparent, and the material is in the shape of a tetrahedron, which contributes to the law that at an angle of 45 degrees the imaging result is the best. Of course, the video playback equipment is completely flexible and depends on the usage scenario; for example, a projector and a mobile phone should be used in a large-scale scene and a small-scale scene, respectively. Basically, the contents of a holographic projection can be divided into three types: virtual character animation, computer-synthesized images, and videos for science and technology.

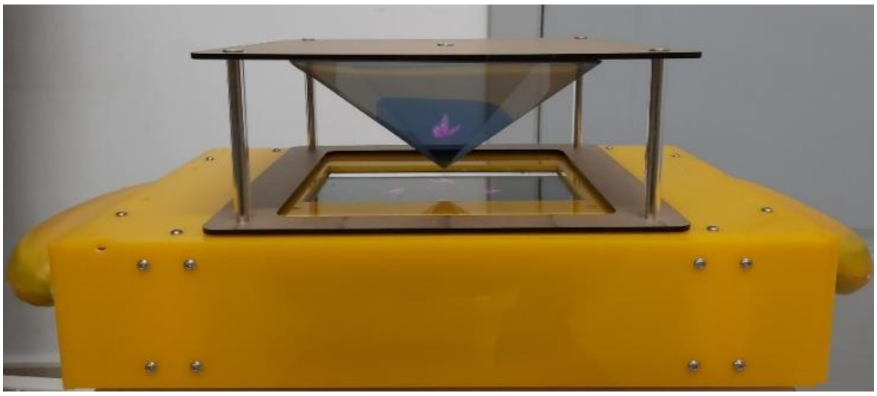


Figure 3. Outward appearance of the holographic imaging system.

Holographic imaging technology stems from Pepper's Ghost. In short, with the use of smooth glass and a special light source, objects appear as if out of nowhere, vanish into thin air, and then change places with other objects. The glass or transparent materials used in Pepper's Ghost have a different refractive index from air, and light propagates at different speeds in these two media. When light reaches the boundary of materials with different refractive indices, some of the light is usually emitted, and the rest will be refracted at a certain angle, which is also called transmission. The amount of reflected and transmitted light is able to be calculated by the Fresnel equation, so that the increase or the decrease can be controlled. Additionally, the angle of incidence as well as the polarization of light have an influence on this procedure. According to the plane mirror imaging principle, the position of virtual images formed by the reflection of the plane mirror and the position of the lamp is symmetrical with respect to the glass surface. On account of their relative locations, the audience cannot see the mirror and the source, and they will suppose that the scene that they perceive with their eyes issues from the location of the virtual image.

It follows from the above that holographic imaging technology based on the principle of plane mirror imaging and the reflection on the surface of a tetrahedron that is made of holographic film suspends the three-dimensional virtual models in midair above the cabinet. As shown in Figure 4, the lamp located above projects the animation onto the imaging position in the center of the holographic film at an angle of 45 degrees. The holographic imaging system is based on the principle of spectroscopic imaging, which shoots objects and produces three-dimensional images after preprocessing. Then, images of photographed objects or computer-made three-dimensional models are added to the construction of the scene, which displays a system with the combination of static and dynamic items.

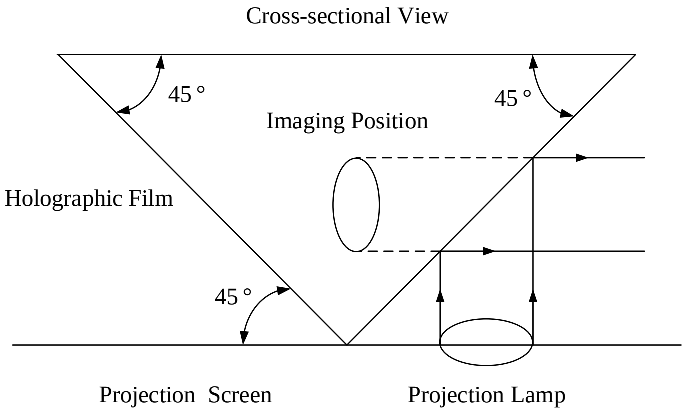


Figure 4. Schematic diagram of the holographic imaging system.

4. Design of Materials

4.1. System Architecture

The proposed design, derived from IoT-based voice command technology, image recognition technology, and 3D display technology, is a platform that supports children who have problems with learning English. Discarding old manual controls, the platform sets a voice trigger action that can keep children away from the display screen. In addition, in order to protect children's eyesight and adjust the study-time allocation strategy, the platform will only guide children to learn new words or review previous words at the corresponding time by means of the display indicator and audible alarms. On the other hand, it is well known that audio–visual integration has become a rule of modern education technology. The combination of words' pronunciation and holographic videos that contain knowledge achieves the purpose of understanding the abstract scientific concepts as well as related comparisons from the three aspects of hearing, oral practice, and vision. The following picture depicts the overall architecture of the design, including the hardware and software, of the early education platform.

The entire monitoring system can be separated into four layers, including the main control layer, the data collection layer, the cloud server layer, and the application layer (from the bottom to the top). The data in the system flow in both directions, so the object layer uploads real-time learning data to the cloud server for monitoring and visualization. Additionally, the user application layer can upload images and audio to the cloud server to remotely control the database, which replace the outdated materials in the main control layer.

As shown in Figure 5, the intelligent early education platform uses a camera to acquire real-time images for object recognition and collects audio data using a CB5654 voice module. Each platform uploads the progress information to the cloud server through the WIFI module to realize data analysis, management, and storage. Additionally, a host computer, a cloud database, and a WeChat applet are supported by the cloud server layer.

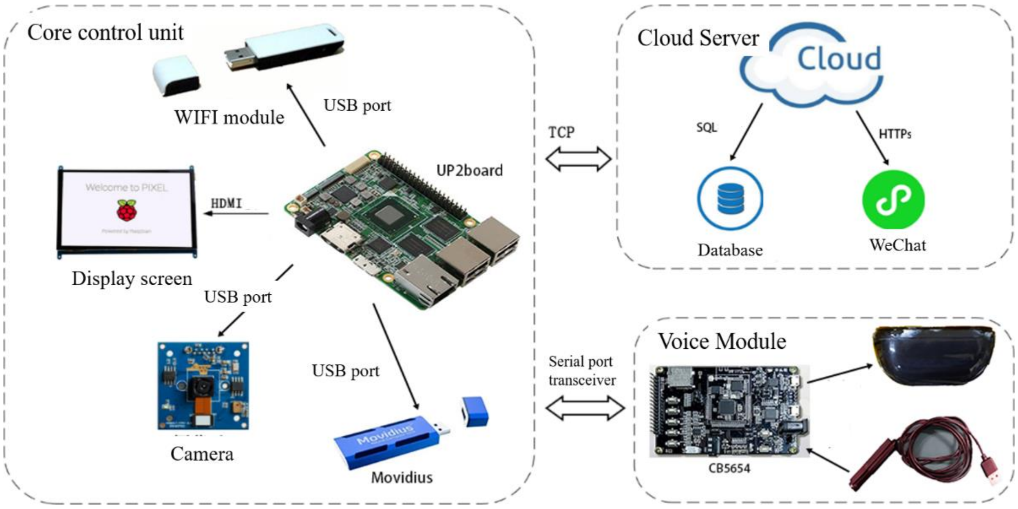


Figure 5. The architecture of the proposed system.

4.2. Hardware

The traditional structure of an early education system is retained in the integrated platform, and it has been improved and innovated upon. The mechanical design is aimed at integration, which requires a flexible structure in order to facilitate the search for the optimal angle during the testing phase. Additionally, a safe and reliable structure that is easy to transport was taken into consideration. By adding a holographic projection device, the playback of three-dimensional video helps children better understand some of the abstract words encountered while learning English so that the novelty will not soon wear off.

Children's safety is another factor that needs to be considered. The requirement of non-toxic materials that guarantee products' safety and make people feel relaxed, comfortable, and warm led us to select polylactic acid. The ability to recognize different colors carries children to healthy growth, which resulted in the usage of seven basic colors in the platform. The appearance is smooth, avoiding any sharp edges as well as small parts that children may choke on or swallow by mistake. The risk of a child pinching their fingers, limbs, torso, or head prevented us from using movable parts. The structural strength should leave a margin for error to withstand the weight and impact of a certain load and maintain the stability. Solidworks software was used for the computer-aided design in this study, and a three-dimensional model was established. At the same time, industrial-grade SLA 3D printing technology was selected to process the shell. The mechanical structure of the system is shown in Figures 6 and 7.

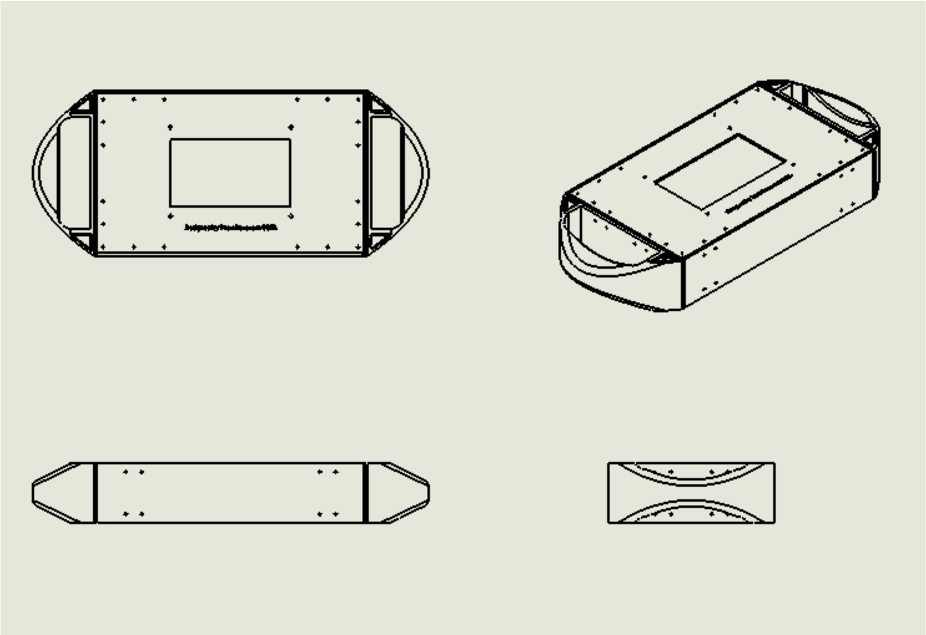


Figure 6. Entire mechanical structure of the early education platform.

