Computer Systems Science & Engineering
DOI: 10.32604/csse.2022.024967
Article

Tech Science Press

Speak-Correct: A Computerized Interface for the Analysis of Mispronounced Errors

Kamal Jambi[1,*], Hassanin Al-Barhamtoshy[1], Wajdi Al-Jedaibi[1], Mohsen Rashwan[2] and Sherif Abdou[3]

[1]Faculty of Computing & Information Technology, King Abdulaziz University, Saudi Arabia
[2]Faculty of Engineering, Cairo University, Egypt
[3]Faculty of Computers, Cairo University, Egypt
*Corresponding Author: Kamal Jambi. Email: kjambi@kau.edu.sa
Received: 06 November 2021; Accepted: 09 December 2021

Abstract: Any natural language may have dozens of accents. Even though the equivalent phonemic formation of the word, if it is properly called in different accents, humans do have audio signals that are distinct from one another. Among the most common issues with speech, the processing is discrepancies in pronunciation, accent, and enunciation. This research study examines the issues of detecting, fixing, and summarising accent defects of average Arabic individuals in English-speaking speech. The article then discusses the key approaches and structure that will be utilized to address both accent flaws and pronunciation issues. The proposed SpeakCorrect computerized interface employs a cutting-edge speech recognition system and analyses pronunciation errors with a speech decoder. As a result, some of the most essential types of changes in pronunciation that are significant for speech recognition are performed, and accent defects defining such differences are presented. Consequently, the suggested technique increases the Speaker's accuracy. SpeakCorrect uses 100 h of phonetically prepared individuals to construct a pronunciation instruction repository. These prerecorded sets are used to train Hidden Markov Models (HMM) as well as weighted graph systems. Their speeches are quite clear and might be considered natural. The proposed interface is optimized for use with an integrated phonetic pronounced dataset, as well as for analyzing and identifying speech faults in Saudi and Egyptian dialects. The proposed interface detects, analyses, and assists English learners in correcting utterance faults, overcoming problems, and improving their pronunciations.

Keywords: Speech recognition; computerized interface; arabic dialects; accent defects; acoustic error

1 Introduction

As english has become a global language, several people are learning and speaking it. Among the most prevalent, but extremely complicated, activities to address when learning English as an additional language is

the use of English in verbal communication. This is because English proficiency has become a need, particularly for those wishing to develop in specific domains of human effort. Speaking English as a foreign language is a significant issue for several people throughout the world. This occurs for a variety of reasons. When learning a foreign language, you must grasp that it employs a wide variety of sounds as well as orthographic norms than your native speech. Learners frequently try to mimic the sounds by using ones they are already acquainted with and pronouncing texts as if they were composed in their home languages.

Any speech processor is made up of two or three components. The first is a word recognizer, which turns any activities with respect into fundamental word sequence. The next is a phoneme recognition system based on any paradigm, such as HTK, that turns the provided utterance into a phoneme series. This series is then evaluated with a matching algorithm to identify the most relevant keywords.

The acoustic input O is treated as a sequence of individual "symbols" or "observations", represented by symbols: O = o1, o2, o3, …, ot. Likewise, a sentence/word will be preserved as a string of words/phonemes: W = w1, w2, w3, …, wn. Various general terminologies used throughout this document, are explained in this section [1–4].

This research paper is written specifically for practising non-specialist learners/students who really need to communicate In english appropriately. The suggested approach uses the fewest linguistic terms possible and seeks to deliver straightforward evaluation results in visual form. This is especially true in the case of speech, where unnecessary technical information can be perplexing to non-specialists. Furthermore, we consider that the explanations and technical aspects provided here are accurate and sufficiently detailed within the context of this technique.

The objectives of this study include designing, building, and assessing a prototype system which can detect and analyse mispronounced mistakes in adult learners' teaching-based operations. This system is not designed to be a total replacement for the actual class teacher, but rather to serve as an additional tool to assist him/her in teaching the fundamental skills of learning and practising in the various fields of advancement.

2 Related Works

To comprehend signs communicated by a voice signal, "Spoken Language Understanding (SLU)" as well as "Natural Language Understanding (NLU)" are utilised. The derived conceptual description of natural language sentences is contributed by SLU and NLU [5]. Signs can be used for understanding and can be encoded into signals that provide extra information. Moreover, the suggested two systems feature an Automatic Speech Recognition (ASR) component and should be sound sensitive, depending on the pattern of spoken language as well as ASR failures.

Dialog categorization and automated segmentation are critical for interpreting SLU. This research proposed a system for segmenting and classifying multiparty meetings based on contextual speech as well as prosodic traits. "Contextual elements are better for recognising, but prosodic features are better for identifying base processes and backchannels," they discover [6]. In data screening, voice is employed. A phonetic matching method has been provided, and the given application is used in the music sector; search errors on text and spoken queries have been reduced [7].

Heracleous et al. (2009) offer consonant and vowel identification in French using HMM in their work. Speech hand-shapes with lip-patterns (as a graphic communication medium) build all oral language sounds clearly, particularly for deaf as well as hearing-impaired individuals. The goal of this study is to overcome the difficulty of lip reading and so enable deaf children and adults to completely perceive spoken language [8,9].

Two ways are used to improve comprehensibility for those who are deaf or deafeningly. The first technique is utilised in the setting of discussion for hearing-impaired users; the second strategy tries to improve speaking-impaired people's intelligibility. The results revealed that an improvement in intelligibility was not attained, and listeners preferred the altered speech of an alternative approach.

Linguistic knowledge is commonly employed in ASR to improve mistake prediction. Tsubota et al. modelled 79 different types of pronunciation error patterns to distinguish Japanese students' English, and the paper proposes a straightforward way to following the pitch of two active presenters. It used HMM to track frequency over time. To illustrate experimental findings, a statistical approach might be utilised. The research demonstrated that the suggested technique outperformed the multi-pitch tracking algorithm [10].

To enable students to acquire Japanese, a Computer-Assisted Language Learning (CALL) approach has been designed [11]. The study provides a model for detecting lexico-grammatical problems in pronunciation, as well as inputting sentence faults. The researchers of [12] presents a technique for generating acoustic sub-word entities from a spoken term recognition system, which can be used to replace standard phone models. The system uses an unsupervised method to construct a series of speaker models that are independent of the training data. Another work [13] presents a decision tree that may be used for error categorization in automatic voice recognition to discover critical and redundant errors.

