MagicABC: A Parent-Child Tool to Modify Preschoolers' Pronunciation With Augmented Reality

Mingcheng LI
44181638-1

Master of Engineering

Supervisor: Prof. Jiro TANAKA
Graduate School of Information, Production and Systems
Waseda University

July 2020

Abstract

With the rapid development of computer science, we widely use ubiquitously adopting Augmented Reality (AR) technologies to enhance our perception and help us see, hear, and feel our environments in new and enriched ways.

In education, especially early education has attracted more and more attention. To increase users' satisfaction degree and provide a better user experience, many companies and research institutions aim to develop better early education system.

We intend to improve current early education system by giving the system the feature of "parent-child interaction" and "accurate pronunciation teaching". The new system named "MagicABC" that we think might be one of the best solutions for current early education idea in the field of learning a second language.

We have designed an early education card system based on AR and natural language processing technology. In addition, a software is designed to test the similarity between pronunciation and standard pronunciation. Different from any other existing works, the "MagicABC" combining AR faecal expression part and pronunciation assessment part. Before using, a parent chooses an English alphabet card to let the preschooler to learn, and the preschooler chooses his favorite puppet to join. After the system scans the card and records the pronunciation in response, the system compares the child's pronunciation with the standard pronunciation of the card content and gets a score. Based on the pronunciation score results, MagicABC will display a series of AR and voice feedback on the puppet to encourage preschoolers and modify their pronunciation. The system uses augmented reality content to stimulate children's interest in learning, and uses speech technology to correct pronunciation. In such a system, it can establish better parent-child interaction through cooperative tasks.

We have invited some participants to test the usability and efficiency of our system. We got a positive feedback through the preliminary user study.

Keywords: Augmented Reality, Pronunciation, Parent-child, Preschooler, Interactive

Acknowledgements

At the beginning, I am very glad to express my sincere gratitude and appreciation to my supervisor Prof.Jiro TANAKA, a very respectable and responsible mentor with rigorous academic attitudes and creative ideas. He has given me lots of help and encouragement, and is always very patient for guiding me to solve problems. Thanks a lot to him for the many valuable and earnest equipment during every stage of preparation of this thesis.

Besides, it is grateful to all my dear lab-mates. They discussed a lot with me and always give me lots of inspiration. Thanks so much for their many helps in experiments so that our experiments could be performed successfully and the good results could be get.

Last but not least, I would like to thank my parents and all love ones. For my parents, they give me financial support and warm mental support throughout my master study life. I will keep going on and making progress to return all those people.

Contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Introduction

Augmented Reality (AR) brings digital information and virtual objects into the physical space. Over the past decades, Mobile Augmented Reality (MAR) has attracted interest from industry and academia. MAR complements the real world of mobile users with computer generated virtual content [1].With the help of AR technology, information about the real world around becomes interactive and digital. Augmented Reality system has three prominent features [2]:

- Combines real and virtual objects in a real environment
- Registers real and virtual objects with each other
- Runs interactively, in three dimensions, and in real time

We widely use Augmented Reality (AR) technology to enhance our perception and help us to see, hear and feel our environment in a new and rich way. AR also support us in education, maintenance, design and reconnaissance [3]. With the help of AR, we can find some new solutions for traditional problems including education.

Early education is very important in the eyes of Asian parents. Especially in East Asia, including Japan, China, and South Korea, learning to write a second foreign language is an

indispensable subject for early education. If you search for the advantages of AR education in Google, the following will be the high-ranking answers:

- High interest: Because the content presented by AR is all 3D, it is very vivid, intuitive and vivid, which is helpful for students to understand and remember. With the help of AR technology, students' classroom experience jumped from 2D to 3D. It is no longer the flat content presented in books or blackboards, but the vivid three-dimensional content.
- Reduce costs: With AR technology, the physical materials needed for many courses are not needed.
- Teamwork: When students use AR technology to learn, they no longer learn by rote, but to experience the learning content, and can participate in teaching in a teamwork manner.
- Time saving: Unlike traditional education, AR technology combining virtual and real, many digital teaching content can be directly integrated into physical teaching items, bringing great convenience to teachers, without the need to repeatedly switch between virtual and reality, thus saving A lot of time has improved the efficiency of teaching.
- Remote support: AR allows teachers and students from different regions to gather in a virtual classroom and achieve real-time and real-time interaction.
- Improve safety: Chemistry, physics and other disciplines need to do experiments in the teaching process, which has certain risks. With the help of AR technology, it is possible to conduct virtual experiments and obtain the same effect.

In general, it could enhance children's interest in learning and improve learning efficiency. In this study, we plan to use pronunciation assessment technology and AR technology to solve the problem of early childhood education. And study the influence of AR on early education.

1.2 Organization of the Thesis

The rest of this thesis is organized as follows: Chapter 2 introduces the background about the thesis. Chapter 3 will tell the research goal and also the approach will be told briefly. Chapter 4 will describe the related work. Chapter 5 is the system design part, where the design concept and ideas will be introduced in detail. Chapter 6 presents the system implementation part which is about the detailed devices, environment and implementation. Chapter 7 will be about the experiments; we will talk about the performance of our approach and the comparison of different approached will also be done. Chapter 8 will be conclusion and future work part, where we will conclude the previous content and talk about the future possibilities.

Chapter 2

Background

2.1 Augmented Reality

There are three types of Augmented Reality: marker-based AR with the digital world anchored to the real world, unmarked AR with users moving virtual objects, and position-based AR with the virtual world in physical space.

2.1.1 Definition

Augmented Reality (AR) was defined in 1997 by a researcher who is called Ronald T. Azuma [4]. Differ from Virtual Reality, Augmented Reality (AR) allows the user to see the real world, with virtual objects superimposed upon or composited with the real world, as shown in Fig. 2.1. From his view, he defined AR as systems that have the following three characteristics:

- Combines real and virtual objects in a real environment;
- Registers (aligns) real and virtual objects with each other;
- Runs interactively, in three dimensions, and in real time.

AR also known as Mixed Reality, aims to combine virtual and real scene together to achieve that virtual ones are belong to the real world. Being characteristic of integration of

virtual and real scene, many applications of Augmented Reality are emerging, such as in field of education, medical treatment and entertainment [5].

Fig. 2.1 AR demo picture

Milgram defined a continuum of real-to-virtual environments, in which AR is one part of the general area of mixed reality [6], as shown in Fig. 2.2.

Fig. 2.2 Milgram's reality–virtuality continuum(Adapted from Milgram and Kishino)

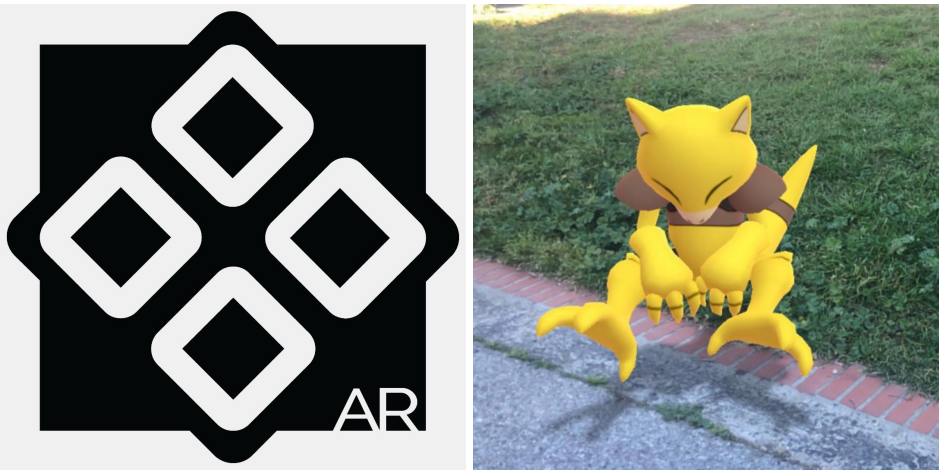
At present, there are two definitions of Augmented Reality. According to Ronald Azuma of the University of North Carolina, Augmented Reality technology includes three aspects: the combination of virtual objects and reality, real-time interaction and three-dimensional. Paul Milgram Fumio Kishino proposed another definition in 1994, which is called Milgram's Reality-Virtuality continuum [7].

2.1.2 Types of Augmented Reality

For Augmented Reality, there are two types of simple Augmented Reality: marker-based which uses cameras and visual cues, and marker less which use positional data such as a mobile's GPS and compass [5].

Different types of Augmented Reality (AR) markers, as shown in Fig. 2.3a, are images that can be detected by a camera and used with software as the location for virtual assets placed in a scene. Most are black and white, though colours can be used as long as the contrast between them can be properly recognized by a camera. Simple Augmented Reality markers can consist of one or more basic shapes made up of black squares against a white background. More elaborate markers can be created using simple images that are still read properly by a camera, and these codes can even take the form of tattoos.

In marker-less Augmented Reality the image is gathered through internet and displayed on any specific location (can be gathered using GPS), like Fig. 2.3b. The application does not require a marker to display the content. It is more interactive than marker based augmentation.



(a) A simple AR Marker (b) Marker-less AR

Fig. 2.3 Types of Augmented Reality

In this research, we focus on the marker-based AR system and try to enhance it.

2.2 Second Language Learning for Preschool Children

English is the international language of business, politics and diplomacy. English is also the first language in colleges, computers and the Internet. A statistics [Statista 2019] [8] statistics show that English is the most used language in the world, and its dominance is expected to continue. The number of English learners in China exceeds 300 million, including 210 million public school students and 90 million professionals.

Studies have shown that reading children aloud can have a positive effect at any age. Jim Trelease, the author of *The Read-Aloud Handbook* [9], tells us that reading aloud to children will stimulate their interest, their emotional development, their imagination and their language. " Language learning is hard work. It takes effort at any time and must be maintained for a long time. Games help and encourage many learners to stay interested and work. By playing games, students can learn English like a child's mother tongue without realizing they are learning. Vocabulary acquisition is increasingly regarded as essential to language acquisition [10]. The use of games is considered very important for presenting and modifying vocabulary [11].

For these reasons, a parent-child interaction link has been inserted into MagicABC to perform phonetic assessment by learning the words of the card. Use AR expression feedback as a reward to promote children's learning. As children learn, they make progress by reading out English and constantly correcting pronunciation.

Chapter 3

Research Goal and Approach

In the previous chapter, we have listed the existing research and analyzed the existing problems. In this chapter we will present our research goals and the progress made so far.

3.1 Research Goal

our research tries to use pronunciation assessment to modify non-English spoken preschoolers' pronunciation and use AR and voice feedback to encourage them.

- Use pronunciation assessment to evaluate the pronunciation of non-spoken children, especially to optimize our system for early education.
- Use children's favorite puppets, and project AR facial expressions on the puppets' faces and use AR content to display cards to improve immersion.

3.2 Research Approach

As shown in the Fig. 3.1, by using our system. After scanning the card and puppet model, the software can record the children's pronunciation through the mobile phone, and analyze the children's pronunciation through the built-in voice evaluation module. Then, the system will display different expressions of different scores according to the score scoring

system. Finally, according to the type of puppet and the child's performance, show different expressions and voices to puppets to motivate children. And correct the child's pronunciation by playing the correct pronunciation.

(a) Not using the system (b) Using MagicABC system

Fig. 3.1 Research approach of MagicABC

In addition, I also built a program for the voice module separately shown in Fig. 3.2. Using this program, you can read specific words and sentences, and get accurate scores for evaluation.

Fig. 3.2 Pronunciation evaluation module

Chapter 4

Related Work

4.1 English Learning Software Designed for Children

Statista 2019 [8] show that English is the most used language in the world, and its dominance is expected to continue. The number of English learners in China exceeds 300 million, including 210 million public school students and 90 million professionals. In non-English speaking countries, more and more parents try to teach their children English as early as possible. This requires learning a second foreign language. However, due to the limitation of parents' knowledge and the scattered learning time. The child's parents tried to use the latest computer technology to solve this problem.

