Hindawi
Journal of Sensors
Volume 2021, Article ID 5946228, 12 pages
https://doi.org/10.1155/2021/5946228

Research Article

Design of an Automatic English Pronunciation Error Correction System Based on Radio Magnetic Pronunciation Recording Devices

Zhang Shufang

Anyang Vocational and Technical College, Anyang Henan 455000, China

Correspondence should be addressed to Zhang Shufang; 2011020103@st.btbu.edu.cn

Received 8 November 2021; Accepted 3 December 2021; Published 27 December 2021

Academic Editor: Gengxin Sun

In this paper, a system for automatic detection and correction of mispronunciation of native Chinese learners of English by speech recognition technology is designed with the help of radiomagnetic pronunciation recording devices and computer-aided software. This paper extends the standard pronunciation dictionary by predicting the phoneme confusion rules in the language learner's pronunciation that may lead to mispronunciation and generates an extended pronunciation dictionary containing the standard pronunciation of each word and the possible mispronunciation variations, and automatic speech recognition uses the extended pronunciation dictionary to detect and diagnose the learner's mispronunciation of phonemes and provides real-time feedback. It is generated by systematic crosslinguistic phonological comparative analysis of the differences in phoneme pronunciation with each other, and a data-driven approach is used to do automatic phoneme recognition of learner speech and analyze the mapping relationship between the resulting mispronunciation and the corresponding standard pronunciation to automatically generate additional phoneme confusion rules. In this paper, we investigate various aspects of several issues related to the automatic correction of English pronunciation errors based on radiomagnetic pronunciation recording devices; design the general block diagram of the system, etc.; and discuss some key techniques and issues, including endpoint detection, feature extraction, and the system's study of pronunciation standard algorithms, analyzing their respective characteristics. Finally, we design and implement a model of an automatic English pronunciation error correction system based on a radiomagnetic pronunciation recording device. Based on the characteristics of English pronunciation, the correction algorithm implemented in this system uses the similarity and pronunciation duration ratings based on the log posterior probability, which combines the scores of both, and standardizes this system scoring through linear mapping. This system can achieve the purpose of automatic recognition of English mispronunciation correction and, at the same time, improve the user's spoken English pronunciation to a certain extent.

1. Introduction

Language is the most natural tool for human communication, and the automatic processing of speech-language information is an important research area in information science. Among them, the more important research directions include large-scale continuous speech recognition natural language understanding, speech synthesis, and machine translation [1]. Human-machine speech interaction is a human-machine dialogue technology based on speech recognition, natural language understanding, and speech synthesis. Speech synthesis is one of the cores of human-computer interaction. It is involved in many disciplines, such as acoustics and natural language processing, artificial intelligence, and signal processing. In recent years, speech information processing has developed more rapidly, and spoken pronunciation detection is one of the important research directions. Speech interaction is the most direct, natural, and effective way people use to convey information, and with the rapid development of mobile phones and other intelligent terminal products in recent years, new human-computer interaction has become a hot spot in scientific research of computers, linguistics, and communications [2]. Human-computer speech interaction is a human-computer dialogue technology based on speech recognition, natural language understanding, and speech synthesis. Speech synthesis is one of the cores of human-computer interaction, which is involved in several disciplines, such as acoustics, natural language processing, artificial intelligence, and signal processing. A lot of English learning is carried out in the "vacuum" of the "nonlinguistic environment." Once it enters the communicative state, it is more susceptible to the inevitable pronunciation defects or poor pronunciation in the real context. Its purpose is to enable computers or other hardware devices to make natural sounds like people. In this environment, to make mobile phones, computers, and other intelligent terminal devices be completely like people that can "speak" and "listen," can understand the natural language of humans, and can get some feedback or according to the instructions to complete the corresponding operation is one of the goals of the current scientific artificial intelligence field [3]. However, for the current TTS (Text-To-Speech) system, it is the main research direction of many enterprises and universities to make it produce clear, understandable, fluent, and natural voices in different scenes to better meet the personalized needs of users.

Language learning generally includes four areas: listening, speaking, reading, and writing, each of which has its method of learning. For example, listening comprehension skills can be improved by listening to various foreign language multimedia resources, including news, movies, and audiobooks. We can also regularly read foreign language newspapers, professional papers, world famous books, and other textual resources to practice reading skills and to obtain information in other languages [4]. Many people tend to be fluent in reading and writing, but their oral English is poor, which further affects the improvement of listening. English writing can be practiced through journaling, translating English resources, etc. There is no good way to improve the ability to "speak"; although it can be practiced in "English corner" or similar places, it is very limited. In addition, a lot of English learning is done in the "vacuum" of a "nonlinguistic environment," and once you enter the real communication situation, you are more likely to be affected by the inevitable pronunciation defects or inappropriate pronunciation factors in the real context. Because "speaking" is an interactive process, it cannot be trained alone but must be interacted with. For these reasons, "speaking" often becomes a bottleneck for language learners.

Computers have brought great convenience to humans due to their powerful information processing, computing, and storage capabilities. Speech recognition technology has been developed over the years and is now starting to gradually come into different applications. Research on language learning and spoken pronunciation detection has received increased attention in recent years, and the application of speech recognition in computer-assisted language learning has become an important research direction. Especially, audio as an information medium plays an important role in the process of human-computer interaction [5]. Therefore, the study of English pronunciation monitoring and automatic correction is not only of theoretical significance but also of great help to the language learning of nonnative learners; through the detection of learners' pronunciation in language learning, it can help learners understand their pronunciation accuracy and improve their speaking level [6].

2. Related Works

The purpose of oral pronunciation testing is to provide a mechanism for learning foreign languages such as English to automatically correct country pronunciation. Many people tend to read and write fluently but speak poorly, which affects listening even more [7]. The key to improving listening and speaking skills is to practice speaking and to receive guidance and correction from English-speaking teachers, but the lack and high cost of English-speaking teachers in China leave many learners without opportunities to practice and improve, and often, after eleven years of study, they are still unable to communicate with foreigners, neither speaking nor understanding. Make use of the error rules in the learner's pronunciation and integrate these rules into the speech recognition to detect and diagnose the possible error categories in the learner's phoneme pronunciation. Given the importance of oral practice, the main means of practice for learners is to play tapes and other recording media repeatedly, and the advent of the repeater has introduced electronic English learning products to the market. The development of teaching aids using electronic and computer technology has become a key step in the transition from basic research to products, and many learning machines have received strong support and input from national education and science and technology departments [8].

