Article

An Empirical Performance Analysis of the Speak Correct Computerized Interface

Kamal Jambi ^{1,*}, Hassanin Al-Barhamtoshy ¹ , Wajdi Al-Jedaibi ¹, Mohsen Rashwan ² and Sherif Abdou ³¹ Department of Computer Science, Faculty of Computing & Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; hassanin@kau.edu.sa (H.A.-B.); wajdaibi@kau.edu.sa (W.A.-J.)² Electronics and Communication Department, Faculty of Engineering, Cairo University, Cairo 12613, Egypt; mrashwan@rdi-eg.com³ Department of Information Technology, Faculty of Computers and Artificial Intelligence, Cairo University, Cairo 12613, Egypt; sheriff.abdou@rdi-eg.com

* Correspondence: kjambi@kau.edu.sa

Abstract: The way in which people speak reveals a lot about where they are from, where they were raised, and also where they have recently lived. When communicating in a foreign language or second language, accents from one's first language are likely to emerge, giving an individual a 'strange' accent. This is a great and challenging problem. Not particularly, because it is a part of one's personality that they do not have to give up. It is only challenging when pronunciation causes a disruption in communication between an individual and the individuals with whom they are speaking. Making oneself understandable is the goal of perfecting English pronunciations. Many people require their pronunciation to be perfect, such as those individuals working in the healthcare industry, where it is rather critical that each term be read precisely. Speak Correct offers each of its users a service that assists them with any English pronunciation concerns that may arise. Some of the pronunciation improvements will only apply to a specific customer's dictionary; however, in some cases, the modifications can be applied to the standard dictionary as well, benefiting our whole customer base. Speak Correct is a computerized linguist interface that can assist its users in many different places around the world with their English pronunciation issues due to Saudi or Egyptian accents. In this study, the authors carry out an empirical investigation of the Speak Correct computerized interface to assess its performance. The results of this research reveal that Speak Correct is highly effective at delivering pronunciation correction.

Keywords: empirical assessment; automatic speech recognition; computerized interface; pronunciation correction; speech processing

Citation: Jambi, K.; Al-Barhamtoshy, H.; Al-Jedaibi, W.; Rashwan, M.; Abdou, S. An Empirical Performance Analysis of the Speak Correct Computerized Interface. *Processes* **2022**, *10*, 487. <https://doi.org/10.3390/pr10030487>

Academic Editor: Chien-Chih Wang

Received: 31 January 2022

Accepted: 17 February 2022

Published: 28 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Speech recognition is the process of using computer software to turn speech into a series of words. Speech recognition's overall goal is to enable individuals to interact more easily and efficiently, because it is the most intuitive mode of interaction for individuals. Whereas long-term goals necessitate seamless integration with several natural language processing (NLP) components, there are a number of new applications that can be quickly deployed using the fundamental voice recognition module. Voice calling, call forwarding, record keeping and dictation, coordination and control, as well as computer-aided multilingualism are just a few of the common applications. The technique of turning human sound impulses into words or commands is known as speech recognition. Automatic speech recognition (ASR), computer voice recognition, and speech to text (STT) are some of its other names. It includes computer engineering, linguistics, as well as technical knowledge and research [1–8].

Speech correction is another form of speech recognition that is based on the sound of a person's voice. It is a subset of pattern classification and an attractive research area

in voice speech accuracy enhancement. Speech correction research spans a wide range of disciplines, including computer science, machine intelligence, sensing applications, cognition, acoustics, linguistics, as well as information psychology. It is an interdisciplinary, all-encompassing research area. Distinct research topics have arisen as a result of the different research demands and restrictions. These domains can be separated into individual words, linked words, and persistent speech recognition algorithms, depending on the needs of the speaker's method of speaking. These domains can be separated into voice recognition algorithms for specific people and generic people based on the degree of reliance on the speaker. They can be classified into short vocabulary, moderate vocabulary, huge vocabulary, as well as infinite vocabulary voice recognition systems based on the size of their vocabulary [9–15].

People employ different recognition strategies and procedures for different voice recognition systems, but the essential concepts are the same. Feature extraction is applied to the gathered voice signals. The speech features acquired by the module are sent to the modeling library component. The speech pattern matching component finds speech segments based on the model library, and then calculates the recognition performance. Some speech recognition systems necessitate “training” (also known as “enrollment”), in which a single speaker delivers text or isolated vocabulary into the program.

The technology analyses the person's unique voice and utilizes it to fine-tune speech recognition, resulting in a higher accuracy. “Speaker-independent” systems do not require training. “Speaker dependent” systems rely on training. The complexities of human communication have made growth difficult. It is one of the most difficult disciplines of computer science to master, as it combines linguistics, arithmetics, and analytics. The speech input, extraction and classification, vectorization, decoder, and word result are all elements of speech recognizers. To identify the proper output, the decoder uses acoustic models, a pronunciation dictionary, as well as language models.

