university of groningen / campus fryslân

Chinese-speaking English learners' Vowel Pronunciation Error Detection

Yining Lei

university of groningen / campus fryslân

University of Groningen - Campus Fryslan

Chinese-speaking English learners' Vowel Pronunciation Error Detection

Master's Thesis

To fulfill the requirements the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Matt (Voice Technology, University of Groningen)
with the second reader being
Dr. (Voice Technology, University of Groningen)

Yining Lei (S5558506)

June 6, 2024

Acknowledgments

With the completion of this master's thesis, my graduate career is about to come to an end. Looking back on this time, I feel lucky and grateful because so many people have given me selfless help and support. Here, I would like to express my most sincere gratitude to them.

First of all, I would like to thank my mentor, Professor Matt. On the academic road, Professor Matt has set an example for me with his profound academic attainments and rigorous scientific research attitude. From the selection of the topic, conception to writing of the thesis, Professor Matt has given me careful guidance. He not only gave me guidance in theory, but also supported me in practice and provided me with data sets for recording. These experiences not only allowed me to gain rich academic knowledge, but also taught me how to think independently and solve problems. Here, I would like to express my sincere respect and gratitude to Professor Matt.

Secondly, I would like to thank my classmate Brandi. In the process of recording the data set, Brandi gave me great help. Her selfless help made me feel warm.

In addition, I would like to thank my colleague Zou Xinyu. In the process of model construction and optimization, Zou Xinyu gave me great advice and guidance. I have benefited a lot from his rich practical experience and solid professional knowledge. His help has saved me from many detours and made me more determined in my research direction.

At the same time, I would like to thank my family and friends. They have always been my strong backing and given me endless support and encouragement. When I encountered difficulties and setbacks, they always gave me care and help at the first time, which made me feel endless warmth and strength.

Finally, I would like to thank the University of Groningen and all the teachers who have taught me, providing me with a good learning environment and rich academic resources, so that I can continue to grow and progress.

Abstract

English is the main international language in the world today, and its scope of use is very wide. The study of English has always been valued by all countries. In my country, although many college students have good English grades, their oral ability is relatively weak. With the rapid development of Internet information technology, the use of multimedia to promote the learning of English oral pronunciation has become a hot topic in the field of assisting English oral learning. As an important element of clear and authentic pronunciation, vowels are mostly learned through classroom demonstration by teachers and imitation by students, lacking personalized and targeted pronunciation guidance. There is a lack of timely feedback, and the teaching resources and teaching methods are single.

Therefore, this paper designs and implements a hybrid neural network model combining CNN and LSTM, and further improves the model performance by introducing a complex network architecture to obtain an English vowel phonetic pronunciation error recognition model based on speech recognition technology. The Mel Frequency Cepstrum Coefficient (MFCC) is used to extract the characteristics of the speech signal using speech signal processing technology. By matching it with the vowel data of the native speaker, the gap between the test speech and the standard speech is obtained as the scoring result. At the same time, the formant is extracted and compared with the vowel data of the native speaker to give specific vowel pronunciation tongue position suggestions.

The English vowel phonetic mispronunciation recognition model constructed in this paper can quickly analyze a large amount of speech data and accurately identify the specific location and type of pronunciation errors. At the same time, the model can also provide personalized feedback and suggestions based on the learner's pronunciation characteristics to help learners improve their pronunciation in a targeted manner. This personalized teaching method, compared with traditional teaching methods, can better meet the needs of different learners and improve teaching effectiveness.

The relevant audio and model has been uploaded to the Github (https://github.com/LeiYining/Vowel-Pronunciation-Error-Detection).

Keywords: English vowel pronunciation; speech signal processing; speech recognition; MFCC; formant; ASR

Content

1. Introduction

Today, with globalization, the importance of English, as the universal language of international communication, is becoming more and more prominent. In China, English education has become an important part of basic education, which is pivotal to improving national quality and cultivating internationalized talents. However, Chinese English learners still face challenges in pronunciation, especially in the pronunciation of vowels, although they have made remarkable progress in vocabulary, grammar and listening.

Vowels are the core of English pronunciation and play a decisive role in shaping clear and natural spoken pronunciation. However, due to the huge difference in phonological systems between Chinese and English, Chinese learners of English often encounter difficulties in vowel pronunciation. While the pronunciation of vowels in Chinese is relatively simple and straightforward, the pronunciation of vowels in English is more complex, requiring learners to control the shape of the mouth, tongue position and lip shape to achieve accurate pronunciation. This difference makes Chinese learners feel incompetent in imitating standard English vowel pronunciation, and they are prone to problems such as inaccurate pronunciation position and inappropriate control of pronunciation duration, thus affecting the accuracy and naturalness of overall oral pronunciation.

In order to improve Chinese English learners' vowel pronunciation problems, the existing English pronunciation teaching methods, although helpful to a certain extent, still have many limitations. Although the traditional approach of teacher modeling, student imitation and error correction can provide some basic pronunciation guidance, it is difficult for teachers to provide personalized feedback because it is difficult for them to continuously and accurately monitor and assess each student's pronunciation.

At the same time, it is often difficult for students to accurately perceive where their pronunciation problems lie, and they are unable to make targeted improvements. In addition, traditional pronunciation teaching methods tend to focus too much on the accuracy of pronunciation and neglect key factors such as the shape of the oral cavity, tongue position and lip shape, which are also crucial to the accuracy and naturalness of vowel pronunciation.

In addition to traditional classroom teaching, there are some basic pronunciation applications available on the Chinese market, such as KeDa Xunfei and NetEase YouDao Dictionary, etc., but most of them focus on general pronunciation practice, giving model readings and pronunciation scores, and lack in-depth analysis and guidance on the pronunciation of specific phonemes.

To overcome these limitations, this study aims to develop an English pronunciation analysis system based on automatic speech recognition (ASR) technology. The system is able to analyze English vowels pronounced by Chinese learners in real time and provide detailed feedback about tongue position and lip shape. Through advanced ASR algorithms and speech processing technology, the system is able to accurately identify the vowels pronounced by learners and generate personalized feedback and suggestions based on the analysis results. These feedbacks and suggestions will cover all aspects of pronunciation, from tongue position to lip shape, helping learners to improve their English pronunciation in a comprehensive way.

The development of this system not only helps to solve Chinese English learners' difficulties in vowel pronunciation, but also has important practical significance and far-reaching historical significance. First, by providing more accurate and personalized pronunciation feedback, the system can help learners correct pronunciation errors more quickly and improve their oral communication skills. Secondly, the system can provide English teachers with an effective tool to help them

more accurately assess students' pronunciation levels and provide more targeted teaching guidance. Finally, the development of the system can also promote the application and development of ASR technology in the field of English pronunciation education, provide more possibilities and innovative ideas for the future of English pronunciation education, and provide a better learning experience for English learners around the world.

The thesis is structured as follows: Section 2 provides a literature review on the difficulties faced by Chinese-speaking English learners in vowel production, followed by a discussion of the significance of formant values and MFCC in vowel articulation, and application of ASR system in CAPT. Section 3 describes the methodology and experimental setup, while Section 4 presents the results and analysis. Finally, Section 5 discusses the implications and limitations of the study, as well as future directions for research.

2. Literature review

As globalization advances, learning English, the primary language of international communication, becomes increasingly important for native Chinese speakers. However, the differences in the phonological systems of Chinese and English present significant challenges in mastering English vowel pronunciation. This study aims to explore these specific difficulties, their underlying causes, and potential solutions through an in-depth analysis of related lit.

In the literature review, we will first focus on the specific challenges Chinese learners face in English vowel pronunciation, including accuracy problems in vowel pronunciation and difficulties in perception and output of vowel pronunciation. Then, we will compare the differences between the Chinese and English phonological systems, with a special focus on how these differences affect the perception and

output of English vowels by native Chinese speakers. In addition, we will explore the effects of native language interference on learners' ability to perceive and produce new phonemes in the target language, in order to further our understanding of what makes it difficult for native Chinese speakers to learn English vowels. Subsequently, we will discuss in depth the importance of resonance peaks in vowel pronunciation and elaborate on their specific roles in vowel pronunciation. Last, we will explore the potential application of automatic speech recognition (ASR) systems in computer-assisted pronunciation training (CAPT). Through in-depth analysis and summarization of these literatures, we will provide a solid theoretical foundation and literature support for subsequent empirical studies.

2.1. Challenges in English vowel pronunciation for Chinese learners

In oral communication, the accuracy of pronunciation plays a crucial role in the effectiveness of communication, in which vowel pronunciation is particularly crucial to the improvement of secondary school students' overall phonological level. However, in the process of teaching English phonetics, vowel pronunciation is often a major challenge for Chinese students. The fundamental reason is that the correct pronunciation of vowels is highly correlated with the opening of the mouth and the position of the tongue, which makes it difficult for students to master them accurately. In addition, the difference in vowel systems between English and Chinese is also a major challenge. Due to the negative transfer of mother tongue, Chinese students tend to assimilate English vowels with their mother tongue pronunciation habits, which makes it difficult for them to accurately master the English pronunciation (Zhang Jinsheng, 2002).

Chen Xiaoli (2010) conducted an in-depth study on English learners' front vowel pronunciation. Fifteen Chinese university students and eight native English speakers were selected to participate in the experiment, and the subjects were asked to read

aloud and tape-record the sentence "I will say __." The subjects were asked to read aloud and record 30 vowel-containing words in the sentence "I will say __.". Three native English speakers acted as judges. The results showed that Chinese students had a better grasp of the pronunciation of /i:/, ɪ, and e/, but showed significant confusion in distinguishing /i:/ from /ɪ/ and /ɜ:/ from /æ/. This finding verifies the phenomenon of native language transfer due to the differences between the English and Chinese vowel systems, especially the loose and tight differences in the pronunciation of the pairs of phonemes /i:/ vs. ɪ and /ɜ:/ vs. /æ/, which do not exist in Chinese and are therefore easy to confuse for Chinese students.