Additionlly, in many technology-driven societies, smart media has been deeply integrated into the daily lives of young people [12]. Currently, online English courses and AR novel books for preschoolers are the two latest solutions. They can provide children with a way to learn anytime, anywhere. Using the current technological trend, there are already some tools and software for solving existing needs.

4.1.1 Online English Courses

Online course is a new course opened with the development of the Internet. In addition to traditional course materials (such as film lectures, reading materials, and problem sets), many

online courses also provide interactive courses for user forums or social media discussions to support community interaction between students, professors, and teaching assistants, as well as quick Feedback for immediate feedback. Testing and homework. Online courses are a new development that has been extensively studied in the field of distance education in recent years.

The online English course for preschoolers connects teachers and students through the Internet, and displays the facial expressions of both sides through cameras. The online English platform for preschool children will broadcast the lessons on the screen, and send the pronunciation, pronunciation and feedback of preschool children to teachers through recording and the Internet. Finally, the teacher will grade and interact.

During the learning process, the online education software will play various pre-recorded multimedia courses, and obtain teacher guidance and feedback through video calls.

Fig. 4.1 Online English courses(Adapted from LAIX® Application)

4.1.2 AR Books

For AR novel books, they use the latest technology Augmented Reality. Augmented reality is a rapidly developing field not only in the entertainment field but also in the education field. Augmented reality books are a subset of the field and provide a wide range of opportunities for new developments in entertainment and educational models.

Fig. 4.2 AR books sample from Little Hippo® Books

These books use technologies such as smartphones or game consoles to merge traditional text with digital content, and use applications that play videos, create models based on content, or allow interaction with text. A large number of AR books have been printed, but the possibilities for publishing (and self-publishing) such books are almost limitless.

AR books are physical or digital copies of traditional books, including text and illustrations, which are then linked to other non-traditional content through the use of technology. When a technology with a display and a camera (such as a smartphone, a tablet, or a computer with a webcam) is pointed to a page in a traditional book for which other content is created, an application is installed on the technology to read the page and Display other content on the screen of the device. This content may be as simple as another picture or video file or

audio clip, or as complex as an entire animation sequence or even a game or activity related to traditional media.

4.1.3 Existing Problem

The existing early education tools on the market are conducive to children's self-learning, and little attention is paid to children's participation. In an era when children cannot learn by themselves, the market is blank. For children's learning, it is still in a traditional learning state. There are no advanced tools to help parents, especially in the field of second language learning.

According to a survey of American families with children eight years and younger (Rideout, 2013), the number of young children using mobile media platforms and applications ("apps") has surged in the past few years. For families from 0 to 8 years old, three-quarters of families with young children in the United States have mobile devices (such as smartphones, tablets) [13].

Unlike traditional media (such as TV), this new type of portable smart touch screen media is becoming more and more available in various situations. The use of smart media as an educational tool is becoming more and more widespread, which also complicates parents' adjustment of smart media time. Although the use of smart media has many positive effects [14].

Therefore, in the field of early education, a new tool is needed to solve the problems of using control of smart devices and lack of self-learning ability in children's education.

4.2 Research on Pronunciation Evaluation System

English pronunciation affected by variety of factors, ability to accurately perceive the speech sounds of language [15], and different dialects of the same language [16]. English dialects are frequently used in different domains; business, education, governmental, etc [17]. The presence of English need to understand it, and there are increasing to interact with governmental and business sectors using English [18]. Three main factors are playing

in English perception [15]: listeners' native language, 920 English dialects, and phonetic context of speech sounds are presented.

At present, the mainstream scoring method of oral follow-up in the industry is mainly based on the hidden Markov model (HMM) speech recognition engine. The likelihood score and other relevant information were used as the basis of scoring.

Among them, the most classic GOP (goodness of pronunciation) method was proposed by Mike Witt of MIT in his doctoral dissertation.

The GOP algorithm proposed by Witt [19] occupies a very important position in voice evaluation.

The later scoring methods are mostly similar to or derived from GOP algorithm. The scoring algorithm used in LAIX application is also based on GOP algorithm.

With the improvement of computing power and the rapid development of computer-based speech processing and the creation of advanced speech recognition methods (including dialects and accents), it is now possible to apply modern speech technology to "Computer-aided Speech Teaching" (CAPT) [20]. There are currently several systems that can measure the quality of pronunciation of students by analyzing a few minutes of speech, and show a system as reliable as a trained human expert [21]. Although a high level of global pronunciation scores may be sufficient to meet the requirements of oral proficiency and pronunciation assessment, overall, the training objectives are not sufficiently detailed [22].

The technology in the field of speech evaluation is mainly reflected in the method of speech feature extraction and speech model matching. In the feature extraction stage of voice evaluation and diagnosis, the main task is to extract feature parameters related to voice quality from the original voice signal, while removing or reducing other redundant information. Most scholars directly use existing spectrum features or make minor improvements, and MFCC and PLP features are usually preferred. Obviously, the short-term spectral characteristics are based on the independence of each signal frame, while ignoring the time-varying characteristics of the signal. The time-frequency feature considers the energy change trajectory in the frequency domain of each subband, contains time-varying details that are

more important to the speech quality characteristics, and enhances the distinction between phoneme pronunciation and phoneme category.

The TRAP (TEMPORAL Patterns) feature proposed by Hermansky [23] is a more classic time-frequency feature, and has achieved better performance than traditional spectral features in phoneme recognition and other fields [24]. Li Hongyan [25] introduced the TRAP feature of the long time domain into the field of pronunciation error detection, which was used to characterize the pronunciation quality characteristics of phonemes, which greatly improved the error detection effect.

Chapter 5

System Design

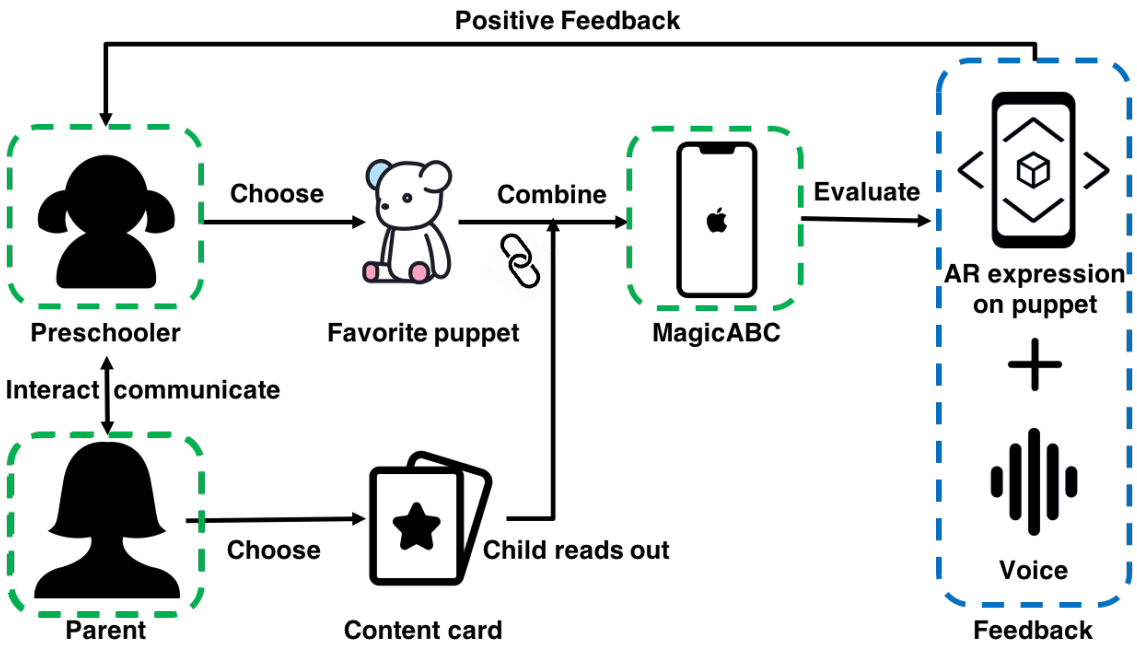


Fig. 5.1 System overview of MagicABC

In this chapter, we will introduce our system design and each important pieces of our approach. Fig. 5.1 shows the overall structure of our system.

Summarizing the framework, it mainly includes 3 parts. Preschoolers, parents and MagicABC system. First, the parents choose a learning card for the child. Secondly, the preschooler chooses her favorite puppet to join and read the content of the card. Third,

MagicABC system will use the built-in voice evaluation module to evaluate his pronunciation. The final MagicABC system will use AR facial expressions on puppets as rewards to encourage children to study hard.

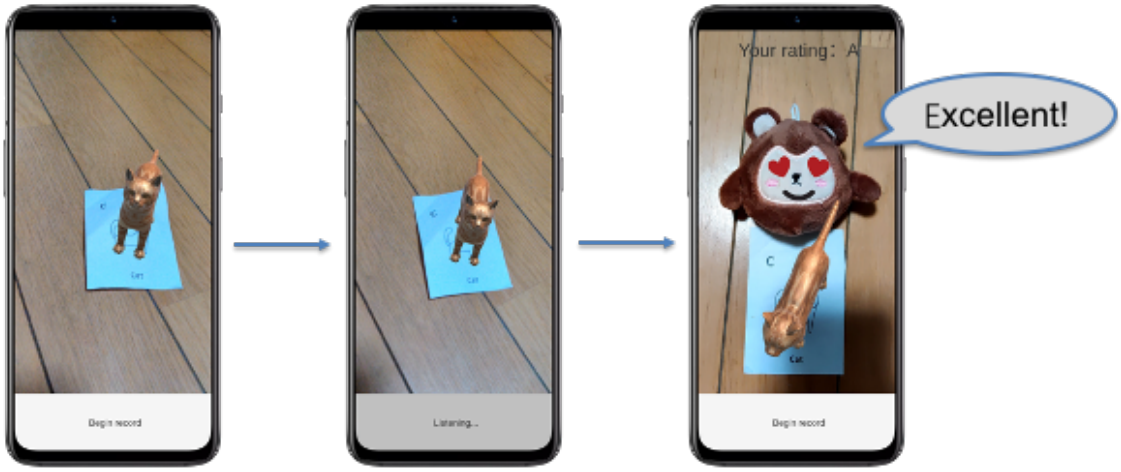


Fig. 5.2 System pipeline

For our system, the input of the system is to select a puppet and a card and the pronunciation of preschoolers. The output of the system is the music score on the puppet, AR facial expressions and voice feedback.

The following subsections provide detailed information about the key parts of this research.

5.1 AR Facial Expression Part

To operate MagicABC, parents should first run MagicABC on their smartphone. Then, when parents use smart media to scan the 2D content card tags and puppets, after inputting the pronunciation, facial expressions will cover the puppet's face, as shown in Fig. 5.2. The type of facial expression depends on the child's pronunciation score. If the child's

pronunciation is close enough to the standard pronunciation, the child should be given a high score and positive feedback, otherwise negative feedback should be given.

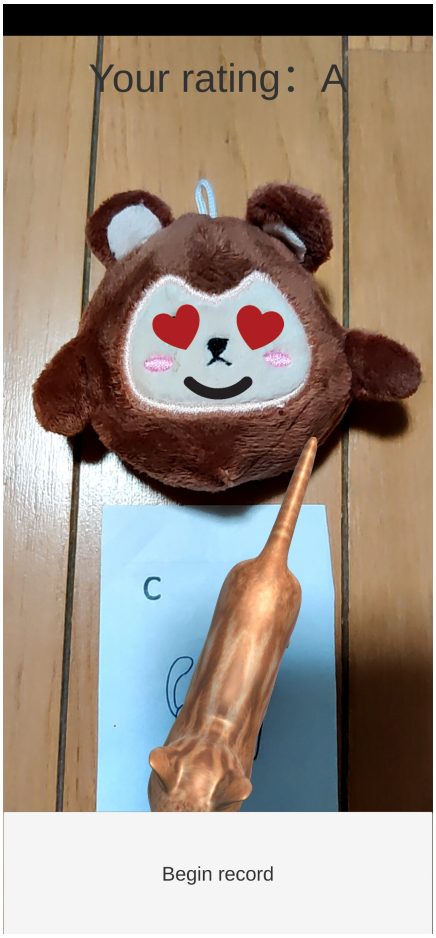


Fig. 5.3 AR facial expression sample

The specific results and corresponding feedback results of MagicABC. For example, when a preschooler's pronunciation is good, the AR program superimposes a smiling face and sweet voice. The scoring criteria and system are shown in the Fig. 5.4.