This paper discusses Speak Correct that is a computerized linguist interface. Speak Correct can assist its users in many different places around the world, when they have English pronunciation issues due to their Saudi or Egyptian accents. Further, in this study, the authors conduct an empirical evaluation of the Speak Correct computerized interface for an efficiency assessment.

The remainder of the paper is organized as follows: Section 2 discusses the recent works related to the similar research area; Section 3 presents the overview of the Speak Correct computerized interface; Section 4 presents the interactive experiments for the efficiency analysis of the Speak Correct computerized interface; Section 5 discusses the findings of the interactive experiments; Section 6 presents the discussion on the findings; and, finally, Section 7 concludes the research work.

2. Related Works

Spring and Tabuchi [16] demonstrated how ASR technology could be utilized in an electronic curriculum of English as a foreign language (EFL) to assist L1 Japanese learners to develop better pronunciation. They used a combination of pre- and post-records as well as survey responses to figure out how learners would react to the ASR program, if they would advance, and which courses would be most beneficial to them. The findings indicated that the learners were mainly happy about the ASR-assisted practice and that they considerably enhanced their intelligibility, particularly those who started with lesser competence.

Evers and Chen [17] evaluated how learning methods (visual/verbal), including how the use of ASR software impacts adult learners, improved performance in English as a second language throughout a 12 week curriculum concentrating on pronunciation. According to the findings, the learning strategies produced a large disparity in the pronunciation effectiveness of the reading process across all teams. In the reading process, visual learners outscored the verbal learners. Throughout both the reading activities and live discussions, the mixture of ASR and peer correction resulted in a significant performance.

Eskenazi [18] discussed how ASR could be used to train learners to improve their accents in a foreign tongue. First, the elements of effective language instruction were mentioned, as were the limitations of using ASR, and how to cope with them. The author also used the Carnegie Mellon FLUENCY mechanism as an example to demonstrate how such an approach would work. Phonetics and prosody training were highlighted. Eventually, using the FLUENCY process as an instance, the author emphasized the importance of having a platform that adjusts to the user.

Evers and Chen [19] studied the comparison in adults' pronunciation effectiveness with the help of an ASR platform with peer assessment as well as individual practice. The respondents were Taiwanese working individuals. During the weekly learning experience, respondents dictated a document to the ASR application, Speechnotes, on its online platform, and then practiced misinterpreted phrases on their own (the comparative group) or with responses from team members (the experimental group). Following the initial intervention, the pronunciation of the trainees was assessed through extensive reading as well as spontaneous discussion activities.

Cao and Hao [20] designed a spoken English assistant pronunciation training model based on the Android mobile application. They proposed a lip movement judgement algorithm focusing on ultrasonic identification. This was used to support the conventional voice recognition method in the double feedback judgement, based on an in-depth research and assessment of the verbal English speech correction algorithm and speech feedback process. A dual standard scoring technique was also designed in the smart speech training feedback method to exhaustively assess the verbal trainer's utterance as well as to correct the presenter's speech in real time. The experimental findings demonstrated that the platform's speech precision reached 85 percent, which enhanced the standard of oral English lecturers to some significant degree.

Moxon [21] explored whether the automated assessment of pronunciation accuracy employing speech recognition advanced technologies could enhance the pronunciation abilities of 105 Thai undergraduate learners learning English in Thailand (88 female, 17 male). A pre-test, post-test architecture was used with intervention and control sampling methods that were reversed over two six week timeframes. Participants in the treatment team were provided entry to an online system where they could record as well as publish their speech for electronic analysis and feedback using SpeechAce, a voice recognition functionality aimed at evaluating pronunciation as well as fluency.

García et al. [22] experimented with the use of synthesized voice as an asset for formative assessment. A group of learners used a Computer-Assisted Pronunciation Training (CAPT) platform to complete a battery of minimal pair discriminatory practice tasks; some of those who declined the production procedures were given the option of undergoing additional training by using synthesized voice as a prototype in a round of interaction workouts. Respondents who used this resource significantly outperformed by simply repeating the previously failed activity. The findings showed that the text-to-speech processes provided by the current Android operating system should be regarded as a useful feedback source of information for the pronunciation instructions, particularly when paired with teaching techniques.

Although there are many literatures available on the pronunciation correction tool, very few focus on the correction of Saudi or Egyptian accent defects. The proposed Speak Correct computerized interface is efficiently capable of correcting the Saudi or Egyptian accent defects in mispronounced English statements, even if they are spoken incorrectly, and it then uncovers the speaker's pronunciation faults.