Zhu Wenjuan (2012), on the other hand, explored the phenomenon of negative phonological transfer among students in the Central Plains through a questionnaire survey and phonological testing method. The survey revealed that the teacher's level of phonological teaching had a significant effect on students' phonological acquisition. Many students reported that teachers only gave a brief introduction to phonemes and failed to teach phonics knowledge systematically. Analysis of the phonological recordings revealed that at the segmental level, students tended to confuse phonetic pronunciation with Chinese pronunciation; at the suprasegmental level, backward shifting of stress was common, and they did not have sufficient knowledge of phonological changes in speech streams, such as weak reading, ellipsis, legato, augmentation, imperfective bursting, assimilation, and dissimilation.

Xie Jing's (2014) study focused on the negative transfer effects of the Henan dialect on English phonology. At the level of bursting sounds, students often append the /ə/ sound after pronunciation; the distinction between long and short vowels is not obvious, and there is the phenomenon of replacing diphthongs with single vowels. At the suprasegmental level, syllable structure, stress and intonation are all affected by the negative transfer of native language, such as the insertion of vowels in consonant concatenation, and insufficient variation in sentence intonation.

Cao Xiaojin (2016) explored the problems in middle school students' phonological learning through a questionnaire survey. The results showed that students' problems were particularly prominent at the level of suprasegmental segments, including deficiencies in stress, incomplete bursts, alliteration, and phonological intonation. In vowel pronunciation, about half of the students had difficulty in distinguishing long and short sounds and failed to adjust their muscle relaxation according to the pronunciation requirements; consonant pronunciation was interfered by dialects and Chinese pronunciation habits.

2.2. Comparison of vowels between Chinese and English

2.2.1. Comparison of two inventories

Phonetics, as an important branch of linguistics, is devoted to the study of the production, characteristics and changing law of speech sounds in human languages. In phonetics, a phoneme is defined as the smallest phonetic unit that distinguishes the meaning of a language. According to Kenyon, an American phonetician, there are 44 phonemes in American English, of which 19 are vowels, covering the categories of monophthongs, diphthongs and even triphthongs, while there are 25 consonants. However, different scholars disagree on the exact categorization and number of English phonemes. Structuralists Trager Smith and Gleason argue that there are thirty-three phonemes in English, including nine vowels and twenty-four consonants, while Jones proposes a system of fifty-two phonemes, which contain twenty vowels and twenty-eight consonants.

In contrast, the phonological system of Mandarin Chinese presents very different characteristics. In Chinese, morphemes with tones are the smallest units of meaning, and each Chinese morpheme has a fixed tone. Therefore, when defining Chinese phonemes, one faces a more complex challenge than that of English. According to some authoritative documents, the number of morphemes in Mandarin Chinese is

1,382 and 1,644, but the exact number of phonemes is difficult to define precisely.

The most significant difference in the phonological system between Mandarin Chinese and English is reflected in the vowel system. The English phonological system contains twelve vowels, whereas Mandarin Chinese has only six vowels. This difference in number reflects that Chinese vowels have richer functions than English vowels. In addition, English vowels have a series of characteristics that Chinese vowels do not have, such as duration, tense, and tightness. These characteristics play an important role in the pronunciation of English vowels, and sometimes even lead to changes in word meanings, such as the distinction between "it" and "eat", "full" and "fool", which depends on the length of vowels.

It is worth mentioning that some vowels in English and Chinese do not have their direct counterparts in the other language. For example, the /u/ sound in Mandarin Chinese does not have a direct counterpart in English, and the /æ/ sound in English does not exist in Mandarin Chinese. This lack of phoneme correspondences undoubtedly increases the difficulty of Chinese learners in pronouncing English vowels.

2.2.2. Definition and importance of formants

In the process of speech generation, the resonance or amplification effect of the vocal tract structures (including the larynx, the oral cavity and the nasal cavity, etc.) on the sound waves, these specific resonance points are collectively referred to as resonance peaks.

The definition of vowel sound quality is strongly influenced by the resonance peaks. Specifically, the first resonance peak (F1) is usually closely related to the height (openness) of the vowel, with high vowels having lower F1 values and low vowels having higher F1 values. The second resonance peak (F2), on the other hand, primarily reflects the front-to-back position of the vowel, with higher F2 values for front vowels and lower F2 values for back vowels (Ohala, J., 1989).

For example:

Vowel /i/: In the word "heed", this vowel has a low F1 and a high F2, reflecting the forward and elevated tongue position during pronunciation.

The vowel /æ/: in the word "had", the F1 value is high and the F2 value is relatively high, revealing a low and forward tongue position during pronunciation.

Vowel /u/: in the word "who", both F1 and F2 values are low, which represents the high lift and retraction of the tongue during articulation.

2.2.3. The relationship between formant values and the accuracy of phonetic pronunciation

There is a close correlation between the quality of vowels and tongue position, which He (2002) elaborated in Comparative Studies of English and Chinese Languages and analyzed in depth based on the tongue position diagram and the subtle differences in similar phonemes between Mandarin and Standard English.

In the comparison between the English vowels /i:/, /ɪ/ and the Mandarin vowel (i)[i], all three belong to the category of high front vowels, but the tongue position of Mandarin (i)[i] is higher and more forward than that of the English vowels, and the corresponding F1 value is lower, while the F2 value is higher. For the English vowels /u:/, /ʊ/ and the Mandarin vowel (u)[u], although they are all high back vowels, the English /u:/ has a slightly higher and more anterior tongue position, while the Mandarin (u)[u] is located in a higher and more posterior tongue position and therefore has lower F1 and F2.

In the comparison of the English vowels /ɔ:/ and /ɒ/ with the Mandarin vowel (o)[o], /ɔ:/ and (o)[o] have similar tongue positions and are both mid-back vowels, while /ɒ/ is located in a low and back tongue position. The small difference in front and back position between these three vowels may mean that they have similar F2 values.

In the comparison between the English vowels /æ/, /ɑ:/ and the Mandarin vowel (a)[A], all three are in low lingual position, but /æ/ is in front lingual position, while /ɑ:/ and (a)[A] are in back lingual position. Although /ɑ:/ and (a)[A] are similar in anterior and posterior tongue position, Mandarin (a)[A] has a slightly lower tongue position and a higher F1 value than English /ɑ:/.

Combining the two tongue position charts, the range of tongue position fluctuations for Mandarin monophthongs is wider compared to English, and their limit positions exceed those of English in terms of high and low, anterior and posterior. In addition, there is no high-mid vowel in English, while Mandarin lacks mid-front, low-front and high-mid vowels.

In summary, the values of F1 and F2 directly reflect the tongue position of vowels, and the analysis of these two enables us to clearly identify the similarities and differences between English and Mandarin in terms of tongue position. Therefore, the F1 values of Mandarin (i)[i] are lower than those of English /i:/ and /ɪ/, while the F2 values are higher than them; the F1 and F2 values of English /u:/ and /ʊ/ are higher than those of Mandarin (u)[u]. Mandarin (o) [o] has similar F1 and F2 values to English /ɔ:/; while Mandarin (a) [A] has lower F2 values and higher F1 values than English /æ/, /ɑ:/.

2.3. Influence of native language on second language phoneme acquisition

2.3.1. Theoretical frameworks

Native language interference plays an important role in the process of second language acquisition, and its influence is mainly reflected in learners' ability to perceive and produce new phonemes in the target language. The causes of vowel pronunciation errors are complex and usually involve both internal and external factors. Internal factors include age differences, brain processing mechanisms, and

auditory perception; while external factors mainly focus on the influence of mother tongue and dialect transfer, and the quality and frequency of second language input. In this paper, we focus on the effects of external factors, especially native language interference, on Chinese native speakers' (CN speakers) learning of English (EN) vowel pronunciation.

For Chinese native English learners, native language interference has a significant effect in perceiving and producing new phonemes in English. At the perception level, learners are susceptible to the influence of the native phonemic system, which biases their perception of English vowels, thus affecting their pronunciation accuracy. At the production level, interference from the native phonemic system makes it difficult for learners to escape from the influence of native pronunciation habits when attempting to produce English vowels, resulting in inaccurate pronunciation.

2.3.2. Empirical evidence

Native language interference, as a common phenomenon in second language learning, has received extensive attention from scholars around the world. The significant difference in the articulatory system between Chinese and English is one of the major causes of pronunciation differences. Chen et al. (2001) studied American English vowels produced by Mandarin native speakers through acoustic analysis and found that "Mandarin subjects had a smaller vowel quadrilateral range and lower acoustic diversity in vowel production compared to American native speakers". Jin, Su, and Chang (2014) also pointed out in their study that native Chinese-speaking ELLs have intelligibility problems with vowel pronunciation. In addition, Pillai et al. (2010) study of Malaysian English vowels found that Malaysian female subjects generally faced sound quality and duration problems in vowel pronunciation, again reflecting the effects of native language interference.

In addition, China is a vast country with many dialects, and the pronunciation habits of the dialects vary significantly from place to place, which affects the acquisition of English vowels to varying degrees. By comparing the vowel systems of

Wu and Min with those of American, Jiang (2010) found that "due to the influence of dialects, students in Wu and Min dialects have difficulties in vowel pronunciation". Dai (2019), on the other hand, conducted a study on monophthong articulation errors and their acoustic characteristics for English learners in three dialect areas in Jiangsu Province, China. Gao and Gong's (2011) study also showed that native language transfer had a significant effect on Chinese ELLs' English vowel pronunciation, and Wang (2015) and Kong (2019) conducted studies on the acoustic characteristics of vowel pronunciation for Chinese ELLs in Chifeng and Zaozhuang regions, respectively, further confirming the prevalence of native language interference in second language acquisition.