The study [14] concentrates on non-native accent issues in uninterrupted speech recognition. It attempts to investigate the transformation principles of non-native speech expressed in Mandarin by local speakers. As a result, a corpus in Mandarin is used to train the HMM algorithms and check the effect of voice recognition. As a consequence, the results show that the collected information is useful for adjusting a native speaker ASR method to model nonnative accented content.

Another article, titled "Vowel Effects on Dental Arabic Consonants Based on Spectrogram" [15], investigated the influence of Arabic vowels on Arabic consonants with three easy diacritics by Malaysian youngsters. Malaysian children add these vowels to the fundamental consonants using three simple diacritics. The location of articulation is essential in dental consonants and formant frequencies, according to the report.

Kensaku Fujii et al. [16] cancelled acoustic echo by replacing the differential between adaptable filter coefficients for the prediction error in the former. Nevertheless, Juraj Kac and Gregor Rozinaj et al. [17] investigated the effect of replacing a number of speech parameters with voiced/unvoiced data in approximated pitch value. Investigations were carried out on the data for smartphone platforms, employing context dependent as well as independent phonemes from HMM models.

3 General Terminology

This general terminology is intended to help readers comprehend commonly used terminology and phrases when researching, interpreting, and assessing scholarly method of investigation. Basic words or phrases are also discussed in the perspective of how they relate to commence research for the SpeakCorrect computerised interface.

3.1 Phoneme

A phoneme is the smallest element of speech; it is used to differentiate meaning. The phoneme is the most essential element in the word; every word is made up of phonemes, therefore changing them changes the essence of the word. When the sound [b] in the word "pin" is substituted by the sound [p], the word becomes "bin." As a result, /b/ is a phoneme [18]. The phoneme is characterized purely in terms of changes in its immediate phonetic context that constrain allophonic variations, with no regard for higher-level language structures. To describe a phoneme, just remark that two essentially identical

utterances (in practise, words because they are the shortest utterance length in a naming task) vary due to the presence of two separate sound components.

3.2 Phone

It is the lowest physical sound part. Phones are thus the physical manifestation of phonemes. Allophones are sometimes defined as phonic variations of phonemes [19].

3.3 Phonetics

Phonetics is the analysis of human voice, specifically the qualities of spoken tones [20].

3.4 Phonology

Phonology is the study of audio systems and the abstraction of sound units, such as phonemes as well as phonological principles. As a result, phonetics concepts are universal, whereas phonology is language specific. [21] represents the phonetic of a sound, while // represents the phoneme.

3.5 Syllable

The term "syllable" refers to a component of pronunciation. It is greater than a single sound but smaller than a word. The syllable begins and finishes with consonants and also is made up of vowels [22].

4 Common Pronunciation Errors

One of the major issues with spoken English is the preference for informal over formal communication. In informal communication, ellipsis, contractions, and relative clauses lacking relative pronouns are increasingly common. Formal speech follows traditional grammatical standards and is typically employed for strangers or in writing. "He is my first brother," for instance, sounds more formal than "He's my first brother," which sounds more informal. Fig. 1 depicts almost all of the pronunciation problems that must be considered in effective training and evaluation models [23]. It can be divided into two sorts of errors, as illustrated in the Fig. 1: phonemic and prosodic.

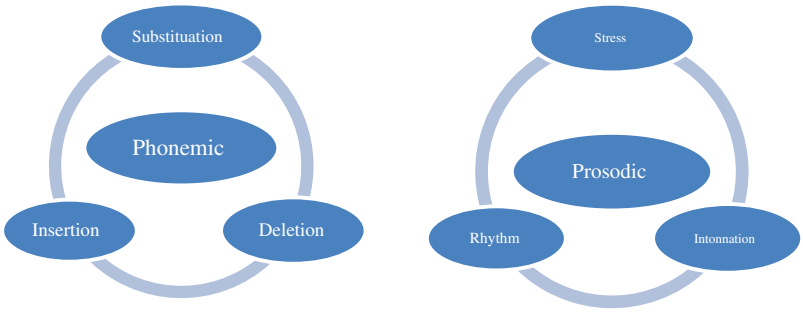

Figure 1: Classification of pronunciation errors

- In this work, phonemic errors are classified as substituted, removed, or introduced. There are even minor errors "when the proper phoneme is greater or fewer being pronounced" [24].
- Stress, rhythm, and intonation are the three types of prosodic faults.

As a result of these two types of faults, pronunciation becomes a multi-dimensional issue. As a result, a wide variety of measures are used to quantify these characteristics [24,25].

We identified a considerable body of research outlining the usual patterns of error committed by Korean Learner Segmental Errors (KLEs) while developing SpeakCorrect. A pilot corpus of suggested English speech data from a variety of various forms of material is collected and phonetically labeled. Short sections of text, phrase prompts, and words with terribly challenging consonant clusters are included in the *corpus* (e.g., refrigerator). In total, 25,000 speech specimens were taken from 111 Korean learners for the pilot *corpus*. The *corpus* allows for direct comparisons of realised phone sequences to anticipated canonical patterns from native speakers. A description of Korean for English speakers likewise does not distinguish between fricative /f/ and /v/, instead substituting /p/ and /b/. Other commonly reported errors involve aspirated /t/ for / as well as un-aspirated /t/ for /. In addition, the report demonstrates the most common segmentation errors encountered. The current work focuses on mistake substitution, removal, and addition in phonemics.

5 SpeakCorrect Error Detection Processing Engine

The proposed computerised interface is divided into three major steps. The first stage is utilised for pre-processing or feature training collection, development of pronunciation hypotheses, and HMM adaptation. The second stage employs a decoder that detects the user's supplied speech, as well as a confidence measurement and a pronunciation fault detector. This step's output is the third stage, which comprises analysis of the respond to changes and generates recognised errors, advice to the repaired errors, and an evaluation method. As a result, phone estimation use statistical techniques (such as neural networks or Gaussian models) to detect particular speech sounds such as f or s. This step creates a vector of possibilities over phones for every frames. The last stage contains numerous sub-modules, the most significant of which is the decoding module, which is used to discover the sequence of words with the greatest chance given the acoustic occurrences.

