Doi: 01.0005/j.Art.2023.06.001

Application of Speech Recognition in English Pronunciation Correction

Kaibo Zhang

Abstract

In the context of economic globalization, countries are beginning to conduct ordinary trade transactions, so English, as the most frequently used language, has been incorporated into the education system of each government and students are being taught English systematically, but the fastest way to improve English must be for a student to have a one-on-one lesson with a native English speaker. However, with the world's advanced educational resources, teaching English to students on a one-to-one basis is obviously not practical. Therefore, implanting a speech recognition system into educational software to help the software track or correct students' English learning in real time is possible. However, replacing a real English teacher with the current speech recognition accuracy rate is still a challenge. Therefore, this paper explains in detail how to improve speech recognition accuracy through lip motion tracking and MFCC and to give learners the corresponding vocabulary ratings and suggestions for improvement through this model. In the long run, the knowledge tracking model is used to help students find the most suitable learning style.

Keywords: MFCC, Speech recognition, Lip fusion model

1. Introduction

In modern society, English has always been an internationally recognized language, and every country has incorporated English into its education system. With the continuous update of contemporary technology, the technology is now available to help learners learn English through auto speech recognition technology with educational software. However, most of the software focuses mainly on English grammar and the way of writing. Very few software focus on pronunciation correction of spoken language [1]. For a non-native speaker, due to the accent is easy to mispronounce words if they learn English by themselves. For example, in some Asia countries, people like to pronunciation of "three" and "trees "the same. This happens all over the world. The most straightforward way to avoid this is to have a native-level tutor to correct learners when they mispronounce words. But this is not a realistic thing to do. After all, the world's educational resources cannot support this type of learning. Therefore, the educational software will replace natural educational resources with auto speech recognition technology to provide the most professional help for this group of people with English learning needs. In this software, the voice recognition technology will be used to record the learner's voice when reading words and break it down into four models, which are intonation, rhythm, intonation, and speed issues, and compare these parts with the parts of the first language speaker that are prepared, the closer to the standard parameters of the native reader, the higher the score will be. This paper will focus on how educational software works and how speech recognition and scoring systems intervene [2].

2. Speech recognition model

2.1 System Introduction

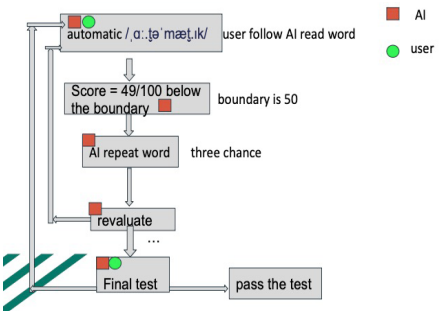

Figure 1. Structure diagram

The learning system collects learners' pronunciation of specified words through speech recognition technology. It then systematically evaluates the pronunciation of the selected words through a scoring system and provides feedback to learners on their scores and the parts that can be improved. Figure 1 shows the flow of the system. First, the system selects a word, reads it with a pre-recorded native-level speech package, and then asks the learner to repeat it [3]. The speech recognition system collects information about the learner's reading during this process. If the score given by the evaluation system is higher than 50, the system will determine that the learner has succeeded in learning and will start learning the next

word. If the score is below 50, the learner will fail, and the system will re-read the comment and ask the learner to repeat the word up to 3 times. If the learner's score is higher than 50 out of three, the system will start the next word, and if the score is lower than 50 out of three, the system will temporarily skip the word and start the next word. This step is designed not to hurt the learner's enthusiasm for learning. If the learner is repeatedly asked to read a word and fails to pass it, the learner may lose confidence. When all the words have been learned, the system will repeat the words that were not passed and let the learner try again, repeating the above steps until all the words have been discovered [4]. The system will test the learner's learning at the end of the learning process. During this process, the speech recognition and scoring systems intervene but do not give feedback to the learner until the learner has completed all the tests and the corresponding word scores and improvement sections are given to the learner [5]. This test is given not only for the words currently learned but also for some of the words learned several times before, referring to the Ebbinghaus 'Forgetting Curve' rule to help learners recall these words more effectively.

2.2 Speech recognition

Before the advent of speech recognition systems, educational software has always existed in the role of supporting human teachers. The main reason for this is that educational software has no way to listen to the user's voice, especially in the area of learning to speak. The most educational software was focused on reciting words. The addition of speech recognition systems has made it possible for educational software to replace human teachers in the future. However, even after the software has speech recognition, the accuracy of the speech recognition is still slow. This is because when humans talk, the high-frequency part is about 800Hz or more and drops by 6db/octave, also the human speech signal is a typical non-smooth signal, and the average power spectrum of the speech signal receives the influence of body organs such as the vocal cords or the mouth when a person speaks. This indicates that the high-frequency part of human pronunciation is more difficult to be captured by the speech recognition system than the low-frequency part, and the influence of environmental sounds around the learner will further lead to a decrease in speech recognition accuracy. After reading Li Chengcheng's English pronunciation similarity study based on speech recognition and Duan Wenting's English conversation robot pronunciation standardization detection method, the above two scholars have achieved certain research results; based on this combined with the above research experience, the accuracy of speech recognition can be further improved by lip angle fusion and by pre-emphasis of speech signals [6].

