Design and Development of College Tourism English Training System Based on Speech Recognition Technology

Wenyu Xu(✉)

Sichuan Vocational and Technical College, Suining, Sichuan, China
182329702@qq.com

Abstract. Based on language recognition technology, the author uses Web Audio API interface technology and Microsoft voice service to complete the construction of college tourism English training system in ASP.NET CORE environment. The practical training function of the system can simulate various sections in tourism activities and provide oral dialogue training in different situations, so as to facilitate the students majoring in tourism English to complete the corresponding knowledge and skills learning and listening, speaking and reading dialogue exercises. In addition, under the system evaluation function, voice recognition technology is used to convert the voice signals of students' users into digital information, which is evaluated layer by layer from sentences, words and phonemes, and students' English pronunciation is accurately analyzed and diagnosed from multiple dimensions such as pronunciation, intonation and fluency, and corresponding scores and guidance are given to facilitate students' improvement. The construction of practical training system can solve the problem that students majoring in tourism English in colleges and universities can't directly participate in teaching practice activities under the current epidemic situation. With the help of online practical training, it can break through the limitations of space and time, ensure that students can get more practical opportunities, and enhance their professional ability.

Keywords: Speech Recognition Technology · Tourism English · Practical Training System · Multi-Dimensional Evaluation

1 Introduction

With the deepening of reform and opening up, the social economy has achieved unprecedented prosperity. China's tourism industry has achieved a leap-forward development from scratch, from weak to strong, and has grown into a strategic pillar industry of the national economy and a modern service industry that satisfies the people. By 2019, the number of inbound and outbound tourists has exceeded 300 million, and such a huge market scale confirms the booming development prospects of the tourism industry. Under this background, colleges and universities at all levels in China set up tourism English major, aiming at cultivating high-quality and compound tourism talents to adapt to the development of foreign-related tourism market. Tourism English majors not only need

C. F. Peng et al. (Eds.): EIMT 2022, AHSSEH 3, pp. 66–75, 2023.
https://doi.org/10.2991/978-94-6463-012-1_9

to master the basic knowledge and theories in the field of international tourism business, but also have certain operational skills and business handling ability in tourism business. They can also use English as their working language to complete oral communication in different communication situations in tour guide practice [7]. In the daily teaching plan of tourism English education, both theoretical knowledge teaching and professional training activities should be taken into account. The professional training activities mostly depend on real scenes such as travel agencies, tourist attractions and hotels, and situational conversation training and task-based practical exercises are the main means to improve the comprehensive application ability of tourism English majors. However, under the current global epidemic situation, inbound and outbound tourism has been hit hard, and the whole tourism industry is facing a major crisis, so that the education and teaching of tourism English in colleges and universities are also facing a severe test, and new measures are urgently needed to realize the reform of professional teaching mode.

The great influence of the epidemic has spread to all industries and fields of the whole society, and the education and teaching activities of tourism English major in colleges and universities have also undergone direct changes. As colleges and universities are crowded with people, network information technology has been introduced into students' daily classroom teaching mode to realize online teaching. Although online teaching mode can't meet the standard of classroom teaching in teaching quality, it can basically ensure the normal learning of theoretical knowledge. As for the practical training courses for tourism English majors, the epidemic almost killed all practical opportunities for tourism English majors. As a result, students can't directly participate in teaching and training activities, and the theory can't be translated into practical operation. Students' professional practice, post practice and future career development will be greatly affected [6]. In view of this, the practical training course of tourism English major can refer to the online teaching mode of theoretical knowledge and integrate the innovation of network information technology into the practical training course, so as to break through the limitations of space and time and ensure that students can get more practical opportunities. Therefore, this paper holds that, based on the language recognition technology, under the network environment, the construction of college tourism English training system is completed by using Web Audio API interface technology and Microsoft voice service, with emphasis on situational dialogue training in all aspects of simulated tourism activities. With the specialized and labeled corpus as the judging standard, the multi-dimensional evaluation of students' spoken pronunciation, intonation and fluency can be completed, so as to strengthen students' professional ability and achieve the teaching goal of practical training courses.