When preschool children have excellent pronunciation, the AR program will overlay the smiling facial expression with laughter. The feedback includes four groups of different facial expressions and corresponding sound effects. The AR expression looks like the picture in Fig. 5.5.

The Vuforia library and Unity 3D are used to achieve two-dimensional label and 3D puppet recognition. More specifically, we followed these three steps.

Level	Grade	Score	Expression
A	Excellent	85~	
B	Amazing	75~85	
C	Great	60~75	
D	Nice	~60	

Fig. 5.4 Scoring system for MagicABC

- First, we created a 3D model database for each puppet through Vuforia Object Scanner so that we can identify the animal type, such as "monkey" or "bear". Then register these databases in a format that Vuforia can read and process;
- Secondly, for each puppet, we determine the position and size of the puppet according to the 3D data of the puppet, and indicate the relative position and size of the puppet head;
- In the process of making AR expressions, we use Unity3D to draw a yellow sphere, which represents the head of the puppet. The yellow sphere will act as a screen, and the facial expressions of the puppets will be projected onto the screen;
- Finally, we will remove the yellow sphere that represents the head but retain the facial features. In this way, these facial features will cover the facial features of the puppet and show the facial expression of the puppet.

After completing these steps, the expression production and projection are basically realized. Whenever the Vuforia library recognizes the marker-less puppet model captured by the smart media camera and gets the pronunciation score, the application will cover the AR facial expressions on the top of the puppet.

For the feedback part, the system will use AR facial expressions and puppet sounds to encourage children. It will also play local pronunciation to modify the pronunciation of preschoolers.

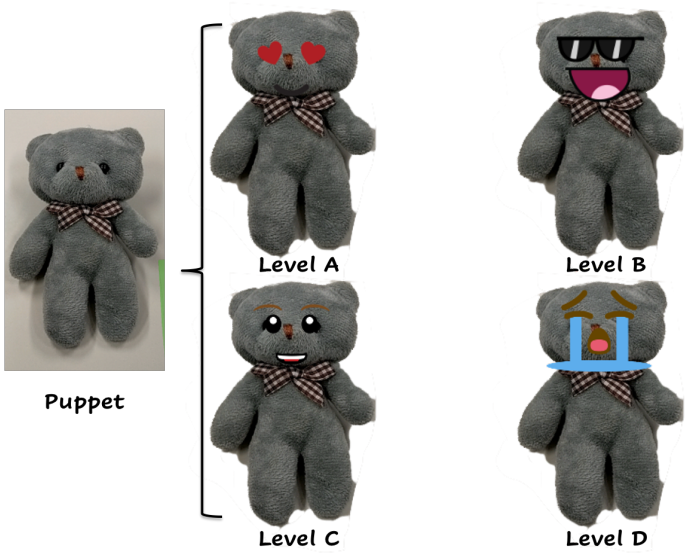


Fig. 5.5 Facial expression examples of MagicABC

AR facial expressions will change with different puppets. It will display different expressions and voice feedback based on the score.

For this part, the system can realize these functions.

- Encourage the children with the AR facial expressions and voice of the puppets;
- Play the local pronunciation of the card to modify children's pronunciation.

5.2 Pronunciation Assessment Part

To get this score, we should do some work about natural language processing.

For example, if we want to evaluate text "about" [əˈbaʊt]. And the actual pronunciation of the user is "aboud" [əˈbaʊd]. There are four steps for assessment.

- First, find phoneme sequence. In this case, "about" consists of [ə][b][aʊ] and [t].
- And then find the boundary of phoneme by analyzing the spectrum.
- After that, comparing with native pronunciation, calculate phoneme posterior probability.

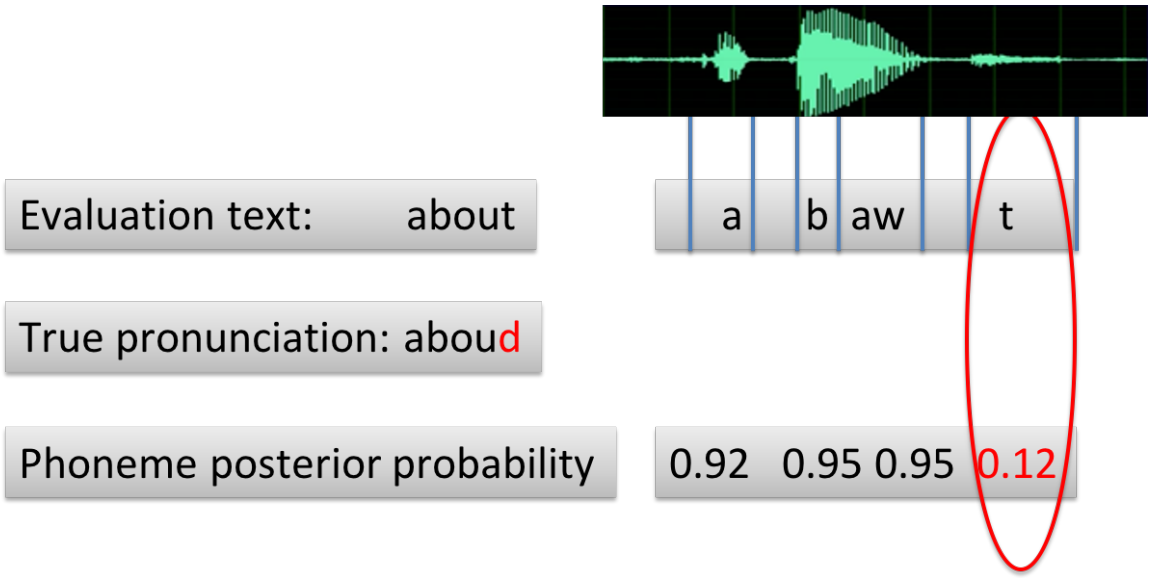


Fig. 5.6 Pronunciation analysis sample

- Finally, according to threshold method, we find the probability of this phoneme [d] is low. At this time, the system found an error of pronunciation.

For details of implementation, we use Goodness of Pronunciation (GOP) algorithm [19] to calculate phoneme posterior probability based on related paper2. And we also use the threshold method to judge whether it is wrong or correct. For scores, calculation is based on comprehensive accuracy and pronunciation duration.

Chapter 6

System Implementation

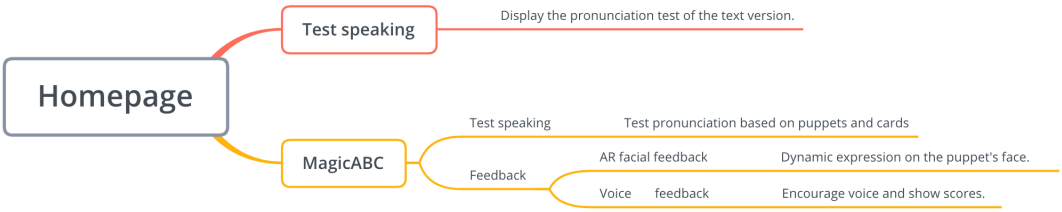


Fig. 6.1 Overall flow

Generally speaking, our system has two parts. They are the test speaking part and MagicABC part. For test speaking page, it can display the pronunciation test in the text version. For MagicABC page, it can realize the functions including test speaking and give AR facial expression feedback and voice feedback.

The related work will be introduced in this part.

6.1 Hardware and Programming Environment

To build a pronunciation assessment and AR feedback system, we used some hardware and programming tools to create the system.

- Hardware Device for using system: Mobile with camera (OnePlus 7);

- Programming Environment: Unity 3D 2018.2.20f1 on, WeChat Mini Programs Tools V1.02.2003250;
- Android SDK with Java for front-end building on smartphone;
- Vuforia for AR-marker detection and AR contents display.

Operation System	Microsoft Windows 10
CPU	Intel® Core™ i7-6500UIntel(R) @2.50GHz 2.59GHz
Graphics Card	Intel® HD Graphics 520
Ram	8 GB

Table 6.1 Information of PC

Fig. 6.2 OnePlus 7 powered by Android 10

We built the system in Unity 3D 2017.4.8f1, built the required 3D model, and used C sharp and JavaScript as the development language. In order to implement the system, Android platform should be Android 7.0 or higher, so we use OnePlus 7 like as shown in Fig. 6.2 as the device, and do some front-end construction for the system. To identify images or tags and add virtual content to them, we use the Vuforia unity SDK and a PC as shown in Table 6.1 to support programming. And I also use Macbook to program and do some support work as shown in Fig. 6.3.

Fig. 6.3 MacBook Pro powered by macOS Catalina

In addition, we need a USB C cable to connect the phone to the computer. We use Unity 3D 2017.4.8f1, or later to select Android build support during installation. You need to use the SDK manager in Android studio to install Android SDK 7.0 (API level 24) or later.

6.2 AR Content Card

The original card is shown in Fig. 6.4. It has the same size as a normal card. Different recognizable patterns are printed on the surface of the card, and users can choose according to their own preferences. Even if some parts are blocked or worn, this photo can still be recognized.

6.3 3D Puppet Scanned and Recognized

In this project, because the preschoolers' favorite puppets are used to display AR facial expression. So we need to use Vuforia engine for 3D object recognition. First we need to download the software, Vuforia Object Scanner, from the Vuforia Developer Portal.

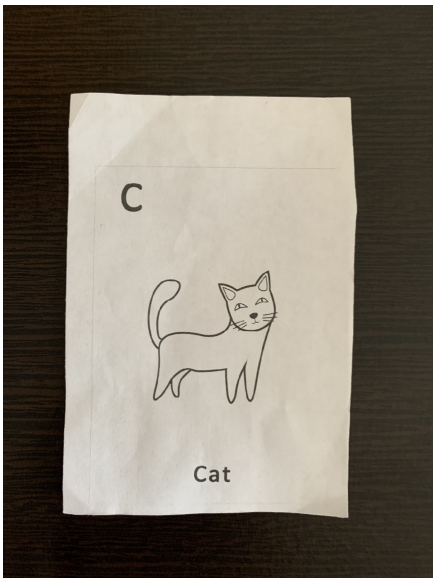


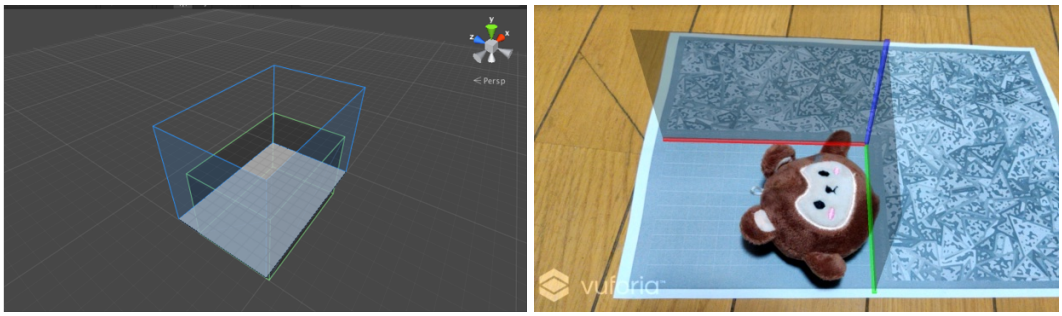
Fig. 6.4 AR content card

Vuforia Object Scanner allows you to create targets by scanning objects using an Android device. Just install the app, place an object on the Vuforia scan target, and start scanning. The app can provide you with real-time visual feedback on scan progress and target quality, and establish a coordinate system so that you can build an immersive experience with precisely aligned digital content. The test mode allows you to evaluate the quality of identification and tracking in the application before starting any development.