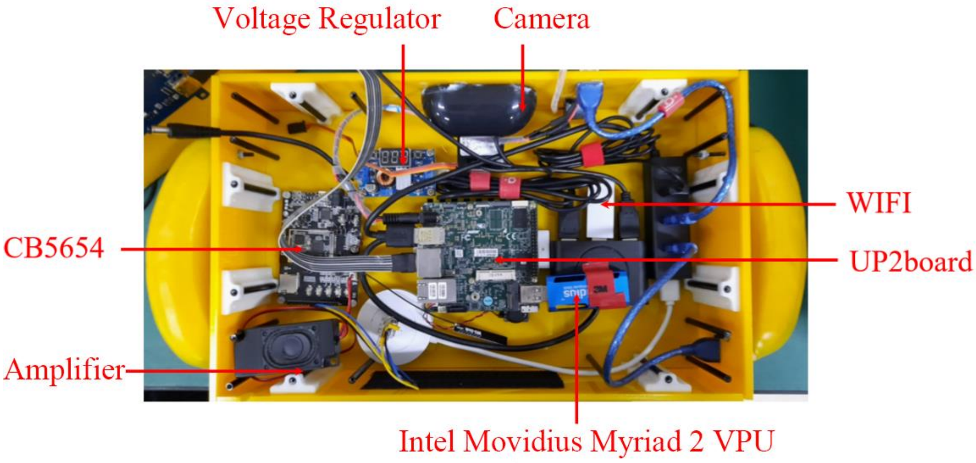


Figure 7. The base of the early education platform.

As shown in Figure 8, a UP Squared (UP^2) board, as the core of the whole system, schedules platform resources and runs deep learning models to realize the functions of single target detection, voice wakeup, and the storage of learning records. The intelligent speech module includes a CB5654 voice module, a microphone, and a speaker, which are used to implement the data collection and push audio information to the cloud speech engine. Thus, speech recognition, semantic analysis, and speech synthesis are able to proceed. The display module includes a touch screen and holographic imaging equipment. The camera is used to obtain graphical information about objects, and the WIFI module is used to provide network support to the platform.

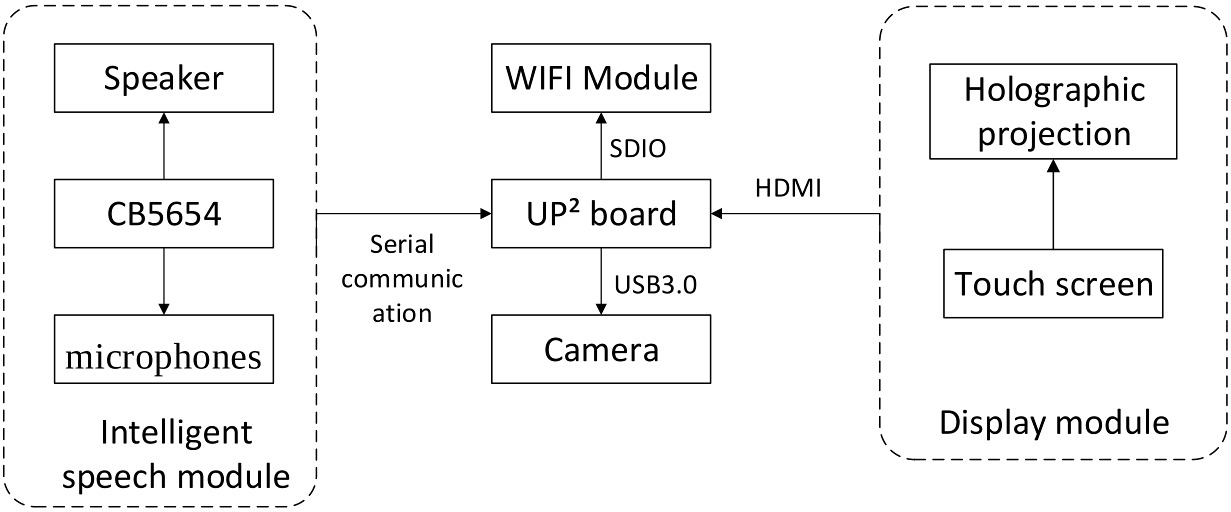


Figure 8. The relationship and interactions between components.

4.3. *Software*

The subroutine construction programming method was used for optimization and the functions of the platform were divided into three applications. The image recognition system and the education subsystem were developed using PyQt5, which integrates the Python programming language and the Qt library. The IDE is python idle. The design of the speech interaction system uses T-Head's Yun-on-Chip (YoC) Cloud Development Kit (CDK). CDK is a cloud-based all-in-one IDE tool, which includes the functions required for development, such as downloading required components, code editing, compilation, burning programs, and debugging. The multi-platform interconnection system uses the PHP scripting language as the backend, and the host computer interacts with the cloud database through the PHP file. The development environment was configured as PHP 7.4.6, IIS, and MySQL 8.0.20.0. Each subprogram module has its own means of manipulating the data flow as well.

As a Graphical User Interface (GUI) helps users control a platform, we were prompted to select a human–computer interaction interface. The flow chart for manipulating the early education platform is shown in Figure 9. The main interface we gravitated toward proposes an independent and modularized idea that favors the simplification of operation procedures. Thus, the image recognition system and the education subsystem respectively constitute a cradle of object classification and the broadcast of educational information composed of text, sound, and video. The functions of automatic speech recognition and synthesis mainly rely on the remote high-performance server, which was put forward for the realization of interactive control over the whole course of communication. Particularly, it features a close association with the AR camera, which centered on the input of voice commands and feedback on displaying objects in three dimensions. Finally, the platform, the cloud server, and the smart terminals form a network of desired results and events in which images, sounds, and the progress information can be accessed and exchanged.

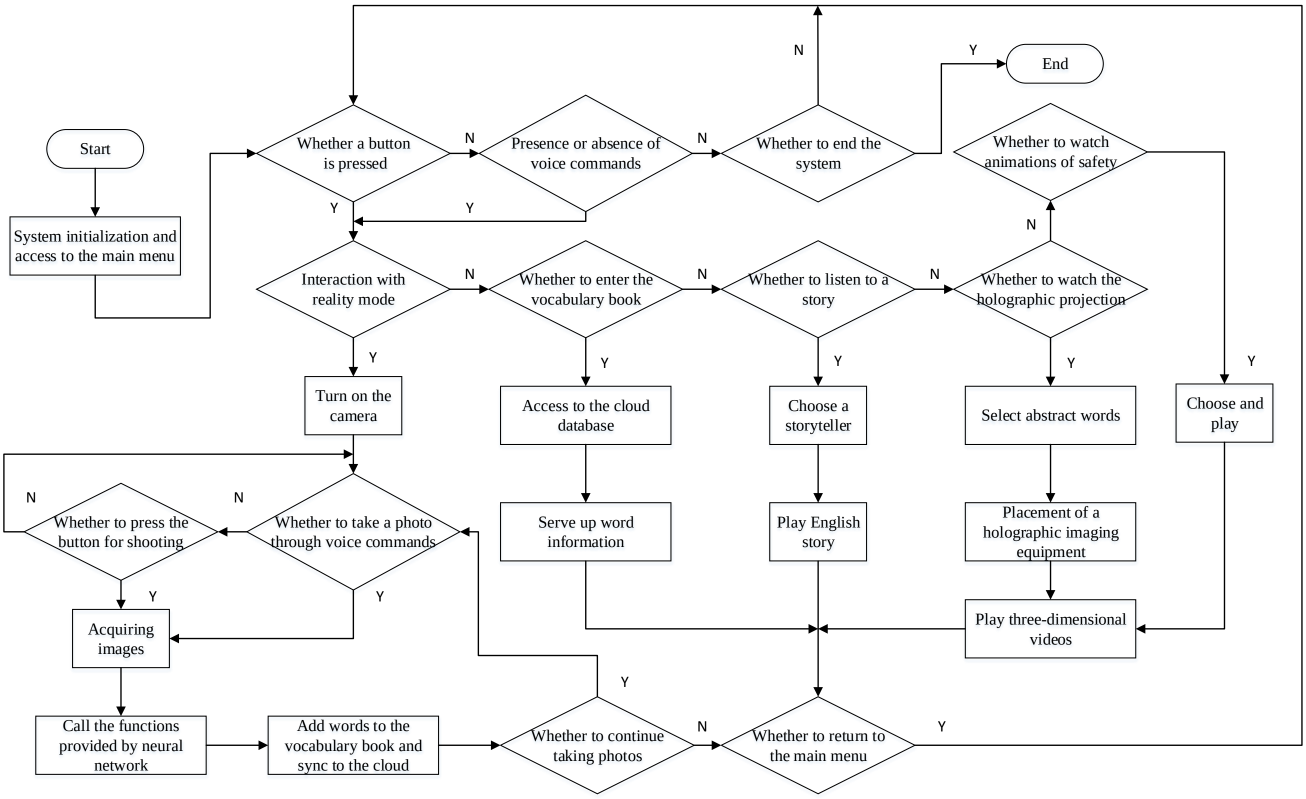

Start
System initialization and access to the main menu
Whether a button is pressed
Presence or absence of voice commands
Whether to end the system
End
Whether to watch animations of safety
Interaction with reality mode
Whether to enter the vocabulary book
Whether to listen to a story
Whether to watch the holographic projection
Turn on the camera
Access to the cloud database
Choose a storyteller
Select abstract words
Choose and play
Whether to press the button for shooting
Whether to take a photo through voice commands
Serve up word information
Play English story
Placement of a holographic imaging equipment
Play three-dimensional videos
Acquiring images
Call the functions provided by neural network
Add words to the vocabulary book and sync to the cloud
Whether to continue taking photos
Whether to return to the main menu
Y
N

Figure 9. The flow chart for controlling the system.