3. Speak Correct Computerized Interface

3.1. Speak Correct Phonetic Editor

The Speak Correct system has four primary phases: the trainer phase, which is used to train speech characteristics; the decoder phase, which is used for speech decoding using pronunciation assumption; the evaluation phase, which is used to assess and create speaker

feedback; as well as the phonetic editor, which is used to enhance the suggested language model using the adaptation of mapping and generation guidelines. The system employs a decoder that detects user input speech and calculates the confidence and mistakes in pronunciation. Furthermore, the system analyzes feedback messages and detects errors, provides help for correcting the faults, and performs an evaluation. To distinguish specific speech sounds, the phonetic editor comprises mapping and generation procedures as well as analytic methods (e.g., neural networks and Gaussian models). Figure 1 shows the data flow diagram (DFD) for the suggested Speak Correct system. Each word is presented to the users in graphs and a lattice form, with a visual indicating different phonetic levels and corresponding teachings (Saudi or Egyptian accent defects). As a result, the Speak Correct provides users with the option of choosing their own levels and examples [23–26].

Figure 1. Speak Correct with phonetic analyzer overview.

At the phonetic phase, skilled annotators insert the utterances, which are then mapped and generated using automatic voice recognition and backed by trained occurrences in the format of a grammar structure (Lattice graph) for the intended word. Mistakes are recognized and evaluated, and the decoder and assessment stages are used to assess the results.

The difficulty of determining the correct “underlying” sequences of symbols/patterns is addressed in the extraction phase. As a result, the Viterbi algorithm provides an efficient technique of addressing the decoding problem by examining all the potential strings and computing the probability of obtaining the observed sequence by employing additional rules (such as the Bayes rule [27]).

The reference speech techniques are commonly subjected to further preprocessing in order to adjust them to the speaker’s speech qualities. In these kinds of circumstances, the speaker adaptability module of the Maximum Likelihood Linear Regression (MLLR) is employed to improve the adapted component.

3.2. The Speak Correct Tool Architecture

The Speak Correct system is depicted in Figure 2 as a block diagram. It detects pronunciation problems in users' speech using a state-of-the-art speech detection method based on the Hidden Markov Model (HMM). The following Table 1 explains the main components:

Figure 2. The Speak Correct system architecture.

Table 1. Different modules of the Speak Correct computerized interface.

Modules	Description
The HMM model trainer	Gather patterns in the data using the training data and save them mathematical analyses for the pronunciation verification process.
Verification of HMM models	The platform's acoustic HMM algorithms.
Pronunciation hypotheses generator	It examines a training activity and produces all potential pronunciation alternatives, which are then sent to the speech recognition system and tested versus a spoken phrase.
The HMM adapter	It utilized to optimize system efficiency by adapting acoustic models to every user's acoustic attributes.
The HMM decoder (ASR)	The speech processor that recognizes the data from the user.
Confidence measure	It takes the decoder's n-best decrypted word sequence and examines their ratings to decide whether or not to transmit the outcome.
The pronunciation errors analyzer	Analyzes speech recognition outcomes and generates feedback messages for the user.
The intonation analyzer	Evaluates the frequency curves of the user's speech and provides suggestions for prosodic and rhythmic faults.
Feedback generator	Map identifies mistakes to feed back comments that describe the users' shortcomings and provide guidance on how to improve their pronunciation.

The tool has two main functions. It firstly analyzes a mispronounced statement, even if it is spoken incorrectly, and then it uncovers the speaker's pronunciation faults at the phoneme stage. The Automatic Speech Recognizer (ASR) as well as the Pronunciation Analyzer (PA) are two modules of the system that perform these two functions. The ASR's job is to record the user's words in the system, whereas the pronunciation analyzer analyzes the ASR's output to determine if the pronunciation is correct or not, as well as to identify prototypically aberrant phonemes (i.e., finding on what part of the utterance the feedback should be focused).

Only the possible pronunciation alternatives that encompass frequent forms of pronunciation problems are analyzed to improve the interface's effectiveness. A method for automatically generating pronunciation hypotheses is employed [28–30]. The pronunciation assumptions are reached using this method, which employs matching criteria to recognize pronunciation patterns and generate potential pronunciation mistakes. The Speak Correct interface adaptor and the confidence score module are thoroughly described in the following sections.

4. The Speak Correct Interactive Experiments

Because of certain complexities in the Arabian accent, exercising the skill of speaking as a procedure of phonemes presents a significant challenge for speech recognition mechanisms, particularly for non-native English-speaking people. To address such issues, the proposed solution provokes users to say phrases and utterances, and then recognizes the phoneme sequence data of entry statements to classify submissions for them. As a result, training the Speak Correct disciplined component is a challenging task; each phrase with pronounced phonemes should be properly trained in order for the suggested scheme to recognize such phrases.

For analysis, the English term “picture” may be pronounced as “/pik-cher/” or “/p k ch r/”; if the user pronounced such a term as it is trained (unique frame order of the lattice chart), therefore, the output is correct. According to the literature [31–33], the speech recognition system is unreliable when the using phoneme recognition approach, and speech recognition achieves about 80% correctness.