2 Related Technical Introduction

2.1 Speech Recognition Technology

Speech recognition technology refers to the understanding and recognition of speech semantics, also known as Automatic Speech Recognition (ASR). As the most natural way of human-computer information interaction, the basic idea of speech recognition technology is to take speech as the research object, recognize and understand the input

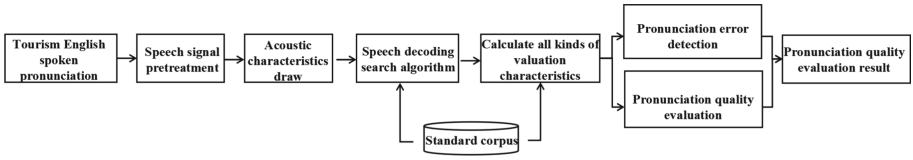

Fig. 1. Structure diagram of overall evaluation model of speech recognition technology (Photo credit: Original)

speech signal by the machine, compare the pronunciation quality model according to the pronunciation and vocabulary in the expert knowledge base, and complete the functions of scoring and error correction. The general workflow of speech recognition technology is: speech signal preprocessing, feature value extraction, acoustic and language model comparison (recognition), comprehensive evaluation of pronunciation quality, and recognition result output. As shown in Fig. 1, the language signal preprocessing module can digitally convert the spoken English pronunciation input by student users, and convert voice files into digital signals. After preprocessing, various noises and other interference factors are eliminated, and then feature parameters are extracted. Then, through the search and decoding module, phoneme segmentation and recognition are carried out on students' spoken pronunciation, so as to calculate various evaluation features needed for pronunciation quality evaluation. After error detection and evaluation, the evaluation results are output and fed back to students' users. There are many methods and technologies in speech recognition technology, including speech signal preprocessing technology, Mel-frequency cepstral coefficients acoustic features, hidden Markov model matching and decoding algorithms of various evaluation features, etc.

2.1.1 Speech Signal Preprocessing Technology

Signal preprocessing technology is an important premise and foundation of speech recognition technology, and it is also the first stage of the overall system construction. The result of speech signal preprocessing has a direct impact on the subsequent feature extraction. Speech processing technology includes five links: signal digitization, endpoint detection, framing, windowing and pre-emphasis [8]. Among them, signal digitization is to convert the analog signal of human sound wave form into digital signal that can be recognized and processed by computer, complete the sampling of human voice through microphone equipment, and form discrete digital signals after systematic quantization processing. Endpoint detection is to confirm the start and end points of words and words in the input speech signal, so as to eliminate useless signals and interference signals in the speech signal. At the same time, it can also improve the working efficiency of subsequent speech recognition. Framing is to divide the speech signal in a very short time range to obtain the steady-state signal. The spectrum characteristics of steady-state signals can be regarded as fixed, and the speech evaluation function can be realized by comparing with the segmented characteristic parameters of standard speech. Windowing is to eliminate the inter-frame distance caused by framing operation, so as to highlight the feature change of speech signal. Pre-emphasis is to pass the speech signal through a filter to improve the high-frequency signal of the speech, eliminate the low-frequency

signal of the speech, make the spectrum of the speech signal flatter and facilitate the extraction of the subsequent frequency features [5].

2.1.2 Mel-Frequency Cepstral Coefficients Acoustic Features

Mel-frequency cepstral coefficients (MFCC) is a way to describe the characteristics of speech signals. MFCC feature parameters make use of the auditory principle and the decorrelation characteristics of cepstrum, and are the most widely used speech features in ASR because of their relatively low computational complexity and low difficulty in realizing functions [4]. The MFCC feature parameter uses Mel frequency, which is obtained by converting the linear frequency of speech signal. As shown in Formula 1, the conversion relationship between linear frequency and Mel frequency is as follows:

$$\mathrm{m} = 2595\lg(1 + \frac{f}{700}) \tag{1}$$

Compared with other linear prediction coefficients, MFCC feature parameters have better anti-noise and robustness, and the speech recognition accuracy of MFCC is relatively high in practical application. The feature extraction of voice signals by MFCC feature parameters can reduce the storage capacity of voice signals in the system, reduce the time spent by the system in identifying voice signals, and improve the work efficiency of system identification and evaluation.