When the system parameters initialize, the TCP client is set up and the TCL server is connected to the host computer. Users enter the main menu and then call the general form of the system's functions. When voice commands are detected by the intelligent speech module, the UP^2 board obtains communication instructions through serial ports as the primary task of speech recognition is to establish the action of converting a voice command to the corresponding text. As soon as the speech recognition module presents itself, the education platform will engage in the interactive mode.

The control strategy can be divided into five parts: the optimization of image recognition and the intelligent speech interaction; a vocabulary book that presents words in the form of question-and-answer drills, allowing them to be recited much more easily; stories specifically chosen for their ability to imitate parents' voices; a holographic projection known for its pleasing design and used as a diffraction technique to overcome difficulties and eliminate bottlenecks when learning abstract words; and an important preventive countermeasure that improves the overall safety levels.

4.3.1. Image Recognition

Due to the particularity of the usage scenario, a dataset for the identification of 36 kinds of objects was made. Its labels are summarized in Table 1. Developing custom labeling each time image preprocessing is conducted is a routine and laborious task. To bridge the gap between large-scale image retrieval and limited computing and human resources, it is necessary to introduce an efficient image annotation and evaluation method that uses a bounding box to implement image matching and speed up the image's transformation in computer vision. In this paper, the object to be identified in samples is enclosed in a rectangular box and all of its properties are transmitted to the host computer for further processing. Figure 10 shows the dataset annotated by LabelImg, which influences the systemic features of the boundary.

Table 1. Information on the selected input dataset.

Activities	Words	Activities	Words
1	movidius	19	computer
2	USB_flash_disk	20	glasses
3	card	21	pen
4	entry_permit_card	22	Mickey_Mouse
5	pie	23	paper
6	tissue	24	earphone
7	Peppa_Pig	25	bag
8	ring	26	carrot
9	mobile_phone	27	chips
10	watch	28	tomato
11	bracelet	29	key
12	hand_cream	30	wallet
13	socket	31	umbrella
14	folder	32	shoe
15	cup	33	drink
16	SpongeBob	34	banana
17	toothbrush	35	mouse
18	Wong_Lo_Kat	36	napkins

On account of giving consideration to both the processing speed and the high-definition images, a KS8A17-AF was selected as the image acquisition device as it has 8 million pixels and supports automatic focus and digital zoom. In cases where overfitting of the model generated problems, an elaborate data expansion strategy was applied. The actions were targeted at being able to feed more invariant image features into the CNN and included rotation, horizontal migration, vertical migration, scaling, tangential transformation, and horizontal transformation. A total of 7200 object images were collected, of which 4320 object images were used as the training set and the remainder were used as the test set. Each object contained about 200 different images. Moreover, in order to speed up the processing of images, the images were resized to 227 (pixels) × 227 (pixels) in the experiment.

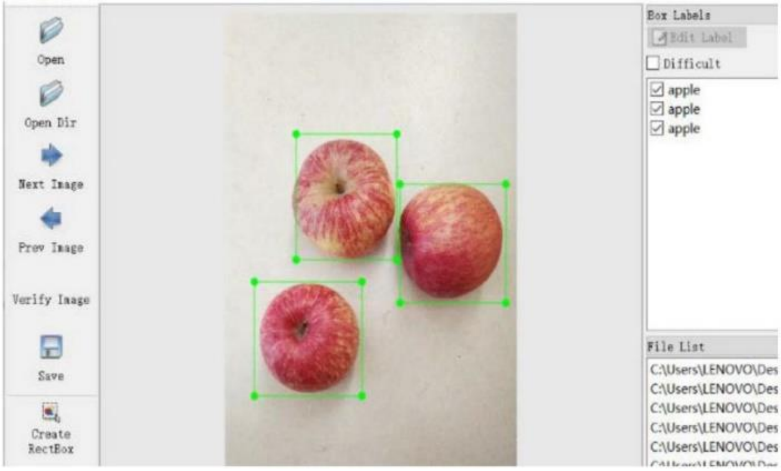


Figure 10. The annotation of the dataset.

Figure 11 shows the self-built library for image recognition.

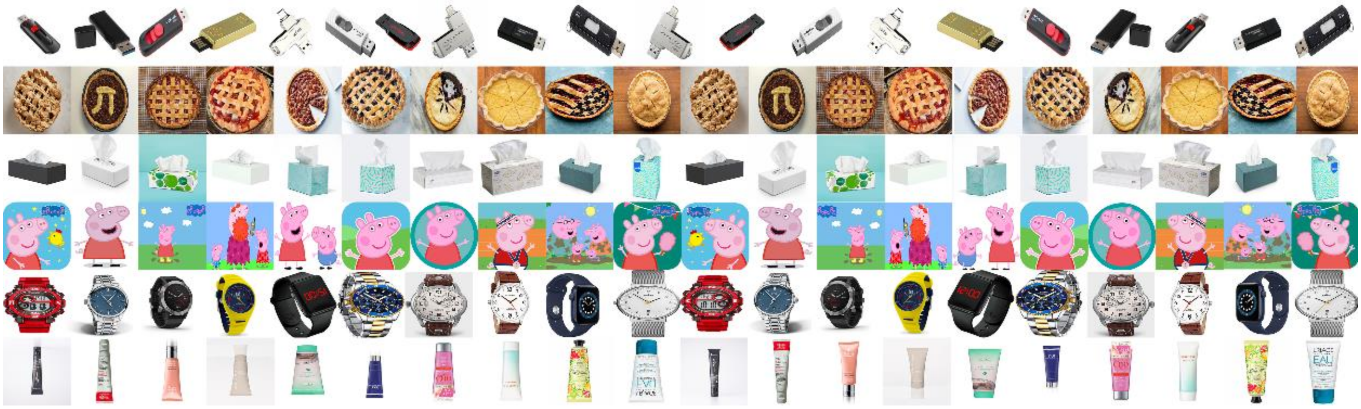


Figure 11. The data on the samples in the database.

The widespread use of deep learning has increased the demands on heterogeneous multi-core processors dramatically. Thus, the UP^2 board, which is an open-source intelligent hardware development platform with high performance and low power consumption, was selected to be the control core in the slave computer, which had an Intel 14 nm Atom x5-Z8350 CPU, 8 GB of LPDDR4, a 128 GB eMMC, and an Ubuntu operating system. Moreover, the UP^2 board supports the AI Core X mPCIe module, which features Intel Movidius Myriad X. Adding the AI Core X module to the UP^2 board creates a powerful and compact deep learning and machine learning solution.

To meet the demanding goal of real-time and fast object recognition of the early education platform, we chose the deep learning framework Caffe, which we found to be convenient, reasonable, and compatible. Specifically, Caffe performs modular processing and functional decomposition, which implements clear regulations and was good for the subsequent optimizations that enhanced the transfer learning database. Additionally, CaffeNet is an efficient model designed for commodity clusters and mobile devices and has streamlined acquisition and supply processes that ensure high recognition accuracy at the same time.

The optimizer was used to iteratively minimize or maximize the loss function E(x) by updating and computing network parameters that play vital roles in research on neural network models. In order to approximate or reach the optimal value, the model training and model output algorithms should be robust against variations in network parameters and use gradient values for each parameter to reduce errors.

In image recognition, accompanied by frequent updates and fluctuations that complicate the convergence to the exact minimum, the overshooting of stochastic gradient descent (SGD) is a serious problem that reduces the calculation precision. A learning rate that is too low will cause the network to converge too slowly, while a learning rate that is too high may affect the convergence and will cause the loss function to fluctuate as soon as it is close to the minimum. Even the gradient will diverge if an appropriate learning rate is not chosen, which makes it difficult to realize a much better and stable convergence. In addition, the same learning rate does not apply to the update of all parameters. If the training set is sparse and the features are very different in frequency, they should not all be updated to the same extent. However, for features that rarely occur, a higher update rate should be taken into account. Therefore, the accuracy of SGD is not high compared with other optimizers. Both the RMSProp and Adadelta algorithms are based on the optimization of Adam's algorithm, solving the problem that the denominator in Adagrad's algorithm will keep accumulating, causing the learning rate to shrink and eventually become very small. In many cases, the results of RMSprop and Adadelta are similar. Because good recognition accuracy for multiple classifications and tracking robustness with high efficiency are achieved by Adadelta, we chose Adadelta as the optimizer algorithm.

The complete model contains four parts: the input layer for importing the target image, the CaffeNet database for extracting features from images, the Adadelta algorithm for classification regression and bounded box regression, and the output layer for exporting the detection results. The CaffeNet–Adadelta model optimized by experts with years of experience in deep learning support was run in the Intel Movidius Myriad 2 VPU and the Intel HD Graphics 400 GPU of the UP^2 board to perform the sustained recognition of objects. GoogleNet is a deep CNN for 1000 classifications trained by ImageNet and used as a pre-training model. The model was modified to deal with practical problems and classifies 36 real objects for the output. The model's hyper-parameters were configured as listed in Table 2.

Table 2. List of selected hyper-parameters.