The authors developed a collaborative system that trains the word utterances and orders of the phonemes. With the help of the proposed Speak Correct computerized interface, users can correct mis-recognized phonemes by constructing improvement utterances according to the wave graph reactions.

The proposed Speak Correct system aims to realize the correction of speaking English words; therefore, the originalities of the utterance evaluations are summarized as follows:

- Word-based error correction. The Speak Correct interface’s envisaged communication component allows the user to demonstrate improper phonemes based on the phoneme discrepancies of the pronounced statements. To successfully complete the speech correction, the platform must first identify the phoneme mistakes and then correct them after presenting the appropriate phoneme patterns. Throughout this engagement, the platform will recognize the users’ statements using pre-defined grammatical structures, and then investigate and demonstrate the wave of such statements.
- History-based evaluation. The presented Speak Correct platform analyses the effectiveness of users’ utterances by using the previous data mistakes of phoneme sequence data that were captured (corrected or not) earlier during testing process experimentation.

4.1. Accent Utterances and Phoneme Errors in Speak Correct

To determine the dependability and subsequent comparison of the users’ statements, the Generalized Posterior Probability (GPP) confidence metric was used. As a result, the phoneme range was evaluated with the help of the phoneme-based confusion/matching matrix. The phoneme confusion/matching matrix was created using the saved Saudi as well as Egyptian dialects information, which included 300 speech sample illustrations from 70 presenters (35 men and 35 women) for each area, totaling 21,000 illustrations. To create the confusion matrix, Nakamura and colleagues employed a phoneme recognition system.

Such a confused matrix $C(\alpha, \beta)$ signifies the β value of phonemes accepted by α phonemes. The phoneme range may be calculated through the following equation:

$$D(\alpha, \beta) = -\log \frac{C(\alpha, \beta)}{\sum C(\alpha_i, \beta)}$$

GPP could also be used to confirm the recognition of the sub-phrase, phrase, and statement identification. It is computed by averaging the probabilities of the various entities (sub-phrase, phrase, or statement). The association among the appreciation efficiency and also the GPP aimed at voice recognition was studied, and it was matched inside the Speak Correct dataset, as illustrated in Figure 3.

Figure 3. Relationship between recognition accuracy and the GPP values for the phonemes in the Speak Correct system.

4.2. User History Evaluation in Speak Correct

The dataset for Speak Correct contains statements from 40 respondents. In the IT Institution, all of the trainees were indigenous Arabic women and men. Every trainee recommended that the Speak Correct interface be tested with ten illustrations for every lesson. Such instances were chosen during the individual liberty selection process to verify the incorporation of the Speak Correct data source (12,756 utterances). The dataset was divided into two sections: calibration and evaluation.

Different particulars of the user history assessment can be demonstrated as arrangements of instances $\{e1, e2 \dots en\}$ that were registered previously during the interaction of Speak Correct. During the interaction of the Speak Correct system, in which some mistakes can be matched, the system ensures that each correction result of the phoneme sequences, p_i , are different from the non-corrected phonemes. The following algorithm of the user assessment was presented in the Speak Correct evaluation algorithm (below). It is word-based correction, not corrected performance, with four different kinds of experimentation used in this proposed algorithm: “start changing the standard of testing/change associated instance,” “stop/end” assessment, and “change the level of verification pertaining illustration”. The original decision, “It is correct”, is used to proceed and recognize the new Speak Correct words. The second determination, “It is not correct,” is used to rectify the speech or phoneme pronunciation. To modify or and choose another standard or instance, use “Modify the standard of checking associated instance.” The final option is “stop/end,” that can be used to finish the Speak Correct assessment algorithm.

Figure 4 shows the evaluation algorithm for the Speak Correct computerized interface.

Begin

1. Initial Step: level = 1; example = 1; correct=0; nonCorrect=0; m = maxNoOfExamples;
 2. Use the recognized phoneme sequences of the training module of Speak Correct (Lattice Graph with learnt rules)
 3. Request the user utterances entry (Level and examples; Speak the displayed word)
 4. According to the user responses Do the following:
 - If the user utterance is recognized with decision “is correct” go to Step 5
 - If the user utterance is recognized with decision “is not correct” go to Step 8; the entry selection has phoneme error(s)
 - If the user selects “Stop/End” go to Step 10
 5. correct = correct + 1
 6. If correct > m Then go to Step 10; Otherwise go to Step 4
 7. Use the recognized phoneme sequences e_i of the word at index i; Store it in the correct set examples; go to Step 4
 8. nonCorrect = nonCorrect + 1; Store the tested example in the non-Correct set
 9. According to the user’s response Do the following:
 - If the user selects “Stop/End” decision, go to Step 10
 - If the user selects another Level/Example, go to Step 4
 - Otherwise display the correct list and the non-correct list;
 - Calculate the user Score
 10. Stop/End
- End**

Figure 4. Speak Correct evaluation algorithm.