The feature region of the target consists of overlapping triangular shapes. This region serves two roles. It enables the scanner to precisely identify the pose of the physical target in the grid region and also defines the culling region of the scanning space. Object features outside the target area will not be incorporated into the object data.

The local origin is represented by (0,0,0) in the lower left corner of the grid of the target scanning target. It corresponds to the local (0,0,0) origin of the bounding box of the object target. The unit scale of the grid is meters. The meter is also used for scene units and estimated proportions of physical objects.

First, we need to print out the scanning target paper and point target puppet on the scanning target paper. Put them in an environment without background noise. Place the



(a) Bounding box for Vuforia Scanner (b) Point at an object on the scanning target

Fig. 6.5 Vuforia Object Scanner and the scanning target

model to be recognized in the specific scanning process, the software will determine the position of the coordinates according to the canning target paper.

Scan the object. The following steps explain how to scan an object. To avoid damaging the results, do not move objects or targets while scanning.

1 Open the Vuforia object scanner;

2 Press the + icon to start a new scanning session

3 Use the axis expansion to confirm that the objects are properly aligned;

4 Press the record button. Remember not to move objects or targets while recording the scan;

5 Use the camera to capture important advantages and provide a user experience for your application. After successfully capturing the surface area, its corresponding facet will turn green as shown in Fig. 6.6;

6 After capturing most of the required surface area, press the stop button to stop the scan.

When the scan is complete, we can view the scan results through the software. The green square in the lower left indicates the area distribution of the captured features, and the file size and number of points are displayed on the scan summary screen on the right. Complete instructions can be found in the user guide of the application.

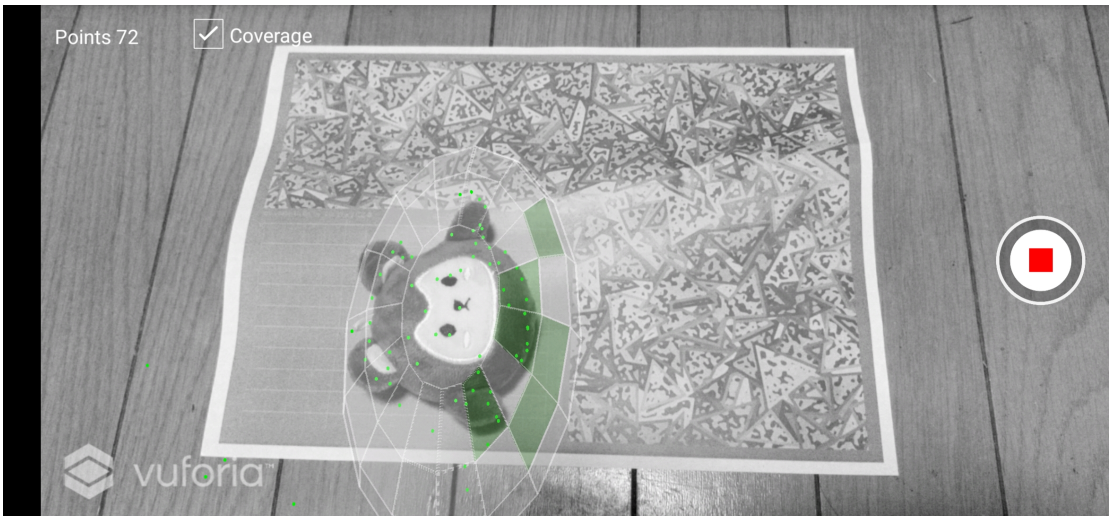


Fig. 6.6 The process of building a model using Vuforia Object Scanner

Once we have scanned an object, a summary screen displays our scan results. If we want to check the effect of model building and test whether Vuforia can recognize the model. We need to click the test button and point the camera at the target object. If the application can display a green cuboid. That shows that Vuforia is working properly and this model is available. Then, we can add the database to Unity3D to scan and identify the physical model. It can be seen from the test results that the two models we constructed can be successfully identified. So we will use this database to construct the system.

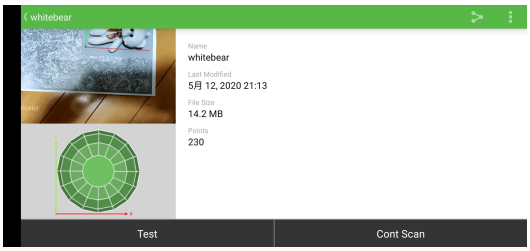
This is the model data and test results of the bear and monkey puppets model we created, as shown in Figure 6.13 and Figure 6.14. Then, through the Vuforia Developer Portal, upload the model data as a 3D object type to complete the creation of the 3D model.

So far, we have completed the construction of Vuforia database including AR content card and 3D puppet model.

6.4 AR Module Implementation Process

6.4.1 Software Environment Settings

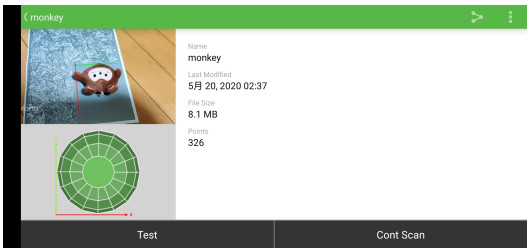
Based on the Vuforia service, our system can preset and identify AR tags in the user view through AR devices or PCs with webcams. To use this engine, we need to get the



(a) Vuforia Object Scanner for bear

(b) Vuforia test for bear puppet

Fig. 6.7 3D bear puppet scanned and recognized



(a) Vuforia Object Scanner for monkey

(b) Vuforia test for monkey puppet

Fig. 6.8 3D monkey puppet scanned and recognized

MagicABC Edit Name Delete License Key

License Key | Usage

Please copy the license key below into your app

```
AdudIQb/////AAABma7FNgjUp0KlkRv4HjC/Kyow7cAqjZWfH1tOSTno/NOLPOc7i68fJG0dows7AXeFXdz0n2XYxvOUCuU/zZHjUQ2FSb
rOJDQBRcyV4AQQbbAcY7dkfiARmXDVjmsRm53vBXH3scq4raSNMxjNFJ80JLKciidV8jxg3ngp0wZLwhRjtHhc4hdTCruIDRFkxh9C4JQa
zp/p+UMqKgALKrsEw0VPS+JRaKRoAkrffMxTzvTbP/9yDM0/Hl/Kv5TSr/I2rmwrd4sxkDZLaCy11TdBT/J5bLaZ7/P+4tmZocgCI19Sws
0AgOvArJnOx025g+GnXryTGOIcy68yCgjZCmQ0FPc/KIcEgfihPK7/42vnCWFz
```

Plan Type: Develop
Status: Active
Created: Mar 08, 2020 21:29
License UUID: 6435c352b6254451b165555b9a163f15

Permissions:

- Advanced Camera
- External Camera
- Model Targets
- Watermark

Fig. 6.9 License Manager for MagicABC

free license from their homepage (https://developer.vuforia.com/legal/license). We applied for the license of Vuforia service on the official website of Vuforia. The license key of MagicABC is Fig. 6.9 and the Vuforia Image Target Database of MagicABC is Fig. 6.10.

☐	Gorilla	Single Image	★★★★★	Active	Mar 30, 2020 11:58
☐	Fish	Single Image	★★★★★	Active	Mar 30, 2020 11:58
☐	Elephant	Single Image	★★★★★	Active	Mar 30, 2020 11:58
☐	Dog	Single Image	★★★★★	Active	Mar 30, 2020 11:58
☐	Cat	Single Image	★★★★☆	Active	Mar 30, 2020 11:58
☐	Butterfly	Single Image	★★★★★	Active	Mar 30, 2020 11:58
☐	Ant	Single Image	★★★★★	Active	Mar 30, 2020 11:57
☐	bearmarker	Single Image	★★★★★	Active	Mar 28, 2020 22:19

Fig. 6.10 Vuforia Image Target Database

After obtaining the license, we uploaded the tags to the database and obtained the analysis data. For best results, we should target image tags with 4 or 5 stars for upload, like Fig. 6.11.

Fig. 6.11 The details of the picture target tag

Because it uses image recognition to capture the target, the camera needs to capture the target quickly and clearly. If the target is not well focused in the camera view, the camera image results may be blurred, target details may be difficult to detect, and the performance of induction, detection, and tracking may be negatively affected. We need to use the appropriate camera focus mode to ensure the best camera focusing conditions.

We can download the Vuforia database from the target manager page and import it into our unified project. Then, we can track the image target mark in the system and generate AR virtual content above it.

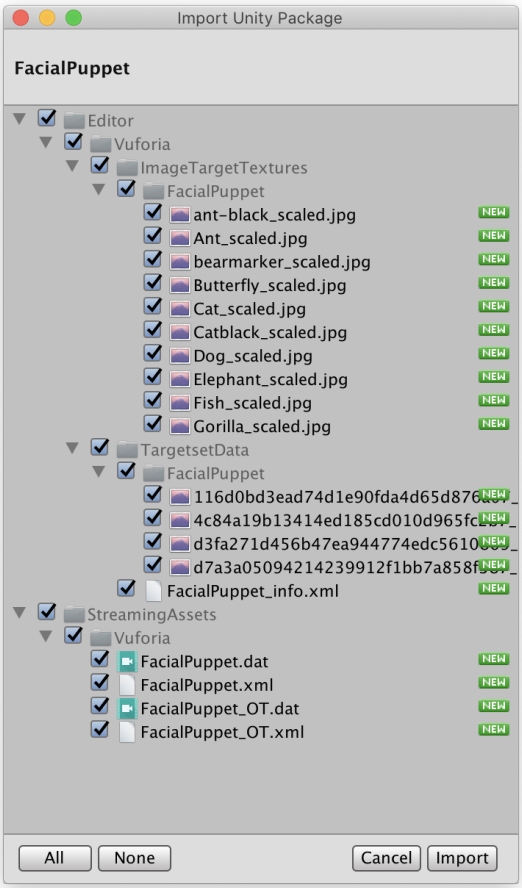


Fig. 6.12 Import the database as Unity package

After downloading the database and importing it, we need to activate Vuforia Augmented Reality in unity's player settings panel, create a new project and import the SDK:

1 Open Unity and create a new 3D project.

2 Unity 2017.2 or later:

1) File > Build Setting > Platform > Android.

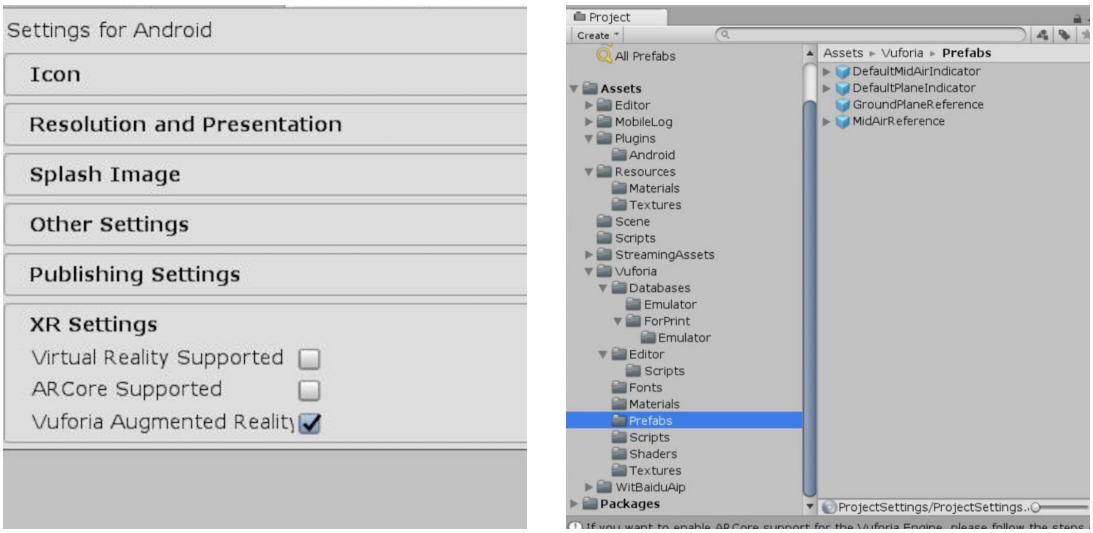
2) Player Settings > XR Settings > Vuforia Augmented Reality.