Fig. 2 depicts the Data Flow Diagram (DFD) of the suggested SpeakCorrect interface. Each word entails the students being depicted as in ATN or Lattice, which is a visual showing several phonetic levels and accompanying training (Saudi or Egyptian accent defects). As a result, Speak Correct gives users the option of choosing their own degrees and instances. After entering the syllables, computerized speech recognition is used, which is backed by trained instances in the manner of a grammar network (Lattice graph) for the specific word. Errors will be discovered and assessed, and feedback statistics will be provided to users, resulting in the evolution of the model.

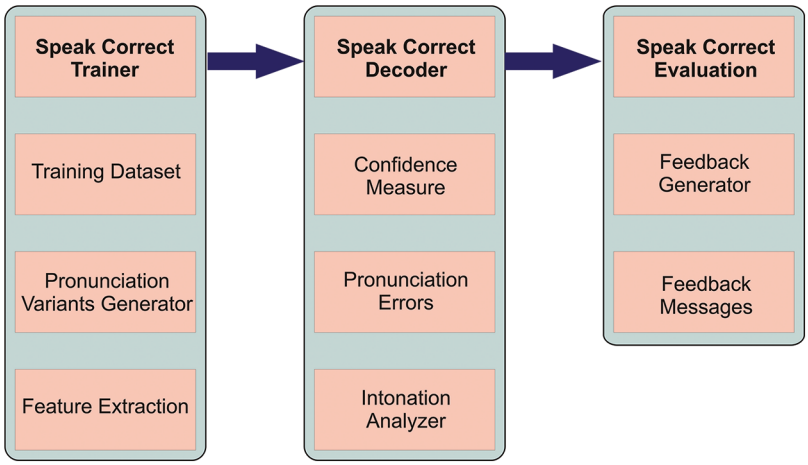

Figure 2: SpeakCorrect error detection overview

To explain the feature extraction procedure, beginning with sound waves as well as finishing with a feature vector. Firstly, an input audio wave is digitised (analog-to-digital conversion), which is accomplished in two steps: Quantification and sampling.

The number of observations taken per second is referred to as the sampling rate. To precisely measure a waveform, at least two samples are required in each cycle: one for the positive half of the wave and one for the negative part, and more than two samples per cycle improve amplitude precision.

There are volume measurements for each second of voice at each sample rate. Quantization refers to the technique of expressing such quantities as integers. The waveform is transformed after digitization to set the spectral characteristics. Any prominent feature set (Linear Predictive Coding (LPC) or Perceptual Linear Predictive (PLP)) can be used directly to view HMM symbols [26–30]. The phones assessment is a fast method for calculating the likelihood of an ordered set given weighted automata. HMM enables us to add together numerous routes that individually account for the very same observation sequence.

The decoding stage is concerned with determining the right "underlying" succession of symbols/patterns. As a result, the Veterbi algorithm provides an efficient method of addressing the decoding problem by evaluating all potential strings and employing addition rules (such as the Bays rule) to determine their probability of obtaining the observed sequence.

The features are frequently subjected to additional processing in order to match the standard speech models to the presenter's speech attributes. As a result, the speaker adjustment module of Maximum Likelihood Linear Regression (MLLR) is employed to refine the adapted module.

5.1 SpeakCorrect Background Architecture

Several techniques used in voice recognition have been developed by different academics. As a result, the terms phone and syllable are formed [31–33]. Furthermore, the N-gram language concept as well as the HMM are described in greater depth.

Initially, HMM were presented as stochastic approaches for modelling temporal pattern classification and analysis systems. As a result, the HMMs can be shown employing finite state machines, with an assessment from a given state at each transition and output symbol emissions for each state. In another words, to select the most likely word given the insight, the single word with the highest P (word | observation). If w is the predicted right terminology and O is the recorded sequence (individual observation), the Eq. (1) for deciding the proper word is:

$$W = \text{argmax}\ P(o|w) \quad P(w) \tag{1}$$

where P(o|w) denotes likelihood, P(w) denotes prior, w denotes vocabulary, w denotes right word, and o denotes observation. Once the probability computations and decoding difficulties for a reduced input comprising of strings of phones have been addressed, feature extraction would be swiftly implicated.

5.2 Acoustic Probabilities Counting

As previously stated, the speech intake can be transformed into a sequence of feature vectors by passing it through signal conditioning transformations, with every vector reflecting one time-slice of the spoken input signal. One frequent method for computing probability on extracted features is to first cluster them into discrete numbered symbols. As a result, the likelihood of a specific cluster can be computed (number of times it occurs in some training set).

This technique is known as vector quantization, which is used to compute observation possibilities or probability density functions (pdf). There really are two typical methods: Gaussian pdfs, which convert the observation vector Ot to a likelihood, and neural networks or multi-layer perspectives, which may be

trained to attribute a possibility to a real-valued feature space in audio. A neural network is a collection of small computer units linked together by weighted connections. When given vector variables, the network calculates a vector of target value.

Mishra et al. [34] proposed a standard model which is founded on a probabilistic neural network that is suited for testing as well as pattern categorization. The number of input voice variables M, the number of recognition patterns required N, and the training instances for each pattern are denoted by S1, S2,… SN comprise the architecture of such probabilistic neural network model. There are four layers: input layer, model layer, summation layer, as well as output layer; the weights among accumulation layer with output layer are calculated as follows (Eq. 2):

$$\mathrm{W}(M) = Si \,/\, \sum_{i=1}^{N} Si \tag{2}$$

As a result, when speaker adaptation occurs, specific qualities such as acoustic gender, dialects, and age would be modelled in speech processing. As a result, the speaker's unique accent should be unaffected.

5.3 SpeakCorrect Principles Modules

Our purpose is to create a model that can find out how it changed this "true" term and therefore recover it. The essential recognition procedures for the complete talk right tasks are as follows.

5.3.1 Main Module

Step 1: Gathering and collecting the speech input samples.

Step 2: Dividing such samples into two parts, one part is for training samples and the second part is for testing.

5.3.2 Training Module

Step 3: Do the following:

3.1 Speech Adaption.

3.2 Confidence measuring.

3.3 Tuning the native Arabic speaker accent.

a. Tuning Saudi accent.

b. Tuning Egyptian accent.