2.1.3 Hidden Markov Model Matching

Hidden Markov Model (HMM) is a statistical model in which the acoustic input of speech signal is associated with the basic unit of speech. In ASR system, hidden Markov model can model the single phoneme and its sequence in speech signal, form observable observation sequence and hidden state sequence, find the optimal hidden state sequence through observable state set and characteristic parameters and use it in the system, and become a statistical model for processing speech signal [2].

In order to improve the fitting distribution of state number and observation sequence probability in the parameter structure of HMM model, GMM is usually used to simulate the probability distribution of acoustic feature vectors extracted from speech frames in each state of HMM phoneme, so as to determine the matching degree between each state of HMM and acoustic feature vectors of speech frames. Therefore, in the current mainstream ASR system, a Gaussian mixture-hidden Markov model, abbreviated as GMM-HMM model, has been formed. Its principle is shown in Fig. 2. The overall training method of GMM-HMM statistical framework is relatively simple, and it is changeable and adaptable, which can improve the application convenience of ASR system.

2.2 ASP.NET Core

ASP.NET Core is a brand-new, open source, cross-platform framework created by Microsoft, which is used to build the Web framework of Web applications, APIs and microservices. ASP.NET Core framework realizes the functional integration of MVC

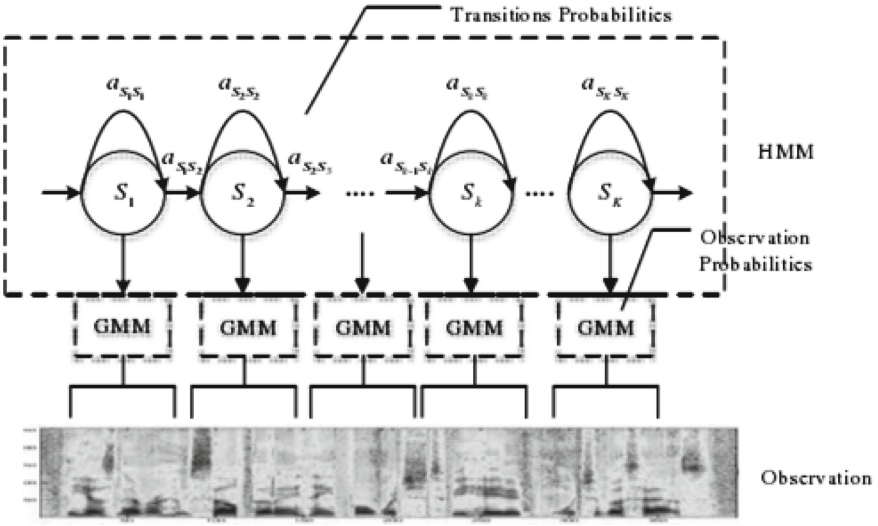

Fig. 2. GMM-HMM acoustic model (Photo credit: https://blog.csdn.net/nsh119/article/details/79496409)

framework and WebApi framework in the original ASP.NET, uses a large number of request processing pipelines composed of dependency injection and Middware, and supports various scripting languages to complete the writing of ASP.NET Core programs, such as C#, Visual Basic and F#. Among them, C# language is the preferred language, which can realize the agile development of ASP.NET Core. The designed ASP.NET Core program can run on .NET Core and the complete .NET Framework.

