Hindawi
Mobile Information Systems
Volume 2023, Article ID 9819704, 1 page
https://doi.org/10.1155/2023/9819704

Retraction

Retracted: The Spoken English Practice System Based on Computer English Speech Recognition Technology

Mobile Information Systems

Received 1 August 2023; Accepted 1 August 2023; Published 2 August 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] C. Gao, "The Spoken English Practice System Based on Computer English Speech Recognition Technology," *Mobile Information Systems*, vol. 2022, Article ID 9033421, 11 pages, 2022.

Hindawi
Mobile Information Systems
Volume 2022, Article ID 9033421, 11 pages
https://doi.org/10.1155/2022/9033421

Research Article

The Spoken English Practice System Based on Computer English Speech Recognition Technology

Chi Gao

School of Applied Foreign Languages, Xinyang Vocational and Technical College, Xinyang, 464000, China

Correspondence should be addressed to Chi Gao; gaochi840705@xyvtc.edu.cn

Received 20 February 2022; Revised 6 March 2022; Accepted 18 March 2022; Published 6 April 2022

Academic Editor: Chia-Huei Wu

Spoken English practice requires a combination of listening, speaking, reading, and writing, among which listening and speaking are the most difficult. In order to improve the speaking ability of the practitioner, the pronunciation of spoken English needs to be corrected in time. However, the workload of manual evaluation is too large, so it is necessary to combine intelligent methods for spoken language recognition. Based on the needs of spoken English pronunciation correction, this paper combines the computer English speech recognition technology to construct the spoken English recognition and correction model and combines the coding technology to study the English speech recognition technology. Moreover, this article constructs the spoken English practice system based on the actual needs of spoken English practice. Finally, this paper verifies the reliability of this system through experimental research, which provides a reliable means for the subsequent intelligent learning of spoken English.

1. Introduction

Spoken English ability is an important standard to measure the language ability of English learners, and practice is an important part of English speech teaching. Therefore, special research on English speech practice is not only an important method to improve the pertinence and scientificity of the practice in the design of English speech practice but also an important way for teachers to use it correctly in teaching practice and improve the efficiency of practice [1]. The text can only show the new language words and sentence patterns once, and the accurate mastery of language knowledge must be achieved through a certain amount of practice. The texts and exercises of oral practice are all set up for the special skill of "speaking", and set up for the purpose of blurting out. Although there are similar parts to the exercises of comprehensive and reading textbooks in form, the fundamental purpose is different, and the directions of the exercises are also quite different [2].

The ultimate goal of English speech practice is to cultivate the learners' language communicative ability. The mastery of this ability is different from the learning of pure English language knowledge. The use of language can be divided into four skills: listening, speaking, reading, and writing. Neurolinguistic research has also proved that the four language skills have corresponding central regions in the human brain mechanism. Therefore, the current listening class, speaking class, reading class, Chinese character class, writing class, and other class settings are based on skill training, which is scientifically based. Since the ultimate goal of learning English for learners who use English as a second language is to communicate, among these four skills, "listening" and "speaking" seem to be particularly important. In particular, for beginner learners of English, in the initial stage of learning a language, they must first learn to "listen" and "read" in order to better prepare for "speaking" and "writing." Therefore, the compilation of primary English speech textbooks is particularly important, which is related to whether second language English learners can have a good start in the process of learning English [3].

Based on the above analysis, this paper combines computer English speech recognition technology to construct

an English speech practice system, which makes the process of English speech practice more intelligent and improves the effect of English speech practice.

2. Related Work

The current research on computer English speech recognition technology mainly includes what are the practice content of oral English class exercises, the effect of a certain practice method, and the classification of oral English class exercises. From the perspective of language skills, literature [4] believes that in the course of oral English class exercises, attention should be paid to the selection and application of words and sentences, the cohesion of sentences, the organization of sentences, the conversion of style styles, rhetorical skills, and speech strategies. Literature [5] takes the lesson type as the starting point, and through the comparison of different lesson type practice methods, it is concluded that the classroom practice content of oral English class should not only have phonetic exercises but also include the usage of word making and sentence making and finally through actual practice. The communicative exercises enable students to truly appreciate the charm of spoken English. The two summarized the content of the English class exercises from different perspectives. In the actual oral English class exercises, we should also pay attention to the comprehensiveness of the exercises, and we should not neglect one of them. Literature [6] analyzes the use of group activities in oral English classroom exercises. Through investigation and analysis, it shows that group activities can enhance the initiative and enthusiasm of students to participate in activities and promote the interaction between students and teachers. Exchange and cooperate to increase the amount of activities students participate in language practice. In the classroom practice of oral English class, group discussion is one of the ways of practice. When using this practice method, we must use our strengths and avoid weaknesses to maximize the advantages of group practice. Literature [7] not only divides the exercises into three types: "imitation memory, association creation, and task communication," but also points out the characteristics of these three exercises. Literature [8] divides classroom exercises into four categories: "understanding, imitating memory, intellectual development, and communicative." The classification results are similar to those of Zhou Jian and Tang Ling, but before the "imitation memory" exercises, more emphasis is placed on "understanding" exercises. Comprehension exercises are easier than other exercise methods. Students only need to do to fully understand it.

Literature [9] pointed out that starting from the needs of English communication, summarizing and teaching communicative grammar can help English learners learn spoken English systematically. That is, the grammar used in the communication process is taught as the content of oral English learning. Because the purpose of learning spoken English is to communicate, using communicative grammar as a teaching content can help spoken learners to master spoken English more effectively. Literature [10] puts forward the teaching process of single sentence-discourse-discourse text when teaching oral English in segments. It is believed that oral teaching is a gradual process, and the content of oral teaching should be from simple to complex. It should follow the process of "single sentence-segment-discourse" to learn spoken English step by step. Literature [11] pointed out that the goal of oral teaching is to improve students' oral communication ability, which is a comprehensive ability. At the same time, it also pointed out that the content of oral learning is very extensive, so it should be phased and focused to complete the goal. In other words, grammar or vocabulary or pronunciation cannot be solely used as the content of oral teaching. Instead, the learning content and learning focus should be divided according to the learning stage of the learner.