2.4. Application of ASR and speech analysis in CAPT

2.4.1. Overview of CAPT systems

In the past decades, the wide application of computer-assisted language learning (CALL) technologies has significantly contributed to the rapid development of language teaching and has brought a lot of fun to language teaching, but it is difficult to ensure the pedagogical value of many of the applications due to the lack of a solid pedagogical and linguistic foundation. In this context, as a key component of the CALL field, due to the specificity of language teaching and the rapid development of language technology, Computer Aided Pronunciation Training (CAPT) has become the most popular and promising application technology, which focuses on teaching and learning the characteristics of segments and suprasegmentals in the phonological system of a language. CAPT has become the most popular and promising application, focusing on the teaching and learning of segmental and suprasegmental features in the phonological system of a language (Lan, 2010). According to Rostron and Kinsell (1995), CAPT technology is defined as a digital sound processing tool designed to enhance articulatory skills.

CAPT, like other CALL systems, provides learners with advantages that

traditional classroom instruction cannot: a personalized, stress-free, self-paced learning environment; rich, multimodal learning materials; targeted feedback; and visualization of the physiological mechanisms of articulation and acoustic-physical features of speech, which are especially needed for learning phonetics.

However, the lack of explicit pedagogical guidelines for CALL, and CAPT in particular, based on second language acquisition (SLA) research has become a central obstacle limiting its further development. As scholars such as Levy (1997), Pennington (1999), and Chapelle (1997, 2001) have emphasized, the design of CALL software often lacks systematic and theoretical pedagogical guidelines despite rapid technological advances. This problem is particularly pronounced in the CAPT system, as the highly specialized nature of pronunciation training requires specific pedagogical methodologies and strategies.

In order to fully realize the educational potential of the CAPT system, it is necessary to establish a set of rational pedagogical norms. In traditional teaching environments, pronunciation training usually follows a set of widely recognized pedagogical standards, which include clear pronunciation goals, effective feedback mechanisms, and sufficient opportunities for practice. These standards also apply to the CAPT system. Specifically, the core elements of the CAPT system - input, output and feedback - should be subject to in-depth analysis and scientific design.

2.3.1.1 Inputs in CAPT

It is widely recognized in the academic community that CAPT instruction has significant potential to provide instructional advantages over traditional classroom environments. This is because such software is able to provide learners with access to a wide range of native speech samples. The CAPT software currently on the market all support users in accessing unlimited second language (L2) input resources. In light of the observation that lip movements have been widely demonstrated to improve articulation and perception (Massaro, 1987; Jones, 1997), some CAPT systems have

also been specifically designed with features to show learners articulatory points through 3D animations of the inside of the mouth. This is sometimes complemented by written explanations (Glearner, 2001; Pro-nunciation, 2002) or presented through video sound of native speakers' pronunciation targets (Glearner, 2001; Nieuwe Buren, 2002; Eurotalk, 2002). As a result, animation and video are widely used as instructional media in these systems.

However, on the other hand, mere exposure to L2 input does not seem to be sufficient to fully improve pronunciation. For example, long-term residents of foreign countries may still have a strong accent in their pronunciation even when they are in the target language environment, making the L2 difficult to understand (Morley, 1991). This suggests that although the CAPT system provides learners with a rich source of L2 input, pronunciation improvement needs to be combined with other factors, such as consistent practice and professional guidance.

2.3.1.2 Output in CAPT

As Hendrik's (1997) research has shown, specialized training must be provided and students need to be given the opportunity to practice in the first place. Particular attention should be paid to creating meaningful, engaging and stress-free environments that encourage even the most non-verbal students to speak and facilitate learning (Morley, 1991; Hendrik, 1997), as speaking is essential for improving pronunciation.

Therefore, most current CAPT systems are learners' speech that can be recorded and played back, and provide learners with the ability to pre-record audio. By doing so, learners can observe their output and improve their pronunciation by comparing it with standard speech.

However, according to these systems, it is up to the learners themselves to determine if and how the audio of their pronunciation differs from that of their mother tongue, which is a major problem with CAPT, as most students may find it difficult to perceive or assess phonological differences in a second language. As Neri (2007) points out, numerous studies have shown that L2 learners are often unable to

comprehend the phonetic differences between L1 and L2 and therefore need external feedback.

2.3.1.3 Feedback in CAPT

Neri's (2007) study shows that there is no consensus on a clear conceptualization of external corrective feedback (ECF) in the field of second language acquisition (SLA). However, it usually refers to information provided by native speakers or teachers about non-targeted discourse-often referred to as negative evidence-but lacks a more detailed definition and a categorization of the different types of feedback and their respective learning effects. In computer-assisted language learning (CALL) systems, the term ECF is often used to describe feedback information related to task performance or errors as part of learning assessment, including grading. In some contexts, ECF even encompasses instructions, explanations, or directions in learning aids (Pujola, 2001). Thus, current computer-generated pronunciation feedback utilizes different techniques and graphics to give more or less rich and explicit feedback on different aspects of pronunciation.

Until recently, CAPT instructional materials offered diverse feedback strategies. However, due to technical constraints, existing materials are far from ideal. For example, modern CAPTs often include recording and playback features that allow learners to record their own speech samples and compare them with native speaker recordings. However, there are limitations to this method of self-assessment because learners often have difficulty accurately recognizing speech differences between their native language (L1) and their second language (L2) (Flege, 1995).

Another approach to CAPT feedback that may circumvent the limitations of self-assessment is the distance learning system. The system requires the user to record voice samples, upload them to an online platform or send them via "voice mail," have them assessed and scored by an authorized trainer, and then provide the results back to the student (Ferrier & Reid, 2000). However, the usefulness of this approach has been questioned as the frequency of feedback is limited by the time and willingness of the assessor.

In addition, some CAPT systems measure the volume, intonation, duration, and frequency of students' articulation through acoustic analysis tools (e.g., Pro-nunciation, 2002; Lambacher, 1999), and present the results using spectrograms and waveform graphs. However, the effectiveness of these systems is questionable because learners have difficulty understanding and interpreting these graphs, and even professional phoneticians may have difficulty deriving from them the critical information needed to improve phonological skills.

In response to these challenges, Automatic Speech Recognition (ASR) technology appears to offer an effective way to optimize CAPT systems. Therefore, an ideal CAPT system should integrate the three key elements of input, output and feedback, especially in conjunction with ASR technology, to maximize its pedagogical benefits.

2.3.2 ASR technology in pronunciation training

The two CAPT systems described above maximize neither cost-effectiveness nor time-effectiveness, and the instruction they provide tends to be generalized and fails to address the uniqueness of each learner. Given the uniqueness of each learner, the optimal training strategy should be a one-on-one learner-mentor interaction model. The CAPT system, which incorporates an Automatic Speech Recognition (ASR) module, is able to support this interaction by detecting personalized errors and providing immediate feedback.

In the late 1990s, the first computer-assisted pronunciation training (CAPT) system was introduced that utilized automatic speech recognition (ASR) technology to provide feedback on pronunciation quality. Since then, second language teachers and researchers have had high expectations for such systems. The core strength of these systems lies in their ability to provide users with automatic, immediate, and personalized feedback on pronunciation. Given that feedback is critical for raising learners' awareness of second language pronunciation errors (Flege, 1995), the time

available for teachers to provide personalized pronunciation instruction to each student is extremely limited in traditional classroom settings. In addition, teachers usually avoid correcting pronunciation errors in the classroom, lest students feel embarrassed leading to reluctance to speak the second language. The ASR-based CAPT system provides real-time feedback to each student through headphones and visual information on a monitor, alleviating classroom time constraints and student stress.

CAPT has evolved rapidly in the early twenty-first century. For example, the BetterAccent software system (http://www.betteraccent.com/) is notable for its unique focus on pitch, stress, and rhythm. Specifically, the system first records a sample of the learner's speech, and then uses algorithms to carefully compare the learner's articulatory rhythmic curves (e.g., changes in pitch and intensity) with those of a proficient speaker. This process not only covers a detailed interpretation of the speech pitch curve of the proficient speaker, but also ensures the accuracy and reliability of the assessment results.

However, the system mainly relies on the comparative analysis of articulatory rhythmic curves between proficient speakers and learners when providing suggestions for pronunciation improvement, and lacks personalized strategies to address the characteristics of individual learners. Second, in terms of the hardware requirements for the system to run, the software may encounter memory management problems, such as "low bandwidth memory" errors, when running on a lower configuration computer (e.g., a PC with 512MB of RAM, a Pentium 4 CPU, and Windows XP), which to some extent limits its ability to be used in a wider range of applications. This limits its use in a wider range of teaching and learning environments.

The Streaming Speech software system (http://www.speechinaction.com/) is an online educational platform developed by Speech-in-action. Its academic value is mainly reflected in the rich learning resources it provides. Specifically, the system is able to continuously provide daily updated speech and pronunciation materials in natural contexts, while covering the areas of passage and suprasegmental learning.

What is more, the system presents sentences in the form of stress distribution and tones, providing learners with a more realistic and comprehensive language learning experience. However, from the assessment point of view, although the system allows learners to record their own voices, the assessment method mainly relies on learners' self-reflection and lacks external or system-generated feedback mechanisms. In addition, the system invests relatively little in pronunciation teaching and focuses more on the development of listening skills.

The Pronunciation Power 1&2 software system (http://www.englishlearning.com), developed by English Computerized Learning Inc. is a set of educational software designed to improve English pronunciation skills. The system provides a full range of support from beginner to advanced levels for English language learners of all levels. Specifically, Pronunciation Power 1 is intended for intermediate English beginners, while Pronunciation Power 2 serves as an ideal choice for advanced English learners. Both have similar core features and are dedicated to improving learners' ability to correctly pronounce the 52 phonemes in the English language. The system uses animated characters and video clips to visually demonstrate pronunciation techniques, and is complemented by lessons and practice questions to ensure that learners are able to correct their pronunciation in real-world situations. In addition, learners can not only hear the standard pronunciation, but also play back their own pronunciation through recordings to self-check and improve. However, it is worth noting that the system is not sufficient for teaching suprasegmentals.