2.2.1 Lip angle fusion model

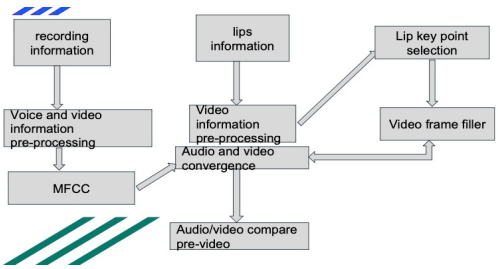

Figure 2. Lip angle fusion model

From Figure 2, It can be seen that the model is divided into three parts: The first part is to collect learners' voice information and video information, and the voice information of the learner when reading the word is recorded through the speech recognition function. The lip information of the learner is recorded through the front camera. The second part is to extract the key points of the lips. Since dynamic lip information is more difficult to extract accurately than static one, the critical issues of the angular features of the lips are mainly extracted here [7]. The third part is to fuse the video key points with the speech and compare the processed audio and video with the pre-recorded standard pronunciation video. After the key point extraction, to ensure the sound frame number and video frame number are consistent, the video frame number needs to be made up to ensure the accuracy of the audio-video fusion stage in the third part of the sound.

2.2.2 how do you use the key points to find the angle at which the lips move?

Figure 3 shows how to use the key points of the lips to capture the movement of the lips. The closer the angle is, the closer the learner's pronunciation is to that of the native speaker.

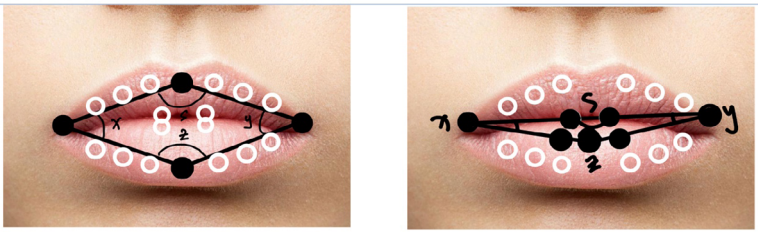

Figure 3. Lip Keys

$$\cos\theta = \frac{\vec{ba} \cdot \vec{bc}}{|\vec{ba}||\vec{bc}|} \quad (1)$$

$$\theta = \arccos\left(\frac{\vec{ba} \cdot \vec{bc}}{|\vec{ba}||\vec{bc}|}\right) \quad (2)$$

X represents a, y represents c, z represents b, and Θ represents the change in lip angle. The value of Θ was calculated using the formula 1 and 2.

2.2.3 How to apply the MFCC in the model

The MFCC (Mel-Frequency Cepstral Coefficients, MFCC) is used to analyze the audio recordings, which can effectively simulate the auditory characteristics of the human ear and has strong noise immunity [8]. The MFCC is extracted by pre-processing the voice information after recording, converting each frame of the voice from time waveform to the frequency domain by FFT, obtaining some frequency features according to the auditory characteristics of the human ear, and then obtaining the MFCC by discrete cosine transform (DCT). After extracting the MFCC, the processed video and MFCC are merged frame to frame and compared with the prepared standard pronunciation video to improve the accuracy of the recognition rate [9]. In the process of imitation, learners will try to simulate the pronunciation habits of native speakers and their pronunciation styles as much as possible.

3. Knowledge tracking Model

3.1 Why use models in the application?

Different students have different learning efficiency, so for students to master as many words as possible in the same learning time, Multiple models can be established to analyze students' learning curves in these mods to determine which words they are more likely to know. For example, when students have successfully learned "too," they can add a word with similar spelling or pronunciation, such as "zoo," to the next word as the learning curve updates("Analytic Comparison of Three Methods to Evaluate Tutorial Behaviors").

3.2 How to apply knowledge tracking to the application?

Knowledge tracking is also used throughout the application to continuously update the student's learning curve as the student learns. Whenever a student successfully reads a word during the first phase of learning, knowledge tracking automatically updates the student's learning curve and uses the current curve to select the next new word that the student may recognize [10]. Suppose the tutor intervenes as the student learns the words. For example, knowledge tracking also updates the student's learning curve when the student's score does not pass the score boundary after reading. In the second stage of learning, students are given a word test, during which any words with a score below the boundary are also recorded in the learning curve.