3.4 Intonation training and teaching the pronunciation effects.

Step 4: Using the feature vector of training samples to train the SpeakCorrect model.

5.3.3 Testing Module

Step 5: Do the following steps:

Step 5.1: Establishing the system with the associated acoustic and language model.

Step 5.2: The feature vectors are used to input test samples into network which has been trained.

Step 5.3: Judging the equivalent speech signal class and the speaker characteristics according to the output values.

5.4 Weighted Finite State and Weighted ATN/Lattice

Computing languages and automata concept were utilised to forecast letter sequences, characterise natural language, apply Context-Free Grammars (CFG), present tree transducer ideas, and parse automatic natural language writing. In the 1970s, voice processing investigators recorded NLP grammar utilizing

weighted Finite State Acceptors (FSAs), which could be trained on machine-readable dictionary, *corpus*, and corpora [35–41].

In the nineties, finite state machines and massive training corpora has become the dominant models in speech processing, prompting the development of software development tools for Weighted Finite State Machines (WFSM) [42]. In the twenty-first century, common tree automata toolkits [42–47] have now been designed to aid research.

Although the single WFST or Augmented Transition Network (ATN) that depicts P(S|E) remains complex, model conversion can be converted into a sequence of transducers as shown in Eq. (3).

$$\text{WFSMa (English text)} \longleftrightarrow \text{WFSMb (English sound)} \tag{3}$$

As a result, a simple model can be utilised to compute 1-, 2-, and n-gram language models of characters. If a *corpus* has 1,000,000 characters, the letter e appears 127,000 times, with a probability P(e) of 0.127.

In the context of a 2-gram system, it can be computed by recalling the preceding letter context- its WFSA condition. The probability P(e|s) can be used to compute the transformation between states s and e, which results in the letter e. The n-gram model produces more word-like elements than the (n-1)-gram approach. The weighted or lattice automaton is a simple automaton whereby each arc is connected with a transition, that can be expressed by a probability level suggesting how that path should be pursued. All arcs escaping a node must have a probability of one. Fig. 3 depicts a weighted ATN trained on actual pronunciation examples for the English word "around." This is an example of a HMM. The behaviour transitioning in the weighted ATN is depicted graphically in this picture. The principle of the transition is as follows:

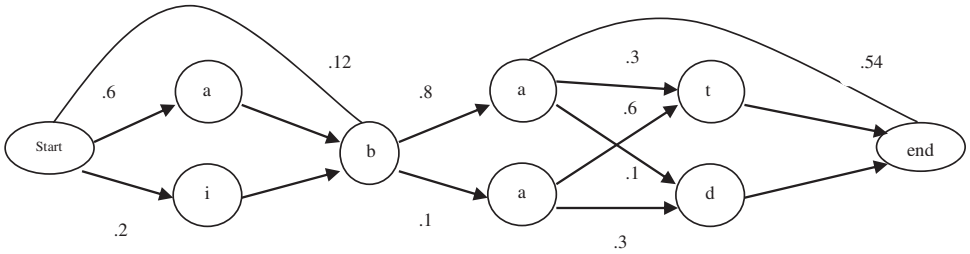

Figure 3: A pronunciation network (weighted atn) for word "about"
Legends: P(w | ax) = .68 P(w | ix) = .20

- Starts in some initial state (start: s1) with probability p(si).
- On each move, goes from state si to state sj according to transition probability P (si, sj).
- At each state si, it emits a symbol w_k according to the emit probability P'(s_i, w_k).

The proposed SpeakCorrect computerised interface is referred to as a hybrid approach since it employs components of the HMM or weighted state-graph model of a word's pronunciation, as well as observation-probability computing through multilayer perception. This network contains a single output unit for each phone; by adding all output unit numbers to one, the SpeakCorrect may be used to calculate the probability of a condition j given an observation vector Ot, P(qj | ot), or P(ot | qj). As a result, when given the series of spoken words which created a specific aural speech, a standard model - such as the one depicted in Fig. 3 is utilised. The model produces P(E|S) given a received speech signal S, and it is described as follows.

- A series of phonemes is detected with different probability for each phonetic in S and can thus be construed as the word.

- A word-to-phone correspondence is created for every phonetic.
- Every phone can be represented by a different set of audio signals.

Once constructed, the input audio sequence and the final communicative approach are weighted using the likelihood technique and detecting probabilities from the training dataset.

5.5 Training the SpeakCorrect

In most ASR systems, a concise summary of the integrated training process is provided. A few of the algorithm's features are discussed in [48,49]. To train the SpeakCorrect system, four probabilistic models are required:

- Language model probabilities: P(wi|wi-1 wi-2).
- Likelihood observation: bj(ot).
- Transition probabilities: aij.
- Pronunciation Lexicon: Lattice or Weighted ATN of HMM state graph structure.

The SpeakCorrect features the following corpora for training the past probabilities element:

- Speech wave document training *corpus*: that is compiled from news websites on the internet, specific individuals and so on. Such speech wave documents are grouped with word-transactions.
- Text *corpus* containing a large number of comparable texts, such as the word-transaction out from speech database.
- A smaller standard training *corpus* of audio that has been phonetically labelled, i.e., frames have been hand-annotated containing phonemes.

An off-the-shelf pronunciation vocabulary is used to build the HMM lexicon architecture. As a result, training begins with running the system on the observations and determining which transitions as well as observations were utilized. Any state can produce one observation symbol, and all observation possibilities are 1.0 [50,51]. The probabilities pij of a specific transition from condition I to condition j can be calculated by computing the number of transitions undertaken; c I j), then normalising such result using the Eq. (4).

$$a_{ij} = c\ (\ \mathrm{i}\ \rightarrow \mathrm{j}\ )/\sum\nolimits_{q\in Q}^{\infty}(\mathrm{C}\ (\mathrm{i}\ \rightarrow \mathrm{j})) \tag{4}$$

Two strategies are utilised for lattice or weighted ATN and HMM. The first is to iterate calculate the numbers and observation possibilities, and then use similar estimated probabilities to create progressively better probability values. The second step is to compute forward probability along all possible paths to obtain predicted probabilities. Considering the automaton A, calculate the forward probabilities in state I after witnessing the first t occurrences (Eqs. (5)–(8)).