In summary, ASR technology is able to convert speech signals into text, while speech analysis technology is able to further analyze the acoustic-physical features in the speech signals, such as pitch, pitch length, and pitch intensity. Through these technologies, the CAPT system is able to automatically detect learners' pronunciation errors and give corrective suggestions accordingly. In addition, these software programs demonstrate certain advantages over traditional teacher-led pronunciation training and seem to fulfill the pedagogical needs of pronunciation training. In addition, the CAPT system provides additional training time and materials, offering

learners the opportunity to practice in a stress-free environment. Combined with ASR technology, the CAPT system can be further leveraged to optimize the learning experience by evaluating the learner's speech and providing real-time personalized feedback, which can provide a more vivid and interesting learning experience for the learner.

2.5. Conclusion to the literature review

2.5.1. Synthesis of findings

After careful sorting and in-depth analysis of a large number of relevant literature, we were able to gain a deep understanding of the challenges faced by Chinese students in learning English phonetics from multiple perspectives and levels, especially in the difficulty of vowel pronunciation. The literature review shows that due to the significant differences between Chinese and English in the phonetic system, Chinese students encounter many difficulties in learning and mastering the pronunciation of English vowels. For example, the study found that Chinese-speaking English learners exhibit significant difficulties in pronouncing English vowels, particularly in distinguishing between /iː/ and /ɪ/, and /ɜː/ and /æ/. This confusion not only affects students' pronunciation accuracy, but also hinders their effective communication in oral English.

In addition, the literature also reveals the essential differences between the English and Chinese vowel systems, such as the pronunciation of /iː/ and /ɪː/ in English is very different from the pronunciation that Chinese students are familiar with. This difference makes it difficult for Chinese students to get rid of the influence of their mother tongue when trying to imitate the pronunciation of English vowels, thus forming incorrect pronunciation habits.

Finally, the author investigated the application of automatic speech recognition (ASR) and speech analysis in computer-assisted pronunciation training (CAPT): CAPT systems use ASR technology to provide learners with personalized feedback, which significantly improves the effect of pronunciation training. Case studies show

that these technologies have wide adaptability and effectiveness in practical applications, such as using ASR and CAPT systems, showing their great potential in improving pronunciation accuracy.

However, these methods and systems still have some limitations in practical applications. First, the accuracy of ASR technology is affected by multiple factors such as background noise, speech speed, and pronunciation clarity, which may cause the system to fail to accurately recognize the learner's pronunciation. Second, although speech analysis technology can analyze acoustic-physical features, it often has difficulty capturing subtle differences in pronunciation, such as phonemic changes in vowel pronunciation. In addition, existing CAPT systems often lack targeted feedback mechanisms, such as being unable to provide very specific and effective correction suggestions for learners' specific vowel pronunciation errors.

2.5.2. Emergent research questions and hypotheses

Vowel pronunciation is an important part of language learning, and its accuracy directly affects learners' voice quality and communication effectiveness. However, most existing CAPT systems focus on evaluating and providing feedback on overall pronunciation, such as stress, intonation, and rhythm, and do not provide specific and accurate feedback on individual vowels. Therefore, we need to build a system that can effectively integrate speech research and speech technology to provide more comprehensive and in-depth feedback on vowel pronunciation.

Therefore, the research question of this paper is: Can an ASR system trained on native English speech data accurately detect pronunciation errors in the vowel productions of Chinese- speaking English learners by analyzing the discrepancies in formant values compared to native English speakers?

At the same time, the following hypotheses are proposed:

Automatic speech recognition technology can effectively identify and correct vowel pronunciation errors of Chinese English learners, significantly improving their pronunciation accuracy. And the phonological structure of the native language has a significant impact on learners' vowel perception and production in the target language,

and this impact can be alleviated through targeted pronunciation training.

The formulation of these hypotheses not only pointed out the direction of our research, but also provided a basis for subsequent experimental design and data analysis. We look forward to verifying the validity of these hypotheses through rigorous experimental design and data analysis in future research, and providing useful references and inspirations for Chinese students' English pronunciation teaching.

3. Speech signal preprocessing and feature extraction

3.1. Speech recognition process

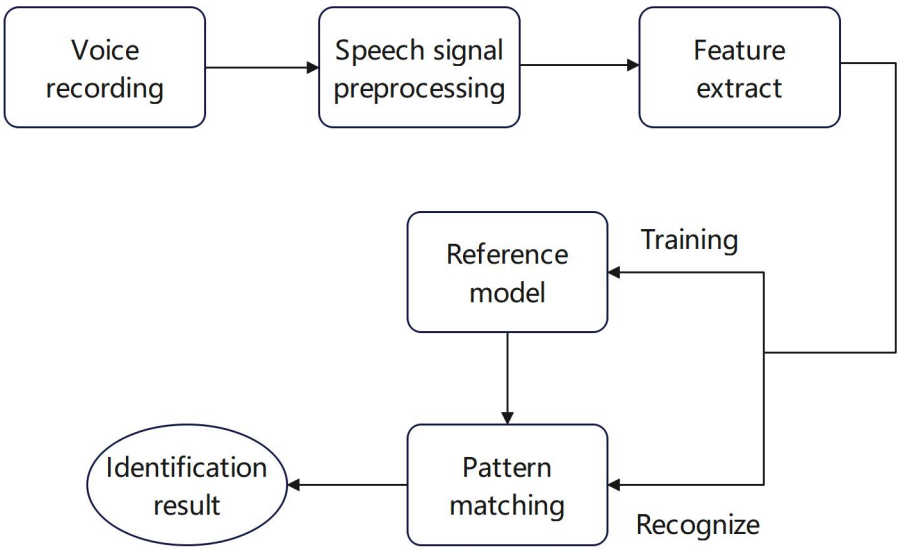

Figure 1 Speech recognition process

The general process of speech recognition is shown in Figure 1. First, the speech signal is captured using a specialized speech analysis tool, such as Praat. This step is crucial because it provides the raw, unprocessed speech data for subsequent signal processing.

In the process of converting analog signals to digital signals, the Nyquist sampling theorem provides the theoretical basis. According to the theorem, when the sampling frequency is higher than twice the highest frequency in the signal, the sampled digital signal can more accurately restore the information in the original signal. Given that the frequency range of human speech signals is usually between 40Hz and 4000Hz, in order to ensure the effective conversion of the signal, this paper chooses 16KHz as the sampling frequency, which not only meets the requirements of the Nyquist sampling theorem, but also takes into account the balance between the computational efficiency and the quality of speech.

Next, the acquired speech signal is preprocessed. Preprocessing is a key part of the speech recognition process, which performs a series of operations on the original signal to reduce noise, improve signal quality, and provide more ideal input data for subsequent feature extraction and model training. The preprocessing steps usually include pre-emphasis, which is used to compensate for the attenuation of the speech signal in the high-frequency part; frame-splitting, which cuts the continuous speech signal into shorter frames for subsequent processing; and windowing, which smoothes each frame of the signal through a window function to reduce the spectral leakage.

After the preprocessing is completed, feature extraction needs to be performed on the processed speech signal. Feature extraction is one of the core steps in speech recognition, which transforms the speech signal into parameters that can characterize speech properties through a series of algorithms. These feature parameters should not only be able to reflect the acoustic properties of the speech signal, but also have enough distinguishability to facilitate subsequent model training and pattern matching. Commonly used feature parameters include Mel Frequency Cepstrum Coefficient (MFCC), Linear Predictive Cepstrum Coefficient (LPCC), etc.

Finally, the extracted feature parameters are used as input objects for model training and pattern matching. Model training is to train a model that can recognize speech through a large amount of speech data; while pattern matching is to match the speech signal to be recognized with the trained model to produce the final recognition result.

3.2. Data collection

3.2.1. Data collection environment

Data collection for speech recognition was conducted in a controlled environment to minimize background noise and interference. The recording was done in a soundproof room with constant temperature and humidity. Participants were seated in front of a microphone at a fixed distance to ensure consistent audio capture. And sampling rate is 16 kHz. During the recording sessions, each participant was instructed to read the pre-defined set of phrases at a normal speaking rate.

3.2.2. Variability in speaker input

Given the inherent variability in human speech, it is important to manage the impact of different speakers on the recognition accuracy. To address this, we ensured that the data set included a diverse range of speakers with different genders, ages, and accents. Additionally, we implemented techniques such as normalization and speaker adaptation to minimize the impact of speaker variability on the recognition performance.

Therefor, the dataset was recorded by two native American-English speakers, Matt (44 years old, male and Northeast variety) , my instructor, and Brandi (23 years old, female), my classmate, both of whom have clear articulation, standardized accents, and good control over their pronunciation, and recorded multiple repetitions of each vowel.

3.2.3. Recording condition

The vowel articulations for this data collection included the following eleven vowel phonemes:

/i:/ as in "beat"

/ɪ/ as in "bit"

/ɔ:/ as in "bought"

/ɔ/ as in "pot"

/u:/ as in "boot"

/u/ as in "put"

/ə:/ as in "bird"

/ə/ as in "about"

/ɑ:/ as in "father"

/ʌ/ as in "but"

/e/ as in "bed"

/æ/ as in "cat"

These vowels cover the main vowel categories in English and are broadly representative.

During the recording process, each speaker was asked to record in a quiet environment to minimize the effects of noise. Each audio sample contained 10 to 20 repetitions of vowel pronunciations to ensure data adequacy and reliability. In addition, in order to "clean up the pronunciation environment" and avoid pronunciation drift, we included a carrier phrase such as "The sound is /i:/" before each vowel pronunciation. This expression helps speakers to better enter the pronunciation state and maintain the accuracy and stability of pronunciation. At the same time, we also noticed that repeating vowel sounds alone may bring problems, such as difficulty in including the initial guttural sound. Therefore, during the recording process, we specifically reminded the articulators to pay attention to these problems and to include these elements in their pronunciation as much as possible.

After completing the recordings, we carefully preprocessed the collected audio samples. First, we converted all audio in MP3 format to wav format for subsequent analysis and processing. Then, we resampled all the audio at a sampling rate of 16,000Hz to ensure consistency and comparability of the audio data.