Literature [12] proposes that applying "communicative method" to oral expression training is also an effective method. This method is aimed at cultivating the communicative competence of the learner, so that the learner can train the communicative competence in a specific language environment. Literature [13] proposed that combined with the characteristics of spoken English, spoken English teaching can be changed from reading to speaking training, recitation training, associative sentence building, speaking training, topic speaking training, and other teaching methods. Many of the methods mentioned here are still widely used in oral English classes, which just proves the practicality of these methods in oral English teaching. Literature [14] incorporates Western task-based teaching methods into oral English teaching and advocates clear communication tasks and the establishment of task-based oral teaching. The so-called "task-based teaching method" is to design some specific and actionable tasks around communication and language projects in the process of teaching, and then, learners complete the tasks through communication, communication, expression, and other methods to achieve learning the purpose of the language.

Literature [15] proposed a method of sentence segmentation and dynamic time planning (DTW) for spoken English recognition. Based on the segmentation of spoken sentences, this method recognizes the repetitive and paired subsequences in the acoustic feature space through comparison with the feature sequence. Then, these similar subsequence sets are grouped into larger sets. These clusters are regarded as linguistic units, and recognizable spoken English translations derived from them are created from the linguistic units. Finding repetitive patterns in continuous English speech requires identifying which part of a pair of speech sequences is similar and which part is different. Literature [16] proposes an algorithm for finding similar speech sequences. The algorithm is also based on the dynamic time planning algorithm. Literature [17] first defines the starting point and ending point and then compares in the predefined area and finds that similar speech sequences are different. Literature [18] calculates all possible similarities and then changes according to the similarity between adjacent elements. Determine the boundaries of the subsequence. In an English dataset containing many phrase combinations, the algorithm can better identify small differences between different language units. Literature [19] proposed a word

and phrase-level semantic unit recognition algorithm in a 13-language recognition system for continuous speech. It is found that all the repeated semantic units in the continuous speech stream require feature comparison of all the semantic units. As the number of similar semantic units detected increases, the computational complexity of this algorithm will become increasingly unacceptable. Literature [20] improves the efficiency of comparison by improving the parallelism of algorithm operation. Through research, it is found that there will be more than a certain period of silence between different sentences. Literature [21] regards continuous spoken language as a series of short sentences separated by silent intervals. At the implementation level, this article uses a large number of GPU computing units to achieve a high degree of parallelization of the algorithm.

In order to improve the accuracy of spoken English speech recognition and improve the robustness of speech recognition, this paper combines the research of computer English speech recognition technology to study intelligent oral practice system and takes corresponding control and adjustment to the energy of the excitation signal according to the type of speech. In order to enhance the robustness of the encoding and decoding, this paper adopts the technology of controlling the contribution of adaptive codebooks at the encoding end.

3. Computer English Speech Recognition Technology

In the process of packet transmission of English voice, the English voice information is encapsulated in P data packets in packet form and transmitted using the transport layer protocol. It has no quality assurance and cannot avoid packet loss. At present, there are many control methods for data frame loss, such as forward error correction, retransmission, and cross-sequencing. They are all based on the sender to compensate for packet loss. Concealment is a packet loss compensation technology based on the receiving end, and it does not need to consider the sending end. The following briefly introduces various frame erasure control methods.

Channel coding uses forward error correction codes to recover bits that have been erroneous during transmission. Now, it is also applied to the packet transmission of English speech to recover the information of lost frames. The parity check code is the simplest forward error correction code. In fact, it is only an error detection code in channel coding. When it is used in packet English speech, as shown in Figure 1, it has error correction capabilities. The advantage of forward error correction is that this type of method has nothing to do with the content of the packet and can completely recover the lost packet. The disadvantage is that it adds additional system delay and bandwidth.

In English speech transmission, the continuous loss of packets is the main reason for the deterioration of English speech quality. The method of cross-sorting is an effective method to eliminate consecutive packet loss. The idea is to disrupt the order of packets before transmission. For example, if the packet length is 20 ms and one unit is 5 ms, then

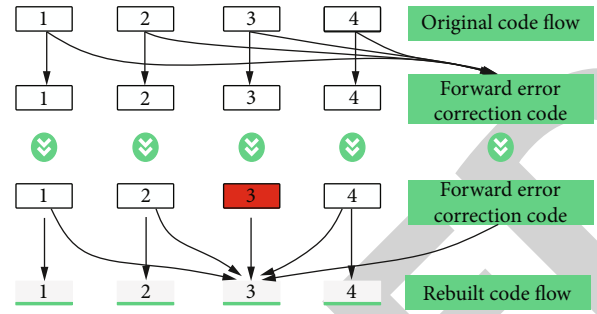

Figure 1: Principle of forward error correction.

the first packet can include 1, 5, 9, and 13 units, and the second packet can include 2, 6, 10, and 14 units, as shown in Figure 2. It can be seen from the figure that a single packet loss will not produce continuous unit loss. Because it is a small unit that is lost, the English speech characteristics will not change basically, so this method can easily realize error concealment. The disadvantage of packet unit crossover is that the delay is relatively large, which restricts its application in real-time transmission.

The basic principle of frame erasure concealment technology is to use a certain method for frame erasure detection on the received signal frame at the receiving end to find out whether the frame is a normal English speech frame or a missing frame. If it is a normal English speech frame, we use the corresponding decoding algorithm to decode the English speech to synthesize speech. If it is a lost frame, we use the corresponding frame erasure concealment algorithm for processing. Generally speaking, frame erasure concealment does not introduce additional delay and bit count.

Early FEC technology usually uses waveform replacement technology, which is aimed at English speech waveform coding (for example, ADPCM). With the development and wide application of English speech parameter coding and hybrid coding (CELP coding), hidden methods such as parameter extrapolation and interpolation for this coding model have been applied. The core layer encoder of this encoder is based on the CELP coding model, so the recovery of the lost frame is achieved by restoring the parameters. The following section will introduce the FEC method of this encoder in detail.

Figure 3 shows the block diagram of the frame erasure concealment method at the decoder end designed in this paper. First, the synthesized English speech of nonlost frames is classified into English pronunciation, and the judgment type is mute, voiced, unvoiced, transition from unvoiced to voiced, and transition from voiced to unvoiced. The parameters used in English phonetic classification include average energy E_s, normalized autocorrelation r_s, zero-crossing rate o_x, and spectral tilt e_x. The one-frame delay in Figure 4 indicates that the type of the current lost frame is estimated with the type of the previous frame's nonlost frame. That is, if the current frame is a lost frame, the type of the frame is the same as the type of the previous nonlost frame. The English phonetic classification process is as follows.