Compared with other Web development frameworks, ASP.NET Core framework has obvious advantages. First of all, the ASP.NET Core framework has a high execution speed, because it is compiled and run. For complex JavaScript control objects in front-end pages, ASP.NET Core can also make special optimization for multi-county and asynchronous tasks, so as to greatly improve the execution speed. Second, the development of ASP.NET Core application depends on the running environment of .NET Core, in which the running environment can be built after the installation of .NET Core SDK 2.1.500 and .NET Core Runtime 2.1.6. ASP.NET Core relies on the .NET runtime library to realize cross-platform operation on Windows, Mac or Linux systems. Finally, ASP.NET Core application can be hosted on IIS, Nginx, Apache, Docker and other Web servers, which improves the system adaptability and compatibility of the application.

2.3 Development Environment

Complete the configuration and deployment of the development environment according to the system development requirements and the use requirements of the above key technologies. The whole system development is based on Windows10.0-64-bit operating system, and the .NET Framework 4.7 development framework and Visual Studio 2019 are used to provide an integrated development environment for C# language development applications. Select SQL Server 2019 as the database platform, and download

```
using System;
using System.IO;
using System.Collections.Generic;
using System.Linq;
using System.Threading.Tasks;
using Microsoft.AspNetCore.Builder;
using Microsoft.AspNetCore.Hosting;
using Microsoft.AspNetCore.Http;
using Microsoft.Extensions.DependencyInjection;
using Microsoft.Extensions.Configuration;
{
    public class Startup
    {
        public Startup()
        {
            var builder = new ConfigurationBuilder()
                .SetBasePath(Directory.GetCurrentDirectory())
                .AddJsonFile("AppSettings.json");
            Configuration = builder.Build();
        }
        public IConfiguration Configuration { get; set; }
https://go.microsoft.com/fwlink/?LinkID=398940
        public void ConfigureServices(IServiceCollection services) { services.AddMvc(); }
        {
        }
```

Fig. 3. Key code for ASP.NET Core to start MVC mode (Photo credit: Original)

SQL Server Management Studio at the same time to complete the configuration and management of the database. After the installation and configuration of the development environment is completed, you can choose to create a new ASP.NET Core Web Application under Visual Studio 2019, and select the Web application template to complete all the settings, in which the version of ASP.NET Core is 2.1. The AspNetCore.Mvc assembly is built in the framework of ASP.NET.Core In order to make the Web server use the MVC mode, it is necessary to add the AddMvc service to the ConfigureServices method in the Startup class. The key code is shown in Fig. 3. In the Startup.cs file, add app.UseMvcWithDefaultRoute () after the app.UseFileServer () statement in the configure () method.

After the design and development on the server side of the system is completed, it will be published on the Visual Studio side. The system supports publishing the ASP.NET Core program to the specified directory in the form of files. After the Publish command is executed, the program is successfully published, and the program files can be found in the established directory, packaged and uploaded to the IIS server to complete the deployment, so as to support the system users to log in and use the system through the Web browser. Through the brief introduction of the above key technology theories, we have determined the overall environment of the system development, the configuration of related software and tools, and the technical feasibility of the overall project of the college tourism English training system.

3 Related Technical Introduction

3.1 Functional Requirements Analysis

The functional modules of tourism English training system in colleges and universities will be divided and designed according to the different roles of students and teachers. In view of the current difficulties in the teaching of tourism English in colleges and universities, with the help of the great advantages of network information technology and speech recognition technology, we have created a virtual training environment for tourism English. Through the vivid presentation of pictures and texts, the system can stimulate learners' brains, eyes, ears, mouth and hands more effectively, so that students can acquire more information, thus promoting the study and consolidation of tourism English and achieving the purpose of practical training for tourism English majors [1]. In addition, using corpus-based speech evaluation technology can accurately analyze and diagnose students' English pronunciation quality from multiple dimensions of language, intonation and fluency, and give corresponding scores and guidance, so as to better help students improve the standard degree of spoken pronunciation, enhance their oral expression ability and improve the efficiency of students' autonomous learning.