$$a_t(i) = \ \mathrm{P}\ (o_1,\ o_2,\ o_3,\ \ldots,\ o_t,\ q_t\ =\ \mathrm{i}\ |\ A) \tag{5}$$

Formally, describe the following iteration:

1. Initialization:

$$\alpha_n(1) = a_{1j} * b_j\ (o_1) \ldots 1\ <\ \mathrm{j}\ <\ \mathrm{N} \tag{6}$$

2. Iteration:

$$\alpha_j(t) = \sum_{i=2}^{N-1} a_i(\mathrm{t}-1) * a_{ij} \,]\, b_j\,(o_t) \ldots \; 1 \; < \; \mathrm{j} \; < \; \mathrm{N}, \; 1 \; < \; \mathrm{t} \; < \; \mathrm{T} \tag{7}$$

3. Termination:

$$p(o|A) = a_N(T) = \sum_{i=2}^{N-1} a_i(\mathrm{T}) * a_{iN} \tag{8}$$

To calculate the alternative phonemes which were most likely given the observational sequence [axe b], with a forward algorithm is performed, and the combination P(o | w) P(w) is produced for each potential word. So, for every word, the probability of ordered set o provided the word w times the previous likelihood of the word is determined, and the word with the greatest value is chosen.

A forward method is an edit distance method, and an intermediary table is utilized to hold the observation sequence's probability numbers. The data can be presented in the table by rows that are orientated; the rows are marked by a state-graph that contains multiple paths from one state to another. With determining the number of each cell from the three cells surrounding it, the table is completed as a matrix. In addition, the forward technique calculates the total of probability for all alternative paths that could yield the observation series. Formally, each cell represents the probability (Eq. (9)):

$$forward\,[t,\, j] \;=\; P(o1,\; o2 \ldots ot,\; qt \;=\; j \mid A)\; P(w) \tag{9}$$

The forward algorithm is implemented to any word is described in the pseudo code below.

```
forwardAlgorithm ( observation, state-graph )
begin
ns = numOfStates(state-graph);
no = length(observation);
/* create probability matrix */
forward [ns + 2 , no + 2];
forward [0,0] = 1.0;
foreach time step t from 0 to no do
foreach states from 0 to ns do
foreach transition s' from s specified by state-graph
forward [s' , t +1] = forward [s, t] * a[s, s'] * b [s', ot];
return sum of the probabilities in the final column of forward;
end.
```

where:

a [s , s'] signifies transition possibility from present state s to subsequent state s'.

b [s', ot] is the observation probability of s' given ot.

b [s', ot] is equal 1 if the observation symbol matches the state, and is equal 0 otherwise.

The part of the forward-backward algorithm is the backward probability. This backward algorithm is almost the mirroring of the forward probability. It computes the probability of the observations from t+1 to the end. Suppose that we are in state j at time t in given automaton A; then (Eqs. (10)–(13)):

$$\beta_i\,(o_t\,) = \;\; \mathrm{P}\;(o_{t+1},\; o_{t+2},\; o_{t+3}, \ldots,\; o_T\,|q_t \;=\; \mathrm{j},\; A) \tag{10}$$

The backward computation is defined as the following:

Initialization:

$$\beta_i(t1) = a_{iN} \ldots\; 1 \;<\; \mathrm{i} \;<\; \mathrm{N} \tag{11}$$

Iteration:

$$\beta_i(t) = \;\; \sum\nolimits_{i=2}^{N-1} \; a_{ij} b_j\,(o_{t+1})\;\beta_j\,(t+1\,) \ldots\; 1 \;<\; \mathrm{j} \;<\; \mathrm{N},\; \mathrm{T} > \mathrm{t} \;\; < \;\; = 1 \tag{12}$$

Termination:

$$p(o|A) = a_N(T) = \beta_1\,(T\,) = \sum\nolimits_{j=2}^{N-1} a_{1i}\; b_j(o_1) * \beta_j \tag{13}$$

Therefore, the transition probability, aij, and observation probability, bi(ot), will be computed from an observation sequence.

5.6 SpeakCorrect Editor and Phonetic-based Approach

SpeakCorrect's conversion principles are a phonetic-based technique that converts text words into pronunciation terms. As a result, an intermediary form among source and the destination words is utilised, in addition to the phonetic-translation principle that captures the pronunciation of the specific words. There would be three phonetic-based principles used: identification, mapping, and generating rules. The identification rule analyses phonemes in the input word(s), and the mapping converts those phonemes into characters in the destination word(s) (orthographic representation). The targeted word is generated as a spoken word by the generating rule (letter-to-sound rule).

The conversion rule principles are founded on a number model and are written as Eq. (14):

$$\mathrm{P}(\mathrm{Ws},\; \mathrm{Wt}\,) \;=\; \mathrm{argmax}\;\; \mathrm{P}(\mathrm{Wt})\;\sum\;\; \mathrm{P}\;(\mathrm{Ws}\;|\;\mathrm{Is}\,)\;\mathrm{P}\;(\mathrm{Is}\;|\;\mathrm{Wt}\,) \tag{14}$$

where P (Ws | Is) is the probability of pronouncing the source word; P (Is | Wt) is the probability of generating the written Wt from the pronunciation in Is ; P (Wt) represents probability of sequence Wt occurring in the target language.

HMM or ATN can be thought of as a conversion rule that takes the source input (Ws) and maps it to the desired response (Wt) utilizing weight for every transition between stages, indicating which output patterns have a higher probability than the others. P(Wt) is a unigram phrase model that may be developed using any *corpus*. Depending on frequency components, P (Ws | Is) can be approximated.

5.7 Dataset of SpeakCorrect

The first sample we used is made up of two distinct categories. The first category we gathered from Aljazeera's internet news website. This dataset contains around 140 collected h, of which 100 h were used to develop the SpeakCorrect linguistic model and 40 h were utilized to test the SpeakCorrect method. The second domain is separated into two regions: Saudi Arabia and Egypt. Tabs. 1 and 2 illustrate the layout of our dataset after it was recorded.

Table 1: Structure of the dataset 1

Dataset 1	Al-Jazeera's news	
Type	Training	Testing
No of hours	100	40

Table 2: Structure of the dataset 2

Dataset 2	Saudi		Egypt	
Type	Male	Female	Male	Female
No of students	39	–	40	30

Both of the training as well as testing datasets are obtained from native speakers, as evidenced by dataset 1. We discovered that during annotating dataset 2, there are only few sounds.