3 Import the Vuforia SDK for 2018.2.20f1 or later:

1) Select Assets > Import Package > Custom Package.

2) Select the vuforia-unity-6-2-10.unitypackage that you downloaded.

3) In the Importing Package dialog, make sure that all package options are selected and click Import.



(a) XR-Reality (b) Vuforia folder

Fig. 6.13 Vuforia SDK import

Then open Vuforia's checker, add our license key, and set the maximum number of simultaneous targets. We also need to check the "Load Marker Database" and "Active" boxes to make the system run normally.

After all the preparations have been down, we need to build an AR camera in the scene and create an image target by Vuforia engine. The AR content we want to show above the image target marker need to be registered below the image target.

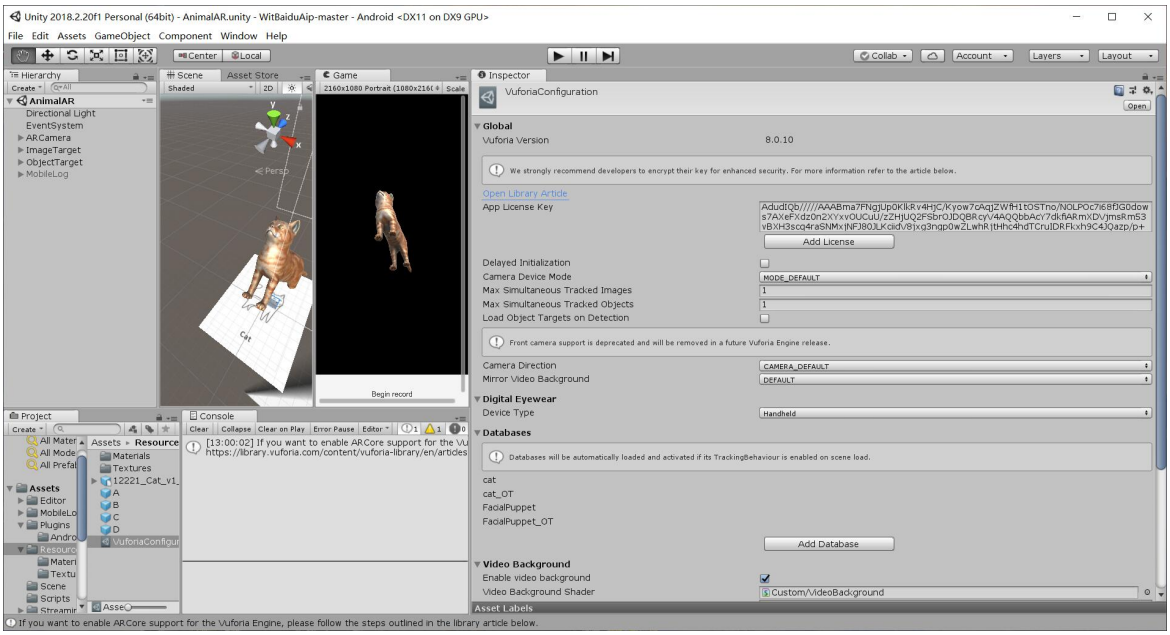


Fig. 6.14 Input the license key

After uploading the designed images and packaging the Image Target database into Unity 3D, and choosing to bind the 3D model to related tags, we can use these tags to display AR content in our project. We can then use the camera to scan the image target mark and view the AR content above it.

6.4.2 AR Content Card Recognition

After completing all the preparations, we need to build an AR camera in the scene and create an image target through the Vuforia engine. The AR content we want to display above the image target mark needs to be registered under the image target. Then, we can use the web camera or smartphone to scan the image target mark and view the AR content above it.

When this step is completed, once the card is scanned in the smartphone's field of vision, it will immediately display the corresponding animal card. For example if we scan a AR Cat Card, the cat model will be displayed on the display as shown in Fig. 6.15.

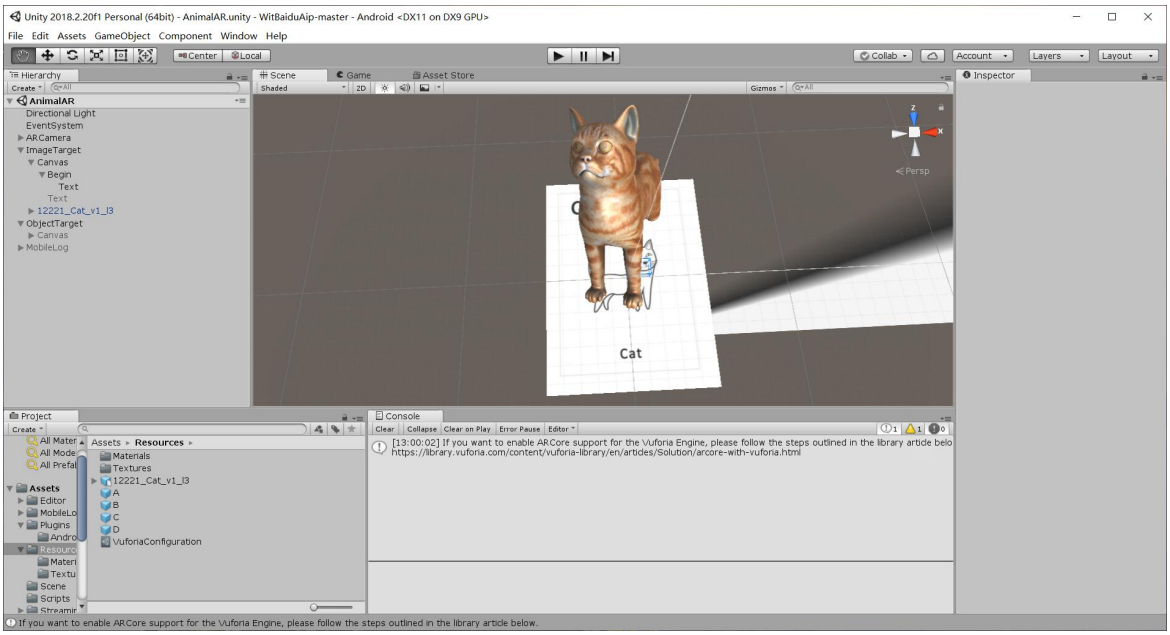
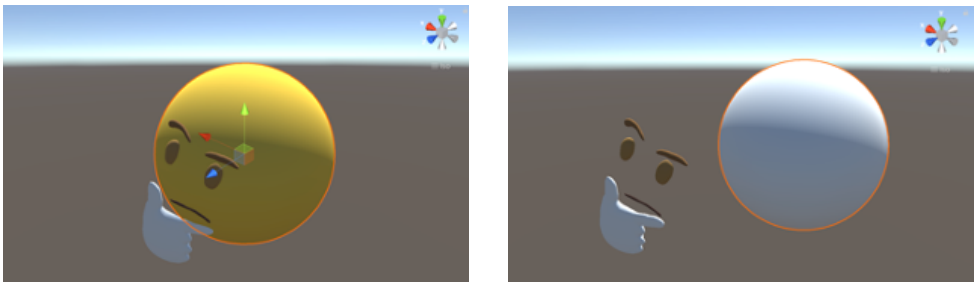


Fig. 6.15 Register the AR content above the marker

6.4.3 3D Animal Puppet Recognition and Facial Expressions

When it comes to 3D puppet model recognition and expression feedback, the first step is to identify the corresponding model, and then display different AR expressions based on the results of the voice evaluation for this model.



(a) Original thinking emoticons model (b) Emoticons in editing

Fig. 6.16 Thinking emoticons model in Unity3D

In completing this goal, we followed these five steps.

Step 1 First, we create a 3D database for each puppet so that we can identify the animal type, such as "monkey" or "bear". This step is completed by the database construction method mentioned earlier.

Fig. 6.17 AR facial expression sample for bear puppet

Step 2 Then, download these databases from the Vuforia Developer Portal and import and register them into Unity 3D. At this time, these databases can be read and processed by Unity3D.

Step 3 After that, for each puppet, we use the data of the position and size of the model collected during modeling, and indicated the relative position and size of the puppet head.

Step 4 Then came the production of a 3D model of expression. we used Unity3D to draw a yellow sphere, which represents the puppet's head. The yellow sphere will act as the screen, and the puppet's facial expression will be projected on the screen. The thinking expression model is shown in Fig. 6.16.

Step 5 Finally, as long as the Vuforia library recognizes a specific puppet thing, it will call the corresponding preset position and size information. Then we will remove the yellow sphere that represents the head but retain the facial features. In this way, these facial features will cover the facial features of the puppet and show the facial expression of the puppet. According to the output result of the voice evaluation module, the corresponding expression is overlaid on the facial expression on the top of the puppet.

When doing this part, first of all, we designed emoticons with different scores. The specific score standard can refer to Scoring system for MagicABC shown in Fig. 5.4.

Project the thinking expression on the puppet, the actual display effect of the software is shown in Fig. 6.17. The system at this time already has the ability to show different expressions to the puppets. When this function is realized. Marked that the AR part of the system has basically been completed.

The next thing to consider is how to accurately evaluate voice evaluations. And how to design a system to better serve parent-child interaction and encourage children to study hard. This is the work content that will be introduced in the next part.

6.5 Pronunciation Assessment Module

At present, the mainstream follow-up oral scoring method adopted by industry is mainly based on the Hidden Markov Model (Hidden Markov Model) speech recognition engine using its likelihood score and other relevant information as the basis for scoring. Among them, the most classic GOP (Goodness of Pronunciation) method was proposed by Silke Witt of MIT in his doctoral thesis. Most of the subsequent scoring methods are similar to or derived from the GOP algorithm.

6.5.1 The Difference With Speech Recognition

Speech evaluation is based on speech recognition, but it is very different from speech recognition tasks.

- In the speech recognition task, the text (content) corresponding to a piece of speech is not known in advance and needs to be "guessed" by the speech recognition system
- In the scoring system, the text corresponding to a piece of speech is known in advance. What the system needs to do is to make a pronunciation evaluation of the speech

The basic idea of the GOP algorithm is to use the text information known in advance,

- Force alignment the speech and its corresponding text
- And compare the likelihood score value obtained by forced alignment with the likelihood score value obtained without knowing the corresponding text
- Use this likelihood ratio as an evaluation of pronunciation

Intuitively, this type of algorithm calculates the likelihood that the input speech corresponds to a known text. If the probability is higher, the pronunciation is more standard.

The calculation of this possibility is based on the acoustic model (acoutisc model) in speech recognition, and the acoustic model is often trained by recording a large number of native speakers. This involves a corpus.

6.5.2 Evaluation Index

The error pronunciation granularity output by the evaluation system may be one of phoneme, syllable, and word. Generally, the phoneme level is used for output. Pronunciation evaluation is to detect errors by comparing reference pronunciations with user pronunciations.

In the following, we will introduce how to model pronunciation based on speech recognition technology, and how to score the pronunciation of each phoneme based on the GOP evaluation algorithm to be introduced in this article.