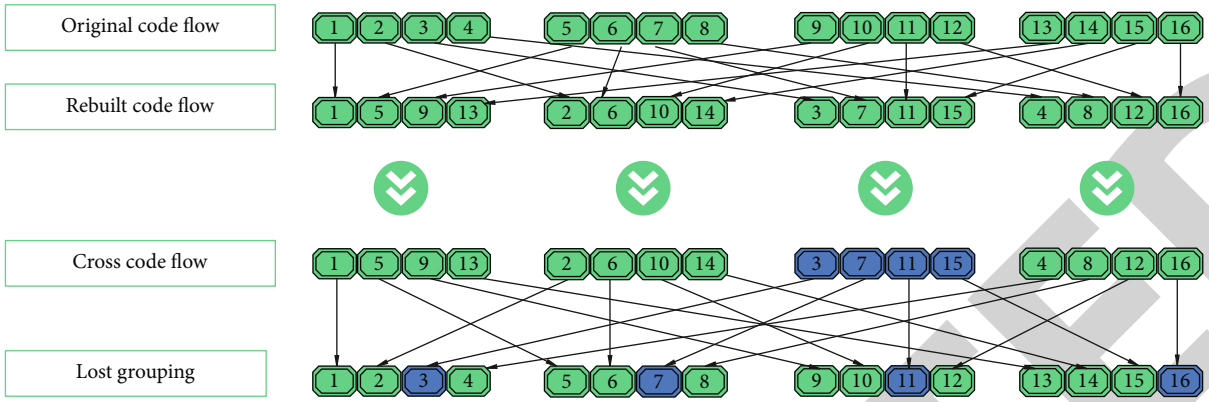

Figure 2: Principle of cross-sorting of grouping units.

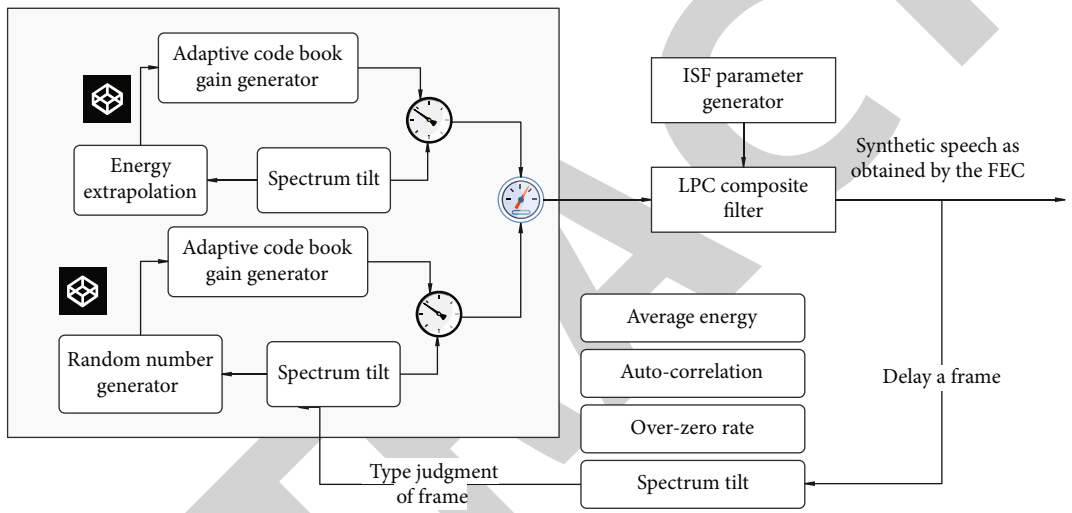

Figure 3: Block diagram of the frame erasure method of wideband embedded encoder.

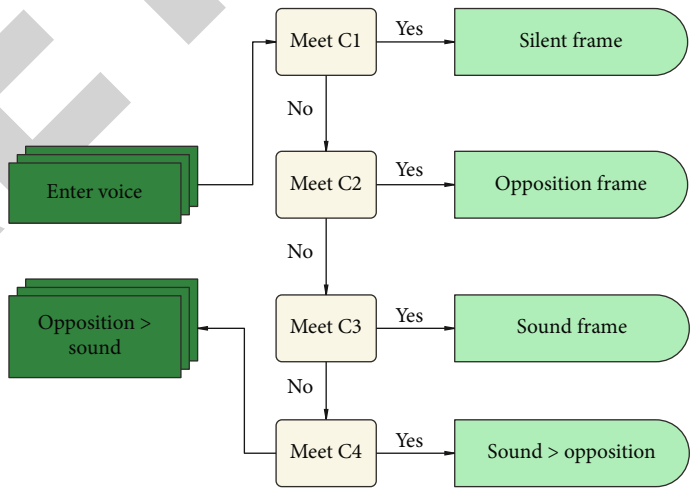

Figure 4: Flow chart of judging English speech type.

According to formula (1), the average energy E_x of the current frame is calculated, where $\hat{s}(n)$ is the synthesized English speech [22].

$$E_x = 10 \log_{10}\left(\frac{1}{256}\sum_{n=0}^{256}[\hat{s}(n)]^2\right). \quad (1)$$

The autocorrelation is calculated according to equation (2), where $\hat{s}(n)$ is the synthesized English speech, T is the integer pitch delay of the fourth subframe, and $t = 256 - T$. If $T > 96$, then T is set to the average value of the third subframe and the fourth subframe. If the pitch delay is less than the length of the subframe ($T < 64$), the normalized autocorrelation must be calculated again. The normalized autocorrelation at this time is the average value of the autocorrelation calculated twice.

$$r_x = \frac{\sum_{n=0}^{T-1}\hat{s}(t+n)\hat{s}(t+n-T)}{\sqrt{\sum_{n=0}^{T-1}[\hat{s}(t+n)]^2}\sqrt{\sum_{n=0}^{T-1}[\hat{s}(t+n-T)]^2}}. \quad (2)$$

The zero-crossing rate o_x is the number of times the waveform of the synthesized English speech in the current frame crosses the zero value.