Hyper-Parameters	Selected Values	Quantity
train_net	"caffenet/train_val.prototxt"	Network files for training
test_iter	30	Number of iterations required during testing
test_interval	200	Tests per thousand training times
base_lr	0.1	Basic learning rates
display	50	Displays debugging information every 10 iterations
max_iter	50,000	Maximum number of iterations
lr_policy	"step"	Learning rate adjustment strategy
gamma	0.1	Learning rate adjustment coefficient
stepsize	1000	Adjusts the learning strategy every thousand iterations
weight_decay	0.0005	Weight attenuation coefficient
snapshot	5000	Take a snapshot every 10,000 iterations
solver_mode	GPU	Adopt the GPU model
test_initialization	false	No test initialization
average_loss	10	Average loss every 10 iterations
type	"Adadelta"	Optimizer

4.3.2. IoT-Based Voice Commands

T-Head's Speech AI Platform was developed in order to build a lightweight terminal-side IoT application, supplemented by a speech algorithm hardware accelerator, that covers various scenarios. SC5654 is a highly integrated audio SoC and a heterogeneous dual-core AI voice chip. It integrates the E803 low-power 32-bit processor as the main control for the system and is equipped with a high-performance, audio-dedicated DSP to process audio codecs and sound effects. The main idea underlying the design and implementation of software is modularization. It defines each function module as an independent service, which realizes the control of audio collection and keyword recognition, the calling of the service interface to push audio information to and obtain the results of speech analysis and synthesis from the cloud, and logical interconnections with the UP^2 board by using the serial port.

The early education platform will only start implementing intelligent speech searches online, which find answers to specific questions, after a trigger action is performed (the user says a wake-up phrase). It is set to monitor and identify voice commands by capturing keywords that provide the fault tolerance ability of the system. The corresponding key events are shown in Table 3.

Table 3. Activities of key events.

Events	Activities	COMMREQ
1	I want to see the animation of instruction in safety	fanSKon
2	I want to know how to wear a mask	fanTMask
3	I want to know what to do when I am alone at home	fanNLone
4	I want to learn about traffic safety	fanTKnow
5	I want to learn about the dangers of drugs	fanNNarc
6	I want to know what I should do in case of fire	fanNFire
7	WeChat applet	fanWChat
8	I want to see the holographic video	fanSHolo
9	I want to see a butterfly	fanHButt
10	I want to see dancing	fanHDanc
11	I want to see the Earth	fanHEart
12	I want to see fireworks	fanHFire
13	I want to see a shark	fanHShak
14	I want to see a jellyfish	fanHJfis
15	I want to see a UFO	fanHUFO
16	I want to see a fish	fanHFish
17	I want to hear my mother tell stories	fanSTMM
18	I want to hear my father tell stories	fanSTDD

Chinese Speech Recognition has a problem in that its independent expression and glossary system is a bottleneck on the progress of the response speed, which leads to the query retrieval system over the network being combined with local control based on an intelligent speech module. The voice commands matched by the cloud are shown in Table 4.

Table 4. Features of cloud events.

Categories	Type of Applications		Response Time (Seconds)
Media Information Search	Music		2.28
	Radio		1.44
	Today’s news		1.20
	Tell a story		1.44
	Animation		1.96
	Weather forecast		1.11
	Date reminders		1.01
	Memorandum		1.48
	Unit conversion		2.26
	Translation		1.31
	Word definitions		2.43
Local Control	Alarms and timers		0.96
	Calculator		1.36
	Volume Adjustment	Set Volume to XX	1.06
		Volume up/down	0.86
	Pause/Resume playback		1.86

As misidentification directly affects the performance of the whole system, the key is initial–final segmentation, which applies an appropriate strategy in the processing of acoustic signals. To dispose of this issue, it is necessary to set up certain limitations, in which too short or too long phrases are not allowed. As a result, we used words of three to four syllables, which dispense an optimal amount of audio information for the load.

As for the uncertain information conditions, making the best decision is the problem that needs to be solved by the proposed system. Based on the Alibaba Cloud, intelligent speech interaction supports the recognition of short speeches that last for less than 1 min. This applies to short speech recognition scenarios such as chat conversations and voice command control. The quantitative meanings of lexical symbols such as “up”, “down”, “high”, and “low” are indefinite, which requires the platform to automate the dynamic interpretation. The prime example is volume control, once a vexed problem of the system. Since the platform should have basic features that protect children’s hearing, the feedback should not be too loud or distracting and the process of decreasing the volume should be fast. Thus, the command “volume up” was set to adjust the volume to move a short distance to account for the limitation of hearing protection. If there is a long way to go to reach the normal level, the corresponding movement will cover a rather large distance, which reduces the repetition of an instruction. Compared with fixed adjustments, a much friendlier operating environment is provided by dealing with uncertain information. Additionally, the above process can be implemented on T-Head’s Yun-on-Chip (YoC) through dedicated software.

4.3.3. Multi-Platform Interconnection

Before planning for how each form of data flow should prioritize efficiently, it is necessary to have a good understanding of the overall design of the early education platform, which divides acquired data into three equal parts. Generally, it is an industrial cloud platform that is oriented towards the practical demands of users, including intelligence in the education industry, three-dimensional digitalization, and networking. The method of maintaining the efficient processing of massive amounts of data and analysis to support the connection and transmission in a ubiquitous environment are utilized in order to build a service system that is stable, and the resources of computing devices are quite similar. Therefore, it brings about challenges and opportunities for the operation and management of systems, which lead to multifunctional and integrative characteristics.

The platform described in this paper has a WIFI module that can transmit records of users’ studying files to the cloud database for storage. The multi-device cloud data management system based on Tencent Cloud was designed to remotely monitor the learning progress of multiple devices in real time and output historical tracks. The Cloud Virtual Machine (CVM) works as a transfer station that promotes data replication across heterogeneous devices that organically connect and beneficially interact with each other. The CVM-based methods for interconnection are shown in Figure 12.

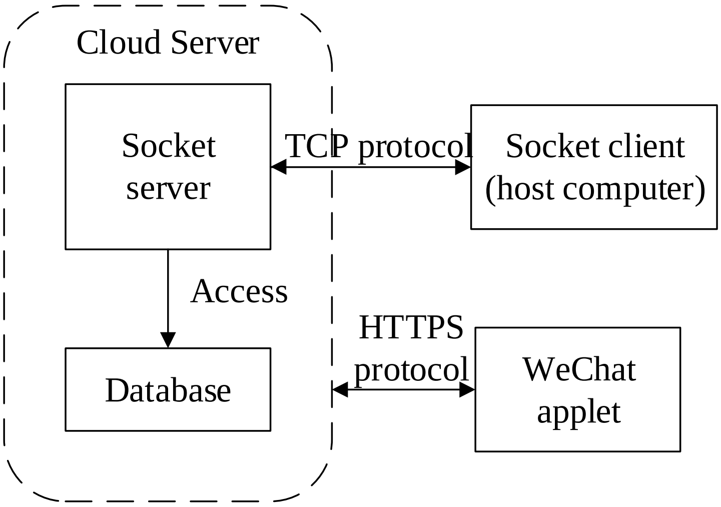
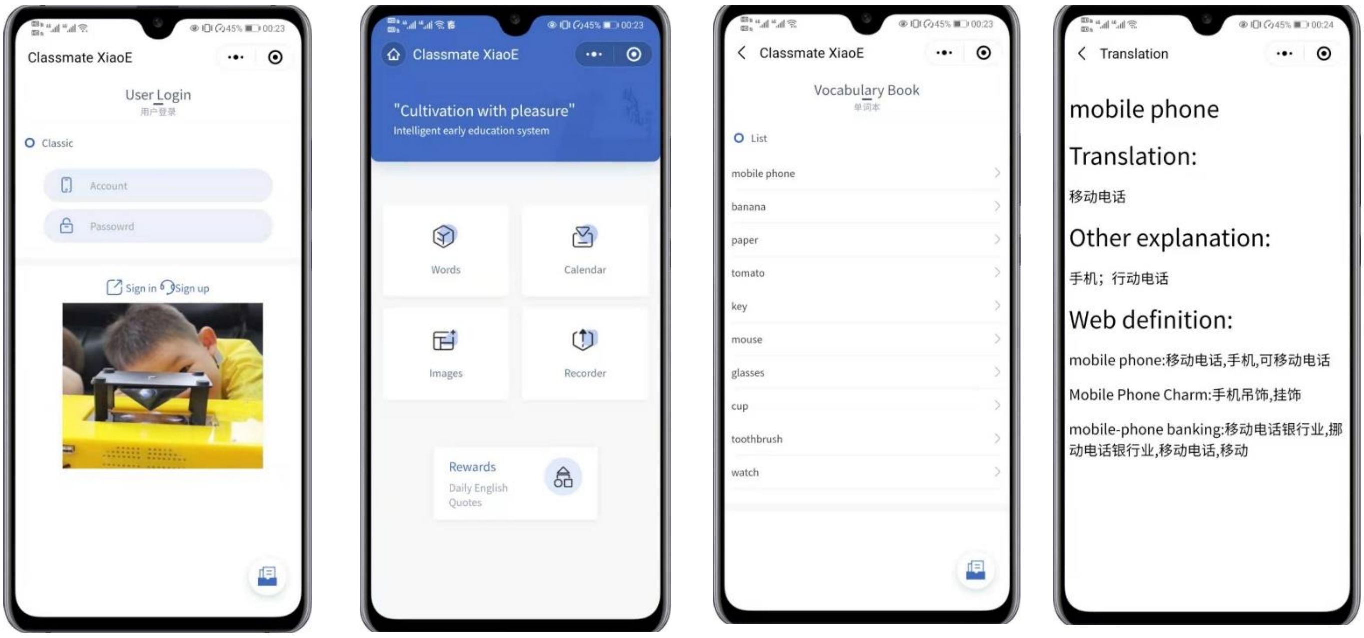


Figure 12. An illustration of executable cloud programs based on computation, storage, and network transmission.