Phoneme

With the reference text, the evaluation system can refer to the pronunciation dictionary to obtain the phoneme sequence [q1, . . . , qi, . . . , qn] corresponding to the text. The follow-up work is to check whether each phoneme qi here is correctly read.

Phonetic Features

The voice audio collected by the microphone is PCM data in wav format. And if it is a compressed format such as mp3, it needs to be decoded to PCM first. This raw data cannot be directly processed. Like most machine learning systems, we need to first extract speech features from raw audio.

In the field of speech recognition, the most commonly used is the Mel frequency cepstrum coefficient-MFCC.

- When extracting this feature, first perform short-time Fourier transform (STFT) on the raw audio to obtain the spectrogram;
- Then calculate the cepstrum coefficient in the Mel domain.

After the speech feature extraction is completed, the original pronunciation audio is converted into a speech feature sequence, which consists of frames.

6.5.3 Speech Recognition Framework Based on HMM

In speech recognition, we think that the pronunciation process of each phoneme is caused by the deformation of the vocal organ, and each shape of the vocal organ corresponds to an implicit state of the HMM model (generally 3 to 5 implicit states are used in speech recognition Modeling), in each shape (implicit state), the vocal organ will produce a specific sound with a certain probability (in speech recognition, it is a speech feature). We cannot directly observe the shape (implicit state) of the vocal organs, but we can see the specific speech features (observed values). Therefore, when the HMM model parameters are known (that is, the HMM model is trained offline for each phoneme), the most probable phoneme state sequence can be calculated according to the speech feature sequence.

In addition, the HMM model can be concatenated, that is, the end state of the current phoneme can jump to the beginning state of the next phoneme, so we can connect a known phoneme sequence, such as the HMM model of 8 phonemes in Hello World [həˈloʊ wɜːrld]. It forms a large HMM model. This large HMM model describes the pronunciation process

of the whole sentence of Hello World. Based on the observed entire speech feature sequence and the model parameters of this HMM, we can calculate the most probable phoneme state sequence of the current speech under the HMM model of Hello World, that is, which phoneme each frame of speech belongs to, and this frame corresponds to the phoneme's Which state, and the corresponding conditional probability.

6.5.4 Evaluation Process

In this part, we complete it through forced alignment and free identification, and GOP evaluation. Pronunciation assessment module implementation process is shown in Fig. 6.18

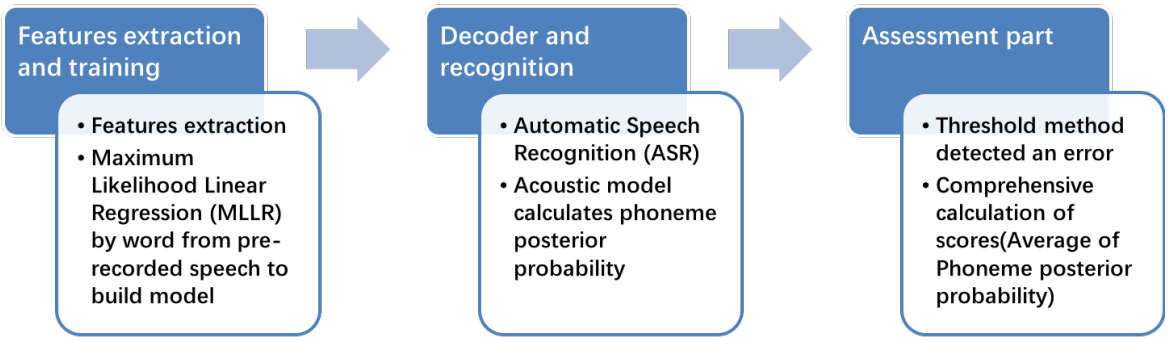


Fig. 6.18 Pronunciation assessment module implementation process

Forced Alignment and Free Identification

In the first step, we want to align the text Hello World with the user's pronunciation audio. This step is usually called "forced alignment".

What about forced alignment? First, we know that the phoneme sequence corresponding to Hello World is [həˈloʊ wɜːrld]. Then we concatenate the 8 phoneme HMM models to form a large HMM model.

According to the above introduction, based on this model and the speech feature sequence extracted from the user's audio, the most probable phoneme state sequence can be calculated, that is, which phoneme in [həˈloʊ wɜːrld] each frame of speech belongs to, and the phoneme Which state. In this way, the correspondence between the audio frame and the phoneme state is realized, and the name of forced alignment comes from this. After the forced alignment, the range of the speech frame interval corresponding to each phoneme (denoted by qi) is known.

If the user's pronunciation is correct, then after forced alignment, it will accurately locate the speech frame interval corresponding to each phoneme, that is, the alignment.

If the user pronounces incorrectly, for example, a phoneme qi is misread, then the aligned qi speech frame interval is actually not the sound of qi. Then free identification will be used. Regardless of the reference text, that is, the phoneme sequence is not limited to [həˈloʊ wɜːrld], and the most probable phoneme state sequence is calculated directly using speech recognition, which is actually equivalent to selecting the phoneme that best matches the user's pronunciation from all phoneme lists sequence.

Because our system is for one word, this step is greatly simplified. We only need to determine the boundary of the word based on the spectrum information to find the user's pronunciation of the word.

GOP Algorithm and Evaluation

After completing the forced alignment and free recognition, we know which section of the speech corresponding to each phoneme in the reference text, and what sound is actually produced by the user's speech.

With this information, we use the GOP algorithm to measure the accuracy of each phoneme of the reference text.

After the forced alignment, the speech corresponding to each phoneme q in the reference text is determined.

With this information, GOP is defined as follows: on the premise of observing this segment of speech, it corresponds to the probability value of the reference text phoneme q. The formula is:

$$GOP(p) = logp(p|o) = log\frac{p(o|p)\,p(p)}{\sum_{q\in O} p(o|p)\,p(q)} \approx log\frac{p(o|p)p(p)}{max_{\{q\in O\}}p(o|p)p(p)}$$

Among them, qi is the phoneme currently to be scored in the reference text, and O is the speech corresponding to qi after forced alignment.

It can be seen that the GOP score is actually a conditional probability, which measures the "probability of this voice corresponding to the phoneme qi when the user's voice O is observed".

The higher the probability, the more accurate the pronunciation. The lower the probability, the worse the pronunciation. It is understood in a physical sense and is a suitable scoring metric.

So far, we have completed the introduction of GOP. As mentioned above, according to the GOP formula, we can score the pronunciation of each phoneme in the reference text. At the same time, according to the result of the forced alignment, we can also know the speech frame interval corresponding to this phoneme, so that we can get the position of the pronunciation error . More importantly, we can know what the user's true pronunciation is based on the results of free identification, and this information can also be returned to the user as a result.

In order to achieve this function, we used the WeChat Mini Program development platform. This is a JavaScript-based development platform supported by Chinese Internet company Tencent.

According to the steps, a prototype software for voice detection was written in the WeChat Mini Program. The software development interface is shown in Fig.6.19.

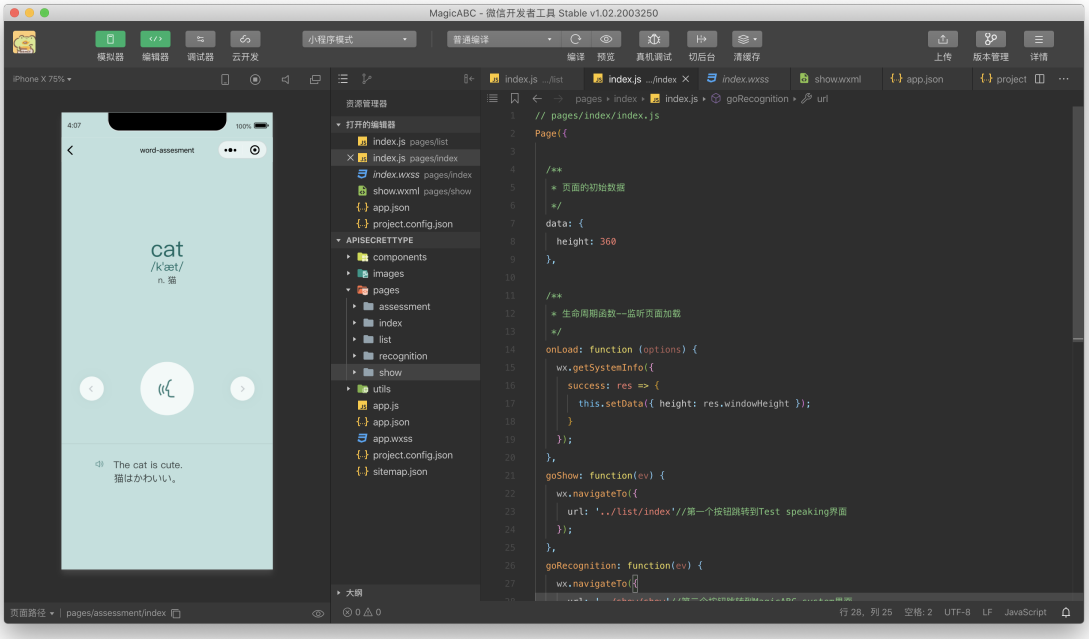


Fig. 6.19 WeChat Mini Program development platform interface

After compiling the code into software. Fig. 6.20 shows homepage and recognition results of the software. Specifically, Fig. 6.20a shows the homepage of prototype application. Fig. 6.20b shows the process of recording and identifying. While Fig. 6.20c shows the recognition results. The result is a score.

(a) Homepage (b) Recording & Identifying (c) Result

Fig. 6.20 Pronunciation assessment module based on WeChat Mini Program

6.6 Fusion Module

When both parts can work independently. Next, we will combine two modules to achieve full functionality to construct a complete system. Before fusing the two modules, we drew the timing diagram of the system, as shown in Fig. 6.21. When the user uses the system, he will go through the following steps.

Step 1 The user opens the software.

Step 2 The user selects cards and puppets. The system scans cards and puppets, recognizes objects, and gets the corresponding preset positions and models.

Step 3 The user long presses the system "record" button. The system records the user's pronunciation data.

Step 4 The user releases the record button. The system displays "Identifying" and transmits the pronunciation data to the voice evaluation module for scoring.

Step 5 The pronunciation evaluation module sends the evaluation score to the AR expression module to display the corresponding feedback. The user obtains appropriate results and can proceed to the next test.

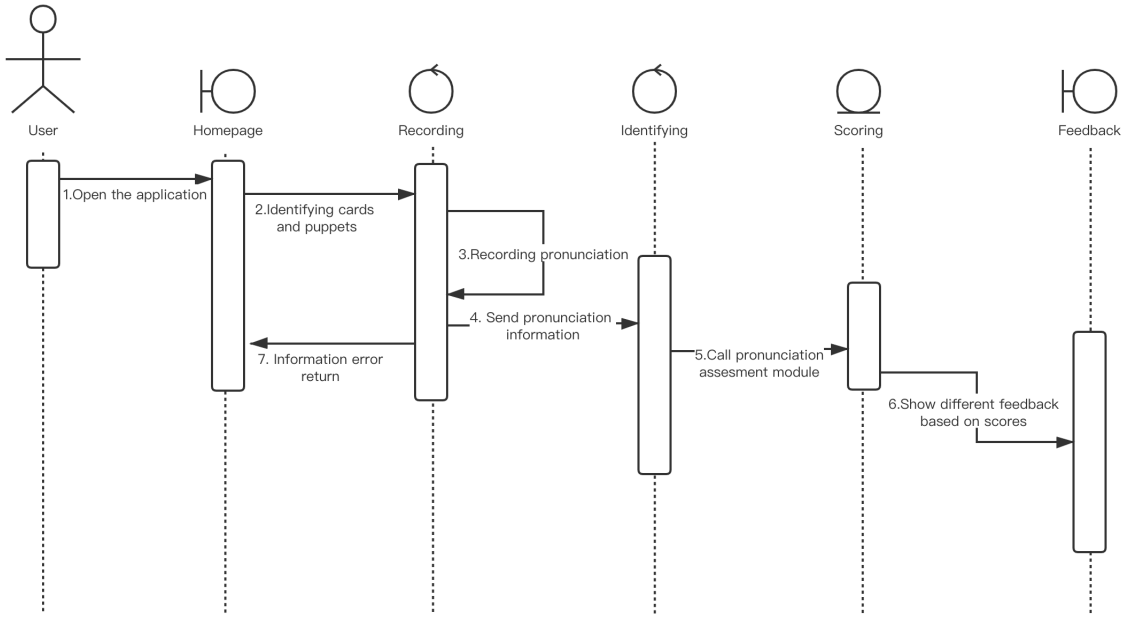


Fig. 6.21 MagicABC sequence diagram

The next job is to write script code as shown in Fig.6.22. In this part, we use Visual Studio Code on the Mac side for prototyping. Then use Thinkpad by win10 to run Unity3D and Macbook powered by macOS Catalina Visual Studio Studio for simulation and compilation.