Next, we utilized the professional audio processing software Praat to further process the audio data. Through the segmentation and labeling functions of Praat, individual vowel pronunciations were extracted from each piece of audio and saved as separate audio files. Each vowel has 30 separate audio samples, totaling 330 audio data of vowel pronunciations for native speakers. These data will be used for subsequent experimental analysis and research.

3.3. Speech signal preprocessing

Before speech signals can be analyzed and processed, they must be subjected to pre-processing operations such as pre-emphasis, frame-splitting, windowing and endpoint detection. The purpose of these operations is to eliminate the effects on the quality of the speech signal due to high harmonic distortions, high frequencies, and aliasing of the human vocal apparatus itself and the speech signal acquisition equipment. Speech preprocessing affects the results of speech feature extraction, and smoother and more homogeneous speech signals can provide better quality parameters for speech feature extraction, thus improving the quality of speech processing.

3.3.1. Pre-emphasis

Influenced by physical properties such as orofacial radiation and vocal gate excitation, speech signals show a significant tendency to attenuate at the high-frequency end of the average power spectrum (above about 800 Hz) at a rate of about 6 dB/oct (octave). In order to improve the spectral characteristics of the speech signal in the high-frequency band, and to ensure that the entire frequency band (from low to high frequencies) can be analyzed efficiently at the same signal-to-noise ratio,

a high-frequency boosting pre-emphasis digital filter with a 6dB/oct characteristic is usually used prior to the analysis of the speech signal. This pre-emphasis filter serves to enhance the high-frequency component of the speech signal to flatten the spectrum.

The response function of the filter can be expressed as:

$$H(z)=1-\alpha z^{-1}, 0.9 \le \alpha \le 1.0$$

Where α is the pre-emphasis coeffic[1]ent, in this paper we set α to 0.9375. After the pre-emphasis process, the relationship between the output signal *y(n)* and the input speech signal *x(n)* can be expressed as:

$$y(n) = x(n) - \alpha x(n-1)$$

This equation demonstrates how the pre-emphasis filter treats the input speech signal by enhancing the high-frequency component by subtracting a weighted value of the signal from the previous moment, thus improving the analyze ability of the speech signal over the entire frequency band.

In this paper, a filter with coefficients [1, -0.99] is applied using the lfilter function. The coefficient of the pre-emphasis filter is -0.99, which is used to increase the weight of high frequency components.

3.3.2. Framing

The speech signal possesses significant time-varying properties, however, due to the inertia of the vocalist's motion, we can assume that over very short time scales (usually about 10 to 30 milliseconds), the speech signal appears to be essentially unchanged or relatively stable. This property allows the speech signal to be approximated as a quasi-steady-state process, i.e., with short-time smoothness, in this time scale. Therefore, in the analysis and processing of speech signals, the method of “short-time analysis” is usually used, which is based on the frame-by-frame processing of the speech signal stream. In general, the number of frames per second is determined by the frame length, the equation as follows:

$$Frames\ per\ second = \frac{1}{t}\ (0.01 < t < 0.03)$$

The framing operation can be performed in a continuous manner or in an interleaved manner, and given the intrinsic correlation between speech signals, a half-frame interleaved framing strategy is preferred in this paper. This processing ensures that there is a certain amount of information overlap between adjacent frames, which improves the accuracy and continuity of the analysis. For the overall speech signal, after the frame splitting process, we get a time series of characteristic parameters composed of each frame, which provides basic data for subsequent analysis and processing.

In this paper, each frame was set to 25 milliseconds (ms) in duration, with a frame shift of 10 ms, ensuring that adjacent frames overlap.

3.3.3. Windowing

In order to enhance the speech waveform in the vicinity of a particular sampling point *(n)* in the speech signal and to attenuate the waveform in the rest of the signal, windowing of the signal is usually required after the frame-splitting process. The windowing process is essentially a specific mathematical operation or transformation of individual short segments of the speech signal, aimed at improving the characteristics of the signal or extracting key information. Specifically, the windowing operation can be expressed by the following equation:

$$Q_n = \sum_{m=-\infty}^{\infty} T[s(n)]\omega(n-m)$$

Where *T[ ]* denotes a generalized transform (which can be linear or nonlinear), *s(n)* is the input speech signal sequence, $Q_n$is the processed time series.

In speech signal processing, the choice of window function is crucial, which determines the characteristics and effects of signal analysis. Commonly used window functions include Hamming window (Hamming), Rectangular window (Rectangular)

and Hanning window (Hanning), which are defined as follows:

(1) Hamming Window

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N \\ 0, \qquad\qquad else \end{cases}$$

(2) Rectangular Window

$$\omega(n) = \begin{cases} 1, 0 \le n \le N-1 \\ 0, \qquad else \end{cases}$$

(3) Hanning Window

$$\omega(n) = \begin{cases} 0.5\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right], 0 \le n \le N-1 \\ 0, \qquad\qquad else \end{cases}$$

The rectangular window has a narrower primary flap and therefore a higher frequency resolution. However, its higher sidelobe can lead to severe interference between neighboring harmonics, which manifests itself as superposition or cancellation of signals in adjacent harmonic intervals, resulting in spectral leakage. In contrast, Hamming windows are more prevalent in many applications due to their smoother spectral characteristics.

In this paper, we choose Hamming windows for windowing speech signals to optimize the spectral characteristics of the signals and reduce unwanted interference.

3.4. Speech feature extraction

3.4.1. MFCCs

Mel frequency cepstrum coefficients (MFCCs) are based on the mechanism of the human auditory system and are designed to model the perceptual response of the human ear to speech signals of different frequencies. Specifically, the human ear has non-uniform sensitivity to speech signals of different frequencies, a property similar to a specific nonlinear system in which the frequency response roughly exhibits a

logarithmic relationship.

The extraction process of the MFCC feature parameters is shown in Figure 2 MFCC feature parameter extraction process, which converts the original speech signal into a series of numerical parameters that can reflect the auditory characteristics of the human ear through a series of signal processing and transformation steps, thus realizing an effective characterization of the speech signal.

The detailed process of extracting Mel-Frequency Cepstral Coefficients (MFCC) speech feature parameters is as follows:

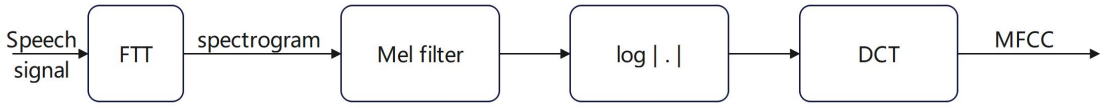

Figure 2 MFCC feature parameter extraction process

1. Fast Fourier Transform (FFT)

To transform the signal from the time domain to the frequency domain, we employ the Fast Fourier Transform (FFT). Specifically, for a discrete speech sequence (x[n]) of length N, where (n = 0, 1, 2, ..., N-1), the FFT is expressed as:

$$X[K] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}nk}, k = 0,1,2,\ldots,N-1$$

Where *X[n], n=0,1,2,...,N-1* is a discrete speech sequence obtained after sampling, and N is the frame length. , *N-1* is a discrete speech sequence obtained after sampling, and N is the frame length. *X[k]* is a complex sequence of *N* points, and then the signal amplitude spectrum *X[k]* is obtained by taking the mode of *X[k]*.

2. Mel-Frequency Conversion

To mimic the human auditory system, we convert the actual frequencies to Mel frequencies. The conversion formula is:

$$Mel(f) = 2597\lg(1 + \frac{f}{700})$$

Where *Mel(f)* is the Mel frequency, and *(f)* is the actual frequency in Hz.

3. Triangular Filterbank and Filtering

A set of triangular filters is configured on the Mel frequency axis, and the filtering output of each filter on the amplitude spectrum |X[k]| is calculated. The filtering output is computed as:

$$F(l) = \sum_{k=f_o(l)}^{f_h(l)} w_l(k)|X[k]| \;\; l = 1,2,\ldots,L$$

where,

$$w_l(k) = \begin{cases} \dfrac{k - f_o(l)}{f_c(l) - f_o(l)}, f_o(l) \le k \le f_c(l) \\ \dfrac{f_h(l) - k}{f_h(l) - f_h(l)}, f_c(l) \le k \le f_h(l) \end{cases}$$

$$f_o(l) = \frac{o(l)}{[\frac{f_s}{N}]}, f_h(l) = \frac{h(l)}{[\frac{f_s}{N}]}, f_c(l) = \frac{c(l)}{[\frac{f_s}{N}]}$$

Within the process, $w_l(k)$ represents the filter coefficient corresponding to filter l, while *o(l)*, *c(l)*, and *h(l)* denote the lower cut-off frequency, center frequency, and upper cut-off frequency of the filter on the actual frequency axis, respectively. f_s represents the sampling rate, *L* is the number of filters, and *F(l)* is the filtered output.

4. Logarithm Operation and Discrete Cosine Transform (DCT)

A logarithm operation is applied to the outputs of all filters, followed by a Discrete Cosine Transform (DCT) to obtain the MFCC features.

$$M(i) = \sqrt{\frac{2}{N}} \sum_{l=1}^{L} log\, F(l)\, cos[(l - \frac{1}{2})\frac{i\pi}{L}] \qquad i = 1,2,\ldots,Q$$

Where *Q* is the order of MFCC parameters (taken as 13 in this paper), and *M(i)* is the resulting MFCC parameter.

3.4.2. Formant extraction

The vocal tract can be regarded as a sound tube with a non-uniform cross-section that acts as a resonator during articulation. When a quasi-periodic pulse generated at the vocal tract is excited into the vocal tract, it triggers its resonance characteristics, resulting in a specific set of resonance frequencies, which are referred to as resonance peak frequencies or simply resonance peaks. The characteristic parameters of the resonance peaks include frequency, bandwidth, and amplitude, all of which are embedded in the envelope of the speech spectrum. Therefore, the core of extracting the resonance peak parameters is to accurately estimate the envelope of the speech spectrum and identify the maximum value as the resonance peak. By applying the Fourier transform to invert the low-frequency portion of the speech spectrum, we can obtain the envelope curve of the speech spectrum. Further, based on the strength of the energy of each peak on the spectral envelope, we can identify the F1 to F4.