The database is designed to require public certificates to secure communications and uses a PHP script to acquire HTTP access to the server [36]. The SQL server controls users' access from the aspects of the authentication of a log-in to an account and explicit access permissions and restricts access to certain rows or columns through views to achieve database security. Devices and users access the database through the HTTP protocol. Users need to select the device to be viewed through the user interface. Then, they can obtain real-time recognition results from the host computer, which uses the OpenAPI provided by Youdao to yield an interpretation as well as example sentences. The fluctuation in old and new words and the check-in status over a two-month period are able to be displayed in the form of data tables or line charts. Additionally, images stored in a mobile phone can be uploaded to the database through the WeChat applet, which makes it possible for the host computer to recognize and generate the corresponding words in English. In addition, to realize the sharing of internal resources, users have the ability to create new audio files on the server. In other words, a Web page was built to provide this service, which requires the username and password to be specified. After a successful operation, the host computer parses the HTML of the page to look for links pointing to audio files. Finally, due to some restrictions on bandwidth and the interference imposed by different network states, the maximum number of simultaneously supported users is limited.

We used the WeChat applet mainly to reach the required daily data storage capacity and expand and configure excess educational resources, which are closely integrated into the business logic of the early education platform. Its interface consists of a login display, a word display, a calendar display, an image display, and a recording display, as shown in Figure 13. Its flow chart is shown in Figure 14.



mobile phone

Translation:

移动电话

Other explanation:

手机；行动电话

Web definition:

mobile phone:移动电话,手机,可移动电话

Mobile Phone Charm:手机吊饰,挂饰

mobile-phone banking:移动电话银行业,挪动电话银行业,移动电话,移动

(**a**) (**b**) (**c**) (**d**)

Figure 13. *Cont.*

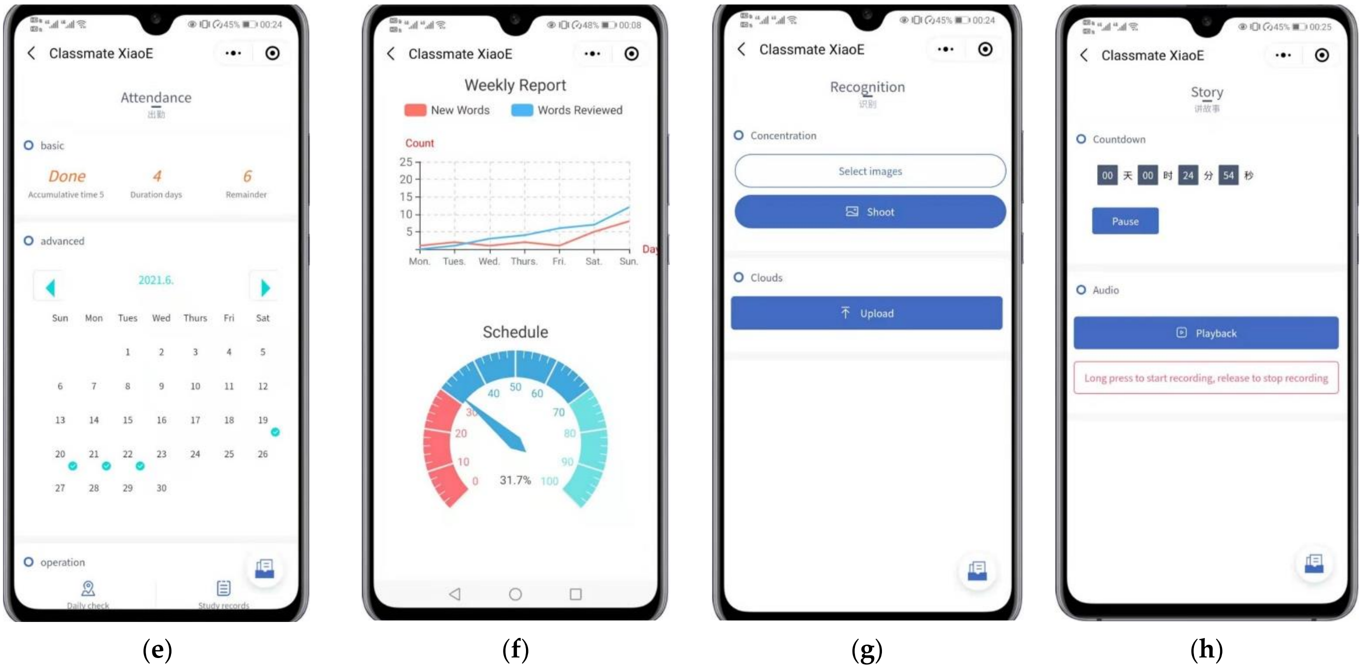


Figure 13. The graphical user interface of the WeChat applet. (**a**) Login; (**b**) Homepage; (**c**) Vocabulary book; (**d**) Words; (**e**) Attendance; (**f**) Summary; (**g**) Images; (**h**) Recorder.

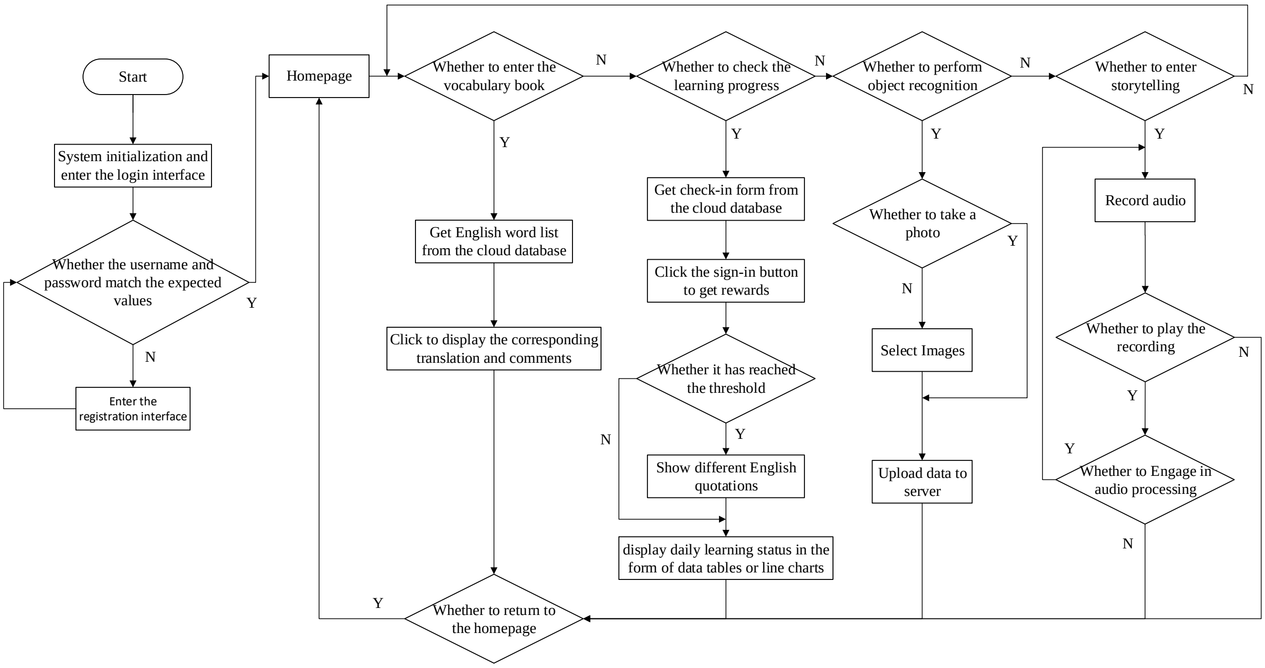

Start
System initialization and enter the login interface
Whether the username and password match the expected values
Y
N
Enter the registration interface
Homepage
Whether to enter the vocabulary book
N
Y
Get English word list from the cloud database
Click to display the corresponding translation and comments
Whether to check the learning progress
N
Y
Get check-in form from the cloud database
Click the sign-in button to get rewards
Whether it has reached the threshold
Y
N
Show different English quotations
display daily learning status in the form of data tables or line charts
Whether to perform object recognition
N
Y
Whether to take a photo
Y
N
Select Images
Upload data to server
Whether to enter storytelling
N
Y
Record audio
Whether to play the recording
N
Y
Whether to Engage in audio processing
Y
N
Whether to return to the homepage
Y

Figure 14. The internal logic of the WeChat applet.

On the other hand, the platform is connected to the Alibaba Cloud, which has emerged as a stronger engine in many industries, such as finance, insurance, e-commerce, and smart homes. The speech development board uses SDK to access the cloud server through a TCP/IP protocol and performs automatic speech recognition (ASR), natural language processing (NLP), and text-to-speech (TTS). Its cloud service component provides the interface for applications to interact with ASR/NLP/TTS services in the cloud. The platform creates speech information and sends a request by calling the corresponding service API and the component automatically finishes the initialization of cloud connection, authentication, and service startup processes. Users only need to put in the audio to be recognized or the string to be synthesized through the interface to retrieve the return value and do something marginally useful with it. The platform responds in predetermined ways according to the results obtained and creates UI elements and displays them on the screen. Users can respond to the information offered by it. After a successful operation, a message is sent to the platform to notify it about the success or failure of the call, and the system will then go to the main menu and wait for the next instruction. The Yun-on-Chip (YoC) defines a unified set of adaptation interfaces, and the application layer can seamlessly switch between different cloud services with the same code, reducing the development cost for users. The flow chart for the intelligent speech interaction is shown in Figure 15.