The main development of pronunciation error detection as part of the CAPT system came after the 1990s. In 1996, a pronunciation scoring algorithm for speech-interactive language learning systems was proposed that combined Hidden Markov Similarity, sentence length, segment length, and segment classification to calculate scores. The important difference between this algorithm and earlier algorithms is that the content of the sentence or phrase to be read aloud by the person to be tested does not have to be specified, making it more flexible to use. The recognition system implemented the algorithm and evaluated French pronunciation in native English speakers, and experimental results showed that the duration score for the same segment was an important indicator of pronunciation fluency and was robust to background noise. In 1997, the pronunciation evaluation algorithm was improved by scoring multiple sentences from a given pronouncer and then averaging them to obtain a higher-level score, while combining different machine scores to obtain a higher correlation coefficient [9]. Experiments show that the improved algorithm requires less speech to be tested on sentence-level scores, increasing the human-machine score correlation coefficient from 0.5 to 0.88, and by combining different machine scores, increasing the human-machine score correlation coefficient by 7%. In 2000, by calculating the confidence (CM) measures derived from the Hidden Markov Model- (HMM-) based ASR system for phoneme measures (CM) for phoneme-level articulation error detection. In 2003, the articulation error detection method was improved based on the GOP algorithm. The improved method shows that the phoneme detection recognizer can determine the correct pronunciation rate, and the lower the CM, the higher the probability of incorrect pronunciation of the speech [10]. The advantage of these CMs is that they are readily available through the ASR system; however, when analyzing individual voices, the correlation between CMs and human judgments is low over a relatively long range of speech sounds. At the phoneme level, it was found that the lack of features of CM resulted in a low correlation between assessment levels and human judgments and that these features were computed algorithmically using similar feature sets of speech sounds and were not suitable for performing pronunciation error detection. In 2004, when studying pronunciation errors in Dutch as a second language, it was found that learners of Dutch had pronunciation problems in terms of vowel length, a problem that suggests that pronunciation errors resulting from pronouncing phonemes different from the expected phonemes can cause deviations in word comprehension [11].

Since the study of automatic pronunciation detection is closely related to the study of linguistics, phonology, etc., the problems faced by different languages when learning another language such as English are different and the solutions must be targeted. In general, one of the mechanisms of oral pronunciation testing is the assessment of phonological accuracy, which has always been an important aspect of research. In early pronunciation tests, an acoustic model was created based on the standard phonetic pronunciation of native speakers, and then, the pronunciation of learners from nonnative speakers was tested [12]. Some studies have added expert speech from nonnative countries to the training data as well to improve judgments of the difficulty of pronunciation of phonemes. Many studies have only evaluated the pronunciation of limited words with limited phonemes, and less research has been conducted on the detection of continuous natural speech with larger vocabularies. We believe that it is difficult to obtain standard phoneme pronunciation scores by merely applying forcing regularization. Since the native Chinese learners of English are too far from the standard English, pronunciation recognition may fail to obtain valid phoneme pronunciation accuracy. Some studies use the output of speech recognition and regularized acoustic model scores as phoneme pronunciation scores, which are useful to reasonably assess the accuracy of phoneme pronunciation [13].

3. Design of an Automatic English Pronunciation Error Correction System Based on Radiomagnetic Pronunciation Recording Devices

3.1. Automatic Calibration System Model Design. In this paper, we propose a kind of automatic system for detecting incorrect phoneme pronunciation for continuously spoken pronunciation for English learners. The core idea of the method is to use the error patterns existing in learners' pronunciation to detect and diagnose the possible error categories in learners' phoneme pronunciation by incorporating these patterns into speech recognition. Three main problems are faced in the methodology. From the phoneme pronunciation, we summarize the typical pronunciation error rules through cross-language phonological comparison and analysis. This law is the form of confusion rules from phoneme to phoneme to predict the learner's possible wrong phoneme pronunciation.

(a) How to summarize the error pattern. It is very difficult and unnecessary to make a summary analysis of each situation one by one. The method of this paper is to summarize the cases of errors that are common in learners' pronunciation and are regular and extended. The confusion rules are represented in the form of confusion rules

(b) How to design a speech recognition system where error laws are effectively combined with speech recognition as a priori knowledge to detect and diagnose mispronunciation. Mispronunciation detection is mainly oriented to users who are nonnative speakers, and recognition requires accuracy down to the phoneme, which places high demands on the system design. The accuracy of speech recognition will be improved if error laws are integrated into speech recognition to reduce the burden on the recognizer and do recognition in a recognition range with a priori knowledge

(c) How to provide corrective feedback information. This is a basic human-computer interaction problem intelligent problem of the wrong pronunciation detection system and an important part of the system. More reasonable and intuitive feedback can make learners understand more quickly and correct pronunciation errors and achieve the purpose of computer-aided pronunciation training

For the different functional structures and approaches of the three problems, this chapter divides the system into three modules extended pronunciation dictionary generation module, speech recognition module, pronunciation detection, and feedback module; the overall system structure design is shown in Figure 1.

The main pronunciation problems faced by language learners are the inaccurate pronunciation of phonemes, inappropriate stress and intonation, and nonfluent and continuous pronunciation. The causes are mainly categorized into the following three types: (1) differences in linguistic, phonological, and phonetic pronunciation structures between learners' native language and the target language and differences in the functioning force of the articulatory organs; (2) the learner's misunderstanding of linguistics, phonology, and phonemes, not knowing the continuum, or misunderstanding the rules of letter pronunciation. For the first two reasons, based on the theory of language transfer, this paper systematically analyzes the characteristics of English mispronunciation of native Chinese speakers from linguistics, phonology, and phonemic and finds that the errors are mainly concentrated in those phoneme pronunciations that are present but not in Chinese, and learners habitually substitute the pronunciation