The empirical assessment of the Speak Correct word pronunciation assignment is discussed in the next section. Initially, authors compared the collected findings before as well as after the training modules to assess Speak Correct’s effectiveness.

4.3. Speak Correct Evaluation Scenario

The Speak Correct method instruction is divided into six stages, each of which comprises between five and fifteen instances. There are around 390 selective words in these levels, which encompass vowels, consonants, as well as clusters. The Speak Correct scores are distributed throughout a series of 160 instructional lessons on aggregate. As samples, each lesson has up to 15 phrases. For non-native individuals, the words constitute a group of similar accent defective terms. Users review the stages and relevant instances of the phrase’s data set prior to attempting to perform.

4.4. Speak Correct Feedback

The inaccuracy categorization is accomplished by correlating the properties of the recorded phoneme sequences of a specific word to features of the Speak Correct system’s previously taught users. As a result, the system provides users with comments on their faults (students). The response is based on categorizing errors by the amount of errors at every stage across all cases.

Textual and visual features are used to deliver Speak Correct feedback. Participants can choose lessons using a variety of categories and examine the pronunciation of any terms in the lesson. Learners might also search for such words by listening to the audio recordings. In order to achieve proper pronunciation, the proposed Speak Correct method converts identification findings into meaningful information by emphasizing mispronounced words as well as offering descriptions employing phonetic characters using articulation characteristics.

Twenty participants were engaged in the assessment process. Participants simultaneously utilize the identical gadgets (for instance, laptops and headsets) and must accomplish all of the courses for every stage after going through 6 stages and 15 lists of phrases. In this stage, two groups of participants are involved (Saudi group as well as the Egyptian group). The majority of the phrases were spoken accurately, as seen in Figure 5.

Figure 5. The levels and associated lessons testing the Speak Correct system.

Pronunciation aid using visual information is depicted in Figure 5. If the conversation is not progressing, the response provides recommendations and clarifications regarding what the user may attempt to communicate. The response tab displays a visual representation of all the objects obtained by the user. Lastly, the evaluation provides students with course summaries at various levels.

4.4.1. Speak Correct Scoring

The below equation is used to determine the penalty value for an inaccuracy, e , of a specific phrase. w :

$$\text{Error} = E_w \times W_{er}$$

where E_w represents the weight related to the phonological error, as listed in Table 2, and W_{er} is the cost related to each possible error pattern. The W_{er} is calculated by considering the frequency of the errors.

Table 2. Phonological errors.

Error Type	Error Description	Weight (E_w)
Substituting	Dental to alveolar, unstressed to stressed vowel, bilabial plosives (p, b), labio-dental fricative ((v, f), (s, z), (t, d), ... , etc.)	4
Deletion	Consonant and vowel deletion	3
Insertion	Consonant and vowel insertion	2

Consequently, such weights' values are calculated and based on the observed frequencies after normalization, as illustrated in the following equation:

$$W_{er} = F_{er} \times 20 + 1.$$

where F_{er} represents the relative frequency of the error phoneme, it is normalized by the number of cases in the training module. The total score for any user can be articulated by the following formula:

$$\text{Score} = \sum_w \text{Error} (E_w, W_{er})$$

4.4.2. Data Set and Dictionary in Speak Correct

In fact, the data set of the Speak Correct system presents varieties, taking into consideration the following characteristics:

1. Gender: males or females.
2. Educational level: university undergraduate group.
3. Age: young speakers, the age is between 17–20 years.
4. Regional country: two regions in Saudi Arabia and another two regions in Egypt.

The dataset is based on the expertise of the linguistic teachers and analysis of the data collected from 150 students (male and female). Therefore, the recordings were made in a clean environment without any vibration. The recorded files have a wave format, with mono channel sampling at 16 KHz and 32 bits.

The total number of recording voices was 2100 samples for the entire data set (Egyptian males and females). Two regions were used for the recordings, the first region was located in the middle of Egypt (Cairo), and the second region in the northern of Egypt (Alexandria). The number of speakers used for training was 70 students (35 males and 35 females). The test corpus included 300 words and sentences to cover all the English pronunciation defects in Saudi and Egyptian accents.

5. Results

5.1. Speak Correct Experimental Results

The correctness using the Word Error Rate (WER) that is obtained from the Levenshtein distance was used to evaluate the effectiveness of the Speak Correct system. The following formula may be used to compute the WER:

$$\text{WER} = (P_{\text{sub}} + P_{\text{del}} + P_{\text{ins}}) / N$$

In which P_{sub} -denotes the amount of substitution phonemes;

P_{del} , the quantity of deletion phonemes;

P_{ins} , the quantity of insertion phonemes;

N , the overall amount of inaccuracies.