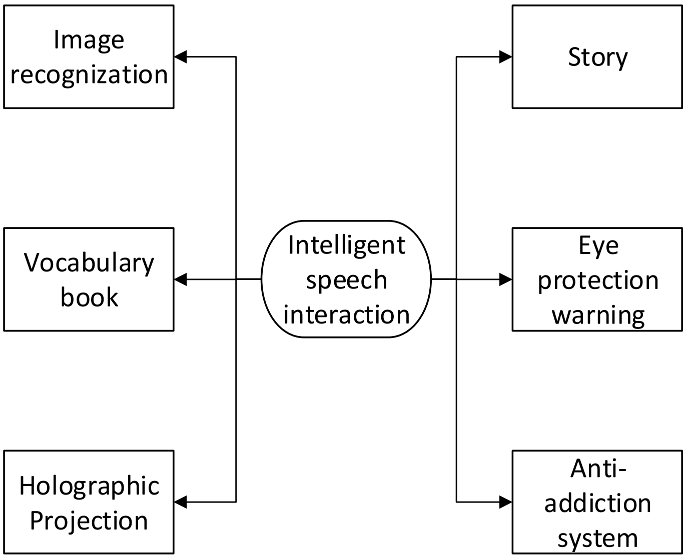


Figure 15. Services and features of the proposed model.

5. Assessment and Analysis

In a series of experiments, the performance was evaluated in terms of reliable message delivery or throughput and by the following numbers as well as diagrams, which help to clear the path to cross-language understanding and communication.

5.1. Evaluation of Datasets

In order to comprehensively address the problem of misrecognition and slow reaction to a target appearing at a location, we introduced a public dataset to investigate and measure the accuracy of the system. To be specific, with a total of 36 objects that respectively contain 500 images, 20,000 data samples were collected through the Internet and during the photoshoot. Additionally, to firmly establish the relationship between the dataset and the test set, images in the training set were not allowed to be selected repeatedly, which effectively avoids the "one-sided" phenomenon of CNNs and better evaluates the entire recognition process. Table 5 shows the classification results obtained by the artificial observation and the instrumental survey when the model was predicted with the test set.

According to the statistics, there were only 139 instances of misclassification, and almost half of the items recognized by the specified graphics filter had a relative error that fell into the range 0%~0.2%. Although the overall accuracy reached 99.24%, the system showed non-ideal discrimination in the choice of USB_flash_disk and Mickey_Mouse. It is appropriate to control the precision in the range of 98%~99%. The main reason for the frequency of errors may be that these two items have similarities in some respects from the perspective of cameras.

Table 5. Accuracy Index.

Types of Items	Precision (%)	Types of Items	Precision (%)
movidius	1	computer	0.993732
USB_flash_disk	0.909091	glasses	0.998757
card	1	pen	1
entry_permit_card	0.993731	Mickey_Mouse	0.909092
pie	1	paper	0.995687
tissue	1	earphone	0.986153
Peppa_Pig	1	bag	0.993698
ring	0.998534	carrot	1
mobile_phone	1	chips	0.99697
watch	1	tomato	0.99658
bracelet	1	key	0.99497
hand_cream	0.99697	wallet	0.997660
socket	0.997669	umbrella	1
folder	0.998377	shoe	1
cup	0.994144	drink	0.998317
SpongeBob	0.992762	banana	1
toothbrush	0.986155	mouse	1
Wong_Lo_Kat	0.998759	napkins	0.994154
Average		0.992455	

5.2. Impact of Dynamic Identification on System Performance

Referring to the flow chart for controlling the system, the user is to press the 'recognize' button to enter the camera mode, point the camera at the item to be shot, and then press the confirm button to complete the capturing operation. The system will be about to jump to the next screen that displays the recognized words and example sentences and mobilizes the speech resources to make users bear in mind the correct pronunciation of the words. Additionally, the user can press the cancel button to finish the process, and the system will return to the homepage, which shows step-by-step instructions on the screen. The example below involves two overlapping procedures, as shown in Figure 16.

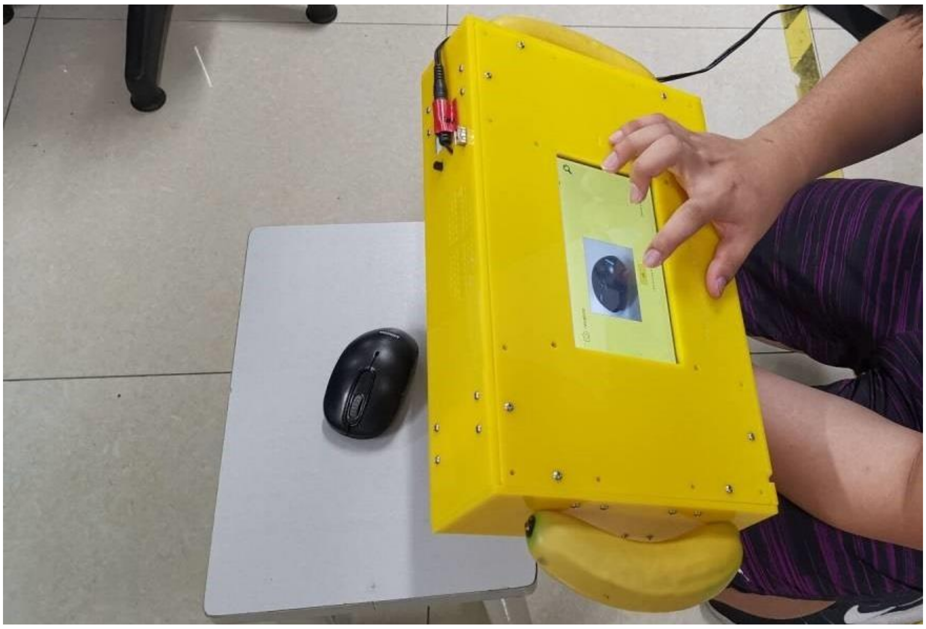


Figure 16. The performance test of image recognition.

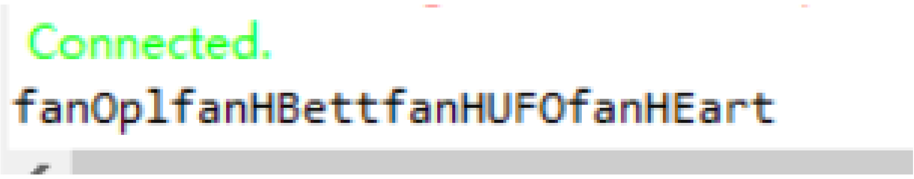
After testing the function, it was found that the interface lagged during the process of acquiring image data. In order to solve this problem, we adopted the following measures:

1. Since the real-time status of objects is displayed on the monitor, the programming requirements are relatively high. After proposing a critical method for optimizing the interface, the video data stream was imported in a multi-threaded way, which effectively improved the fluency and led to a good human–computer interaction experience.
2. The resolution of the monitor was changed to 1024 × 600, which is more in line with the requirements of most people and alleviates the physical and mental fatigue caused by systems with delays.

Through the specific application, it was proven that the optimized system enhances the capacity for accurate data transmission, improves the response speed, and provides more stable working conditions that result in an increase in user satisfaction. After 100 tests, the success rate reached 96% with automatic uploading, which lowers the risk of misrecognition and enhances the strength of the CNN by marking, retraining, and evaluating unrecognized images. On the other hand, if the system encounters blurred images and multiple objects, it will return to the initial stage of recognition. Users can roll the camera and repeat the scene to make up for the previous failure.

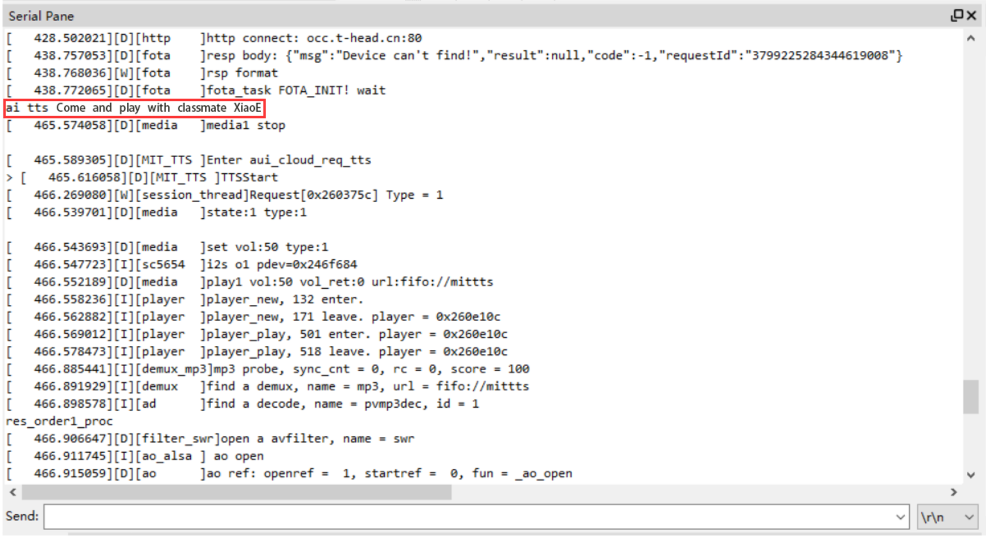
5.3. Impact of Human–Computer Interaction on System Performance

The intelligent speech module is able to make an effective wake-up call, which emits an event notification due to the detection of a change in acoustic sensors that influence the corresponding signal processing. The importance of computing resource management and coordination mechanisms to the early education platform is presented in this section, and we introduce the methods for forming and spreading voice commands. Additionally, there is a mutual relation between the UP^2 board and the CB5654, which involves a simple as well as useful communication protocol that receives and sends data via a serial port. Figures 17 and 18 show their respective differences.



```
Connected.
fanOplfanHBettfanHUFOfanHEart
```

Figure 17. The event notification of a successful transmission.



```
[    428.502021][D][http     ]http connect: occ.t-head.cn:80
[    438.757053][D][fota     ]resp body: {"msg":"Device can't find!","result":null,"code":-1,"requestId":"3799225284344619008"}
[    438.768036][W][fota     ]rsp format
[    438.772065][D][fota     ]fota_task FOTA_INIT! wait
ai tts Come and play with classmate XiaoE
[    465.574058][D][media    ]media1 stop

[    465.589305][D][MIT_TTS ]Enter aui_cloud_req_tts
> [    465.616058][D][MIT_TTS ]TTSStart
[    466.269080][W][session_thread]Request[0x260375c] Type = 1
[    466.539701][D][media    ]state:1 type:1

[    466.543693][D][media    ]set vol:50 type:1
[    466.547723][I][sc5654  ]i2s o1 pdev=0x246f684
[    466.552189][D][media    ]play1 vol:50 vol_ret:0 url:fifo://mittts
[    466.558236][I][player  ]player_new, 132 enter.
[    466.562882][I][player  ]player_new, 171 leave. player = 0x260e10c
[    466.569012][I][player  ]player_play, 501 enter. player = 0x260e10c
[    466.578473][I][player  ]player_play, 518 leave. player = 0x260e10c
[    466.885441][I][demux_mp3]mp3 probe, sync_cnt = 0, rc = 0, score = 100
[    466.891929][I][demux    ]find a demux, name = mp3, url = fifo://mittts
[    466.898578][I][ad       ]find a decode, name = pvmp3dec, id = 1
res_order1_proc
[    466.906647][D][filter_swr]open a avfilter, name = swr
[    466.911745][I][ao_alsa ] ao open
[    466.915059][D][ao       ]ao ref: openref =  1, startref =  0, fun = _ao_open
```

Figure 18. The event notification of a successful reception.

