Development of Automatic Evaluation System for Preschool Children's Speech Level Based on Speech Technology

Fei He[1], Qingling Meng[2]*

[1]*Inner Mongolia University for Nationalities, Tongliao, Inner Mongolia, China*
[2]*Faculty of Education, Hulunbuir College, Inner Mongolia, Hailar, 021008, China.*
739701092@qq.com
**Correspondence: mengql059@nenu.edu.com*

Abstract

Based on speech technology, combined with FFmpeg application, the development and construction of automatic evaluation system for preschool children's speech level is completed in Java Web application development system. In view of the shortcomings of the current standardized tools for evaluating preschool children's speech level, such as poor systematicness and large errors, the system will complete the real-time transcription of preschool children's speech and audio content by the speech transcription engine under the framework of deep full sequence convolutional neural network, and then complete the evaluation of children's language ability through quantitative indicators or manual analysis. This helps doctors or children's speech therapists to improve their mastery of preschool children's speech level by analyzing real and natural children's language samples, and combines the evaluation results of this language sample with the evaluation results of standardized tools to form a brand-new and multi-angle comprehensive evaluation system. It is not only helpful to have a more comprehensive understanding of children's language development, but also to find an appropriate treatment plan for preschool children with language disorders and promote their healthy growth.

Keywords: *speech technology; FFmpeg; JavaWeb; automatic evaluation; children's speech*

1 INTRODUCTION

Language is a set of symbolic systems used by people to exchange information. Human beings rely on language to receive education to grow up, exchange ideas and get to know each other with the help of language, and cultural achievements are also transmitted and accumulated through language, which is an important force to promote the civilization development of human society [5]. The biological basis of human language ability is a unique evolution of human brain and an innate instinct. The cognition, acquisition and use of language will accompany people's life, and different stages have different language tasks and requirements, which require different language abilities. A large number of studies have shown that preschool children aged 0-6 have strong language learning sensitivity, and they can naturally acquire their mother tongue in the living environment after birth. However, after this "critical age", their language learning ability will decline rapidly. Therefore, the language level of preschool children determines the height of cognitive development, which has an important impact on the display of their social and individual values.

According to the survey data, preschool children's language comprehension, processing, integration and output level are often lower than those of their peers in the process of language acquisition, such as pronunciation difficulties, inaccurate pronunciation, lack of vocabulary, too simple words, reluctance to speak, etc., that is, children's language barriers appear. According to statistics, 6% ~ 8% of preschool children can't achieve the expected language development goals [7]. The appearance of language barrier not only directly affects the education of preschool children, but also may cause problems of children's emotional management and social behavior. Therefore, children with language barriers need to receive comprehensive language proficiency screening and evaluation, and receive professional individualized language rehabilitation training, so as to scientifically help preschool children with language barriers master

B. Fox et al. (Eds.): IC-ICAIE 2022, AHCS 9, pp. 1396-1403, 2023.
https://doi.org/10.2991/978-94-6463-040-4_208

language learning ability as much as possible and stimulate their communication potential. At present, in China, the screening of preschool children's language barriers and the assessment of their speech ability are mainly based on foreign standardized assessment tools, such as Peabody Picture Vocabulary Test and Goldman-Fristoe Test of Articulation. The standardized language assessment test procedures and scoring standards are unified, and the test results are presented in the form of standard scores, so that the scores of the whole child and each language item can be seen at a glance [8]. However, foreign standardized assessment tools do not consider the characteristics of Chinese and Putonghua sufficiently, focusing only on vocabulary understanding and pronunciation, and there are differences in grammar and syntax, which leads to single assessment items and deviation of assessment results due to differences in grammar and syntax. In view of this, this paper believes that an automatic evaluation system of preschool children's speech level will be developed and constructed based on speech technology, FFmpeg application and JavaWeb application. The system pays attention to preschool children's daily and natural communicative language, and completes the non-standardized evaluation of preschool children's speech level with many functions such as language sample collection, real-time speech transcription, standard judgment and automatic evaluation, and manual auxiliary evaluation. The evaluation results of the system will well reflect preschool children's abilities in language form, language content, speech speed sequence, organization and application, etc., and can be combined with the evaluation results of standardized tools to form a brand-new and multi-angle comprehensive evaluation system. In addition, with the help of the advantages of efficient and convenient application of Web programs, the system supports users in various roles, such as professional doctors, children's speech therapists, scientific research scholars, children's parents and teachers. Promoting the diversity and comprehensiveness of preschool children's language test contents and scenarios, and increasing the diversity of communicative language samples will help to obtain more comprehensive and true information of preschool children's language development assessment, and also help to implement the follow-up educational intervention strategies.