The spectral tilt e_x is approximated by normalized autocorrelation, and the calculation formula is as follows, and $\hat{s}(n)$ is the synthesized English speech [23].

$$e_x = \frac{\sum_{n=64}^{256}\hat{s}(n)\hat{s}(n-1)}{\sum_{n=64}^{256}[\hat{s}(n)]^2}. \quad (3)$$

Then, the algorithm judges the type of the current frame based on the four parameters calculated above. The specific judgment process is shown in Figure 4. The four judgment conditions are obtained from experience, and the details are as follows [24]:

Condition 1 (C1): $E_x \le 35$.

Condition 2 (C2): $r_x \le 0.77$ and $e_x > 0.885$ or $e_x > 0.96$ and $o_x < 25$.

Condition 3 (C3): $o_x > = 53$ and $e_x < 0.75$.

Condition 4 (C4): The type of the previous frame is silent frame or voiced > unvoiced frame.

Tests show that the classification accuracy of this method is above 90%. As shown in Figure 5, the blue waveform is an English speech (140 frames in total, 320 samples per frame), and the red waveform represents the classification result. The amplitude of "8000" indicates voiced frames, the amplitude of "O" indicates silent frames, and the amplitude of "2000" indicates unvoiced frames. Moreover, an amplitude of "-2000" meansunvoiced ≥ voicedframe, and an amplitude of "-6000" means voiced > unvoiced frame. It can be seen from the figure that this classification method can be correctly judged except for certain transition frames, but voiced frames, silent frames, unvoiced frames, and most transition frames can be correctly judged, which basically meets the needs of the encoder.

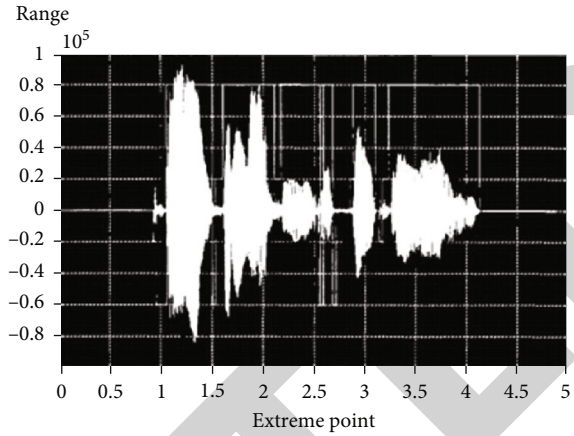

Figure 5: The result of English speech classification at the decoding end.

Synthetic English speech is obtained by stimulating through a synthesis filter. In the case of no missing frames, the decoder decodes the ISF parameters from the received code stream and converts them into ISP parameters and then obtains the ISP parameters of each subframe through interpolation. The coefficients of the synthesis filter are obtained by transforming the ISP parameters into LP coefficients.

The quantization of the ISF parameters of this encoder uses the unequal-coefficient interframe prediction split vector quantization method for each dimension. If the current frame is lost, the ISF parameter of the lost frame is set to be the same as the ISF parameter of the previous nonlost frame, and other processes remain unchanged. The experimental results prove that this method works best.

The adaptive codebook is obtained by interpolating the excitation of the past frame with the pitch period as the delay. The recovery of the pitch period of the lost frame in this encoder uses the method of estimating the pitch period of the lost frame in the G722.2 standard. According to the short-term stationary characteristics of English speech, the pitch period of the lost frame is usually replaced by the pitch period of the fourth subframe of the past frame. The technique for estimating the pitch period of a lost frame in the G722.2 standard is to judge the availability of the pitch period of a subframe in the past. If the voiced and stability are strong, it means that the lost frame is less likely to change compared with the past frame, and the pitch period of the lost frame can be replaced by the past subframe. Otherwise, the pitch period value of the lost frame is randomly generated within a certain range. The process is as follows:

First, the algorithm judges the availability of the pitch period of the previous subframe, denoted by $Q_{\log\_t-1}$ [25]:

$$Q_{\log\_t-1} = \begin{cases} 1, & g^p_{\min} > 0.5 \text{and} T_{\text{dif}} < 10, \\ 1, & g^p(n-1) > 0.5 \text{ and} g^p(n-2) > 0.5, \\ 0, & \text{other}. \end{cases} \quad (4)$$

Among them, $g^p_{\min} = \min(g^p_{\text{buffer}})$, and g^p_{buffer} is the stored adaptive codebook gain of the four subframes of the previous normal frame. g^p_{buffer} is the adaptive codebook gain of the fourth subframe of the previous normal frame, and $g^p(n-2)$ is the adaptive codebook gain of the third subframe of the previous normal frame. T_{dif} is the difference between the maximum value and the minimum value of the pitch period value of the four subframes of the previous normal frame.

Then, the pitch period of this lost frame is estimated as follows based on the availability of the pitch period of the previous subframe [26]:

$$T = \begin{cases} T(n-1), & Q_{\log\_t-1} = 1, \\ \frac{1}{3}\sum(T_{\max} + T_{\max-1} + T_{\max-2}) + \text{RND}(T_{\max} - T_{\max-2}), & Q_{\log\_t-1} = 0. \end{cases} \tag{5}$$

Among them, $T(n-1)$ is the pitch delay of the fourth subframe of the previous normal frame. We sort the stored pitch delays of the four subframes of the previous normal frame from smallest to largest; then, g^p_{buffer} is the largest value, $T_{\max-1}$ is the second largest value, and $T_{\max-2}$ is the third largest value. $\text{RND}(x)$ is to generate a random number in the range of $[-(x/2), x/2]$.

In addition, it is proved through experiments that if the pitch period of a subframe in the past is available, the effect of adding 1 to the pitch period value is better than using the value directly. Therefore, in the G722.2 standard adopted here, the technique for estimating the pitch period of the lost frame is slightly modified. If the pitch period of the past subframe is available, the algorithm adds 1 to this pitch period value and then uses this value as the integer pitch period of the lost frame. The above process obtains the integer pitch period of the lost frame, and the fractional pitch period is set to 0. Then, according to the restored pitch period, the past excitation is interpolated to obtain an adaptive codebook.