The experiments focus on the effect of decibels in the usage scenarios and the application logic that implements an appropriate error correction, which were evaluated by self-comparison and duplicate experiments. With respect to an increasing decibel level, the test of the intelligent speech interaction under actual working conditions was mainly conducted from three perspectives. As shown in Table 6, the noisier the environment, the worse the recognition results obtained by the system with a varying number of decibels (from 0 to 30, from 30 to 60, and above 60). The upper part, the middle part, and the lower part of the decibel scale were created to simulate the roadside, a kitchen in an urban residence, and a classroom that a teacher is speaking in, respectively. The results show that the probability of completing the entire dialogue remains above 85%, which emphasizes the robustness and effectiveness of the interaction.

Table 6. Experiments on different levels of ambient noise.

Sound Intensity (dB)	Test Projects	Testing Times	Number of Errors	Probability
	Wake-up	300	0	100%
≤30	Capture	300	2	99.33%
	Dialogue	300	2	99.33%
	Wake-up	300	0	100%
≤60	Capture	300	5	98.33%
	Dialogue	300	7	97.67%
	Wake-up	300	23	92.33%
60≤	Capture	300	27	91%
	Dialogue	300	36	88%

Additionally, the errors caused by similar pronunciations, especially those collected under the same conditions, may mislead the system to do wrong actions, thus causing energy to be wasted, a cost increase, and a decline in efficiency. To identify the difference between these keywords, a method was applied to filter those phrases that increase the risk of misidentification, and we prepared a list of standard words as a reference. By removing undesired words from the list, the wake-up phrase is absorbed first. In addition, if the trigger word is distinguished, the system will switch from a standby state to a working state, which provides a period of time for users' convenience. Furthermore, there were some problems with ranking the results of the current speech engine with multi-keyword retrieval. Thus, the schemes were limited to handling a single keyword search.

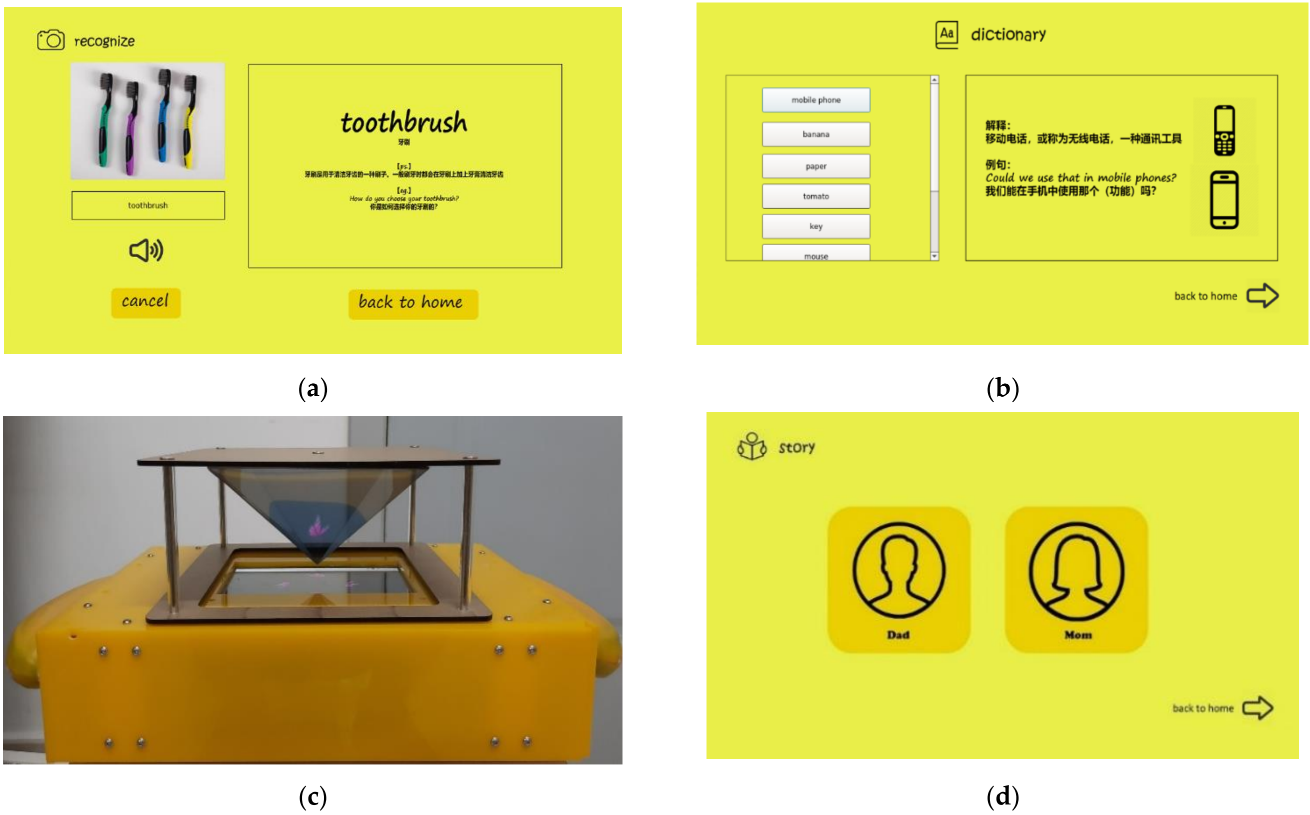
As young children have a limited attention span, which makes it difficult for them to concentrate on one activity for very long, creating an easy-to-use and straightforward user interface that takes full advantage of the interaction between image recognition and intelligent speech recognition is important. Therefore, the use of a short set of voice commands to invoke applications does not fulfil the function of preparing children for school. Due to the correlation between verbal skills, experimental skills, and abstract reasoning skills, the complex working parts of the system consist of the following four blocks:

1. Take a photo. An object is placed in front of the fixed camera of the platform and the monitor shows its three-dimensional display. After the user says "shoot", the calls will be returned by the two subsystems immediately, which provides positive feedback to children.
2. Dictionary. When the user says "learn English", the monitor displays the result of each query. While touching the corresponding position with a hand, the system enters the next stage in order to retrieve the definitions of words and pronounce them automatically. The words learned are recorded, and the data are synchronized to the cloud database, which is available through the WeChat applet and can be used to check current learning efficiencies. Children are expected to practice the use of memory, with the displayed object showing its meanings in a different order.
3. Holographic projection. Children sometimes misunderstand the meanings of words, especially those who have not been mentored in open science. If they have difficulty grasping the connotations of abstract English nouns, they can try again by saying

“holographic video”, with the definition vividly depicted by holographic imaging equipment, which stimulates the ability of children to bring forth new ideas.

4. Audio and video. The quality of synthetic speech depends on phonetic sounds that simulate those of users’ parents. It is much more straightforward for children to pick up information from good-quality synthetic speech, regardless of who they are familiar with. The audio texts can be transformed into a parent’s voice and combined with the animation for education on safety. “Tell a story” and “safety” are distinguished from each other in order to avoid confusion. Children lack experience and have a weak sense of self-protection. By adding comments to stories and videos, children have the chance to understand how to cope with danger, which makes it possible for them to choose correct countermeasures.

The characteristics of the above-described activities are given in Figure 19.



toothbrush
牙刷
【ps.】
牙刷是用于清洁牙齿的一种刷子，一般刷牙时都会在牙刷上加上牙膏清洁牙齿
【eg.】
How do you choose your toothbrush?
你是如何选择你的牙刷的?

解释:
移动电话，或称为无线电话，一种通讯工具
例句:
Could we use that in mobile phones?
我们能在手机中使用那个（功能）吗?

Figure 19. Illustration of the implementation of voice commands. (**a**) Take a photo; (**b**) Dictionary; (**c**) Holographic projection; (**d**) Audio and video.

6. Conclusions

A novel early education platform that combines deep learning with a favorable virtual interactive environment was introduced from the perspective of intelligent speech interaction, object recognition, and 3D object reconstruction. With the aim of promoting the development of children’s ability to think, be creative, and adopt different perspectives, a prototype system was implemented as an all-round teaching tool that focuses on helping children participate actively in an interactive learning process rather than having them remember new words and grammar points by rote. In response to this characteristic, voice-controlled programs were used to replace traditional manual control methods that rely on a mouse or keyboard, which greatly reduces the possibility of maloperation under high load conditions. Moreover, a holographic system was added to strengthen the bonds between three-dimensional objects and images that express complicated life experiences. The reproduction of an English learning scenario was obtained that visualizes the abstract concept and connects children's cognition of an image with concrete perceptions of movement and space. Finally, we developed measures of statistical analysis that provide a distribution of learned words and lower the risk of falling behind schedule. Through smartphone apps, the real-time recording collected by the platform is transmitted and exchanged in the cloud server, which makes it capable of multi-device management and obtaining a better dynamic response to children who are in remote places.