On the student side, the function of the system is mainly divided into two parts. One is the situational training of daily tourism oral English, that is, students can choose different roles to play in different situations independently, and complete the oral English training content through preview-preparation-follow-up. Second, after the students finish the training, they can check the analysis and evaluation results of their pronunciation quality, and make timely adjustments and key improvements according to the corresponding results. For teacher side, the function under the system pays more attention to providing auxiliary guidance and help to students. At the same time, teachers will also take into account the production, uploading and maintenance of tourism English training content. Teachers can check students' training results, training duration and average scores, filter according to conditions, and generate statistical analysis charts, which is more intuitive.

3.2 Global Design

In view of the functional requirements of college tourism English training system, combined with the development and configuration of related technologies mentioned above, we have completed the overall design of the system. The system is designed with B/S architecture, with Web pages as the main presentation form. Users can log in and use the system only through the client browser. After the system is started, the device microphone service will be automatically started to complete the collection of user voice information. In order to reduce the processing burden of voice information flow on the front end of the system and improve the running efficiency of the system, the Web end of the system needs to use the Web Audio API interface of HTML5 to receive the audio stream input from the microphone, and realize the interactive operation of the audio stream between the Web front end and the web server end through the SignalR. On the Web server side, voice recognition, analysis and evaluation are completed by calling Microsoft voice service. The basic structure is shown in Fig. 4. The microphone audio is collected and uploaded in the browser. Therefore, it is necessary to use the Web Audio

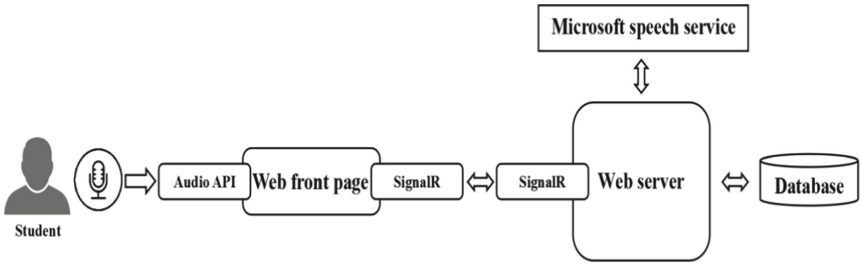

Fig. 4. Basic structure diagram of college tourism English training system (Photo credit: Original)

API under HTML5 standard to obtain the audio stream and use the Audio Context to process the audio stream in real time. Among them, as a real-time communication technology based on Web socket, SignalR can push information between eb client and Web server for a long time [3]. For Microsoft voice service, as the core part of the system, the encapsulation of voice recognition technology has been completed, which can be fully compatible with ASP.NET Core technology stack. Through Nuget Manager under API project, two data packets of voice recognition service Microsoft. CognitiveServices. Speech and Microsoft. AspNetCore. SignalR. Protocols. MessagePack related to SignalR are added, and the development and deployment of voice recognition service can be completed after relevant SignalR service, interface and audio stream configuration are started.

4 Functional Implementation

4.1 Student Side

4.1.1 Travel Daily Oral English Situational Training

In this functional module, student users can choose different sections for oral training independently, and the design inspiration of the sections comes from the common scenes in every link such as eating, living, traveling, traveling, shopping and entertainment in the process of tourist reception service. Such as Train and Taxi, Hotel Services, Taking Orders, Exhibition Travel, Handling Emergencies, Complaint Settlement. After entering any section, students can get the relevant knowledge and content of this section to preview, and get familiar with the situation in which they play their roles to complete the preparation for situational dialogue training. After the training begins, students will follow the dialogue sentences prompted by the system to the end. After the follow-up, students can complete the follow-up exercises, and the exercises system will no longer provide prompt dialogue, which will be done by students freely. After the students finish the practical training, the related voice data will be saved in the database to support the subsequent playback of listening to their spoken pronunciation and practicing the voice content repeatedly.

4.1.2 Speech Analysis and Evaluation

In this functional module, students can view the analysis and evaluation results of spoken tourism English pronunciation in situational training. The analysis mainly evaluates pronunciation errors and pronunciation quality. For pronunciation error detection, Microsoft speech recognition service adopts phoneme error detection algorithm, that is, by segmenting the spoken pronunciation digital signals of students' users into phonemes and comparing them with phonemes in the standard pronunciation corpus, and relying on GOP algorithm to complete factor pronunciation error detection, and finally inputting the detection results and displaying the words or sentences with pronunciation errors.