5.8 Similarity between Two English Words

Given two word phonemes, $W_1(p^1\ p^2\ p^3 \ldots p^n)$ and $W_2\ (p^1\ p^2\ p^3 \ldots p^n)$, three factors are used to describe and evaluate similarity:

- The similarity of pronunciation in each phoneme pair (Pi(W1) , Pj(W2)) between W1 and W2.
- The similarity between the length of W1 and the length of W2, where 0 <= i <= n, 0 <= j <= m.
- The similarity between each character pair (recognized sound character) between W_1 and W_2.

The three factors having different effect in calculating the similarity between two words, we get Eq. (15):

$$S_{w1}(W_2) = \sum_{i=1}^{3} Sw_i\ S_{iw1}\ (W_2) \tag{15}$$

In our case, w_1 is selected to be 0.5, w_2 equals 1 and w_3 is selected to be 1.

5.9 SpeakCorrect Error Correction

As a result, a kernel characteristic model matrix is applied to compute the correlation between two words in order to correct the speech sequence (Syllable words). The confusion rating from Wi to Wj may be derived as the averaged confusion of every speech segment Si labeled as Wi to trained HMM model of Wj, Awj, given two words Wi and Wj.

Consequently, the confusion similarity score from Wi to Wj is estimated as the following equation Eqs. (16)–(18):

$$\text{Sim(Wi and Wj)} = (P(Oi|Ai)\ +\ P(Oj|Aj))/2 \tag{16}$$

The Guassian kernel function can be applied to calculate the confusion score between Wi and Wj.

$$\text{Conf.Score}\ (Wi\ ,\ Wj) = \exp\ ((\text{Sim}\ (Wi\ ,\ Wj))2\ /\ 2\ \sigma\ 2) \tag{17}$$

where σ represents the variance calculated over the distribution Sim (Wi , Wj).

The best correction result can be calculated according to the following equation:

$$C = \text{argmax}\ (P\ (W)\ P\ (E \mid W)\ P\ (W \mid S)) \tag{18}$$

where P(W) represents the word language model for the corrected word sequence W. Fig. 4 illustrates the proposed SpeakCorrect correction system.

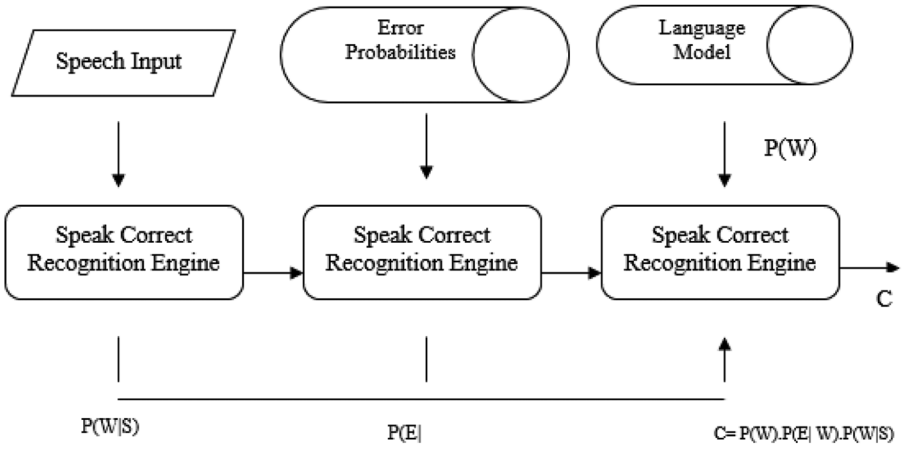

Figure 4: SpeakCorrect errors' correction

6 Implementation and Testing

The following developing model is based on components to detect error for pronunciation analysis and pronunciation adaption.

6.1 The SpeakCorrect Corpus Architecture

The *SpeakCorrect* corpus is based on annotated speech; it will be designed to provide acoustic for supporting the development and evaluation of automatic speech recognition systems.

6.1.1 The SpeakCorrect Structure

SpeakCorrect, such as the Brown *Corpus*, contains a broad range of dialects, speakers, and texts. It has two primary accent zones, each with two dialect localities, 150 men and women presenters ranging in age (18–21 years old) and academic background, and every reads 390 specially selected words. The terms were selected to be phonetically rich as well as to address all of the Arabic participants' pronunciation flaws (substitution, deletion, and insertion) (Saudi and Egypt regions). Furthermore, the design employs numerous speakers saying the same words in order to allow comparison among speakers, as well as having a vast variety of terms to ensure maximum coverage of flaws. As a result of the speakers reading by region (and two locales), 150 hundred captured statements are saved in the *corpus*, and each data file has inner structure, as illustrated in Fig. 5.

Speaker ID	Gender	$Word_1$	$Word_2$	...	$Word_{389}$	$Word_{390}$
001	Female			...		
002	Male	...	...	...	...	...
...	...	...	...	...	...	...
150	Female	...	...	...	...	...

Figure 5: Structure of speakcorrect structure

With the help of Tabs. 1 and 2, and Eqs. (1)–(18), Every item has a phonetic transcription, as well as the associated word tokens, which can be retrieved.

6.1.2 The SpeakCorrect Design Features

As illustrated in Fig. 6, SpeakCorrect incorporates *corpus* design elements. Firstly, such a *corpus* incorporates description at the phonetic as well as orthographic layers, with various labelling techniques at each level. A second property of SpeakCorrect is its balancing across various dimensions of variance, to encompass accent and accent areas and places, which aids later use of the *corpus* for applications such as sociolinguistics, which were not intended when the *corpus* was established.