To sum up, there are advantages to the proposed platform, which showed good results in performance tests and has a wide range of potential applications. However, compared with other early education platforms, there remains room for further improvement, which raises the following prospects for this system. First, the capacity and content of the self-built library are limited and data augmentation will require time and labor. Due to the low thermal conductivity of the shell, its cooling performance may cause the system to overheat. The addition of fans or drilling holes may help to optimize the overall structure and stability. Finally, expanding from the fusion of multiple types of media, including audio, images, and video, to a higher level of immersion that involves the senses of touch, smell, and taste needs to be explored.

7. Patents

The intelligent early education platform can be searched for using the publication (announcement) number CN213716200U.

Author Contributions: K.X. conceived of the key idea and provided guidance during the discussions; X.X. drafted the manuscript; H.F. and H.L. designed the model and performed the experiments; X.X. designed the interface of the mobile app; H.F. collected and analyzed the raw data; H.L. designed the mechanical structure. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tu, Y.-H.; Du, J.; Lee, C.-H. Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2080–2091. [CrossRef]
2. Yan, C.; Zhang, G.; Ji, X.; Zhang, T.; Zhang, T.; Xu, W. The feasibility of injecting inaudible voice commands to voice assistants. *IEEE Trans. Dependable Secure Comput.* **2021**, *18*, 1108–1124. [CrossRef]
3. Muthugala, M.A.V.J.; Jayasekara, A.G.B.P. A review of service robots coping with uncertain information in natural language instructions. *IEEE Access* **2018**, *6*, 12913–12928. [CrossRef]
4. Ansari, J.A.; Sathyamurthy, A.; Balasubramanyam, R. An open voice command interface kit. *IEEE Trans. Hum. Mach. Syst.* **2016**, *46*, 467–473. [CrossRef]
5. Sidenko, I.; Kondratenko, G.; Kushneryk, P.; Kondratenko, Y. Peculiarities of human machine interaction for synthesis of the intelligent dialogue chatbot. In Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Metz, France, 18–21 September 2019.
6. Wang, X.; Zheng, X.; Chen, W.; Wang, F.-Y. Visual human–computer interactions for intelligent vehicles and intelligent transportation systems: The state of the art and future directions. *IEEE Trans. Syst. Man. Cybern. Syst.* **2021**, *51*, 253–265. [CrossRef]
7. Ţucă, L.; Iftene, A. speech recognition in education: Voice geometry painter application. In Proceedings of the 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 6–9 July 2017.
8. Matsane, L.; Jadhav, A.; Ajoodha, R. The use of automatic speech recognition in education for identifying attitudes of the speakers. In Proceedings of the 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 16–18 December 2020.
9. Wei, W.; Wei, L. Design and implementation of early childhood education interactive platform system. In Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 6–9 December 2019.
10. Benson, B.; Arfaee, A.; Kim, C.; Kastner, R.; Gupta, R.K. Integrating embedded computing systems into high school and early undergraduate education. *IEEE Trans. Educ.* **2011**, *54*, 197–202. [CrossRef]
11. Alzubi, T.; Fernández, R.; Flores, J.; Duran, M.; Cotos, J.M. Improving the working memory during early childhood education through the use of an interactive gesture game-based learning approach. *IEEE Access* **2018**, *6*, 53998–54009. [CrossRef]

12. Filgueiras Damasceno, E.; Augusto Nardi, P.; Anastacio Silva, J.; Dias Junior, B.; Cardoso, A. 3D virtual simulation approach in brazilian vocational education for computers network adapted to student knowledge. *IEEE Lat. Am. Trans.* **2017**, *15*, 1917–1925. [CrossRef]
13. Tseng, J.-L. Intelligent augmented reality system based on speech recognition. *Int. J. Circuits Syst. Signal Process.* **2021**, *15*, 178–186. [CrossRef]
14. Muñoz-Cristóbal, J.A.; Jorrín-Abellán, I.M.; Asensio-Pérez, J.I.; Martínez-Monés, A.; Prieto, L.P.; Dimitriadis, Y. Supporting teacher orchestration in ubiquitous learning environments: A study in primary education. *IEEE Trans. Learn. Technol.* **2015**, *8*, 83–97. [CrossRef]
15. Ondáš, S.; Kiktová, E.; Pleva, M.; Oravcová, M.; Hudák, L.; Juhár, J.; Zimmermann, J. Pediatric speech audiometry web application for hearing detection in the home environment. *Electronics* **2020**, *9*, 994. [CrossRef]
16. Risal, M.F.; Sukaridhoto, S.; Rante, H. Web explainer for children's education with image recognition based on deep learning. In Proceedings of the 2019 International Electronics Symposium (IES), Surabaya, Indonesia, 27–28 September 2019.
17. Xia, K.; Huang, J.; Wang, H. LSTM-CNN architecture for human activity recognition. *IEEE Access* **2020**, *8*, 56855–56866. [CrossRef]
18. Yu, C.; Kang, M.; Chen, Y.; Wu, J.; Zhao, X. Acoustic modeling based on deep learning for low-resource speech recognition: An overview. *IEEE Access* **2020**, *8*, 163829–163843. [CrossRef]
19. Ni, C. The human-computer interaction online oral English teaching mode based on Moodle platform. In Proceedings of the 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 14–16 April 2021.
20. Zhang, S.; Chen, A.; Guo, W.; Cui, Y.; Zhao, X.; Liu, L. Learning deep binaural representations with deep convolutional neural networks for spontaneous speech emotion recognition. *IEEE Access* **2020**, *8*, 23496–23505. [CrossRef]
21. Park, S.-W.; Ko, J.-S.; Huh, J.-H.; Kim, J.-C. Review on generative adversarial networks: Focusing on computer vision and its applications. *Electronics* **2021**, *10*, 1216. [CrossRef]
22. Isyanto, H.; Arifin, A.S.; Suryanegara, M. Performance of smart personal assistant applications based on speech recognition technology using IoT-based voice commands. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea, 21–23 October 2020.
23. Jokinen, K.; Nishimura, S.; Fukuda, K.; Nishimura, T. Dialogues with IoT companions: Enabling human interaction with intelligent service items. In Proceedings of the 2017 International Conference on Companion Technology (ICCT), Ulm, Germany, 11–13 September 2017.
24. Shimada, K.; Bando, Y.; Mimura, M.; Itoyama, K.; Yoshii, K.; Kawahara, T. Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio. Speech. Lang Process.* **2019**, *27*, 960–971. [CrossRef]
25. Jayson Baucas, M.; Spachos, P. Fog and IoT-based remote patient monitoring architecture using speech recognition. In Proceedings of the 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France, 7–10 July 2020.
26. Narayanan, A.; Wang, D. Investigation of speech separation as a front-end for noise robust speech recognition. *IEEE/ACM Trans. Audio. Speech. Lang Process.* **2014**, *22*, 826–835. [CrossRef]
27. Cecil, J.; Kauffman, S.; Cecil-Xavier, A.; Gupta, A.; McKinney, V.; Sweet-Darter, M. Exploring human-computer interaction (HCI) criteria in the design and assessment of next generation VR based education and training environments. In Proceedings of the 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Lisbon, Portugal, 27 March–1 April 2021.
28. Sanna, A.; Lamberti, F.; Paravati, G.; Demartini, C. Automatic assessment of 3D modeling exams. *IEEE Trans. Learn. Technol.* **2012**, *5*, 2–10. [CrossRef]
29. Ke, J. The use of stereoscopic display technology under human-computer interaction in navigation simulator. In Proceedings of the 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 14–16 April 2021.
30. Alvarez-Marín, A.; Velázquez-Iturbide, J.Á.; Castillo-Vergara, M. Technology acceptance of an interactive augmented reality app on resistive circuits for engineering students. *Electronics* **2021**, *10*, 1286. [CrossRef]
31. Pérez, J.; Vizcarro, C.; García, J.; Bermúdez, A.; Cobos, R. Development of procedures to assess problem-solving competence in computing engineering. *IEEE Trans. Educ.* **2017**, *60*, 22–28. [CrossRef]
32. Lee, J.; Park, S.; Hong, I.; Yoo, H.-J. An energy-efficient speech-extraction processor for robust user speech recognition in mobile head-mounted display systems. *IEEE Trans. Circuits. Syst. II. Express Briefs.* **2017**, *64*, 457–461. [CrossRef]
33. Petousi, V.; Sifaki, E. Contextualizing harm in the framework of research misconduct. Findings from discourse analysis of scientific publications. *Int. J. Sustain. Dev.* **2020**, *23*, 149–174. [CrossRef]
34. Ge, Y.; Ansari, S.; Abdulghani, A.; Imran, M.A.; Abbasi, Q.H. Intelligent instruction-based IoT framework for smart home applications using speech recognition. In Proceedings of the 2020 IEEE International Conference on Smart Internet of Things (SmartIoT), Beijing, China, 14–16 August 2020.
35. Ma, Z.; Liu, Y.; Liu, X.; Ma, J.; Li, F. Privacy-preserving outsourced speech recognition for smart IoT devices. *IEEE Internet Things J.* **2019**, *6*, 8406–8420. [CrossRef]
36. Xia, K.; Tang, T.; Mao, Z.; Zhang, Z.; Qu, H.; Li, H. Wearable smart multimeter equipped with AR glasses based on IoT platform. *IEEE Instrum. Meas. Mag.* **2020**, *23*, 40–45. [CrossRef]