2 KEY TECHNOLOGY INTRODUCTION

2.1 *Speech technology*

Speech technology refers to automatic speech recognition (ASR) and speech synthesis (TTS), which are the key technologies in the computer field. The practical application significance of this technology is to enable computers to acquire the abilities of listening, watching, speaking and feeling, so as to further improve the effect of human-computer interaction. Automatic Speech Recognition (ASR), as the most natural way of man-machine information interaction, its core idea is to use human speech as an electronic signal, which can be recognized and understood by the machine and transformed into the corresponding text or command. Automatic speech recognition technology can also be referred to as Speech to Text (STT) and Text to Speech (TTS) form a corresponding relationship [2].

During the development of speech recognition technology for more than half a century, many solutions have been proposed and iterated, but their working principles are basically similar, as shown in Figure 1. In speech signal processing, acoustic model, language model and decoder are the three core modules that constitute the technical framework of speech recognition. The whole workflow includes speech signal preprocessing, feature value extraction, acoustic and language model comparison (recognition), recognition result output and other steps.

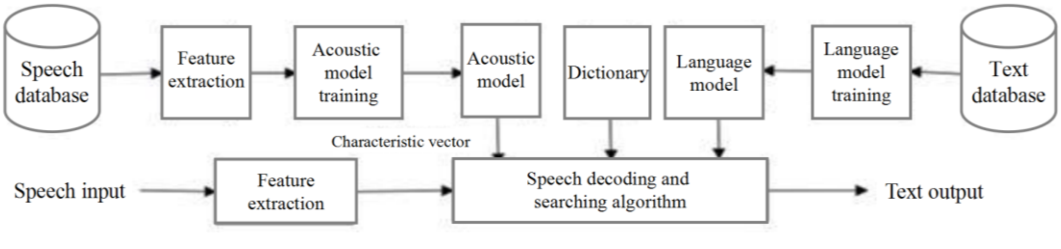


Figure 1: Working principle of speech recognition technology

Among them, speech signal preprocessing is the application foundation of speech recognition technology, and it is also the necessary premise to complete subsequent recognition processing and output results. Speech processing technology includes five steps: signal digitization, endpoint detection, framing, windowing and pre-emphasis [3]. Signal digitization is the use of microphone equipment to convert the analog signal of human voice into discrete digital signal that can be recognized and processed by computer. The purpose of endpoint detection is to distinguish between speech and non-speech areas, accurately locate the beginning and end points of speech, and find the truly effective speech content. After that, the speech signal will be processed by framing, that is, the speech signal will be segmented in frames in a very short time to obtain a relatively stable speech signal. Generally, the time is controlled between 10-30ms, and the continuity between frames is maintained by overlapping segments. Windowing is to improve the periodicity of speech signal after framing by weighting function. The final pre-emphasis is to filter the speech signal for the subsequent feature extraction to increase the convenience of the follow-up work.

At the stage of eigenvalue extraction and acoustic model comparison, the technological transformation and application have promoted the development of speech recognition technology, and the speech recognition rate has been significantly improved with the change of acoustic model. From template matching for small vocabulary in GMM-HMM era, to the application of deep neural network in DNN-HMM era, and then to the rise of end-to-end technology, the industry has released its own new acoustic model structure. However, for the complexity of Chinese speech recognition, the research on acoustic models in China is relatively faster, and the mainstream direction is the integration of deeper and more complex neural network technology with end-to-end technology. For example, in 2018, Iflytek proposed the deep fully convolutional neural network (DFCNN). In the same year, Alibaba proposed LFR-DFSMN model, and Baidu proposed SMLTA model, a streaming multilevel truncated attention model, in 2019 [4]. The proposal and application of a new acoustic model not only optimizes the workflow of speech recognition technology, but also greatly improves the accuracy and speed of speech recognition, thus becoming a research hotspot of speech recognition technology.