The evaluation of pronunciation quality is also based on the running process of Microsoft speech recognition service, which is completed one by one according to the steps of speech information preprocessing, acoustic feature extraction, and phoneme segment recognition. In the recognition stage, different algorithms are selected to evaluate the features of the standard degree, intonation and fluency of spoken pronunciation and complete the scoring. Among them, the spoken pronunciation standard degree algorithm is consistent with the phoneme error detection algorithm, which is completed by GOP algorithm. For the evaluation of pronunciation Pitch, the system mainly measures pitch feature of speech signal. Through feature extraction of pitch features of students' spoken pronunciation and pitch features of corresponding pronunciation in standard corpus, phoneme segments are segmented and aligned, the similarity between them is calculated by DTW algorithm, and the intonation evaluation score is determined according to the similarity mapping relationship. For the evaluation of spoken English pronunciation fluency, the analysis and evaluation pay attention to the change of pronunciation duration of students' travel English sentences and vocabulary to determine the fluency of pronunciation. That is to say, according to the comparison between the length of students' pronunciation segment and the length of pronunciation segment in the standard corpus, it is used as the estimated value of the standard pronunciation segment of the word to determine the fluency of students' spoken pronunciation [9].

4.2 Teacher Side

On the teacher's side, teachers can make, upload and maintain learning resources and contents such as practical training situational dialogues and after-school exercises. And in the score viewing module, the scores of students' online training and comprehensive evaluation scores are viewed. At the same time, teachers can also screen scores according to different key fields, such as training completion, training duration, stage average scores and so on. Teachers can obtain students' practical training results and grasp students' practical training situation in time, and give targeted guidance and help, so as to improve students' practical training ability of tourism English majors and meet their learning needs.

5 Conclusions

The construction of college tourism English practical training system can effectively solve the problem of developing practical education courses for tourism English majors

under the current epidemic situation. Relying on network information technology and speech recognition technology, the system completes the virtual and digital construction of practical content by online practical training, helps the current tourism English majors to effectively improve their mastery of tourism English knowledge and skills, and gradually standardizes their spoken pronunciation, enhances their oral expression ability, improves the efficiency of students' autonomous learning and enhances their comprehensive application ability of tourism English. Moreover, it has achieved the optimization of the teaching effect of tourism English, changed the teaching mode of tourism English, achieved the goal of cultivating interdisciplinary talents of tourism English major, and made a new attempt for the informationization reform of higher education.

References

1. Chen, Mali. 2019. Reform and Practice of Tourism English Training in Higher Vocational Colleges. Contemporary Education Research and Teaching Practice.
2. Shi, Hu, Yi Zhang, et al. 2017. Establishment of Acoustic Model in Speech Recognition System Based on HMM Model. Telecom World 04
3. Yan, Li. 2016. Implementation of Real-time Web Function Based on ASP.NET SignaIR. Computer Knowledge and Technology.
4. Xu, Jiyou. 2016. Research and Application System Design of English Pronunciation Quality Evaluation Combining Phonetic Emotion. Guangdong University of Foreign Studies.
5. Yu, Xiaoming. 2019. Development and Application of Speech Recognition Technology. Computer Era.
6. Yuan, Haibo. 2020. Research on the Demand of Tourism English Majors in Higher Vocational Colleges and the Training Mode of Multi-party Cooperation Talents Based on Epidemic Situation. PR Magazine.
7. Zhang, Qingqing, and Yi Fan. 2019. The Application of "Blue Ink Cloud Class" in the Tourism Oral English Training Course. English Square.
8. Zhao, Hongxia. 2018. Design and Implementation of Oral English Learning System Based on Speech Recognition Technology. Capital University of Economics and Business.
9. Zhu, Hongtao. 2020. Research on the Evaluation Model of English Reading Pronunciation Quality. Guilin University of Electronic Technology.