When the basic functions can be realized, the next thing to do is to design the basic UI of the software. The test machine we use is OnePlus 7 powered by Android 10. The UI design of the software is shown in the Fig.6.23.

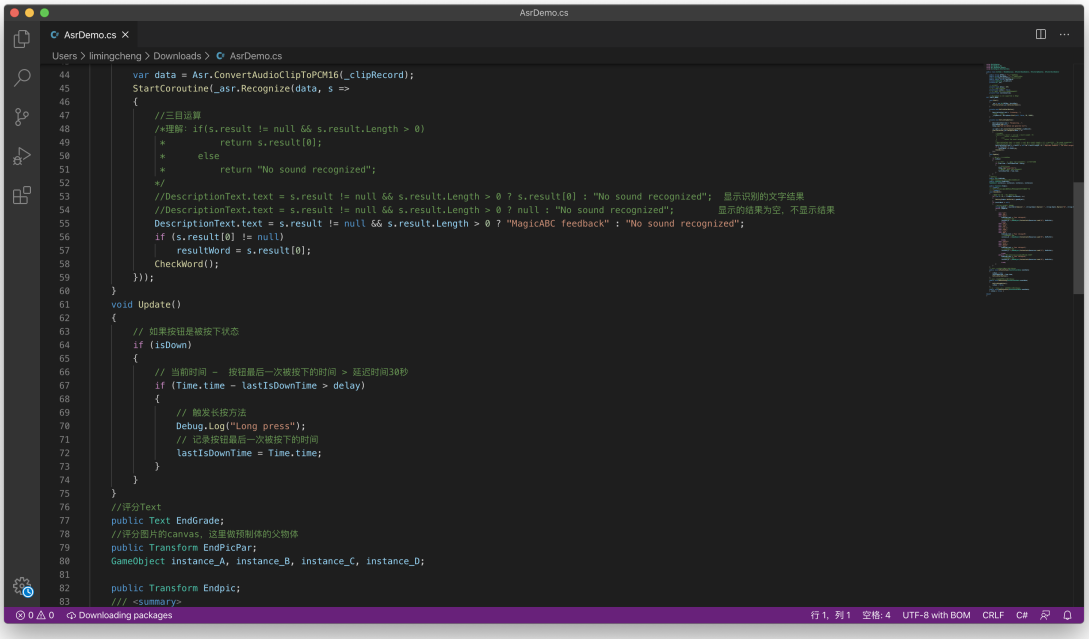
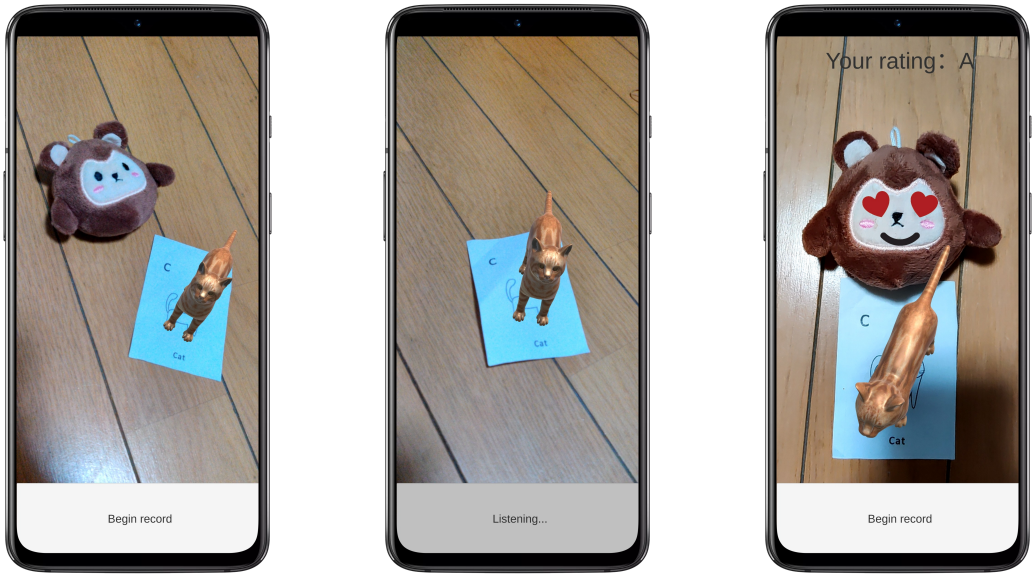


Fig. 6.22 Screenshot of script code for MagicABC



(a) Begin Record (b) Listening (c) Recognizing

Fig. 6.23 UI pages of MagicABC

So far, we have a suitable model and recognition method. The next step is to put the AR content card marker and marker-less puppet model into the real environment. With Unity, we can define interactions. For mobile use, we enable users to scan models and record sounds through the camera. Users can see different scoring results on the screen. As shown on the screen, users can move the model, just like other AR software to appreciate the different expression feedback of puppets.

The actual operation result of the software is shown in the Fig.6.24. At this point, the integration of the two modules has been completed. The functional design and implementation of the software have been completed.

Fig. 6.24 Operation result of the software

Chapter 7

Preliminary Evaluation

In order to evaluate our system, we conducted experiments on 10 users. They were asked to use our system, and then they were asked to answer some questions. Before users use our products, we will tell them how to use our system.

Fig. 7.1 Alphabet Flash Card

In order to make the experience smoother, we import the commonly used models and cards into the database of our system. Compared with traditional traditional Alphabet Flash Card like Fig. 7.1, we provided MagicABC system feedback and the main purpose of which was to test user satisfaction. We divide this problem into four sub problems:

Q1. Can this system solve the problem?

Q2. When using this system, do you like it compared with other similar systems?

Q3. Can this system improve parent-child relationship?

Q4. Can this system improve the level of second language education?

In order to get the answers to the above four questions, we designed the experiment according to the following ideas:

- For users, it is not required whether they must use this function. We just need to see how many users will use it and when they will use it. After the whole test, the interview was conducted according to the user's behavior in the experiment.
- Satisfaction is hard to calculate. In order to get this result, we asked them questions about the Net Recommended Score (NPS), which is widely used by many companies.

7.1 Participants

To evaluate our system, we invited 10 users (8 women and 2 men) to join our experiment. They are all asked to follow the process we provide. All users have taught children to learn English. They are all undergraduates. They are between 22 and 28 years old. They are able to read and write in English and are familiar with the operation of smartphones and laptops.

7.2 Method

As we said before, to make the experience more smooth, we plug our system into a traditional Alphabet Flash Card like Fig. 7.1. In particular, the card is cat content card shown

in Fig. 6.4. The puppet is monkey puppet shown in Fig. 6.8b. The only difference between them is that we provide the system functions of MagicABC in the demo application. Once the participant uses MagicABC system, he can use the all the functions.

In the experiment, users need to follow the following steps:.

1. Use the demo app to educate children, use it at least once. Before testing, we will only tell users how to use our system. The user does not have to use the virtual try-on function. We will observe their behavior while using the demo application, and we will not interrupt or talk to them while they use the demo application.

2. After using the application, each participant will be interviewed and a questionnaire will be filled out. Finally, ask them to give 1 to 10 points to describe how they are willing to recommend the feature to friends.

In the process of user use. We will focus on these behaviors of users. Whether to use this function. Observe whether the entire system surprises users when they use this feature.

After using the demo app, we will interview them below questions;The answer is grading from 1 to 5 (1 = very negative, 5 = very positive).

Question	1	2	3	4	5
Do you think it is helpful for you to educate preschool children?					
Do you think it is easy to use?					
Do you think it can decrease the chance of return?					
Do you think it is better than only pictures?					
Do you think this system is attractive?					

Table 7.1 Investigative questions after using the system

Besides the questionnaire, we also ask the user some open-ended questions, that is do you have any comments for the improvement of this system?

The Net Promoter Score is used as a proxy for gauging the customer's overall satisfaction with a company's product or service and the customer's loyalty to the brand. Based on Net Promoter Score definition, we will let the user give the NPS:

This is a sample of the questionnaire sample for MagicABC, as shown in the Fig. 7.2.

We will divide users into three categories according to their ratings:

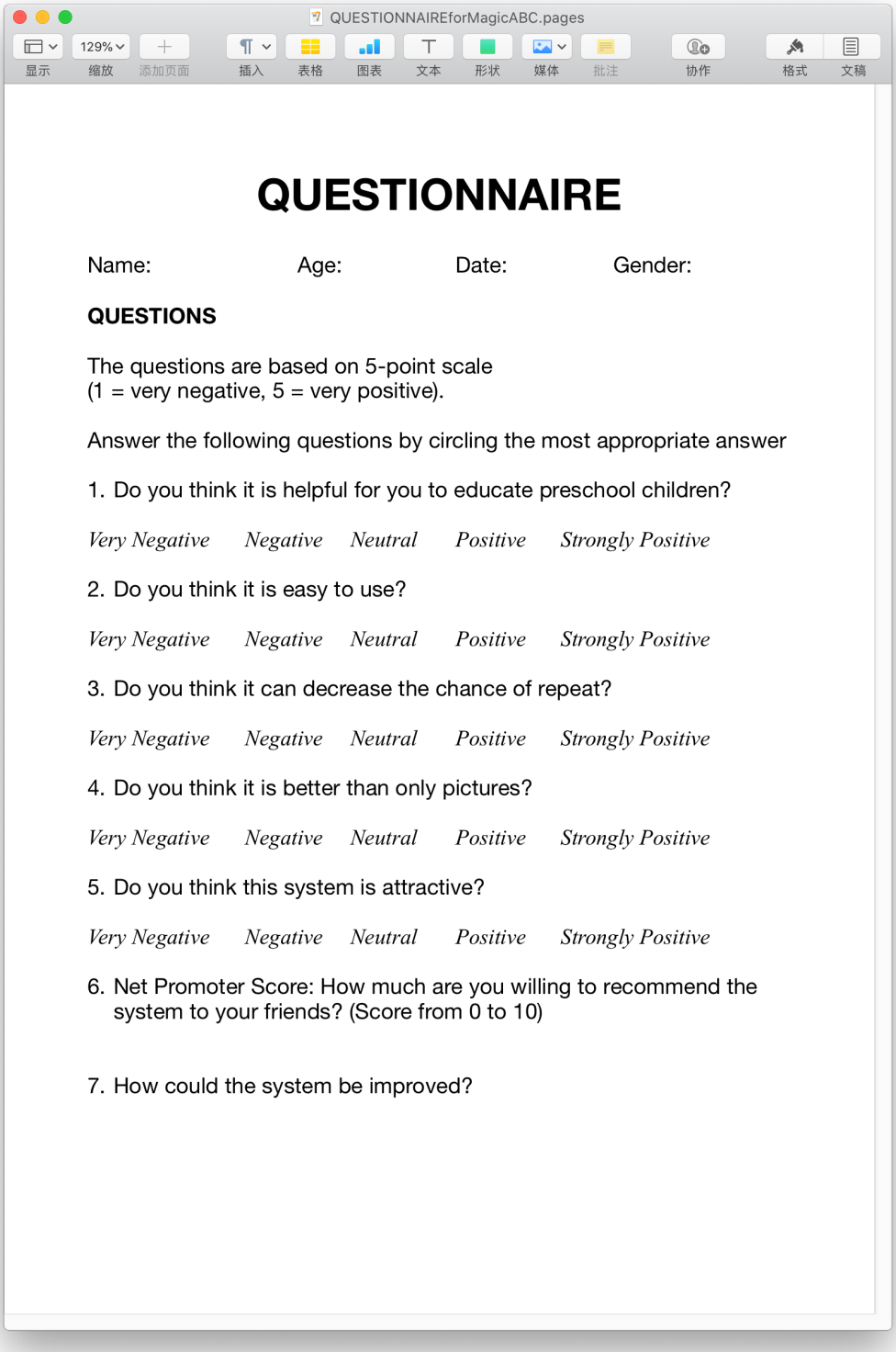


Fig. 7.2 Questionnaire sample for MagicABC

Question	Score
How much are you willing to recommend the system to your friends?	