For the language model and decoder, the mature technology in the past is still used, and there is no big technical improvement. Language model indicates the probability of a certain word sequence, which is the knowledge representation of a group of word sequences. The function of the decoder is to find the word sequence with the highest probability through a certain search algorithm in the search space composed of knowledge sources such as acoustic models, pronunciation dictionaries and language models. Its core lies in the speed, and the decoding method is mostly static decoding.

2.2 *FFmpeg*

FFmpeg is an open source computer program that can be used to record, convert digital audio and video, and convert them into streams. At present, as the most comprehensive open source encoder that supports cross-platform applications, it not only contains audio and video coding development kits, but also provides developers with rich call interfaces for audio and video processing.

FFmpeg is mainly composed of three parts. The first part includes four basic tools, namely audio-video transcoding and converter (ffmpeg.exe), audio-video player (ffplay.exe), streaming media server (ffserver.exe) and multimedia stream analyzer (ffprobe.exe). The second part is SDK that can be used by developers, and compiled libraries for different platforms. [1] It is convenient for developers to use these libraries to develop their own applications according to actual needs. See Table 1 for information and introduction of each class library. The third part is the source code of the whole project. FFmpeg is written in C language, and the development platform is Linux.

Table 1: FFMPEG contains class library information table

libavcodec	Contains the audio and video encoder and the decoder
libavutil	Contains the tools commonly used in multimedia applications to simplify programming, such as random number generator, data structure, mathematical functions and other functions
libavformat	Contains encapsulation and decapsulation tools in various multimedia container formats
libavfilter	Contains filter functions commonly used in multimedia processing
libavdevice	Used for audio and video data acquisition and rendering
libswscale	Used for image scaling and color space and pixel format conversion functions
libswresample	Used for audio resampling and format conversion

2.3 *JavaWeb*

As a distributed network service based on the Internet, Web can support many users to query and browse information through graphical and easy-to-access interfaces. JavaWeb, on the other hand, uses Java technology to solve the technology stack of related Web and Internet fields. It mainly aims at the server side, and uses Servlet technology, JSP technology and third-party framework to complete the design and development of web applications with dynamic resources.

The core idea embodied in the actual development and application process of Java is hierarchical development, that is, Web applications are hierarchically divided from servers to form a unique three-tier framework, as shown in Figure 2. They are presentation layer (WEB layer), business logic layer and data access layer. The presentation layer includes JSP, Servlet, Struts, SpringMVC related technologies or frameworks. The logic layer includes Spring and EJB Session Bean. The data layer includes MyBatis, Hibernate, EJB Entity Bean framework, which completes the encapsulation of database operation details.