As a result, the investigation is developed and evaluated to assess the presented Speak Correct computerized interface, which is premised on a phonetically proposed corpus, and

is therefore independent of the presenter's accent or sexual identity. The computed value of the WER demonstrates that the Speak Correct recognition system performs satisfactorily.

5.2. Arabic Regional Accent

In the Speak Correct system, an assessment method was established. This technique focused on process response and measures the degree of consistency precision. Because of the difficulty of speech procedures and the existence of tunable tolerances and specifications, the assessment of mispronunciation in these kind of assessment processes is critical. The findings show that the Egyptian accent in the middle region (Cairo) has a higher accuracy percentage of acknowledgements than the accent in the northern province (Alexandria). The Speak Correct proposed system recognizes the majority of the phrases pronounced by presenters in this province.

In fact, the samples of spoken dialect in this province are very efficient in improving the system's accuracy. However, in terms of the provincial accent, the effectiveness varies significantly depending on the location. Table 3 depicts the dissemination of speakers (training as well as testing) for each province.

Table 3. Speaking (for training and testing) by regions.

Region/Localities		Training		Testing	
		Male	Female	Male	Female
Egypt	Cairo	17	17	8	8
	Alexandria	18	18	9	9
Saudi Arabia	Jeddah	10	10	3	3
	Rabegh	15	15	5	5

It is necessary to acknowledge verbal communication languages in such regions of the world. There are important cultural discrepancies among the two major areas when the Speak Correct characteristics are considered. In Cairo, Egypt, the Speak Correct effectiveness is very significant (English is considered to be an extremely popular language in this province), and it is commonly used in colleges and universities. Alexandria, the second largest city in Egypt (after Cairo), has a high recognition accuracy. In Saudi Arabia, the Rabegh location received the lowest rating.

Again, for the experimental results of the WER, within substitution, deletion, and insertion, Table 4 shows the results, and Figure 6 and Table 5 represent the WERs for each locality.

Table 4. Word Error Rate (WER) relative to different localities.

Region/Localities		Testing	
		Male	Female
Egypt	Cairo	10%	11%
	Alexandria	11%	12%
Saudi Arabia	Jeddah	16%	17%
	Rabegh	19%	20%

Table 5. Regions and localities for WER.

Region/Localities		Substitution	Deletion	Insertion
Egypt	Cairo	11%	9%	8%
	Alexandria	12%	10%	9%
Saudi Arabia	Jeddah	16%	12%	11%
	Rabegh	21%	13%	15%

Figure 6. Speak Correct errors for each locality.

5.3. Gender-Based Analysis

It is well known that female pronunciations are distinct compared to males' voices. Accordingly, the second variable studied in this paper is the impact of gender on the Speak Correct computerized interface. We know that females have a higher pitch and higher frequency values [30–33]. These characteristics can negatively affect the speech recognition of the Speak Correct by increasing the WER values. At the localities, Cairo and Alexandria, the ratio between the speakers is approximately the same between males and females. However, this difference increases in Jeddah and Rabegh localities in Saudi Arabia. Figure 7 shows the comparison between males and females considering the regions and localities.

Figure 7. Comparison between males and females according to locality.

6. Discussion

The operational rating outcomes of the main test were managed among 0 and 2, with the aggregate rating in the upper range as well as the standard deviation within of 0.5. The Table 4 Word Error Rate (WER) is relative to the different localities; the transitional technical rating outcomes were all dispersed somewhere around 1 and 3, with the aggregate rating around 2.25 and the standard deviation within 0.5. All of the data demonstrated that the practical rating outcomes satisfied the standards of the subject rating, and the ratings were generally focused and consistent. When matched to the subjective scoring findings, the

quantitative approach of the technical rating data revealed a high level of association with the subjective rating, evidence that supports the research's hypothesis. This hypothesis, moreover, requires more evidence; therefore, in this segment, different components of the manual as well as technical rating outcomes are evaluated and contrasted one by one to demonstrate the study hypothesis that speech evaluation technology could indeed accomplish the rating of Speak Correct post-listening repeating concerns.

In this research, an empirical dataset was employed to perform experimental studies on speech quality assessment techniques. The Speak Correct computerized interface introduced in this paper was initially contrasted with specialist scores for correlation experiments; then, the evaluation process in this paper was contrasted with another established speech fluency assessment technique; and, finally, the rating effectiveness of the fluency evaluation procedure in this study was analyzed. Furthermore, this research integrated the phonetic pronunciation bias speech network to recognize phonetic mispronunciation, and the phoneme margin of error of every fluency standard was collected to evaluate the system's erroneous corrective feedback performance. The findings indicate that the high accuracy of the Speak Correct computerized interface described in this work, based on a composite of attributes, has a strong correlation with the real ratings of human operators. The empirical findings reveal that the feature-based combinatorial assessment approach surpasses other fluency assessment methods. This is mostly due to the fact that the approach provided in this work includes information from both the phoneme length and the phoneme acoustical score characteristics, and calculates the most efficient assessment methods by optimizing the regression analysis.