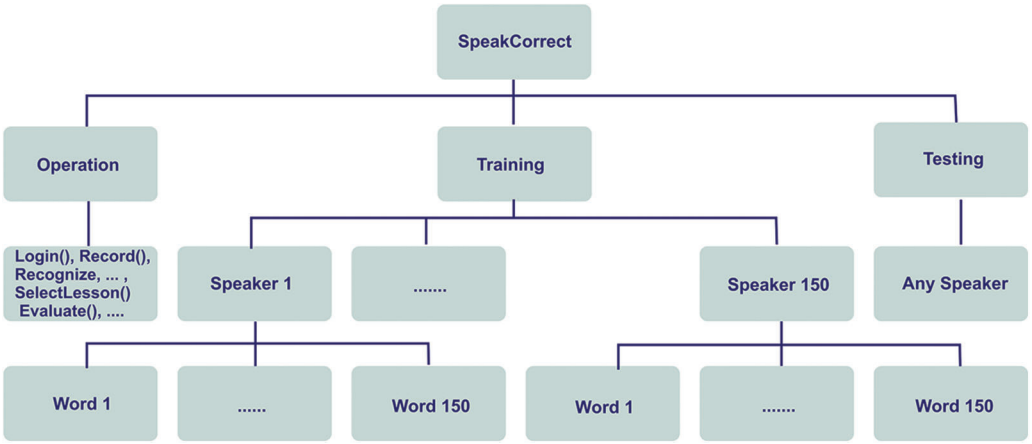

Figure 6: Structure of the implemented speakcorrect *corpus*

6.1.3 The SpeakCorrect Data Acquisition

The internet is a great repository of information for so many natural language processing applications. Consequently, in our scenario, a huge number of data samples are required to obtain. As a result, one of these ways is to get available data from the internet. The benefit of employing such well-defined online data is that it allows for documented, consistent, and repeatable investigation.

6.2 The SpeakCorrect User Interface

Fig. 7 depicts the computerized user interface layout of SpeakCorrect, which is separated into three tiers. The presenting tier is at the top; the logical or commercial tier is in the centre; it begins with registration, in which the login occurs, microphone configuration and adaption, language and speech lessons, and ultimately assessment. The third tier is internal, and it houses all of the attributes, records, files, and so on.

6.3 The Speak Correct Testing

Because of historical issues with English pronunciation in Arabian accents, the accompanying testing methodology is built on components to evaluate as well as assist students for pronunciation assessment, pronunciation adaptability, and pronunciation detect error. Fig. 8 depicts the login information for students.

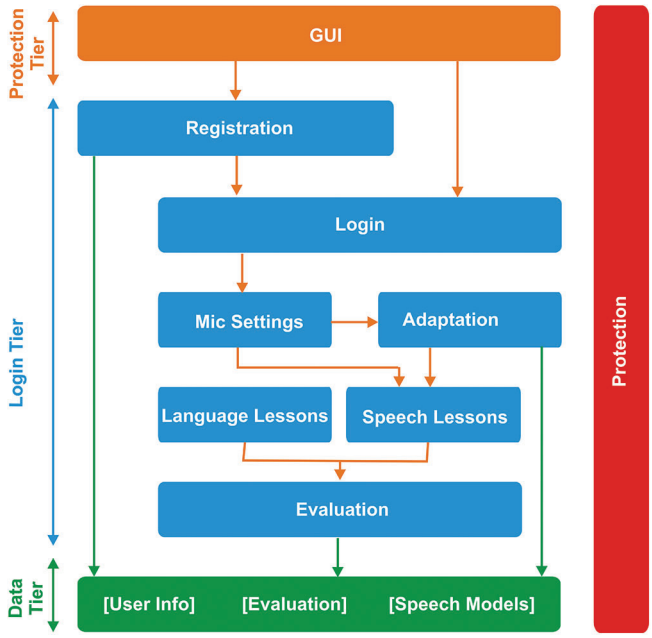

Figure 7: The different tiers of the speakcorrect system

Figure 8: The login interface of the speakcorrect system

As a result, the user interface is built using Silverlight technology. This type of user interface has several visual qualities that allow it to execute basic activities such as going to prior and subsequent demos, streaming a sample (predefined example), human voice, and collecting the user's voice. Fig. 9 depicts device configuration and microphone modification.

The operational code includes a cooperation module that connects C# code (.Net Client/Server), HMM code, and HTK element code. The second element, as illustrated in Fig. 10, aims at comparing the input voice to the predetermined trained voices and, as a result, provides a mistake-if any.

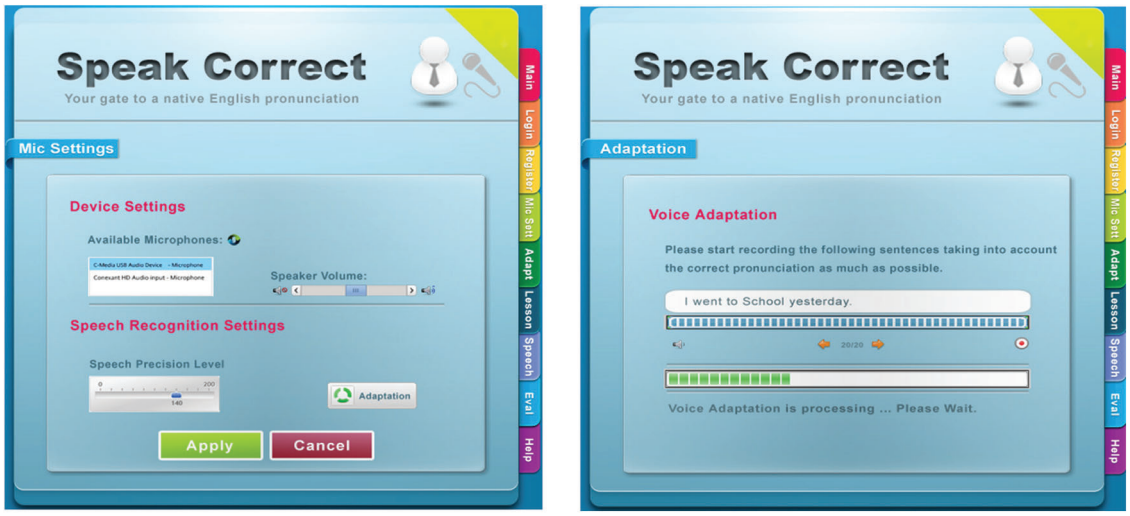

Figure 9: The device setting and microphone adjustment of the speakcorrect system

Figure 10: The levels and associated lessons testing of the speakcorrect system

As a result, the SpeakCorrect user experience is built using Silverlight technology. As seen in earlier drawings, such a user interface comprises tabs for basic operations such as going to previous and next demos, playing a sample (predefined example), user voice, and recording the user's voice. As a result, the MVVM design is employed to facilitate the tabs' connectivity "Click Event." The attributes of such visual elements are constrained in the underlying ViewModel class.