3.3 How does the model help the student to find the most suitable study style?

In education, there are four mainstream learning styles: visual learners, auditory learners, and kinesthetic learners. Even though the current mainstream argument suggests that student's different learning styles do not improve their learning efficiency or even sometimes what students perceive as efficient learning methods do not help them improve their efficiency, however, with the knowledge tracking in the model, we can compare each learning session to the first time the student starts using the application. The software will use a different learning style for each word [11]. The software will let students learn other words according to learning style, visual learners, auditory learners, and kinesthetic learners. For example, when learning the term smartphone, the tutor can provide students with a picture of a cell phone, when learning the word When learning the word "zoo," the software can also help students understand the word by providing a video. By learning to track the student's correct rate of passing and response time to the word, compare which learning method students know the word most efficiently.

4. Application Analysis

The first use scenario is to replace a real teacher to teach students to learn. Who often have problems identifying words (Giving Help and Praise in a Reading Tutor with Imperfect Listening – Because Automated Speech Recognition Means Never Being Able to Say You're Certain.",1997), there is a risk of misuse or mispronunciation of words during long periods of self-study, which often occurs in non-English speaking countries such as Asia, because the teacher's pronunciation is not standardized, resulting in problems with the student's pronunciation as well. Especially when studying alone, you cannot solve your problems in time. The easiest way is to have a native-level tutor correct any word recognition errors with you. A tutor with ASR will listen as the student reads the words and give feedback and corrective guidance if the student's pronunciation is incorrect [12].

The second usage scenario concerns improving students' learning efficiency and self-confidence. When students follow a human tutor, if they make the same mistakes over and over again, they may selectively give up practicing the word out of impatience, thus undermining the student's confidence. However, tutors who use ASR listen to students all the time. Suppose the student does not read

the word in half a minute. In this case, the application tutor will keep teaching the student repeatedly and encouraging them until the student succeeds in learning the word.

The third usage scenario is to track the student's learning progress through speech recognition. For most students, the learning process is painful and tedious, and most feel dull while studying, leading to decreased learning efficiency. To solve this problem, tutors will use ASR to listen to students continuously read words. When the student exceeds 5 seconds, the tutor will cough gently to remind the student not to wander off and to make the student think that the tutor is always following up on his learning [13].

5. Conclusion

This paper explains the importance of speech recognition for educational software and mentions how speech recognition accuracy can be improved by fusing lip movement video and MFCC-filtered speech information. It also helps students to find the most suitable learning style to enhance the efficiency and accuracy of learning through a knowledge-tracking model. It is believed that in the near future, through continuous optimization of the speech recognition model, educational software will change from assisting educators to replacing educators to give more efficient and professional language learning to each learner.

Reference

[1] Ran D, Yingli W, Haoxin Q. Artificial intelligence speech recognition model for correcting spoken English teaching[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(2): 3513-3524.

[2] Liu X, Xu M, Li M, et al. Improving English pronunciation via automatic speech recognition technology[J]. International Journal of Innovation and Learning, 2019, 25(2): 126-140.

[3] Cao Q, Hao H. Optimization of intelligent English pronunciation training system based on the Android platform[J]. Complexity, 2021, 2021: 1-11.

[4] Evers K, Chen S. Effects of automatic speech recognition software on pronunciation for adults with different learning styles[J]. Journal of Educational Computing Research, 2021, 59(4): 669-685.

[5] Kholis A. Elsa speak app: automatic speech recognition (ASR) for supplementing English pronunciation skills[J]. Pedagogy: Journal of English Language Teaching, 2021, 9(1): 01-14.

[6] Wu L, Wu L. Research on business English translation framework based on speech recognition and wireless communication[J]. Mobile Information Systems, 2021, 2021: 1-11.

[7] Terbeh N, Maraoui M, Zrigui M. Arabic discourse analysis: a naïve algorithm for defective pronunciation correction[J]. Computación y Sistemas, 2019, 23(1): 153-168.

[8] Liakina N, Liakin D. Speech technologies and pronunciation training: What is the potential for efficient corrective feedback?[M]//Second Language Pronunciation. De Gruyter Mouton, 2022: 287-312.

[9] Xu Y. English speech recognition and evaluation of pronunciation quality using deep learning[J]. Mobile Information Systems, 2022, 2022: 1-12.

[10] Liu Y, Quan Q. AI recognition method of pronunciation errors in oral English speech with the help of big data for personalized learning[J]. Journal of Information & Knowledge Management, 2022, 21(Supp02): 2240028.

[11] Li J, Hu M. Design and Implementation of the'Dialect'Learning System based on Speech Recognition[J]. International Journal of Social Science and Education Research, 2022, 5(10): 825-830.

[12] Alrumiah S S, Al-Shargabi A A. Intelligent Quran Recitation Recognition and Verification: Research Trends and Open Issues[J]. Arabian Journal for Science and Engineering, 2022: 1-27.

[13] Fadel W, Bouchentouf T, Buvet P A, et al. Adapting Off-the-Shelf Speech Recognition Systems for Novel Words[J]. Information, 2023, 14(3): 179