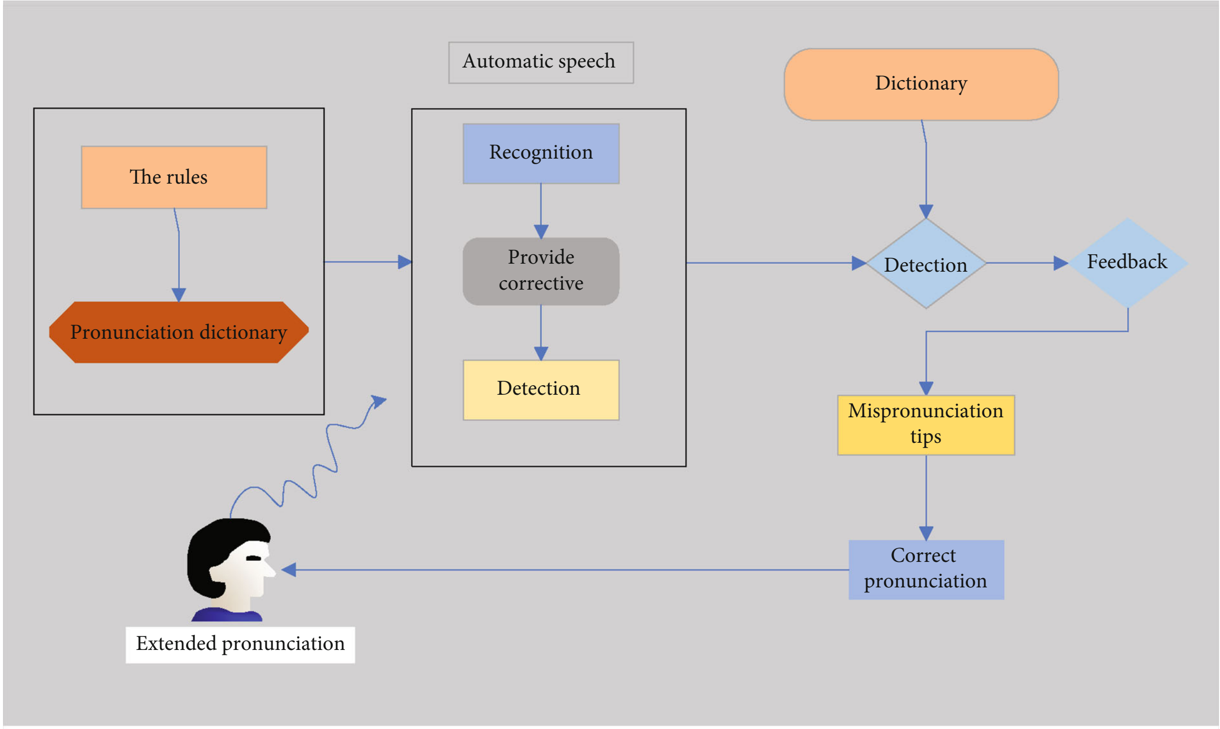

Figure 1: Overall system structure design diagram.

with the phoneme pronunciation of the native phoneme that is similar to this phoneme in linguistics and phonology, which leads to mispronunciation [14]. After the human voice is emitted from the lips, the high-frequency part will be attenuated, so that the energy of the low-frequency part is always higher than the energy of the high-frequency part, which results in a smaller spectral value of the high-frequency part, which is not convenient for analysis and processing. Preemphasis is to let the voice pass through a high-pass filter to enhance the high-frequency part so that the high- and low-frequency ranges are equal. Therefore, we generalize the typical pronunciation error patterns from phoneme pronunciation through crosslinguistic phonological comparative analysis of the differences, and this pattern is a form of phoneme-to-phoneme confusion rule to predict the possible wrong phoneme pronunciation of learners. While the third reason contains too many personal factors of the learner, Chimin's difficult mountain knowledge to predict, therefore, this paper adopts a data-driven approach that does not rely on a priori knowledge, using the identification of the learner's actual pronunciation errors to predict the possible wrong pronunciation by performing phoneme-based automatic speech-to-phoneme recognition on the learner's speech and analyzing the recognition results between the resulting wrong pronunciation phonemes and the standard pronunciation phonemes. The mapping relationships between the mispronounced phonemes and the standard pronunciation phonemes are analyzed.

$$S_n = \cos \frac{3\pi(n+1)}{N+1}. \tag{1}$$

Speech signal preprocessing is the preparation work before speech feature extraction, mainly for the frequency domain processing of speech signal features. After the analog speech signal is sampled and quantized into a digital signal, it needs to be preemphasized so that the high- and low-frequency amplitudes are equal and then, it is framed and windowed to get the speech frame. If the voice data is read directly from an audio file such as a file there is no need for sample quantization processing. Because the human voice is from the lips, the high-frequency part will be attenuated, so that the energy of the low-frequency part is always higher than the energy of the high-frequency part, which leads to a smaller spectral value of the high-frequency part, which is not easy to analyze and process. Preemphasis is to allow the speech to pass through a high-pass filter that enhances the high-frequency part, making the high- and low-frequency amplitudes comparable. Speech signals are slow time-varying signals with short-time smoothness. For a segment of the speech signal, if we take a short enough time (about 6~30 ms), we find that the characteristics of the segment remain the same, but from a long time (0.6 s or more), the speech signal characteristics keep changing, and from that, the content of what the speaker is supposed to express [15]. Because of this characteristic of speech, we need to divide the speech into several short-time segments for analysis, and this process is "framing." There is a certain overlap between two adjacent frames so that the continuity of speech features is maintained by smoothing the comparison between frames. Usually, the overlap is half or one-third of the frame length, and the size of the frame length is between 20 and 30 ms because the characteristics of the speech signal are more stable in this period. Assuming a signal sampling frequency of 16 KHz, a frame length of 25 ms, and a frame rate of 100 frames/second, there is one sample per frame, and the framed speech signal has 40,000 samples per second.

$$S_n = S_{n+1} - aS_{n-1}. \tag{2}$$

The key aspects of text analysis in a TTS are text-to-symbol conversion, including pauses, placement of stress, and hierarchical relationships, as well as standardization of the text, division of words, and determination of the correct pronunciation of the word or phrase in that position. Although the main task in this section is not text analysis, the key to constructing a complete TTS is the transformation of text to symbols, including pauses, placement of stress, and hierarchical relationships, as well as the standardization of the text, the division of words, and the determination of the correct pronunciation of the words and phrases in that position. Although the main task of this section is not text analysis, it is also crucial to sort out the working process of text analysis to construct a complete TTS system. The main function of text analysis is to enable the subsequent synthesis stage to correctly recognize the digital expression converted from text, similar to a code, and perform a shallow analysis of the text to a certain extent according to the corresponding relationship of the text in the sentence and understanding. The main function of text analysis is to enable the subsequent synthesis stage to correctly identify the digital expressions transformed from text, like an electrical code, and to perform a somewhat shallow analysis and understanding of the text based on its correspondence in that sentence, resulting in the determination of how words and conjunctions in the text should sound, what rhyme is needed based on sentence characteristics, the interval based on semantics, and so on. These parameters will be passed on to the back end of the parametric processing process and play a large role in the effectiveness of the synthesis. The study of the whole text analysis can be divided into the following stages:

(1) Standardize the input text, find gaps and typos, and remove illegitimate characters appearing in the corpus and wrong word composition; conversion of letters or Arabic numerals for which Chinese pronunciation exists to their Chinese counterparts

(2) A participle, which can split the text according to verbs, nouns, conjunctions, and other forms

(3) Level the pauses in the utterance, and mark them based on information such as participles and punctuation

The block diagram of the text analysis system is shown in Figure 2.