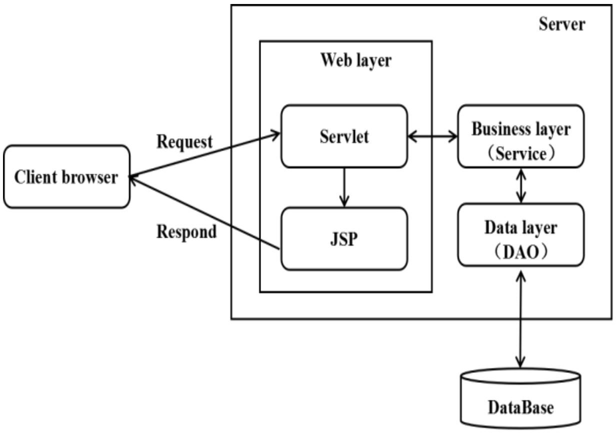


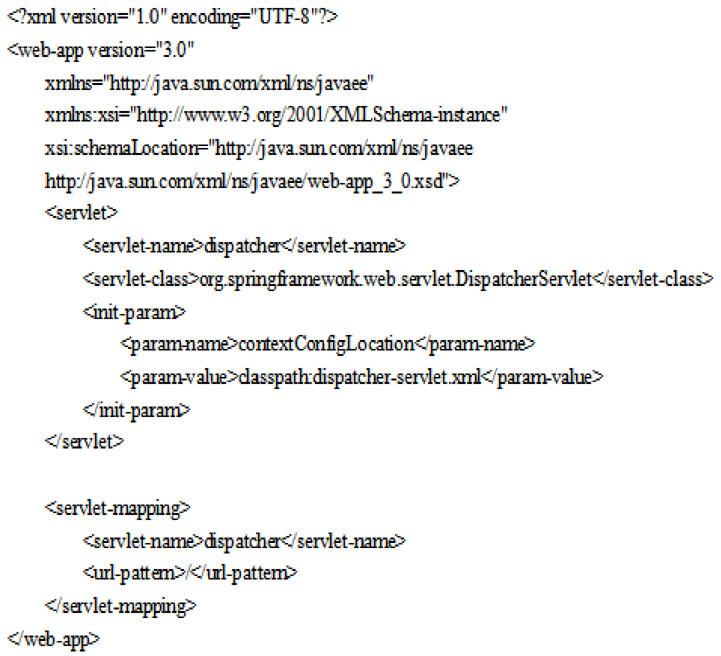
Figure 2: Java Web three-tier framework

2.4 Development environment

According to the system development requirements and the use requirements of the above key technologies, complete the configuration and deployment of the development environment. The overall development of the system is based on Windows Swever Standard operating system. SringMVC 4.1 development framework and MyEclipse2017 CI 3 are used to provide an integrated development environment for Java language development applications. JDK version is above 1.8, MySQL 5.7 is selected as the database platform, and Apache Tomcat 9.0 is selected as the Web server. After completing the installation and configuration of the development environment, you can select New Web Project under MyEclipse, and after completing the settings of J2EE, Java version and server, start to configure SringMVC. Under the option of Sring Facet, select the corresponding version and server to complete the creation, and complete the configuration of SpringMVC under the Web.xml file, including the statement under < servlet > and the configuration of monitoring request of < servlet-mapping >. The key code is shown in Figure 3. Then, through the setting of dispatcher-servlet.xml dispatcher and the setting of com.frank.springmvc.controller control class, the configuration and setting of JavaWeb development environment are basically completed.

As for the speech recognition function, the system will adopt Iflytek's real-time speech transcription service. Under MyEclipse, Iflytek's real-time speech transcription service will be integrated into Web Project by introducing lfasr-sdk-3.0.0.jar package, which is convenient for developers to use programming language to quickly complete the deployment in applications. On the server side, a general HTTP interface is provided to the system through REST API. Based on this interface, users can directly call the speech transcription service of Iflytek Open Platform through the server, and the returned results will be in a unified JSON format.

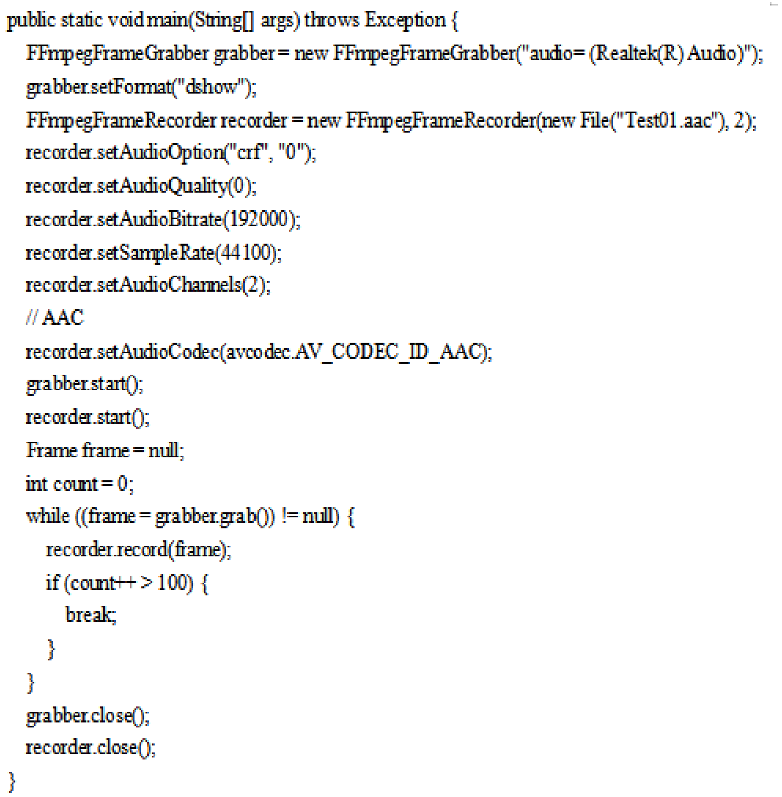
In addition, the JavaWeb server will create an object-oriented interface to realize the definition of system function modules. The FFmpeg.exe tool is introduced under Web Project, and the user's function call through the server is realized through FFmpeg.wasm interface. On the one hand, FFmpeg is introduced into the system to capture preschool children's language audio data by controlling the microphone. The key code is shown in Figure 4. After calling the microphone matrix (Realtek(R) Audio) and using dshow method, it is collected and stored after ACC coding. On the other hand, it is necessary to adjust the attributes and segment the collected audio data to meet the uploading requirements of Xunfei's real-time speech transcription service, such as sampling rate of 16k or 8k, bit length of 8bit or 16bit, format of wav/flac/opus/m4a/mp3, size of 500M and duration of 5-300min [6].



```
<?xml version="1.0" encoding="UTF-8"?>
<web-app version="3.0"
    xmlns="http://java.sun.com/xml/ns/javaee"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://java.sun.com/xml/ns/javaee
    http://java.sun.com/xml/ns/javaee/web-app_3_0.xsd">
    <servlet>
        <servlet-name>dispatcher</servlet-name>
        <servlet-class>org.springframework.web.servlet.DispatcherServlet</servlet-class>
        <init-param>
            <param-name>contextConfigLocation</param-name>
            <param-value>classpath:dispatcher-servlet.xml</param-value>
        </init-param>
    </servlet>

    <servlet-mapping>
        <servlet-name>dispatcher</servlet-name>
        <url-pattern>/</url-pattern>
    </servlet-mapping>
</web-app>
```