The traditional CELP model only has a unique excitation buffer. In the embedded CELP module used in this paper, in addition to the core layer excitation, the decoder also generates the excitation containing the additional layer information. Based on this special structure, this paper presents a lost frame adaptive codebook recovery method as shown in Figure 6. If the current frame is a lost frame, the corresponding excitation should be selected for interpolation by judging the previous frame rate. That is, if the rate of the previous frame is 8 kb/s, the past excitation of the core layer is selected for interpolation. If the rate of the previous frame is 12 kb/s, the past excitation of the enhancement layer is selected for interpolation. If the rate of the previous frame is greater than 12 kb/s, the past excitation interpolation of enhanced layer 2 is selected.

Usually, the fixed codebook of lost frames is replaced by a randomly generated sequence, which is also handled by this encoder. It can also be seen that FEC only restores the basic information of the English speech, while the details of the English speech (information on the enhancement layer) cannot be restored.

The adaptive codebook gain g^p and the fixed codebook gain g^c of the lost frame are obtained from the value of the past subframe [27]:

$$\begin{aligned} g^p &= \text{median5}(g^p(n-1), \cdots, g^p(n-5)), \\ g^c &= p^c * \text{median5}(g^c(n-1), \cdots, g^c(n-5)). \end{aligned} \tag{6}$$

Among them, g^p_{buffer} is the adaptive codebook gain of the past normal frame subframe. $g^c(n-1), \cdots, g^c(n-5)$ is the fixed codebook gain of the normal frame subframe in the past. When the type of lost frame is voiced frame, $P^c = 0.5$, otherwise $P^c = 1$. At the same time, the adaptive codebook gain should be limited; that is, when the g^p obtained above is >0.95, $g^p = 0.95$.

In addition, the energy extrapolation method is used to adjust the gain of the adaptive codebook of the lost frame. This method can be understood as using the average energy ratio of the two subframe excitations before the lost frame to estimate the gain of the adaptive codebook of the current lost frame. Here, E is the average energy ratio of the first two subframe excitations of the current lost frame. $E_1 = 0.75E_1^{-1} + 0.3E$, E_1 is the interframe smoothing value of E, and E_1^{-1} represents the interframe smoothing value of the previous subframe. $T^{(-n)}$ is the pitch period of the previous nth subframe, and $EXC^{(-n)}$ is the excitation of the previous nth subframe.

$$\begin{aligned} E &= \frac{E_v^{(-1)}}{E_v^{(-2)}}, \\ E_v^{(-n)} &= \frac{\sqrt{\sum_{i=1}^{T^{(-N)}} \left|EXC^{(-n)}(-i)\right|^2}}{T^{(-n)}}. \end{aligned} \tag{7}$$

The obtained E_1 is the adjusted value of the adaptive codebook gain of the lost frame, but in the first two cases

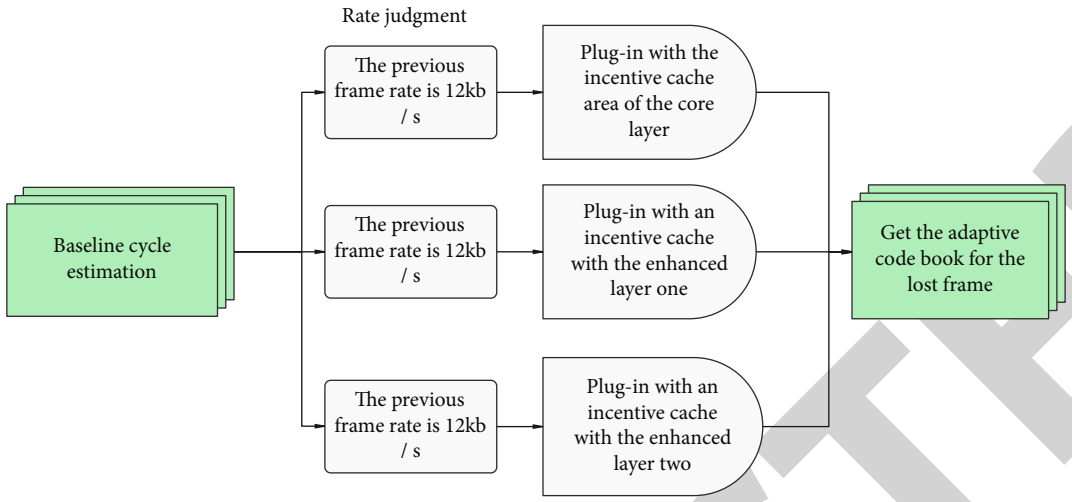

Figure 6: Block diagram of recovery principle of adaptive codebook for lost frames.

of equation (8), E_1 cannot replace the originally calculated adaptive codebook gain.

$$g_p = \begin{cases} 0.98, & E_1 > 0.98, \\ g_p, & \mathrm{E}_1 < g_p, \\ E_1, & \text{other}. \end{cases} \tag{8}$$

Normally, the excitation is the adaptive codebook vector multiplied by its gain, plus the fixed codebook vector multiplied by its gain. The vector sum obtained is an incentive. Usually, for lost frames, the fixed codebook is replaced by a random sequence. Experiments show that if the lost frame is voiced, the synthesized English speech containing this fixed codebook will have obvious noise. At the same time, the fixed codebook will destroy the waveform of the voiced excitation signal and affect the English speech synthesis of the normal frame after the lost frame. Therefore, this article adjusts the fixed codebook energy of lost frames according to different types of English speech, as follows:

(1) If the current frame is a voiced frame, each sample point of the fixed codebook is attenuated by 0.5

(2) If the current frame is a "clear ≥ turbid" transition frame, the samples of the fixed codebook of the 3rd and 4th subframes will be attenuated point by point, and the attenuation coefficient will gradually change from 1 to 0.5:

(3) If the current frame is a transitional frame of "turbid ≥ clear", the fixed codebook samples of the first and second subframes will be attenuated point by point

The attenuation coefficient is gradually changed from 0.5 to 1.

After the fixed codebook is adjusted, the excitation is obtained, and finally, the lost frame English speech that is recovered through the synthesis filter is excited.

The pros and cons of the frame erasure concealment technology are how effective it is to restore voiced frames. The adaptive codebook is the most important component for expressing voiced sounds, and the adaptive codebook is generated by interpolating the past excitation with the pitch period as a delay. Therefore, for lost frames, if the pitch period parameter can be effectively restored to make it close to or equal to the value when no frame loss occurs, the synthesis quality of the lost frame can be greatly improved.