In phonetic error recognition detection, each mispronounced phoneme may be due to the insertion, substitution, and deletion of new phonemes. And most of the mispronunciation problems of nonnative learners also arise due to phoneme confusion. The phonetic features of the phonemes corresponding to the associated phoneme strings, phoneme onset and end times, and ratings were obtained after recognition and forced alignment by the Sphinx system. With this resultant data, phonemes are detected for errors. Witt classifies articulation errors into two types, namely, phoneme errors and rhyme errors, and further classifies phoneme errors into three categories: mispronunciation, omission, and addition of phonemes. Generally, the recognizer only performs one recognition detection process for phoneme sequences, and after phoneme alignment, the recognizer performs one recognition process for phonemes from left to right and outputs the recognition results, but the problem of missed and false detection often occurs. To improve the correctness and accuracy of phoneme recognition detection, this paper proposes a phoneme cyclic recognition detection strategy, which converts the speech to be tested into a feature vector after feature extraction and then expands the phoneme bias pronunciation network into each phoneme recognition state, and the Sphinx recognizer performs the cyclic detection task twice for the phoneme feature vector and phoneme bias network to obtain the recognition results. The phoneme sequence SIL, K, AE, T, and SIL is obtained after phoneme alignment for the single word "cat," and the duration d and acoustic score a feature vectors of the phoneme are generated. Then, for each, three T and two D phoneme groups are identified several times, respectively, and the aligned sequences are subjected to phoneme substitution, insertion, and deletion. And to further determine the error type of the phoneme, the duration feature of the phoneme is also identified. The examination process of the cyclic recognition strategy is shown in Figure 3.

3.2. Research on English Pronunciation Detection Methods. The accuracy of pronunciation detection as the basis of speech intelligibility evaluation in this paper is directly related to the effectiveness of the speech intelligibility evaluation system. Detecting pronunciation errors and providing feedback on the error information can help learners to improve the intelligibility level of speech. In the current pronunciation recognition detection, due to the diversity of learners' pronunciation errors, similar phonemes are easily confused, which is likely to cause the situation of missing and false detection in recognition [16]. The detection methods based on mispronunciation networks are being intensively researched and applied, and this chapter proposes a phoneme recognition detection strategy based on the construction of phoneme biased pronunciation networks, which uses recognizers randomly in a cycle and performs pronunciation error differentiation detection by SVM. This chapter focuses on improving the phoneme recognition detection method for nonnative learners to improve the recognition rate and accuracy of phoneme phonetic features and to provide sufficient and accurate phoneme recognition features for subsequent speech intelligibility evaluation. To improve the correct rate and accuracy of phoneme recognition and detection, this paper proposes a phoneme cycle recognition and detection strategy. After feature extraction, the voice to be tested is converted into feature vectors, and then, the phoneme error pronunciation network is expanded into each phoneme. In the recognition state, the Sphinx recognizer performs two rounds of detection tasks on the voice feature vector to be tested and the phoneme error network to obtain the recognition result. Pronunciation error

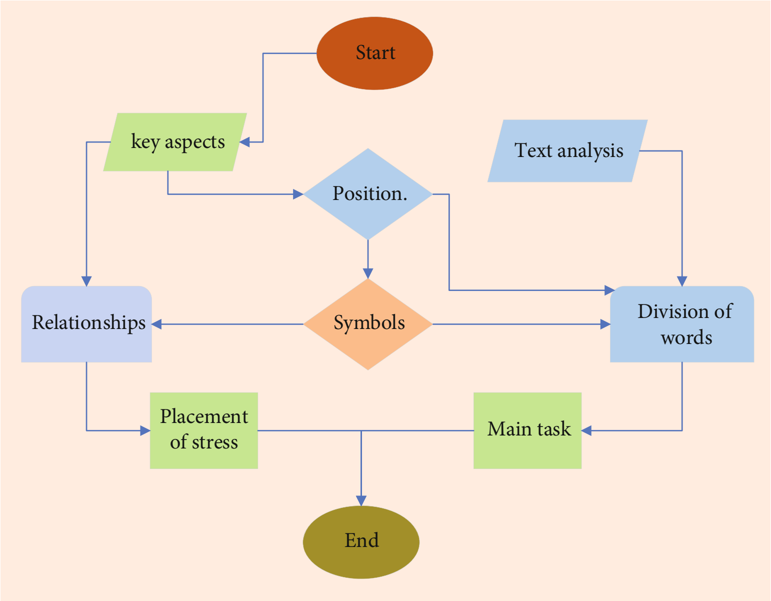

Figure 2: Flowchart of text analysis.

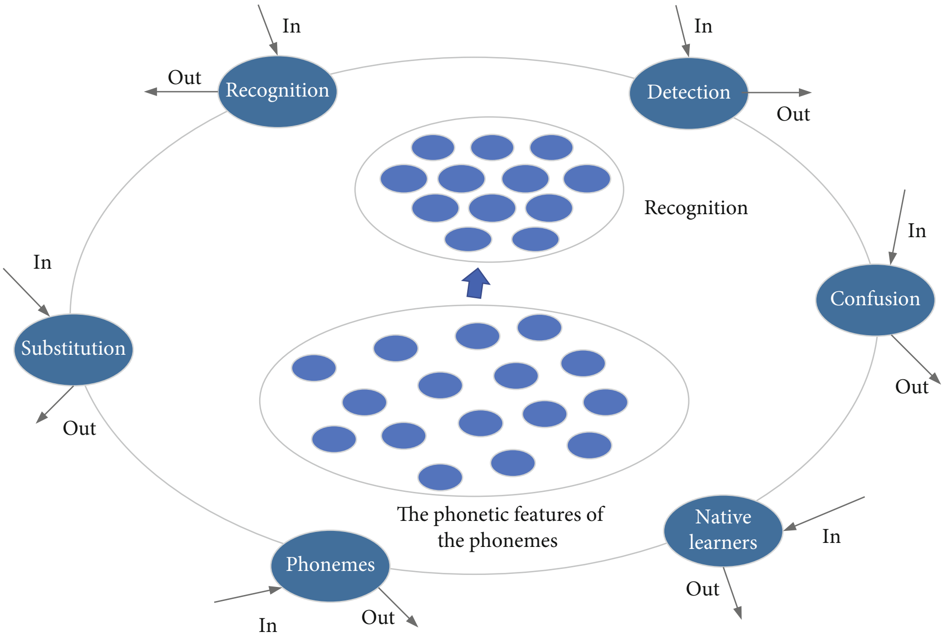

Figure 3: Schematic diagram of the process of phoneme identification test strategy.