Figure 3: The configuration key code for the SpringMVC



```
public static void main(String[] args) throws Exception {
  FFmpegFrameGrabber grabber = new FFmpegFrameGrabber("audio= (Realtek(R) Audio)");
  grabber.setFormat("dshow");
  FFmpegFrameRecorder recorder = new FFmpegFrameRecorder(new File("Test01.aac"), 2);
  recorder.setAudioOption("crf", "0");
  recorder.setAudioQuality(0);
  recorder.setAudioBitrate(192000);
  recorder.setSampleRate(44100);
  recorder.setAudioChannels(2);
  // AAC
  recorder.setAudioCodec(avcodec.AV_CODEC_ID_AAC);
  grabber.start();
  recorder.start();
  Frame frame = null;
  int count = 0;
  while ((frame = grabber.grab()) != null) {
    recorder.record(frame);
    if (count++ > 100) {
      break;
    }
  }
  grabber.close();
  recorder.close();
}
```

Figure 4: Key code of collecting microphone data by FFmpeg

Through the introduction of the above key technical 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 preschool children's speech level automatic evaluation system.

3 REQUIREMENT ANALYSIS

3.1 System requirements analysis

The automatic evaluation system of preschool children's speech level is an online application solution that cooperates with children's doctors, children's speech therapists and children's parents to track, diagnose and scientifically evaluate preschool children's speech level. The system will take preschool children's daily, natural communicative language as a sample to realize the non-standardized evaluation of speech level, aiming at the shortcomings existing in the current evaluation process of preschool children's speech level and combining the needs existing in the current practical application process.

The system will support different users to use various types of terminal devices to log in and use the system. When the system is started, children's speech can be collected directly through the microphone of the terminal equipment, and audio data can also be uploaded through the uploading function of the system. Under the function of speech recognition, the system can call the real-time speech transcription service under the open platform of Iflytek, return the recognition results in the form of text content, and automatically evaluate the preschool children's speech level according to the corresponding standards. In addition, the system also supports users to complete manual evaluation to improve the accuracy of system evaluation. After the evaluation, professional doctors or speech therapists will confirm whether preschool children have language barriers according to the evaluation results, so as to facilitate the subsequent development of educational intervention and treatment programs.