In classical speech signal processing models, the resonance peaks are usually modeled as complex pairs of poles of the channel transfer function. For a male vocal tract with an average length of about 17 cm, there are roughly three or four resonance peaks in the 3 kHz frequency range, while the 5 kHz range contains four or five. The energy of the speech signal decreases significantly as the frequency exceeds 5 kHz. For turbid signals, their three most critical resonance peaks are usually the first three. Therefore, all resonance peak estimation methods rely directly or indirectly on a careful analysis of the envelope of the speech spectrum, and the key is to accurately estimate the envelope of the speech spectrum and determine the maximum of it as the resonance peak. In this paper, the following two approaches are used for resonance peak estimation: cepstrum method resonance peak estimation, LPC method resonance peak estimation.

1. Cepstrum method for resonance peak estimation

The input speech signal x(i) is pre-emphasized. Subsequently, the signal is windowed and sub-framed in order to analyze the local characteristics of the speech signal within a short time window. Fast Fourier Transform (FFT) is performed on each frame of the signal to obtain the spectrum $X_i(k)$, where k denotes the frequency index and i denotes the frame index. For each frame of the spectrum, calculate its inverse spectrum. The inverse spectrum is obtained by taking the logarithm of the amplitude of the spectrum and then performing an inverse Fourier transform.

$$X_i(k) = \sum_{n=1}^{N} x_i(n) e^{-\frac{2\pi knj}{N}}$$

Where N is the length of the FFT and j is the imaginary unit.

In order to further reduce the influence of noise and irrelevant information, the cepstrum signal $\widehat{x_i}(n)$ is windowed. The window function *h(n)* is usually chosen with respect to the resolution of the cepstrum (i.e., sample rate and FFT length). The windowed signal can be calculated by the following equation:

$$h_i(n) = \widehat{x_i}(n) \times h(n)$$

$$h(n) = \begin{cases} 1, n \le n_0 - 1 \quad n \ge N - n_0 + 1 \\ 0, n_0 - 1 < n < N - n_0 + 1 \end{cases}, n \in [0, N-1]$$

where n_0 is the length of the window function, usually determined according to specific needs.

Where the window function $h(n_0)$ is defined as follows:

$$\widehat{x_i}(n) = \frac{1}{N} \sum_{k=1}^{N} \lg|X_i(k)| e^{-\frac{2\pi nj}{N}}$$

Perform Fourier transform on the windowed signal to get its envelope $H_i(k)$. The envelope reflects the distribution of the main frequency components in the speech signal and is the key to the extraction of the resonance peak parameters.

$$H_i(k) = \sum_{n=1}^{N} h_i(n) e^{-\frac{2\pi knj}{N}}$$

Find the extreme points on the envelope which correspond to the resonance peaks in the speech signal. By measuring the frequency, bandwidth and amplitude of these extreme points, the resonance peak parameters can be precisely extracted.

2. Resonance peak estimation by LPC method

A simplified model of speech generation is one that reduces the full effects of radiation, vocal tract, and gate excitation to be equivalent to a time-varying digital filter with a transfer function:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum a_i z_i^{-1}}$$

The power spectrum *P(f)* is:

$$P(f) = |H(f)|^2 = \frac{G^2}{|1 - \sum a_i \exp(-j2\pi if/f_s)|^2}$$

The FFT method can be used to obtain the power spectrum amplitude response for any frequency and find the resonance peak from the amplitude response, and there are two corresponding solution methods: parabolic interpolation and linear prediction coefficients for the complex root method.

4. Model neural network architecture

We design a unique model architecture that fuses the advantages of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with the aim of building an efficient and robust speech classification system. By integrating the CNN's ability in extracting spatial features and the RNN's expertise in processing sequence data, a model that can adapt to complex speech signals successfully constructed.

4.1. Convolutional Neural Network (CNN) module

In the CNN part of the model, a multilayer convolutional network structure is designed to capture the spatial features of the input speech signal. The part consists of multiple convolutional, pooling, normalization, and activation layers that work together in order to progressively extract and compress a high-dimensional representation of the input features.

The input layer of the model receives shape-specific feature vectors that typically contain acoustic features extracted from the original speech signal. In order to fit the input requirements of the convolution operation, the input features are dimensionally extended through the *tf.newaxis* operation.

We use a combination of multiple sets of convolutional layers (Conv2D) and BatchNormalization layers. Each set of convolutional layers contains a different number of filters, the size of which depends on the task requirements. The BatchNormalization layer is used to normalize the output of the convolutional layers, thus speeding up the model training process and improving the stability of the model.

After each convolutional layer, a ReLU activation function is introduced, which introduces nonlinearities by truncating the negative values, which helps to improve the expressiveness and learning ability of the model.

In order to reduce the dimensionality of the feature maps and reduce the computation, we set up a maximum pooling layer (MaxPool2D) after a specific convolutional layer. In addition, in order to prevent the occurrence of overfitting phenomenon, a Dropout layer is add after each convolutional layer, which randomly discards some neurons by setting an appropriate Dropout rate.

4.2. Recursive Neural Network (RNN) module

After the CNN module, the Reshape layer reshapes the feature maps into 2D

feature vectors and passes them to the subsequent RNN module for further processing.The RNN module consists of a fully connected layer and a Bidirectional Long Short-Term Memory Network (Bidirectional LSTM).

The reshaped feature vector is first passed through a fully connected layer (Dense) which has a certain number of neurons and uses the ReLU activation function. The role of the fully connected layer is to further map the features extracted by the CNN to a higher dimensional feature space so that the subsequent RNN module can better capture the sequence information. The recursive part of the model adopts a bi-directional LSTM structure, which is able to capture both forward and backward information of the input sequence. By combining the outputs of the forward and backward LSTMs, we are able to understand the temporal dependence in the speech signal more comprehensively.

Finally, the output of the LSTM is passed through a fully connected layer for classification prediction. The number of output units of this fully connected layer is equal to the number of classes to be categorized and the probability distribution is predicted using the Softmax activation function.

4.3. Innovations in integrating Res2Net and Conformer

To further improve the performance of the model, a complex network architecture called *res2net_plus_edit* is implemented . This architecture integrates the features of Res2Net and Conformer by introducing multi-scale feature representations and self-attention mechanisms in order to capture the complex structure and temporal dependencies in speech signals more efficiently.

Res2Net module implements multi-scale feature representation through slice and concatenate operations. The output of the initial convolutional layer is sliced into multiple sub-feature maps, which are then concatenated after different convolutional operations to form information-rich multi-scale features. In each residual block of the

Res2Net module, there is a Self-Attention mechanism. The self-attention mechanism is able to capture long-range time-dependent and global contextual information, which helps the model to better understand the complex structures in the speech signal.

Convolutional module use a combination of depth-separated convolution and pointwise convolution as the convolutional module. Depth-separated convolution is used to capture local features, while pointwise convolution is used for dimensionality transformation. By combining these two types of convolution, we are able to extract feature information from speech signals more efficiently.

In the convolution module, the Gated Linear Unit (GLU) enhances the nonlinear representation of the model by introducing a gating mechanism, enabling the model to better adapt to complex speech signal data.

5. Model building

In the process of model construction and training, a series of well-designed steps ensure that the model can learn and generalize efficiently. This process includes several key aspects such as data preprocessing, model initialization, loss function, optimizer settings, and model training.

5.1. Data preprocessing phase

Audio data first undergoes pre-processing treatments such as pre-emphasis, windowing, and Framing (3.3.3 above) and feature extraction (3.3.4 above). Using Librosa library, the MFCC of each vowel audio was extracted and resonance peaks were extracted using two methods: cepstrum method resonance peak estimation, and LPC method resonance peak estimation.

In addition, due to the different lengths of the vowel audio data, the strategy of

feature alignment and zero complement is used. The length of all audio features is unified to a fixed value by random intercept and zero complement operation, so as to facilitate the model for batch processing.

5.2. Model initialization

In the model initialization phase, the input and output layers of the model as well as the parameters and structure of each intermediate layer are defined in detail (4 above). This includes the setup and initialization of key components such as the convolutional layer, pooling layer, batch normalization layer, and LSTM layer. By carefully designing the network architecture, we ensured that the model was able to fully utilize the information in the input features and efficiently learn the connection between the resonance peaks and the individual vowels.

5.3. Loss function definition

In order to evaluate the difference between the model prediction results and the real labels, we adopt the Cross-Entropy Loss function (CEL). This function can intuitively reflect the degree of deviation between the probability distribution predicted by the model and the true label. Specifically, the formula for the Cross-Entropy Loss function is as follows:

$$L = -(y log\hat{y} + (1-y) log(1-\hat{y}))$$

where y is the true label of the sample and $\hat{y}$ is the label predicted by the model. When y=1, it means that the sample belongs to the positive category; when y=0, it means that the sample belongs to the negative category. By minimizing the loss function, we are able to push the model to constantly approximate the true label,

thus improving its classification performance.

5.4. Optimizer setting

In the optimizer setting stage, the Adam optimizer is chosen and the learning rate strategy of exponential decay is adopted.The Adam optimizer combines the advantages of the two optimization algorithms, AdaGrad and RMSProp, and is able to adaptively adjust the learning rate of each parameter so as to accelerate the convergence process of the model. At the same time, the initial learning rate was set to 0.0006, the number of decay steps was 10000, and the decay rate was 0.98.

5.5. Model training

In the model training phase, the gradient descent method was used with the following parameters:

batch_size = 8

num_epoch = 20

process_num = 3

lr = 0.0006

feature_dim = 80

6. Vowel pronunciation error detection

This section discusses the method of accurately detecting and optimizing vowel mispronunciation by combining Mel-frequency cepstral coefficients (MFCC) and formant analysis.