detection requires a higher level of detail than pronunciation grading and is typically based on more phonological features such as temporal features, speech rate, articulation rate, and segmental duration that can be computed relatively more easily, and such detection of phonological temporal features is more reliable when measured for longer speech segments and has a greater correlation with human judgments of pronunciation quality. Pronunciation detection grading is often used to calculate the level of pronunciation scores at the speaker or discourse level and can also be a weighted average of native phoneme scores. And the simplest method of pronunciation error detection is to use posterior probability algorithms or GOP algorithms to define error detection by setting bounds.

$$\mathrm{Thresh} = u_p + a\varepsilon_p + \beta. \tag{3}$$

The purpose of speech signal endpoint detection is to detect speech signal segments and noise segments from the digital signal obtained by continuous sampling. Accurate speech endpoint detection not only reduces the computational effort but also improves the recognition rate of the system. Therefore, endpoints, as important features of speech segmentation, largely affect the performance of speech recognition systems, and thus, how to design a robust endpoint detection algorithm in a noisy environment is still a very tricky problem. Traditional endpoint detection algorithms rely on only one feature, such as signal energy, overzero rate, duration, and linear prediction energy error. These methods have good performance at high signal-to-noise ratios, but poor performance at low signal-to-noise ratios. Speech processing systems usually operate in different noise environments, and the endpoint detection methods used in the system should apply to various adverse situations to enable the system to achieve good performance [17]. First, a higher threshold amp1 is chosen based on the speech short-time energy, which is above this threshold in most cases. This allows for a coarse judgment: the speech start and endpoints lie outside the time interval corresponding to the intersection of this threshold and the short-time energy envelope. Then, a lower threshold amp2 is determined based on the average energy of the background noise, and the two points C and D where the short-time energy envelope intersects the threshold amp2 for the first time are searched forward from point A and backward from point B, respectively, so that the CD segment is the speech segment determined by the double threshold method based on the short-time energy, completing the first level of judgment. The second level of judgment is then performed, this time employing a threshold determined by the short-time excess zero rate. From points C and D, we search forward and backward, respectively, to find the two points E and F where the short-time average zero rate is below the threshold for the first time, which are the starting and ending points of the judged speech segment. This is shown in Figure 4.

The standard pronunciation models and grading models of computer-aided spoken English learning systems are obtained by corpus training. The system usually needs two types of speech databases, the standard pronunciation corpus and the nonstandard pronunciation corpus. The former of them is mainly used to train the standard pronunciation model, and the training corpus should be made to ensure the main training of the pronunciation content of spoken English learning as much as possible, and the content of the corpus is mainly obtained from several famous international corpora. The nonstandard pronunciation corpus is used to train the grading scoring model by experts manually and to test the system performance and should be widely representative. The proposed speech intelligibility evaluation method is compared with expert scores for a correlation experiment. After that, the evaluation method in this article is compared with other existing speech intelligibility evaluation methods, and finally, the intelligibility of this article is analyzed. The scoring performance of the degree evaluation method is analyzed. The content of its corpus is given by the experts, and the targets for grading scoring judgments differ according to the learning priorities of the users at different learning stages. In the study of speech recognition-based English-speaking learning systems, some focus on the common pronunciation errors of beginning pronouncers, such as various similar pronunciations and nasal sounds; some focus on pronunciation skills or difficulties specific to English speakers, such as intonation, alliteration, and stress. There is also one that focuses on a whole system of learning spoken English, following the phonetic teaching method combined with computers to make the system user-friendly and optimize its performance. Of the above, it makes sense to conduct an in-depth study of a particular problem in learning spoken English, for example, synchronic pronunciation, intonation, and intonation. Simply solving one of these problems applied to a spoken language learning system can make the system function optimally. Speech recognition is the key to performing pronunciation learning, but it is not fully suitable for English spoken pronunciation learning and many improvements are needed.

4. Analysis of Results

4.1. Automatic Calibration System Implementation. While the three raters were scoring manually, the author proposed to score the same speech documents by using speech evaluation technology. Based on the analysis of speech evaluation technology principles and speech evaluation cases, I found that Xunfei is the most advanced in the field of Chinese speech evaluation and provides free technical support for speech evaluation to the researcher, so I finally decided to use the speech evaluation function of Xunfei Open Platform to achieve the scoring of all test speech samples (hereinafter referred to as "technical scoring"). The Xunfei Open Platform provides speech evaluation technology [18]. The Xunfei Open Platform provides the speech evaluation technology SDK and explains the format of test questions, evaluation results, and frequently asked questions in the developer documentation, which provides great convenience for setting up the technical scoring environment. In the process of testing the technical scoring environment, it was found that the assessment results were on a five-point scale, which did not match the scoring requirements of HSKK Repeat After Listening (2 points for beginners and 3 points for intermediates). Therefore, the technology scoring results were processed as follows: beginner test technology scoring results = speech scoring results $* 0.4$; beginner test technology scoring results = speech scoring results $* 0.6$. The final descriptive analysis of the beginner and intermediate technology scoring results was conducted, and the results are shown in Table 1.