7. Conclusions

The goal of this work was to discuss and analyze the efficiency of the Speak Correct system, which is used to rectify non-native English speakers' speech. This paper conducts an empirical assessment of the Speak Correct computerized interface to show its effectiveness. Saudi Arabia and Egypt are the two main countries from which the data was collected. There are two geographical regions in each country. An interactive suggestion system is included in the proposed system to encourage users to enhance their language capabilities. The proposed Speak Correct computerized interface may automatically choose a recording event to use as a reference, define statistical judgment parameters for compensation, as well as correct sections from recordings as appropriate. After rectification, the frequency of speech signals is excellent. The Speak Correct's computerized recommendations are associated with the outcomes of subjective auditory assessments. A calculation of the correction probabilities and a matching listening test were used to verify the performance of this system. The Speak Correct computerized interface presented in the present paper enhances the pronunciation for the impact of the native tongue on the pronunciation of words in foreign languages. While it achieves an improved outcome, there are still some deficiencies that can be analyzed further in specific instances of using numerous speech functionality criteria for comprehensive assessment and connectivity in the design procedure in future research directions.

Author Contributions: Conceptualization, K.J., H.A.-B., W.A.-J., M.R. and S.A.; methodology, K.J., H.A.-B., W.A.-J., M.R. and S.A.; software, K.J., H.A.-B., W.A.-J., M.R. and S.A.; validation, K.J., H.A.-B., W.A.-J., M.R. and S.A.; formal analysis, K.J., H.A.-B., W.A.-J., M.R. and S.A.; investigation, K.J., H.A.-B., W.A.-J., M.R. and S.A.; resources, K.J., H.A.-B., W.A.-J., M.R. and S.A.; data curation, K.J., H.A.-B., W.A.-J., M.R. and S.A.; writing—original draft preparation, K.J., H.A.-B., W.A.-J., M.R. and S.A.; writing—review and editing, K.J., H.A.-B., W.A.-J., M.R. and S.A.; visualization, K.J., H.A.-B., W.A.-J., M.R. and S.A.; supervision, K.J., H.A.-B., W.A.-J., M.R. and S.A.; project administration, K.J., H.A.-B., W.A.-J.; funding acquisition, K.J., H.A.-B., W.A.-J. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH)—King Abdulaziz City for Science and Technology (KACST)—Kingdom of Saudi Arabia—project number (10-INF-1406-03).

Institutional Review Board Statement: Not available.

Informed Consent Statement: Not available.

Data Availability Statement: Not available.