3.2 *Global design*

The automatic evaluation system of preschool children's speech level is designed with B/S architecture, with Web page as the main presentation form. Users can log in and use the system only through any client browser connected to the Internet. The overall system design is divided into presentation layer, business logic layer and data access layer according to SpringMVC framework. The overall framework structure is shown in Figure 5 [9]. Servlet is the controller and JSP is the view in the presentation layer. In the business logic layer, JavaBean is used as the data model, and through two interfaces, the association among all levels and data flow are completed. When the system is started, the device microphone service will be turned on automatically with the help of FFmpeg class library to complete the collection of children's speech information. For Iflytek real-time speech transcription service, as the core part of the system, it completes the encapsulation of speech recognition technology, which can be fully compatible with JavaWeb technology stack in the form of REST API. The system API includes several interfaces, such as preprocessing, file fragment uploading, file merging, query processing progress, and obtaining results, to complete the function call.

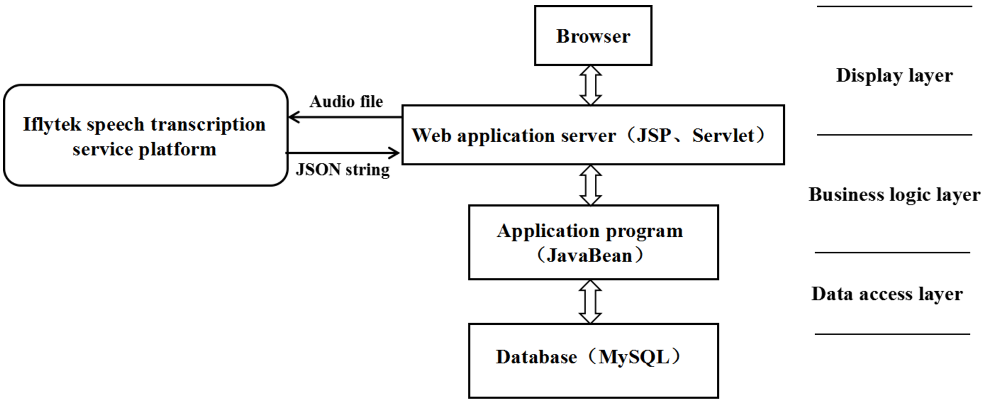

Browser
Display layer
Iflytek speech transcription service platform
Audio file
JSON string
Web application server（JSP、Servlet）
Business logic layer
Application program（JavaBean）
Data access layer
Database（MySQL）

Figure 5: Overall frame structure diagram of the system

4 FUNCTION IMPLEMENTATION

4.1 *Collection of language samples*

When users log in and use this system to evaluate preschool children's speech level, they can collect language samples in two ways. First, direct acquisition, that is, direct acquisition of children's communicative vocabulary or sentences in the current testing environment through microphone equipment. After the user clicks Start Recording, the system will automatically capture and save the audio content. The second system supports the audio source files recorded by other devices to be uploaded and saved in the database.

4.2 *Real-time speech transcriptio*n

There are some differences between the uploading requirements of audio files collected directly or uploaded manually and those of Iflytek real-time speech transcription service. Therefore, before real-time speech transcription, it needs to go through basic processing, including format transcoding, audio cutting and so on. After the user clicks format conversion, the system will automatically create an instance of Ebcoder, and convert the parameters of the audio file to be transcoded, such as duration, sampling rate, bit rate and volume, into the required parameter values to complete format transcoding. The cutting of audio duration will also be done by different class libraries according to different formats of source files. For example, mp3 can be cut by simple FileInputStream class and FileOutputStream class.

After audio file preprocessing, users can click upload and get the real-time speech transcription result back. Iflytek real-time speech transcription service platform starts the processing flow after receiving the system API interface call, and the detailed flow is shown in Figure 6. After the transfer is completed, the server will implement active callback, that is, send the transfer result to the callback address configured by the user.

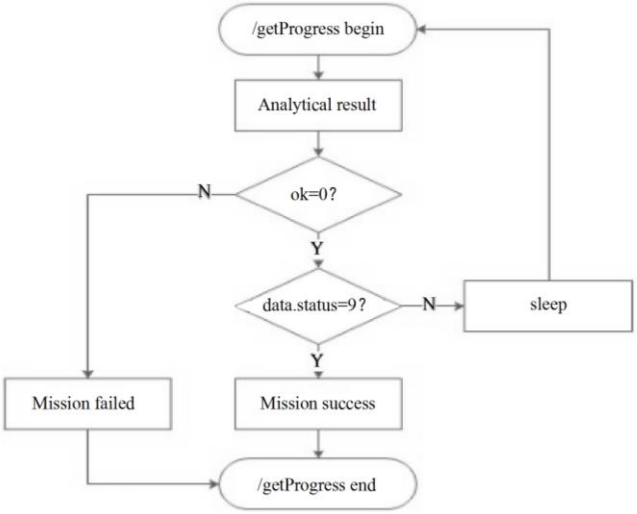


Figure 6: Processing flow of Iflytek real-time speech transcription

The result returned by Iflytek speech transcription service is JSON data. The system will complete the conversion of JSON data through two classes: Sentence class and Word class, form Chinese characters and vocabulary and display them on the page.