The data fitting method is based on the mutual relationship between the data, draws a mathematical formula between them, and draws an approximate curve to reflect the general trend of the given data. The English voice characteristics of the voiced frames in the English speech are slowly changing. Here, since the pitch period of the future frame cannot be obtained, the pitch period of the past frame can only be used to estimate the change trend of the pitch period of the current lost frame, that is, to predict it. Methods are as below.

The past five pitch periods are $T(i), i = 0, 1, \cdots, 4$, where $T(0)$ is the earliest pitch period. Then the prediction model can be defined as

$$T'(i) = a + bi. \tag{9}$$

The pitch period of the current lost frame is

$$T'(5) = a + 5b. \tag{10}$$

a and b are the prediction coefficients, and through $\partial E/\partial a = 0$ and $\partial E/\partial b = 0$, formula (11) is minimized [22]:

$$E = \sum_{i=0}^{4} \left[T'(i) - T(i)\right]^2 = \sum_{i=0}^{4} [(a + b \times i) - T(i)]^2. \tag{11}$$

Thus, $a = (\sum_{i=0}^{4} T(i) - \sum_{i=0}^{4} iT(i))/5$ and $b = (\sum_{i=0}^{4} iT(i) - \sum_{i=0}^{4} T(i))/10$ are obtained.

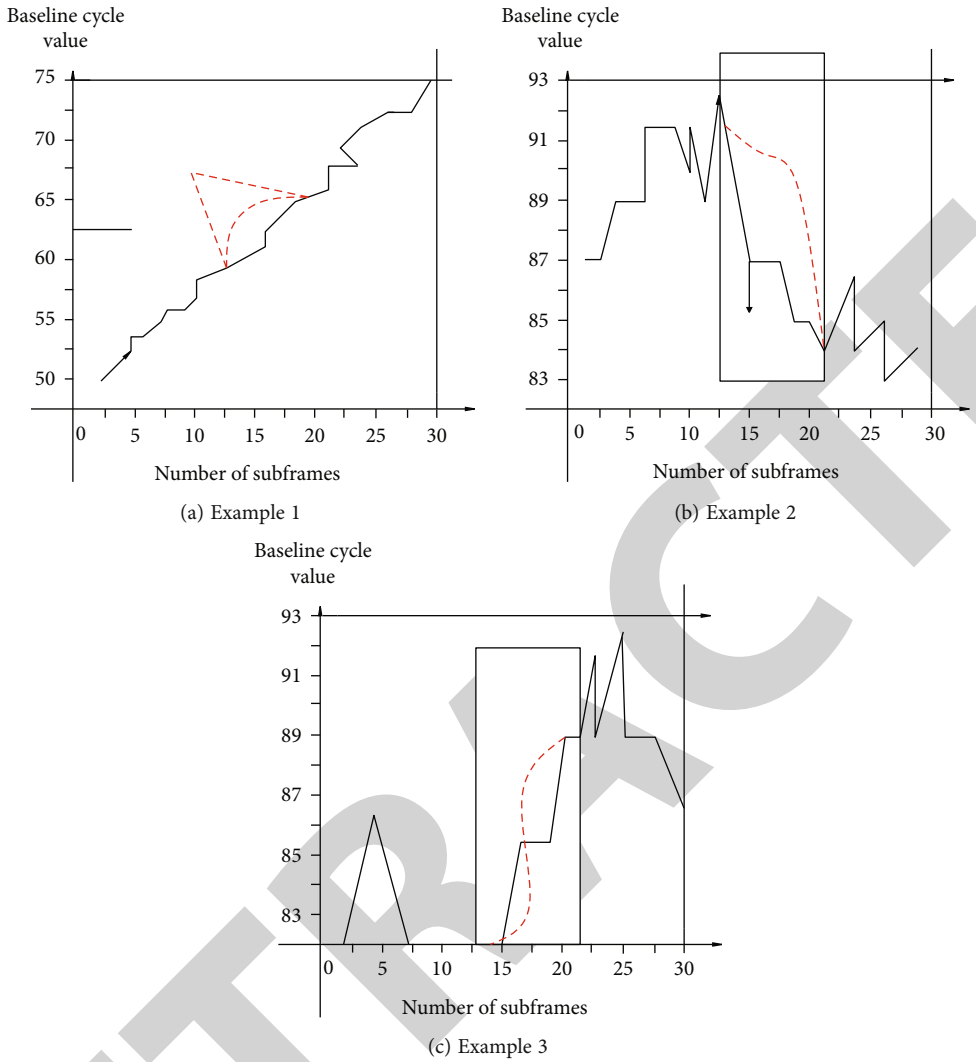

Figure 7: An example of the effect of interpolating the pitch period with the past frame and the future frame.

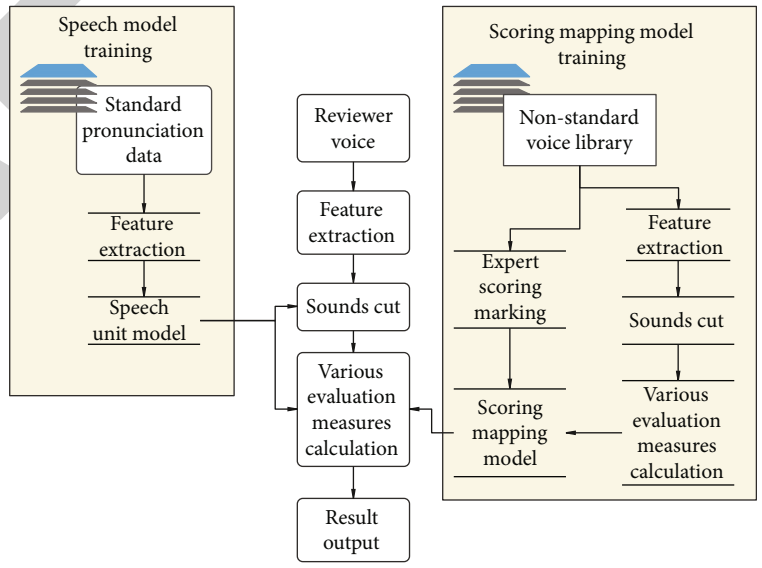

Figure 8: The basic framework of computer English speech evaluation.