6.1. Overall pronunciation accuracy evaluation based on MFCC

In the field of speech recognition, MFCC is recognized as one of the effective features that can characterize the characteristics of speech signals. To evaluate the accuracy of vowel pronunciation, we extract MFCC features of vowel audio clips, which are used to quantify the similarity between English learners and native speakers of vowel pronunciation. The overall scoring process is shown in the Figure 3.

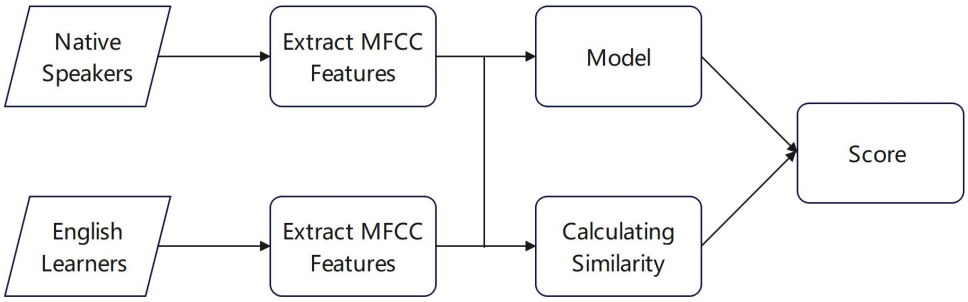

Figure 3 Pronunciation Accuracy Evaluation

Specifically, 80-dimensional MFCC features are extracted from audio clips of native speakers' vowels. This MFCC is used as input for model training.

Subsequently, the MFCC features extracted from the English learner vowel audio are evaluated using a pre-trained neural network model to generate a probability value that reflects the similarity between the English learner and native speaker vowel pronunciations. The probability value output by the model is converted into a percentage to provide quantitative feedback on pronunciation accuracy.

This method can keenly capture the subtle differences in vowel pronunciation and provide a reliable quantitative evaluation of vowel pronunciation accuracy.

6.2. Tongue position suggestion based on formants

Formants are important frequency components that characterize phoneme characteristics in speech signals, and are particularly critical for the identification of vowel pronunciation. Since the formant frequency directly reflects the changes in vocal tract shape and tongue position, specific optimization suggestions for vowel

pronunciation are provided by analyzing formants.

Accurately calculate the formant frequency of each vowel audio. Take the maximum and minimum formant values of the same vowel of native speakers as the standard formant range of each vowel. The relevant code is shown in Figure 4.

```
formant_dict = defaultdict(list)
for data in tqdm(train_data):
    audio_file = data[0]
    file_label = label[data[1]]
    audio_formant = get_formant(audio_file)[1]
    formant_dict[file_label].append(audio_formant[:3])

for key in formant_dict:
    sub_data = np.array(formant_dict[key])
    sub_data_max = np.max(sub_data, 0)
    sub_data_min = np.min(sub_data, 0)
    formant_dict[key] = np.array([sub_data_min, sub_data_max])
```

Figure 4 The standard formant range of vowel

Perform formant extraction analysis on the vowel pronunciation of English learners, and compare the results with the standard range. Based on the comparison results, generate specific tongue position adjustment suggestions. For example:

(1) If F1 is lower than the standard value, lower the tongue slightly.

(2) If F1 is higher than the standard value, raise the tongue slightly.

(3) If F2 is lower than the standard value, stretch forward the tongue .

(4) If F2 is higher than the standard value, retract the tongue.

The relevant code is shown in Figure 5.

```python
for audio_file in test_data:
    file_name = os.path.splitext(os.path.split(audio_file)[1])[0].split('_')[0]

    audio_feature = extract_feature(audio_file)
    infer_result = infer_model(audio_feature[np.newaxis, :, :])[0].numpy()

    audio_formant = get_formant(audio_file)[1][:3]
    audio_formatn_label = formant_dict[file_name]

    result = []
    for i in range(3):
        if i == 0:  # F1
            if audio_formant[i] < audio_formatn_label[0][i]:
                result.append('Suggest to lower the tongue slightly.')
            elif audio_formant[i] < audio_formatn_label[1][i]:
                pass
            else:
                result.append('Recommend that the tongue be slightly raised.')

        elif i == 1:  # F2
            if audio_formant[i] < audio_formatn_label[0][i]:
                result.append('It is recommended that the tongue be stretched forward.')
            elif audio_formant[i] < audio_formatn_label[1][i]:
                pass
            else:
                result.append('It is recommended that the tongue be retracted.')

    print(f"{audio_file}    grade = {infer_result[label.index(file_name)] * 100}    propose = {','.join(result)}")
```

Figure 5 Tongue position suggestion

Through the above analysis, we can provide accurate tongue position adjustment suggestions for vowel pronunciation based on the changes in formant frequency. These suggestions help learners better master the correct pronunciation techniques and thus improve their vowel pronunciation.

7. Discussion

7.1. Experimental results

The use of ASR was core to this study. The ASR system allowed for the precise analysis of pronunciation errors by extracting MFCC and formant frequencies from the speech samples. This technological approach provided detailed insights into the acoustic characteristics of vowel pronunciation errors that would be challenging to obtain through traditional phonetic analysis alone.

Our hybrid neural network model, combining CNNs and LSTM networks, demonstrated high efficacy in detecting pronunciation errors. The CNNs effectively captured spatial features of the speech signal, while the LSTMs handled the temporal

dependencies, providing a framework for accurate error detection. This model's integration of Res2Net and Conformer architectures further enhanced its performance by leveraging multi-scale feature representations and self-attention mechanisms, allowing for more nuanced detection of pronunciation discrepancies. Then, the model provides personalized pronunciation improvement suggestions based on the learners' pronunciation data. This helps learners correct pronunciation errors more accurately, understand their own pronunciation situation in a timely manner, and conduct targeted practice.

The vowel data of English learners comes from the pronunciation samples recorded by myself. It is recorded using the professional audio analysis tool Praat with a sampling rate of 16khz, and then uploaded to the system for analysis. The following scores and suggestions are obtained (Figure 6).

Vowel	Score	Suggestion
/i/	84.57	Recommend that the tongue be slightly raised. It is recommended that the tongue be stretched forward.
/ɔ:/	98.23	Suggest to lower the tongue slightly.
/ɔ/	91.68	It is recommended that the tongue be retracted.
/u:/	95.45	Suggest to lower the tongue slightly. It is recommended that the tongue be stretched forward.
/u/	96.34	Suggest to lower the tongue slightly. It is recommended that the tongue be stretched forward.
/ə:/	89.5	Suggest to lower the tongue slightly.
/ə/	92.12	Suggest to lower the tongue slightly.
/ɑ:/	87.33	It is recommended that the tongue be retracted.
/ʌ/	84.22	Suggest to lower the tongue slightly. It is recommended that the tongue be stretched forward.
/e/	97.33	Recommend that the tongue be slightly raised.
/æ/	80.45	Suggest to lower the tongue slightly.

		It is recommended that the tongue be retracted.
/ɪ/	84.15	It is recommended that the tongue be retracted.

Figure 6 The score and suggestion of English learners' vowel pronunciation

From the experimental results, we can see that the pronunciation scores of most vowels are above 80 points, indicating that the learners' pronunciation is generally accurate. However, for some vowels such as /i:/, /ɪ/, /æ/ and /ʌ/, the learners' scores are relatively low, indicating that the pronunciation of these vowels needs further improvement. The study found that Chinese-speaking English learners exhibit significant difficulties in pronouncing English vowels, particularly in distinguishing between /i:/ and /ɪ/, and /ɜ:/ and /æ/. It confirm our hypothesis that the differences in phonological systems between Chinese and English contribute to pronunciation errors among Chinese learners. Our results align with Smith (2000), who also noted confusion between /i:/ and /ɪ/ among Chinese learners. However, unlike Jones (2001), who found minor issues with /æ/, this research revealed significant challenges with this vowel.

However, since this experiment only provides my personal pronunciation suggestions, it may be a unique error in my pronunciation and cannot represent the common problems of Chinese English learners.

7.2. Limitations

7.2.1. Limitations in the data set

Due to experimental time constraints, the dataset used in this study is relatively limited and small in scale. Specifically, the dataset only contains samples from two recorders. This extremely limited sample size has a significant impact on the generalization ability and performance of the model. In addition, both recorders are Americans, lacking diversity and representativeness, which may make the model unable to accurately capture and predict different dialects, ages, genders and other characteristics, thereby failing to fully evaluate the actual impact of these variables on

model performance. In order to more comprehensively evaluate the performance of the model, future research needs to construct a dataset containing more diverse samples to cover factors such as different dialects, ages and genders, so as to more accurately analyze the impact of these variables on model performance.

7.2.2. Incomplete coverage of vowels and consonants

The current speech learning model has obvious limitations in the recognition of vowel pronunciation errors. Although the model can cover and detect vowel pronunciation errors, its recognition range is limited to 11 single vowels, and it fails to extend to the recognition and correction of diphthongs and consonants. In speech learning, diphthongs and consonants also play an important role, and their correct pronunciation is crucial to the fluency and clarity of speech communication. Therefore, the model cannot provide learners with more comprehensive phonetic pronunciation guidance when vowels and consonants are not fully covered.

7.2.3. Lack of pronunciation duration suggestion

In terms of model structure, the existing model fails to fully consider the differences in pronunciation duration of different vowels. In speech learning, the pronunciation duration of short vowels and long vowels is one of the important features to distinguish between the two. However, the current model does not involve the comparison of short vowels and long vowels in the recognition process, so it is impossible to detect and suggest pronunciation duration. This leads to the fact that learners may not be able to accurately grasp the pronunciation duration of vowels when pronouncing, thus affecting the accuracy and naturalness of pronunciation. Therefore, in future model improvements, the function of detecting pronunciation duration should be added to provide more accurate pronunciation guidance.