4.3 *Standard judgment and automatic evaluatio*n

After obtaining the real-time speech transcription results, the system will automatically evaluate the results of the test children according to the pronunciation accuracy and similarity of preschool children with normal speech ability of all ages. The simulation test results are shown in Table 2. In order to improve the accuracy of the evaluation results, the system will also support manual assistant evaluation, that is, through the professional doctors or speech therapists' understanding of children's speech, the text content can be manually compared. And calculate the effective word count and similarity.

Table 2: Similarity table of preschool children's speech recognition

Type	Age of the moon	Number	Average range	Crest value	Lowest value
Normal children	36-48	5	54.1%-55.8%	55.8%	53.2%
	48-60	10	65.0%-73.1%	76.1%	63.0%
	60-72	15	75.5%-80.3%	88.3%	70.9%
		Total: 30			
Children with language barrier	36-48	8	30.3%-40.7%	40.7%	30.3%
	48-60	7	34.3%-41.5%	43.3%	31.8%
	60-72	5	45.7%-55.4%	58.4%	44.0%
		Total: 20			

5 CONCLUSIONS

The construction of an automatic evaluation system for preschool children's speech level can effectively solve the shortcomings of the current standardized evaluation tools for preschool children's speech level. With the help of network information technology and Iflytek real-time speech transcription service, the collection, analysis and evaluation of real, natural and communicative children's language samples are realized, and the purpose of comprehensively mastering the development level of preschool children's speech level is achieved. After testing, the system can meet the needs of practical application, save a lot of manpower and time costs, and at the same time, it plays a very good supplementary role for standardized evaluation, and promotes the combination of the two to form a brand-new and multi-angle comprehensive evaluation system. It is not only helpful to have a more comprehensive understanding of children's language development, but also to find an appropriate treatment plan for preschool children with language disorders and promote their healthy growth.

ACKNOWLEDGEMENTS

This paper is the research results of two projects, the research results of the "13th Five-Year Plan" project of national education science "Research on the Innovative Development of National Excellent Traditional Culture in the Curriculum Construction of Ethnic Kindergartens in Western China" (Project No.XMA180294), the research results of the discipline construction project of Hulunbuir College of Inner Mongolia University "Research on the Construction of 'New Liberal Arts' in Preschool Education in Colleges and Universities from the Perspective of Transformation and Development" (Project No.2021XKPT035).

REFERENCES

[1] Chen Tianxi, Liu Liming, Chen Kai (2016). Research on Cross-platform Video Codec Based on FFMPEG. Industrial Technology Innovation.04.

[2] Han Yuan (2017). A Study on the Objective Evaluation of Phonetic Recognition Technology for the Initial Pronunciation of Mandarin Children. Nanjing Medical University.05.

[3] Li Xuelin (2018). Overview of Speech Recognition Technology Based on Human-computer Interaction. Electronics World.11.

[4] Ma Han, Tang Roubing (2022). A Review of Speech Recognition Studies. Computer Systems & Applications.01.

[5] Mo Guiming (2020). Research on Automatic Evaluation System of Preschool Children's Speech Function Based on Speech Technology. University of Chinese Academy of Sciences.09.

[6] Qin Fengzhi (2020). Interpreting the Application of Artificial Intelligence Speech Transcription Technology in Conference. Electronics World.11.

[7] Si Boyu (2014). Development of Automatic Evaluation System for Articulation and Speech Disorders Based on Speech Recognition. East China Normal University.03.

[8] Xue Lei, Zhang Chi et al (2019). Application System for Automatically Evaluating Chinese Children's Speech Development Level. Industrial Control Computer.05.

[9] Zhao Lin, Wang Hongxia (2017). Research and Application of Web System Based on Spring MVC+JDBCTemplate. Software Engineering.01.