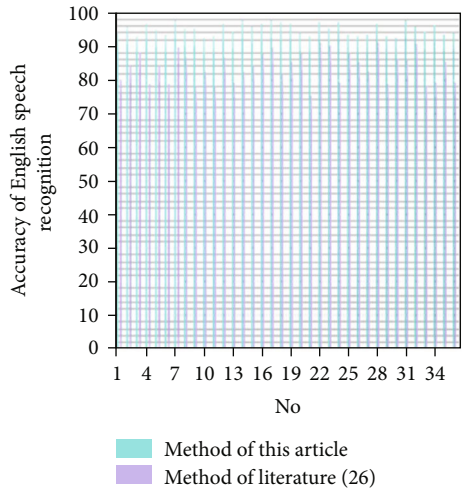

Figure 9: Accuracy of English speech recognition.

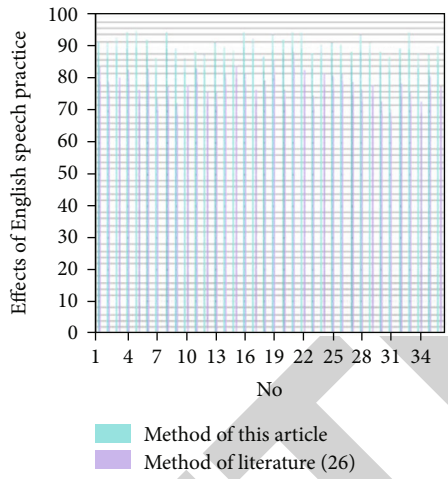

Figure 10: Effects of English speech practice.

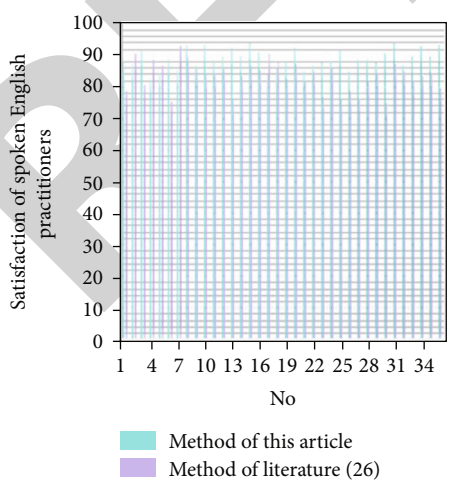

Figure 11: Satisfaction of spoken English practitioners.

Here, it is necessary to judge whether to use the prediction method. If the voice is strong and stable, it is used for prediction; otherwise, the pitch period value is randomly generated.

Experimental observations show that if the pitch period of the current missing frame has a linear relationship with the pitch period value of the previous and subsequent frames, the prediction is effective. However, the dynamic range of the pitch period of wideband English speech is large, and even some voiced frames may not change linearly. Especially, when the frame is a transitional frame, the prediction method shows more shortcomings. Later, by adding weights to equation (11), that is, the closer the pitch period to the current frame is, the greater the weight, but the result is not significantly improved.

However, through observation, it is found that the effectiveness of the forecasting method depends on the accuracy of judging under what circumstances the forecasting method is adopted. Another idea is to add "smoothing the pitch period curve" to the pitch period search process of the encoder to improve the prediction effect.

The current lost frame has the greatest correlation with adjacent English speech frames, so the pitch period of adjacent frames (past frame and future frame) is used to estimate that the pitch period of the current lost frame should be closer to the true value (the pitch period when no loss occurs). The disadvantage of this method is that one frame delay will be introduced.

This article tried to use the interpolation method, using the pitch period P_4^{n-1} of the fourth subframe of the past frame and the pitch period P_1^{n+1} of the first subframe of the future frame to restore the pitch period $\widehat{P}_1^n$ of the current lost frame. The method is as follows: P_{diff} is the difference of the pitch period of the two frames, as in the following formula:

$$P_{\text{diff}} = P_t^{n+1} - P_4^{n-1}. \tag{12}$$

Then, the pitch period of the ith subframe of the current lost frame is estimated as

$$\widehat{P}_1^n = P_t^{n-1} + \left\lfloor \left(\frac{P_{\text{diff}}}{4} \right) \cdot i \right\rfloor. \tag{13}$$

The symbol "$\lfloor \rfloor$ "means rounding.

Experimental observations show that not all pitch periods of lost frames are effective with this interpolation method. The effect is not good if the difference between adjacent pitch periods is too large, so the interpolation method should be restricted. If $|P_{\text{diff}}| \le 10$ is obtained through experiments, the algorithm uses formula (13), otherwise the pitch period obtained by the original method is used, so that a better pitch period restoration effect can be obtained. Figure 7 shows the pitch period curve of several segments of English speech. The pink box indicates that frame loss has occurred, the red dashed line indicates the correct pitch period curve without frame loss, the green dotted line indicates the pitch period curve obtained by the original method, and the blue solid line indicates the pitch

period curve obtained by the interpolation method. It can be seen from the figure that the pitch period obtained by the interpolation method is closer to the correct value. The same subjective listening also shows that the English speech recovered by the interpolation method is better than the original method.

4. Application of Computer Speech Recognition Technology in English Speech

The process of English speech evaluation is to first extract the evaluation feature parameters after preprocessing and segmenting the testee's English speech. Then, the feature parameters to be tested and the pretrained or statistic corresponding standard evaluation model are calculated in a certain manner to obtain the measurement of each feature parameter. Finally, the evaluation measure of each feature is mapped into a score through a certain way of calculation through a model that has been trained in advance and then output, as shown in Figure 8.

This paper designs experiments to verify and analyze the performance of the model in this paper. This article mainly analyzes and evaluates the accuracy of English speech recognition, the effect of spoken English practice, and the satisfaction of spoken language practitioners of this model and compares the experimental research results with the method in the literature [26], and the experimental research results shown in Figures 9–11 below are obtained.

From the above research results, the spoken English practice system based on computer English phonetic recognition technology proposed in this paper has good practical effects.