Table 7.2 Investigative question for NPS

1. PROMOTERS: "Promoters" answered 9 or 10. They love the this system and service. They are the repeat users, are the enthusiastic evangelist who recommends the system and service to other potential users;

2. PASSIVES: "Passives" gave a score of 7 or 8. They are somewhat satisfied but could easily switch to a competitor's offering if given the opportunity. They probably wouldn't spread any negative word-of-mouth, but are not enthusiastic enough about this system or service to actually promote them;

3. DETRACTORS: "Detractors" gave a score lower or equal to 6. They are not particularly thrilled by the system or the service. They, with all likelihood, won't use this system again, could potentially damage the system's reputation through negative word of mouth.

The final score is the ratio between the difference of PROMOTERS and DETRACTORS and the total number of the participants. The logic of the NPS formula is that the PROMOTERS will continue to use and recommend to others to accelerate your growth, while the DETRACTORS may damage your reputation and stop the growth using negative word-of-mouth. A score of more than 50% is considered good. If the NPS score is between 70-80%, your company has a group of loyal customers. According to the survey, most companies still have a NPS value between 5% and 10%.

7.3 Result

In our assessment, all participants gave positive feedback. In addition, we also put forward some useful suggestions for the improvement of the system. Let me first show you the behavior that we observe when users use the system.

We collected the results given by the participants, as shown in Table 7.3. In Table 7.3, U1 represents user 1 while Q1 represents question 1 in the Fig. 7.2. Other abbreviations are similar to this explanation.

Result	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	AVE
Q1	5	4	5	5	5	5	4	5	3	5	4.6
Q2	4	4	5	3	5	4	4	5	5	5	4.4
Q3	5	5	5	5	5	5	5	5	5	4	4.9
Q4	5	3	5	5	4	5	3	5	5	5	4.5
Q5	5	4	5	4	5	5	4	5	4	5	4.6
Q7	10	9	10	9	10	8	10	8	9	8	9.1

Table 7.3 Questionnaire result for MagicABC

We collected the results given by the participants. The survey results are shown in the Table 7.3. As Table 7.3 shows, all users used the function. By visualizing the table content data, we have made Fig. 7.3. These table and figure show the scores and average scores of ten users. From the questionnaire survey, users' evaluation of the system is very high.

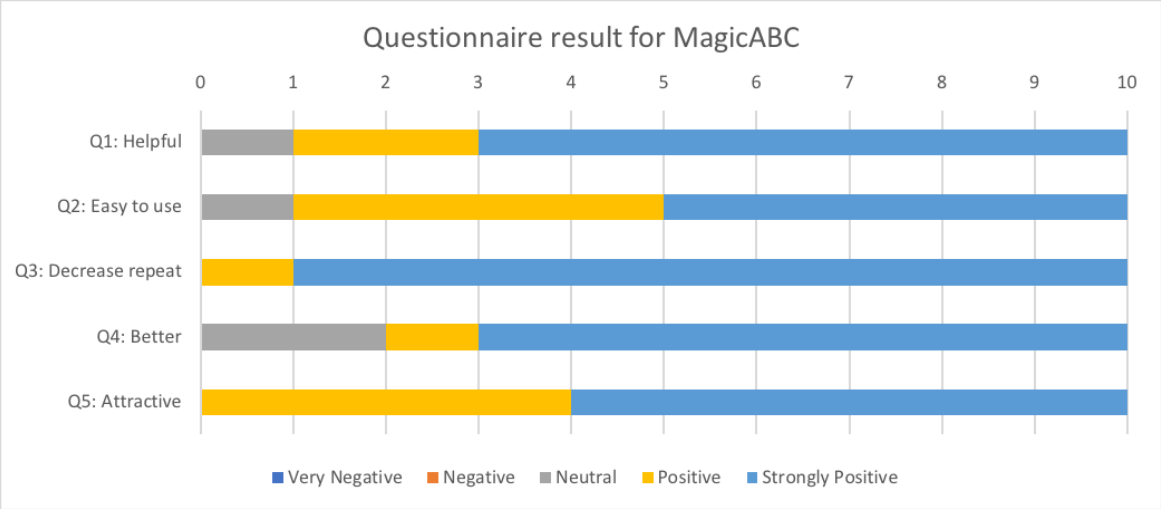


Fig. 7.3 Questionnaire result for MagicABC

For investigative question for NPS(Q7) in Table 7.3, four users gave 10 points, 3 users gave 9 points and 3 users gave 8 points. The average score was 9.1. It shows all users gave us a positive feedback. But this is not enough. The average score of 9.1 shows that our system still has a lot of room for improvement.

Fortunately, according to the previous classification, all users in this experiment are recommenders and promoters, which means that our system can improve users' satisfaction in early childhood education. In the experiment, we get some good suggestions:

1. I hope I can use the puppet not stored in the system. I hope you can provide it as soon as possible.

2. In smart home, TV and iPad are very important. I hope to enhance the adaptability of different devices in your UI design. The use of these large screen devices can be very good hands-free to interact with children.

3. I hope I can also use other languages and cards of this system, and hope to provide OCR function for children to learn by themselves. He can use the system to scan any word in the environment.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

Based on our preliminary assessment results, we believe that our system can better solve the problem of early childhood education in learning a second foreign language.

Generally speaking, we have designed a framework to build a second language learning system with parent-child cooperation. The system uses Augmented Reality technology and a pronunciation assessment system to detect second foreign language learning, and uses puppet's AR expression feedback as an incentive to complete parent-child cooperative language learning. We have converted traditional cards into a more intuitive and easier to understand digital system. In addition, we have designed a unique feedback system to display children's pronunciation.

Technically, we have designed an algorithm to detect children's pronunciation performance. The algorithm compares the children's pronunciation with the standard pronunciation through analysis to obtain accuracy score. In addition, we have also implemented voice feedback, AR expression feedback, AR content display and other functions.

For parents who are not good at second language, our system provides learning tools for parent-child cooperation. The tool can use advanced natural language processing to correct and evaluate children's second foreign language pronunciation. For preschool children, our

system can realize learning motivation through AR technology, thus ensuring long-term learning interest

In order to test the usability and efficiency of the proposed system, we included some participants in the evaluation. Feedback is positive.

8.2 Future Work

Although we have proposed a prototype of the digital alphabet flash card early education system, there are still some limitations and future possibilities to improve its efficiency.

In this system, the system can only identify alphabet flash cards and puppet models stored in the database. In the future system, we hope it can recognize any text and puppet model. And accurately project the expression on the face of the new puppet.

In addition, we are considering the possibility of involving machine learning capabilities in speech evaluation algorithms. Because the pronunciation habits of babies of different nationalities are different, if only one standard American pronunciation is adopted, it is easy to cause frustration in children's pronunciation. By improving the scoring system through machine learning methods, we can not only provide more appropriate incentives to encourage children to learn well.

References

[1] Dimitris Chatzopoulos, Carlos Bermejo, Zhanpeng Huang, and Pan Hui. Mobile augmented reality survey: From where we are to where we go. *Ieee Access*, 5:6917–6950, 2017.

[2] Syed Mohsin Abbas, Syed Hassan, and Jongwon Yun. Augmented reality based teaching pendant for industrial robot. In *2012 12th International Conference on Control, Automation and Systems*, pages 2210–2213. IEEE, 2012.

[3] DWF Van Krevelen and Ronald Poelman. A survey of augmented reality technologies, applications and limitations. *International journal of virtual reality*, 9(2):1–20, 2010.

[4] Ronald T Azuma. A survey of augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4):355–385, 1997.

[5] Anuroop Katiyar, Karan Kalra, and Chetan Garg. Marker based augmented reality. *Advances in Computer Science and Information Technology (ACSIT)*, 2(5):441–445, 2015.

[6] Paul Milgram and Fumio Kishino. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12):1321–1329, 1994.

[7] Ronald Azuma, Yohan Baillot, Reinhold Behringer, Steven Feiner, Simon Julier, and Blair MacIntyre. Recent advances in augmented reality. *IEEE computer graphics and applications*, 21(6):34–47, 2001.

[8] The most spoken languages worldwide. url="https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/".

[9] Jim Trelease. *The read-aloud handbook*. Penguin, 2013.

[10] Agnieszka Uberman. The use of games for vocabulary presentation and revision. In *English Teaching Forum*, volume 36, pages 20–27, 1998.

[11] Diego Giuliani, Ornella Mich, and Marianna Nardon. A study on the use of a voice interactive system for teaching english to italian children. In *Proceedings 3rd IEEE International Conference on Advanced Technologies*, pages 376–377. IEEE, 2003.

[12] Eugene Geist. Using tablet computers with toddlers and young preschoolers. *YC Young children*, 69(1):58, 2014.

[13] Common Sense Media. *Zero to Eight: Children's Media Use in America, 2013*. 2013.

[14] Heather L Kirkorian and Tiffany A Pempek. Toddlers and touch screens: Potential for early learning?. *Zero to Three*, 33(4):32–37, 2013.

[15] Valeriy Shafiro, Erika S Levy, Reem Khamis-Dakwar, and Anatoliy Kharkhurin. Perceptual confusions of american-english vowels and consonants by native arabic bilinguals. *Language and speech*, 56(2):145–161, 2013.

[16] Cynthia G Clopper and Ann R Bradlow. Perception of dialect variation in noise: Intelligibility and classification. *Language and speech*, 51(3):175–198, 2008.

[17] Matthew Clarke. Beyond antagonism? the discursive construction of 'new'teachers in the united arab emirates. *Teaching Education*, 17(3):225–237, 2006.

[18] Richard Rupp. Higher education in the middle east: Opportunities and challenges for us universities and middle east partners. *Global Media Journal*, 8(14), 2009.

[19] Silke Maren Witt. Use of speech recognition in computer-assisted language learning. 1999.

[20] Rodolfo Delmonte. Exploring speech technologies for language learning. *Speech and Language Technologies*, page 71, 2011.

[21] Jared Bernstein, Alistair Van Moere, and Jian Cheng. Validating automated speaking tests. *Language Testing*, 27(3):355–377, 2010.

[22] Olov Engwall. Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. *Computer Assisted Language Learning*, 25(1):37–64, 2012.

[23] Frantisek Grézl and Hynek Hermansky. Local averaging and differentiating of spectral plane for trap-based asr. In *Eighth European Conference on Speech Communication and Technology*, 2003.

[24] Michel Galley and Christopher D Manning. A simple and effective hierarchical phrase reordering model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 848–856, 2008.

[25] Hongyan Li, Shijin Wang, Jiaen Liang, Shen Huang, and Bo Xu. High performance automatic mispronunciation detection method based on neural network and trap features. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.