Acknowledgments: The authors thank the Science and Technology Unit, King Abdulaziz University for the technical support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wiggers, P.; Rothkrantz, L.J.M. Automatic speech recognition using hidden Markov models, Course IN4012TU. *Real-Time AI Autom. Spraakherkenning* **2003**, *70*, 1–20.
2. Noormamode, W.; Rahimbux, B.G.; Peerboccus, M. A speech engine for mauritian creole. *Inf. Syst. Des. Intell. Appl.* **2019**, *36*, 389–398.
3. Kemble, K.A. An introduction to speech recognition. In *Voice Systems Middleware Education*; IBM Corporation: Endicott, NY, USA, 2001; Volume 16, pp. 154–163.
4. Bracha, O. The folklore of informationalism: The case of search engine speech. *Fordham Law Rev.* **2013**, *82*, 1629–1633.
5. Tur, G.; de Mori, R. Spoken language understanding: Systems for extracting semantic information from speech. *Fordham Law Rev.* **2011**, *82*, 1629–1633.
6. Laskowski, K.; Shriberg, E. Comparing the contributions of context and prosody in text-independent dialog act recognition. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 5374–5377.
7. Song, Y.I.; Wang, Y.Y.; Ju, Y.C.; Seltzer, M.; Tashev, I.; Acero, A. Voice search of structured media data. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 3941–3944.
8. Heracleous, P.; Aboutabit, N.; Beutemps, D. Lip shape and hand position fusion for automatic vowel recognition in cued speech for French. *IEEE Signal Process. Lett.* **2009**, *16*, 339–342. [[CrossRef](#)]
9. Osberger, M.J. Speech intelligibility in the hearing impaired: Research and clinical implications. *Intelligibility Speech Disord.* **1992**, *74*, 233–265.
10. Tsubota, Y.; Dantsuji, M.; Kawahara, T. An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors. *ReCALL* **2004**, *16*, 173–188. [[CrossRef](#)]
11. Almekhlafi, G. The effect of computer assisted language learning on United Arab Emirates English as a foreign language school students' achievement and attitude. *J. Interact. Learn. Res.* **2006**, *17*, 121–142.
12. Huijbregts, M.; McLaren, M.; Leeuwen, D.V. Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, 22–27 May 2011; pp. 4436–4439.
13. Wang, H.; Waple, C.J.; Kawahara, T. Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition. *Speech Commun.* **2009**, *51*, 995–1005. [[CrossRef](#)]
14. Vu, N.T.; Wang, Y.; Klose, M.; Mihaylova, Z.; Schultz, T. Improving asr performance on non-native speech using multilingual and crosslingual information. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014; pp. 200–203.
15. Kadir, N.A.A.; Sudirman, R. Vowel effects towards dental arabic consonants based on spectrogram. In Proceedings of the 2011 Second International Conference on Intelligent Systems, Modelling and Simulation, Kuala Lumpur, Malaysia, 27–28 January 2011; pp. 183–188.
16. Spring, R.; Tabuchi, R. Assessing the Practicality of Using an Automatic Speech Recognition Tool to Teach English Pronunciation Online. *J. Engl. Teach. Movies Media* **2021**, *22*, 93–104.
17. Evers, K.; Chen, S. Effects of Automatic Speech Recognition Software on Pronunciation for Adults with Different Learning Styles. *J. Educ. Comput. Res.* **2021**, *59*, 669–685. [[CrossRef](#)]
18. Eskenazi, M. Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Lang. Learn. Technol.* **1999**, *2*, 62–76.
19. Evers, K.; Chen, S. Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. *Comput. Assist. Lang. Learn.* **2020**, 1–21. [[CrossRef](#)]
20. Cao, Q.; Hao, H. Optimization of Intelligent English Pronunciation Training System Based on Android Platform. *Complexity* **2021**, *2021*, 5537101. [[CrossRef](#)]
21. Moxon, S. Exploring the Effects of Automated Pronunciation Evaluation on L2 Students in Thailand. *IAFOR J. Educ.* **2021**, *9*, 41–56. [[CrossRef](#)]
22. García, C.T.; Escudero, D.; Ferreras, C.G.; Arenas, E.C.; Cardenoso-Payo, V. Evaluating the Efficiency of Synthetic Voice for Providing Corrective Feedback in a Pronunciation Training Tool Based on Minimal Pairs. In Proceedings of the 7th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2017), Stockholm, Sweden, 25–26 August 2017.

23. Fujii, K.; Saitoh, N.; Oka, R.; Muneyasu, M. Acoustic echo cancellation algorithm tolerable for double talk. In Proceedings of the 2008 Hands-Free Speech Communication and Microphone Arrays, Trento, Italy, 6–8 May 2008; pp. 200–203.
24. Kacur, J.; Rozinaj, G. Adding voicing features into speech recognition based on HMM in Slovak. In Proceedings of the 2009 16th International Conference on Systems, Signals and Image Processing, Chalkida, Greece, 18–20 June 2009; pp. 1–4.
25. Yopp, H.K.; Yopp, R.H. Supporting phonemic awareness development in the classroom. *Read. Teach.* **2000**, *54*, 130–143. [[CrossRef](#)]
26. Treiman, R. Onsets and rimes as units of spoken syllables: Evidence from children. *J. Exp. Child Psychol.* **1985**, *39*, 161–181. [[CrossRef](#)]
27. Hazen, T.J.; Hetherington, I.L.; Shu, H.; Livescu, K. Pronunciation modeling using a finite-state transducer representation. *Speech Commun.* **2005**, *46*, 189–203. [[CrossRef](#)]
28. Liu, G.Z.; Liu, Z.H.; Hwang, G.J. Developing multi-dimensional evaluation criteria for English learning websites with university students and professors. *Comput. Educ.* **2011**, *56*, 65–79. [[CrossRef](#)]
29. Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; Valtchev, V.; et al. *The HTK Book*; Cambridge University Engineering Department: Cambridge, UK, 2002; Volume 3, pp. 1–12.
30. Kurian, C.; Balakrishnan, K. Continuous speech recognition system for Malayalam language using PLP cepstral coefficient. *J. Comput. Bus. Res.* **2012**, *3*, 1–24.
31. Ansari, M.T.J.; Khan, N.A. Worldwide COVID-19 vaccines sentiment analysis through twitter content. *Electron. J. Gen. Med.* **2021**, *18*, 1–15. [[CrossRef](#)]
32. Mishra, S.; Bhende, C.N.; Panigrahi, B.K. Detection and classification of power quality disturbances using S-transform and probabilistic neural network. *IEEE Trans. Power Deliv.* **2007**, *23*, 280–287. [[CrossRef](#)]
33. Ansari, M.T.J.; Pandey, D.; Alenezi, M. STORE: Security threat oriented requirements engineering methodology. *J. King Saud Univ.-Comput. Inf. Sci.* **2018**, *54*, 1–18. [[CrossRef](#)]