5. Conclusion

With the rapid development of information technology, computer technology and artificial intelligence technology have been widely used in all aspects of social production and life, and the role of computer-assisted language learning has become more and more obvious. The automatic assessment of spoken language is the automatic assessment and diagnosis of spoken language quality based on the physiological and behavioral characteristics of speech signals. Spoken language automatic assessment and diagnosis technology is based on human voice and language characteristics, uses information processing technologies such as signal processing and pattern recognition as means, and integrates multidisciplinary theories and knowledge of phonetics, linguistics, and pedagogy. Compared with traditional manual methods, it can significantly improve the objectivity and fairness of the evaluation test, greatly reduce the cost of manpower and material resources, and make large-scale oral proficiency testing and evaluation possible. Based on the above analysis, this paper combines computer English speech recognition technology to construct an English speech practice system, which makes the process of English speech practice more intelligent and improves the effect of English speech practice. The experimental research results show that the spoken English practice system based on computer English speech recognition technology proposed in this paper has good practical effects.

This paper proposes a frame erasure masking method for broadband embedded codec, which controls and adjusts the energy of the excitation signal according to the voice type. In order to enhance the robustness of the codec, this paper also adopts the technique of controlling the contribution of the adaptive codebook on the encoding side. In order to enhance the robustness of the codec, this article also adopts the method of controlling the contribution of the adaptive codebook on the encoding side.

In order to require the encoder to process narrowband speech and wideband speech at the same time, if the post-processing scheme can process these two signals separately, the effect should be better. Since there is no time to change the design structure of the codec, some frame erasure concealment technology is not used in this encoder.

Data Availability

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

Conflicts of Interest

The author declares no competing interests.

Acknowledgments

This study is sponsored by (1) project name: research and practice on the hybrid teaching method of integrating the great anti epidemic spirit into the foreign language "curriculum thinking and politics", project category: general research project of Humanities and Social Sciences in Colleges and Universities of Henan Province (No. 2022-ZDJH-00509), and (2) project name: design, development and application of online open course teaching resources in Higher Vocational Colleges, project category: 2020 Xinyang Vocational Education and Teaching Reform Research Project (No. ZJB2018).

References

[1] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[2] J. Al-Tamimi, "Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: implications for formal representations," *Laboratory Phonology*, vol. 8, no. 1, pp. 28–40, 2017.

[3] H. N. Choi, S. W. Byun, and S. P. Lee, "Discriminative feature vector selection for emotion classification based on speech," *Transactions of the Korean Institute of Electrical Engineers*, vol. 64, no. 9, pp. 1363–1368, 2015.

[4] T. Haderlein, M. Döllinger, V. Matoušek, and E. Nöth, "Objective voice and speech analysis of persons with chronic hoarseness by prosodic analysis of speech samples," *Logopedics, Phoniatrics, Vocology*, vol. 41, no. 3, pp. 106–116, 2016.

[5] C. T. Herbst, S. Hertegard, D. Zangger-Borch, and P. Å. Lindestad, "Freddie mercury—acoustic analysis of speaking fundamental frequency, vibrato, and subharmonics," *Logopedics, Phoniatrics, Vocology*, vol. 42, no. 1, pp. 29–38, 2017.

[6] A. K. Hill, R. A. Cárdenas, J. R. Wheatley et al., "Are there vocal cues to human developmental stability? Relationships between facial fluctuating asymmetry and voice attractiveness," *Evolution and Human Behavior*, vol. 38, no. 2, pp. 249–258, 2017.

[7] C. C. Hsu, K. M. Cheong, T. S. Chi, and Y. Tsao, "Robust voice activity detection algorithm based on feature of frequency modulation of harmonics and its DSP implementation," *IEICE Transactions on Information and Systems*, vol. E98.D, no. 10, pp. 1808–1817, 2015.

[8] T. G. Kang and N. S. Kim, "DNN-based voice activity detection with multi-task learning," *Ieice Transactions on Information & Systems*, vol. E99.D, no. 2, pp. 550–553, 2016.

[9] P. H. Kumar and M. N. Mohanty, "Efficient feature extraction for fear state analysis from human voice," *Indian Journal of Science and Technology*, vol. 9, no. 38, pp. 1–11, 2016.

[10] A. Leeman, H. Mixdorff, M. O'Reilly, M. J. Kolly, and V. Dellwo, "Speaker-individuality in Fujisaki model f0 features: implications for forensic voice comparison," *International Journal of Speech Language and the Law*, vol. 21, no. 2, pp. 343–370, 2015.

[11] F. L. Malallah, S. Knymg, S. D. Abdulameer, and A. W. Altuhafi, "Vision-based control by hand-directional gestures converting to voice," *International Journal of Scientific & Technology Research*, vol. 7, no. 7, pp. 185–190, 2018.

[12] M. Woźniak and D. Połap, "Voice recognition through the use of Gabor transform and heuristic algorithm," *Nephron. Clinical Practice*, vol. 63, no. 2, pp. 159–164, 2017.

[13] G. Mohan, K. Hamilton, A. Grasberger, A. C. Lammert, and J. Waterman, "Realtime voice activity and pitch modulation for laryngectomy transducers using head and facial gestures," *Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2302–2302, 2015.

[14] M. Sleeper, "Contact effects on voice-onset time in Patagonian Welsh," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3111–3111, 2016.

[15] Q. K. Ngoc, H. T. Duong, and H. T. Duong, "A review of audio features and statistical models exploited for voice pattern design," *Computer Science*, vol. 3, no. 2, pp. 36–39, 2015.

[16] S. S. Nidhyananthan, K. Muthugeetha, and V. Vallimayil, "Human recognition using voice print in Lab VIEW," *International Journal of Applied Engineering Research*, vol. 13, no. 10, pp. 8126–8130, 2018.

[17] S. Orlandi, C. A. Garcia, A. Bandini, G. Donzelli, and C. Manfredi, "Application of Pattern Recognition Techniques to the Classification of Full- Term and Preterm Infant Cry," *Journal of Voice*, vol. 30, no. 6, pp. 656–663, 2016.

[18] R. Rhodes, "Aging effects on voice features used in forensic speaker comparison," *International Journal of Speech Language & The Law*, vol. 24, no. 2, pp. 177–199, 2017.

[19] M. Sarria-Paja, M. Senoussaoui, and T. H. Falk, "The effects of whispered speech on state-of-the-art voice based biometrics systems," in *2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 1254–1259, Halifax, NS, Canada, 2015.

[20] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315–1329, 2016.

[21] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[22] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[23] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[24] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: a survey," *Speech Communication*, vol. 56, no. 3, pp. 85–100, 2014.

[25] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[26] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, no. 3, pp. 535–557, 2017.

[27] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.