As can be seen from the above table, the technical scoring results of the primary test were controlled between 0 and 2, with the average score in the high range and the standard deviation within 0.5; the intermediate technical scoring results were all distributed between 1 and 3, with the average score around 2.3 and the standard deviation within 0.5. All the data showed that the technical scoring results met the requirements of the topic scoring and the scores showed a concentrated and stable state in general. When compared with the manual scoring results, the results of the descriptive analysis of the technical scoring results were found to show a high degree of agreement with the manual scoring, a finding

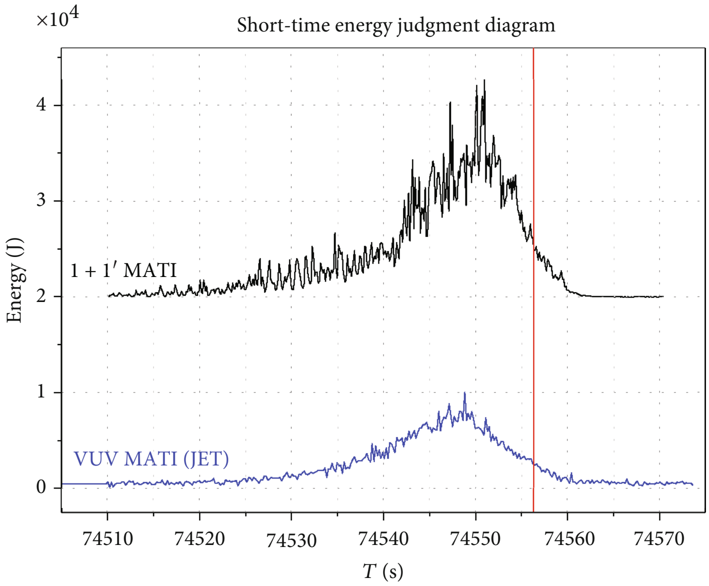

Figure 4: Short-time energy judgment diagram.

Table 1: Descriptive analysis of the results of scoring speech assessment techniques.

Level	Numerical	Minimal values	Maximum value	Average	Variance
Primary	200	0.01	1.50	0.075	0.243
Intermediate	210	0.05	1.00	0.050	0.233
Advanced	230	1.00	2.00	1.500	0.253

that well supports the conjecture of this study. However, this hypothesis needs further proof, so in Section 5, various aspects of the manual and technical scoring results are analyzed and compared to prove the research hypothesis one by one that speech assessment technology can complete the scoring of HSKK postlistening repetition questions.

$$D_j = \frac{L_j^{\mathrm{PL}} + L_j^{\mathrm{FA}}}{d_j}. \tag{4}$$

In the process of conducting experimental tests on speech intelligibility evaluation methods, an experimental database is used in this paper. Firstly, the speech intelligibility evaluation method proposed in this paper is compared with expert ratings for correlation experiments; after that, the evaluation method in this paper is compared with some other existing speech intelligibility evaluation methods, and finally, the scoring performance of the intelligibility evaluation method in this paper is analyzed. Meanwhile, this paper combines the phoneme bias pronunciation network to detect phoneme mispronunciation, and the phoneme error rate of each intelligibility level is counted to verify the effectiveness of the system's error correction feedback. The correlation between the proposed method in assessing intelligibility scores and each expert English teacher in the nonnative test set and TIMIT set is analyzed. Accurately detecting the endpoint of the voice signal can also reduce the amount of calculation for subsequent processing, and improving the utilization of communication equipment will help improve the recognition performance of the system. The experimental results show that the intelligibility evaluation method based on the combination of features proposed in this paper has a high correlation with the actual scores of human experts. The experimental results show that the combination feature-based evaluation method proposed in this paper outperforms the GOP scoring method and the AI index-based intelligibility evaluation method. This is mainly because the method proposed in this paper combines the information of both phoneme duration and phoneme acoustic score features and makes the most effective evaluation method calculation by optimizing the linear regression model. As shown in Figure 5.

For regression analysis in speech characteristics, in probability statistics, regression is the process of studying to estimate the relationship between different variables. Regression analysis studies the process by which the independent variable changes with the dependent variable and describes the trend of the dependent variable change by the characteristics of the probability distribution. And in speech signal analysis, regression analysis of feature vectors is used to explore the relationship between independent variables and dependent variables, which is generally applied to the attributes or parameters of speech for evaluation and prediction. By the

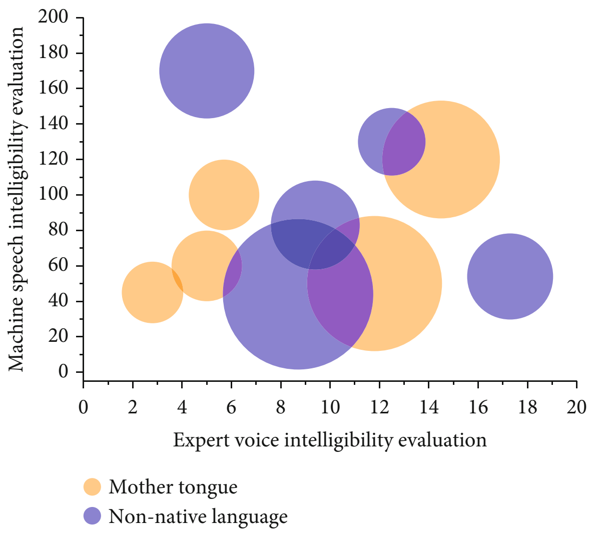

Figure 5: Machine speech intelligibility evaluation concerning expert evaluation.

relationship of different speech features, the correlation between their linear combination of features and phonology is sought, and this statistical perspective is useful for studies with sufficient correlation data. To estimate intelligibility scores, this paper uses a linear regression model, which is derived from the relationship between acoustic/phonological measures and expert scores. In this paper, independent variables (x_i) and variables Y are defined for each expert score, and the linear regression model is defined as shown in the following equation:

$$Y = \sum_{x=1}^{i} (\alpha_x + y_x) + \omega. \tag{5}$$

4.2. Analysis of Simulation Test Results. The endpoint detection is performed before feature extraction and recognition of the input speech. Experiments show that the accuracy of endpoint detection has a very important impact on the recognition rate of the speech signal; in addition, the accurate detection of the endpoints of the speech signal can also reduce the amount of computation for subsequent processing and improve the utilization of communication equipment to help improve the recognition performance of the system [19]. Endpoint detection is used to delineate the articulation and silence zones. The popular endpoint detection methods at this stage are based on short-time energy, based on short-time average overzero rate and pattern recognition, based on inverse spectral distance, based on wavelet transform, and based on other methods. In the prototype English-speaking learning system implemented in this paper, a double threshold endpoint detection method is used, which firstly sets two closed values each by short-time energy and overzero rate and obtains the endpoint detection method of speech signal content by a certain operation with the closed values. Exceeding the high threshold can basically determine the beginning of the voice, and the low threshold is used to determine the true endpoint of the voice. Exceeding the low threshold may not be the beginning of speech, and it may also be a short-term noise. Since the interval between the start of recording and the start of vocalization is generally considered to be the first 100 ms of the speech signal as a silent segment, the average energy and the average overzero rate of this segment of the speech signal can be extracted as the characteristic parameters when making a rough judgment. For the calculation of the threshold, a lower energy threshold is used, which is taken as two times the average energy of the background noise, and a higher energy threshold is taken as the average energy of multiple frames of speech data [20]. Exceeding the higher threshold identifies the beginning of the speech, and the lower threshold is used to determine the true endpoint of the speech. The low threshold being exceeded may not necessarily be the beginning of the speech but may also be short-lived noise. When the high threshold has determined the beginning of the speech, go back, and use the low threshold to determine the true beginning of the speech, and the end of speech is discriminated similarly. Sometimes, the noise is also quite energetic and may exceed the high threshold, but the noise which is generally of short duration can be used to determine whether it is noise or speech, as shown in Figure 6.

Determining the most effective error correction feedback decisions to help improve learners' phonetic intelligibility levels is the final issue to be addressed in this paper's evaluation system. In general, the overall level of pronunciation of nonnative learners when learning English pronunciation falls short of the standard pronunciation. However, for CAPT, if all the feedback is given to the learners without considering the impact value of the phonetic intelligibility of the wrong phoneme pronunciation, all the learners' key pronunciations will be judged as mispronunciations, which will weaken the learners' confidence in learning. The learner pronounces the phoneme sequence through the given lexical text, and after matching and forcing alignment with the phoneme biased pronunciation network, the possible biased pronunciation recognition sequence of the learner is obtained, where {} denotes the possible mispronunciation of phonemes. To determine which mispronounced phonemes should be improved, we define the priority $\pi(j, i)$ of the mispronounced phoneme j on intelligibility level i as the difference between the learner's error rate and the average error rate of the learners at level i as follows:

$$W(\mathrm{j}) = \sum_{j}^{2} \pi(i, j) \cdot l(i, 1). \tag{6}$$

Based on the results of phoneme detection, all potentially misleading phoneme reading detection rates are ranked, and by adjusting the priority of each phoneme to determine which phonemes are most in need of improvement, the best adjustment is made to obtain the intelligibility rating level for the entire word pronunciation. The problematic phoneme of the target word that has the greatest impact on phonological intelligibility is also given as positional feedback to the learner, informing the learner that improving the

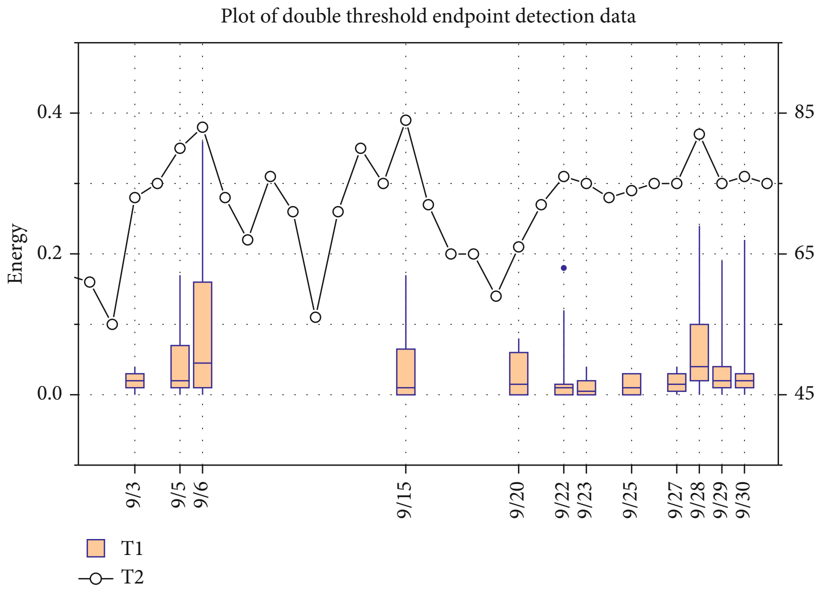

Figure 6: Plot of double threshold endpoint detection data.

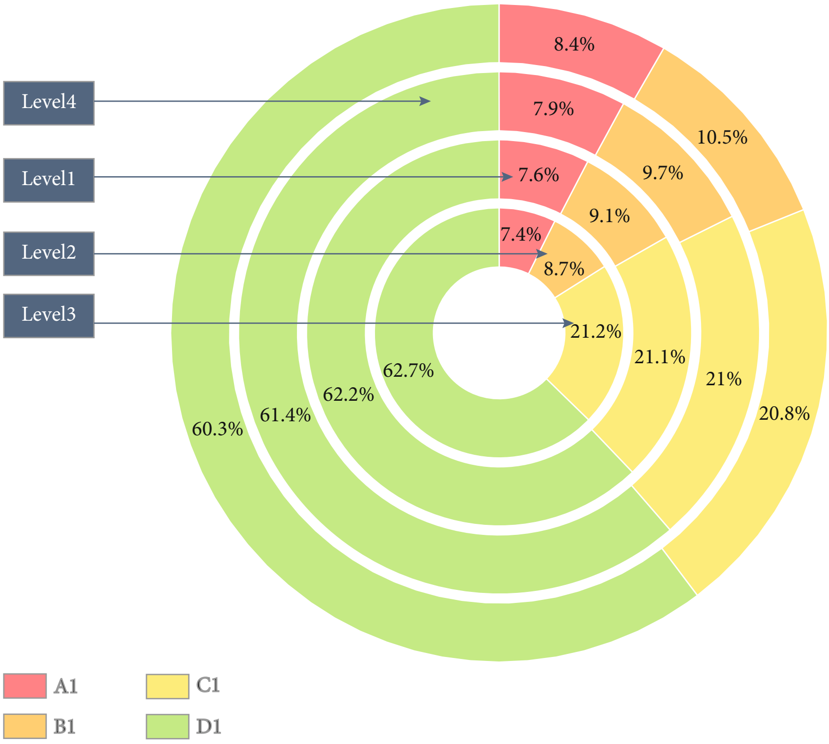

Figure 7: Phoneme recognition error rate.

offending phoneme will most effectively improve the intelligibility score of the word pronunciation.