7.3. Future work

7.3.1. Design and implementation of interactive interface

The current speech learning model is still in the basic stage, and its functionality is mainly focused on audio recognition. Although certain results have been achieved,

there is still a lot of room for improvement in user experience and learning efficiency. In order to overcome these limitations, it is particularly important to develop a user-friendly interactive interface. Through this interface, learners can easily record their pronunciation and get immediate feedback from the model. This instant interactive learning method can stimulate learners' interest, improve their learning motivation, and thus be more actively involved in learning.

In order to more intuitively demonstrate pronunciation skills, video learning functions can be integrated into the interface. Learners can watch video materials, observe the correct pronunciation posture and mouth shape, and understand the muscle movements during pronunciation. This audio-visual learning method can help learners understand the pronunciation essentials more deeply and better master pronunciation skills.

After mastering the pronunciation skills, learners can also practice pronunciation through the recording function in the interface. At this time, the model will analyze the learner's pronunciation in real time and compare it with the standard pronunciation. Based on the analysis results, the model will give instant feedback and suggestions to help learners find and correct pronunciation errors in time. This instant feedback can quickly improve the learner's pronunciation accuracy and speed up the learning process.

7.3.2. Application of multimodal recognition technology

The accuracy of vowel pronunciation is crucial for speech learning. However, traditional audio recognition methods often have limitations when dealing with vowel pronunciation errors. To overcome this challenge, exploring the application of multimodal recognition technology in vowel pronunciation error recognition is one of the future development directions.

By analyzing the learner's video data, lip shape features such as the roundness of the lips can be extracted. These features can provide more specific guidance for vowel pronunciation. For example, the pronunciation of some vowels requires a specific roundness of the lips. By comparing the learner's lip shape with the standard lip shape, we can give more precise pronunciation suggestions.

Through the application of multimodal recognition technology, we can help learners master pronunciation skills more accurately, thereby improving the accuracy and naturalness of pronunciation. This is of great significance to improving learners' speech communication ability.

7.4. Model evaluation

The evaluation of the system in this paper can be comprehensively evaluated from both objective and subjective levels.

7.4.1. Objective evaluation

In order to objectively evaluate the performance of the automatic vowel mispronunciation recognition model, we used F1 score, accuracy (acc) and recall as indicators.

The F1 score of the model reached 92.7%, indicating that the model has high accuracy and reliability in identifying vowel mispronunciations, and can balance precision and recall well. The accuracy is close to the F1 score, which is 92.5%, further verifying its effectiveness in the vowel mispronunciation recognition task. The recall rate is 92.4%, indicating that the model can more comprehensively identify pronunciation errors and reduce the number of missed reports.

7.4.2. Subjective evaluation plan

In order to have a more comprehensive understanding of the performance of the model in practical applications, it is planned to establish a multi-expert scoring

mechanism to evaluate the scores and suggestions given by the system.

Establish scoring criteria: First, we will develop detailed scoring criteria, including the type and severity of pronunciation errors, as well as the accuracy and practicality of the suggestions. This will provide a clear basis for experts to score.

Recruitment of expert team: Next, we will recruit a certain number of native English speakers or phoneticians to form an expert team. They need to have rich experience in teaching vowel pronunciation and sensitivity to pronunciation errors.

Conduct scoring experiments: After the expert team is formed, we will ask them to score the vowel pronunciation errors identified by the system and evaluate the suggestions given by the system.

Summarize and analyze the scoring results: After completing the scoring experiment, we will summarize and analyze the scoring results of the expert team and calculate statistics such as the average score and standard deviation of each indicator. By comparing the expert's scoring results with the system's score, the system performance can be evaluated.

By evaluating the automatic vowel pronunciation error recognition model at both objective and subjective levels, we can fully understand its performance. The objective evaluation results show that the model has high accuracy and reliability in identifying pronunciation errors; while the subjective evaluation plan will provide more information about the performance of the system in actual applications. Combining the evaluation results of both aspects, the model can be targeted and optimized to improve its performance and practicality in the vowel pronunciation error recognition task.

8. Conclusion

As the core of English pronunciation, vowels play a decisive role in clear and natural oral pronunciation. However, due to the huge differences between Chinese and

English in the phonetic system, Chinese English learners often encounter difficulties in vowel pronunciation. Although the existing English pronunciation teaching methods are helpful to a certain extent, there are still many limitations. Although the traditional method of teacher demonstration, student imitation and error correction can provide certain pronunciation guidance, it is difficult for teachers to provide personalized feedback and to accurately correct the specific problems of each student.

Therefore, the improvement strategy for the vowel pronunciation problems of Chinese English learners should pay more attention to personalization, systematicness and scientificity. Therefore, this study uses MFCC and formant features to establish a model, automatically output the students' vowel pronunciation scores and precise tongue position suggestions, and provide personalized feedback and guidance.

At the same time, the model also has good scalability and adaptability. With the continuous advancement of technology and the continuous accumulation of data, the model can be continuously optimized and improved to improve its recognition accuracy and feedback quality. At the same time, the model can also be applied to more teaching scenarios and fields, such as interactive oral teaching platforms, to provide English learners with more comprehensive and efficient pronunciation learning support.

Reference

[1] Chapelle, C. (1997). CALL in the year 2000: Still in search of research paradigms? Language Learning and Technology, 1(1), 19-43. Retrieved March 2, 2007.

[2] Chappelle, C. A. (2001). Innovative language learning: Achieving the vision. ReCALL, 13(1), 3-14.

[3] Cao, X. (2016). A study on the pronunciation problems of junior high school English [Master's thesis, Chongqing Normal University].

[4] Chen, X. (2010). A study on the pronunciation of English vowels by Chinese learners [Master's thesis, Shanghai Jiao Tong University].

[5] Dai, C., Gu, M., & Miao, X. (2019). Errors in pronunciation of English monophthongs in three dialect regions of Jiangsu and their acoustic characteristics. Journal of Huaiyin Normal University (Philosophy and Social Sciences Edition), (06), 635-640.

[6] En-Minh, L. (2010). A study on the attitude of English-majored students toward computer-assisted pronunciation learning at a technological university. Effects of an on-line syntactic analysis strategy instruction on university students' reading comprehension of English science texts, 16, 155.

[7] Eurotalk. (2002). Retrieved February 27, 2002, from http://www.eurotalk.co.uk.

[8] Ferrier, L., & Reid, L. (2000). Accent modification training in The Internet Way®. Proceedings of InSTILL 2000, 69-72.

[9] Flege, J. E. (1987). Effects of equivalence classification on the production of foreign language speech sounds. Sound Patterns in Second Language Acquisition, 9-39.

[10] Gao, Y., & Gong, H. (2011). An experimental study on the native language transfer in Chinese students' acquisition of English vowels. Journal of Bohai University (Philosophy and Social Sciences Edition), (01), 126-132.

[11] Gleamer. (2001). Retrieved May 10, 2001, from http://www.gleamer.com.

[12] Jiang, Y. (2010). An acoustic experimental study on the vowel pronunciation of English learners in Min and Wu dialect areas. Foreign Languages Research, (04),

36-40.

[13] Jones, R. H. (1997). Beyond “listen and repeat”: Pronunciation teaching materials and theories of second language acquisition. System, 25(1), 103-112.

[14] Kendrick, H. (1997). Keep them talking! A project for improving students' L2 pronunciation. System, 25(4), 545-560.

[15] Lambacher, S. (1999). A CALL tool for improving second language acquisition of English consonants by Japanese learners. Computer Assisted Language Learning, 12(2), 137-156.

[16] Levy, M. (1997). Computer-assisted language learning: Context and conceptualization. Oxford University Press.

[17] Massaro, D. W., & Simpson, J. A. (2014). Speech perception by ear and eye: A paradigm for psychological inquiry. Psychology Press.

[18] Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. TESOL Quarterly, 25(3), 481-520.

[19] Nieuwe Buren. (2002). Retrieved February 26, 2002, from http://www.nieuweburen.nl.

[20] Neri, A. M. B. R. A., Cucchiarini, C., & Strik, H. (2006). Improving segmental quality in L2 Dutch by means of computer assisted pronunciation training with automatic speech recognition.

[21] Pennington, M. C. (1999). Computer-aided pronunciation pedagogy: Promise, limitations, directions. Computer Assisted Language Learning, 12(5), 427-440.

[22] Pillai, S., Mohd. Don, Z., Knowles, G., & Tang, J. (2010). Malaysian English: An instrumental analysis of vowel contrasts. World Englishes, 29(2), 159-172.

[23] Pro-nunciation. (2002). Retrieved February 26, 2002, from http://users.zidworld.com.au/pronunce/products.html.

[24] Pujolà, J. T. (2001). Did CALL feedback feed back? Researching learners’ use of feedback. ReCALL, 13(1), 79-98.

[25] Rogers, C. L., & Dalby, J. M. (1996). Prediction of foreign-accented speech intelligibility from segmental contrast measures. The Journal of the Acoustical Society of America, 100(4_Supplement), 2725-2725.

[26] S. H., & Liu, C. (2014). Intelligibility of American English vowels and consonants spoken by international students in the United States. Journal of Speech, Language, and Hearing Research, 57(2), 583-596.

[27] Xie, J. (2014). Negative transfer of Henan dialect on English pronunciation acquisition and teaching strategies. Journal of Inner Mongolia Normal University (Education Science Edition), (12), 100-102.

[28] Yang, C., & Robb, M. (2000). Acoustic features of vowel production in Mandarin speakers of English. In The Proceedings of the 6th International Conference on Spoken Language Processing (Volume II) (pp. 669-672).

[29] Zhang, J. (2002). A comparison of English and Chinese vowels and English phonetics teaching. Journal of PLA University of Foreign Languages, (01), 56-59.

[30] Zhu, W. (2012). The negative transfer effect of Central Plains Mandarin in northern Anhui on English pronunciation acquisition. Journal of Inner Mongolia Agricultural University (Social Sciences Edition), (04), 380-381.