To examine the phonological error rates of nonnative data at different intelligibility levels, this paper conducts recognition detection experiments on pronunciation. The average pronunciation error rate is used as a reference for the average error rate of subjects at each intelligibility level. The error rate of phonemes in the highest intelligibility level is only 0.22, while the error rate of phonemes in intelligibility level 1 reaches 0.53. Because of the difference between the error rate of each intelligibility level and the overall average phoneme error rate, this paper further analyzes the average error rate of phonemes in each intelligibility level and finds that, for the pronunciation of each phoneme in the lexicon, the higher the intelligibility level is, the lower the average. This is in line with the human perception of phonological intelligibility. Also, based on the results of phoneme error detection, we calculated the influence between the trend of phoneme error rate and intelligibility and verified that the design of the online assessment system in this paper needs to consider the feedback to learners those correction suggestions that can most effectively help them improve their speech intelligibility level. This is shown in Figure 7.

5. Conclusion

Automatic assessment of pronunciation is a complex subject involving knowledge from many disciplines such as linguistics, acoustics, signal processing, and pattern recognition. The language pronunciation rules are also very complex, and it is very difficult to perform an automatic assessment of pronunciation. In this paper, we propose a system that uses automatic speech recognition technology to effectively detect incorrect phoneme pronunciations in continuous Japanese pronunciation by English learners. The research focuses on how to effectively generate an extended pronunciation lexicon to predict possible mispronunciations in learners' pronunciation, combine speech recognition to detect erroneous phoneme categories in pronunciation, and provide corrective feedback to learners to help them improve their pronunciation. This paper introduces error elimination calculation for speech recognition and proofreading, which can effectively improve the recognition ability of spoken English, avoid the data progression error in traditional recognition and proofreading methods, and optimize the feedback control system to improve the system's ability to recognize speech and fundamentally solve the speech recognition confusion problem. The system in the paper improves the pronunciation model for the influence of the native language on the pronunciation of the second language, and although it achieves better results, there are still some shortcomings, and future research can be further explored in terms of using multiple speech feature parameters for comprehensive evaluation and networking in the implementation method.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. Xu, "From sonic models to sonic hooligans: magnetic tape and the unraveling of the Mao-era sound regime, 1958–1983," *East Asian Science, Technology and Society: an International Journal*, vol. 13, no. 3, pp. 391–412, 2019.

[2] X. Li and M. Mills, "Vocal features: from voice identification to speech recognition by machine," *Technology and Culture*, vol. 60, no. 2S, pp. S129–S160, 2019.

[3] M. Caprolu, S. Sciancalepore, and R. Di Pietro, "Short-range audio channels security: survey of mechanisms, applications, and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 311–340, 2021.

[4] Y. Li, H. Cheng, Z. Alhalili, G. Xu, and G. Gao, "The progress of magnetic sensor applied in biomedicine: a review of non-invasive techniques and sensors," *Journal of the Chinese Chemical Society*, vol. 68, no. 2, pp. 216–227, 2021.

[5] "Reviews of acoustical patents," *The Journal of the Acoustical Society of America*, vol. 144, no. 4, pp. 2022–2029, 2018.

[6] M. Romera, P. Talatchian, S. Tsunegi et al., "Vowel recognition with four coupled spin-torque nano-oscillators," *Nature*, vol. 563, no. 7730, pp. 230–234, 2018.

[7] O. S. A. Chronicle, "66th open seminar on acoustics Boszkowo, Poland, September 18–20, 2019," *Archives of Acoustics*, vol. 44, no. 3, pp. 603–622, 2019.

[8] D. Liu, S. Guo, Y. Yang, Y. Shi, and M. Chen, "Geomagnetism-based indoor navigation by offloading strategy in NB-IoT," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4074–4084, 2019.

[9] O. Marshall, "The maniac-making machine: a media history of delayed auditory feedback," *Technology and Culture*, vol. 62, no. 3, pp. 839–860, 2021.

[10] H. Dehra, "Acoustic signal processing and noise characterization theory via energy conversion in a PV solar wall device with ventilation through a room," *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, no. 4, pp. 130–172, 2018.

[11] J. E. Huggins, C. Guger, E. Aarnoutse et al., "Workshops of the seventh international brain-computer interface meeting: not getting lost in translation," *Brain-Computer Interfaces*, vol. 6, no. 3, pp. 71–101, 2019.

[12] H. Rayes, G. Al-Malky, and D. Vickers, "The development of a paediatric phoneme discrimination test for Arabic phonemic contrasts," *Audiology Research*, vol. 11, no. 2, pp. 150–166, 2021.

[13] H. Charaf, G. Harsányi, A. Poppe et al., "BME VIK annual research report on electrical engineering and computer science 2016," *Periodica Polytechnica Electrical Engineering and Computer Science*, vol. 61, no. 2, pp. 83–115, 2017.

[14] C. DeLaurenti, "Imperfect sound forever: a letter to a young phonographer," *Resonance: The Journal of Sound and Culture*, vol. 2, no. 2, pp. 125–167, 2021.

[15] J. Pinkl, E. K. Cash, T. C. Evans et al., "Short-term pediatric acclimatization to adaptive hearing aid technology," *American Journal of Audiology*, vol. 30, no. 1, pp. 76–92, 2021.

[16] Y. Huang, X. Guan, H. Chen, Y. Liang, S. Yuan, and T. Ohtsuki, "Risk assessment of private information inference for motion sensor embedded iot devices," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 3, pp. 265–275, 2020.

[17] A. F. T. Borges, A. L. Giraud, H. D. Mansvelder, and K. Linkenkaer-Hansen, "Scale-free amplitude modulation of neuronal oscillations tracks comprehension of accelerated speech," *Journal of Neuroscience*, vol. 38, no. 3, pp. 710–722, 2018.

[18] S. A. Fulop and L. Rice, "Reviews of acoustical patents," *The Journal of the Acoustical Society of America*, vol. 142, no. 1, pp. 124–133, 2017.

[19] V. G. Sardegna, J. Lee, and C. Kusey, “Self-efficacy, attitudes, and choice of strategies for English pronunciation learning,” *Language Learning*, vol. 68, no. 1, pp. 83–114, 2018.

[20] H. Vančová, “Current issues in pronunciation teaching to non-native learners of English,” *Journal of Language and Cultural Education*, vol. 7, no. 2, pp. 140–155, 2019.