7 Conclusions

The purpose of this paper is to discuss the development of the SpeakCorrect computerised interface, which could be used for speech correction for non-native English speakers. The data set includes information for two major countries: Saudi Arabia as well as Egypt. Every region is divided by two locality regions. An interactive suggestion system is included in the planned system to encourage individuals to enhance their linguistic skills. SpeakCorrect would be utilised in the future to train and assess pronouncing individuals. As a result, an analytical component should be put in place to examine

the phonetic characteristics of the unidentified word and the missing phonemes. As a result, post-testing as well as a comparative study would be conducted. As a result, independent of the speaker's accent or sexual identity, the experiment is constructed and performed to assess the proposed resolution of the SpeakCorrect interface based on phonetically trained database. The performance of the SpeakCorrect interface indicates that the recognition system performed satisfactorily.

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

Funding Statement: 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).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] I. P. Wiggers and L. J. M. Rothkrantz, "Automatic speech recognition using hidden Markov models," *Course IN4012TU-Real-time AI & Automatische Spraakherkenning*, vol. 70, no. 8, pp. 1–20, 2003.

[2] W. Noormamode, B. G. Rahimbux and M. Peerboccus, "A speech engine for mauritian creole," *Information Systems Design and Intelligent Application*, vol. 36, no. 7, pp. 389–398, 2019.

[3] K. A. Kemble, "An introduction to speech recognition," *Voice Systems Middleware Education-IBM Corporation*, vol. 16, no. 8, pp. 154–163, 2001.

[4] O. Bracha, "The folklore of informationalism: The case of search engine speech," *Fordham Law Review*, vol. 82, no. 5, pp. 1629–1633, 2013.

[5] G. Tur and R. De Mori, "Spoken language understanding: Systems for extracting semantic information from speech," *Fordham Law Review*, vol. 82, no. 5, pp. 1629–1633, 2011.

[6] K. Laskowski and E. Shriberg, "Comparing the contributions of context and prosody in text-independent dialog act recognition," in *Proc. of the 2010 IEEE Int. Conf. on Acoustics, Speech and Signal Processing Newyork*, USA, pp. 5374–5377, 2010.

[7] Y. I. Song, Y. Y. Wang, Y. C. Ju, M. Seltzer, I. Tashev *et al.,* "Voice search of structured media data," in *Proc. of the 2009 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Delhi, India, pp. 3941–3944, 2009.

[8] P. Heracleous, N. Aboutabit and D. Beautemps, "Lip shape and hand position fusion for automatic vowel recognition in cued speech for French," *IEEE Signal Processing Letters*, vol. 16, no. 5, pp. 339–342, 2009.

[9] M. J. Osberger, "Speech intelligibility in the hearing impaired: Research and clinical implications," *Intelligibility in Speech Disorders*, vol. 74, no. 6, pp. 233–265, 1992.

[10] Y. Tsubota, M. Dantsuji and T. Kawahara, "An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors," *ReCALL*, vol. 16, no. 1, pp. 173–188, 2004.

[11] A. G. Almekhlafi, "The effect of computer assisted language learning on United Arab Emirates English as a foreign language school students' achievement and attitude," *Journal of Interactive Learning Research*, vol. 17, no. 2, pp. 121–142, 2006.

[12] M. Huijbregts, M. Mclaren and D. V. Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *Proc. of the 2011 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Biging, China, pp. 4436–4439, 2011.

[13] H. Wang, C. J. Waple and T. Kawahara, "Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition," *Speech Communication*, vol. 51, no. 10, pp. 995–1005, 2009.

[14] N. T. Vu, Y. Wang, M. Klose, Z. Mihaylova and T. Schultz, "Improving asr performance on non-native speech using multilingual and crosslingual information," in *Proc. of the Fifteenth Annual Conf. of the Int. Speech Communication Association*, LA, USA, pp. 200–203, 2014.

[15] N. A. A. Kadir and R. Sudirman, "Vowel effects towards dental arabic consonants based on spectrogram," in *Proc. of the 2011 Second Int. Conf. on Intelligent Systems, Modelling and Simulation*, NY, USA, pp. 183–188, 2011.

[16] K. Fujii, N. Saitoh, R. Oka and M. Muneyasu, "Acoustic echo cancellation algorithm tolerable for double talk," in *Proc. of the 2008 Hands-Free Speech Communication and Microphone Arrays*, NY, USA, pp. 200–203, 2008.

[17] J. Kacur and G. Rozinaj, "Adding voicing features into speech recognition based on HMM in Slovak," in *Proc. of the 2009 16th Int. Conf. on Systems, Signals and Image Processing*, Bhopal, India, pp. 1–4, 2009.

[18] H. K. Yopp and R. H. Yopp, "Supporting phonemic awareness development in the classroom," *Reading Teacher*, vol. 54, no. 2, pp. 130–143, 2000.

[19] R. Treiman, "Onsets and rimes as units of spoken syllables: Evidence from children," *Journal of Experimental Child Psychology*, vol. 39, no. 1, pp. 161–181, 1985.

[20] T. J. Hazen, I. L. Hetherington, H. Shu and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," *Speech Communication*, vol. 46, no. 2, pp. 189–203, 2005.

[21] G. Z. Liu, Z. H. Liu and G. J. Hwang, "Developing multi-dimensional evaluation criteria for English learning websites with university students and professors," *Computers & Education*, vol. 56, no. 1, pp. 65–79, 2011.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw *et al.,* "The HTK book," *Cambridge University Engineering Department*, vol. 3, no. 175, pp. 1–12, 2002.

[23] C. Kurian and K. Balakriahnan, "Continuous speech recognition system for Malayalam language using PLP cepstral coefficient," *Journal of Computing and Business Research*, vol. 3, no. 1, pp. 1–24, 2012.

[24] M. T. J. Ansari and N. A. Khan, "Worldwide Covid-19 vaccines sentiment analysis through twitter content," *Electronic Journal of General Medicine*, vol. 18, no. 1, pp. 1–15, 2021.

[25] S. Mishra, C. N. Bhende and B. K. Panigrahi, "Detection and classification of power quality disturbances using S-transform and probabilistic neural network," *IEEE Transactions on Power Delivery*, vol. 23, no. 1, pp. 280–287, 2007.

[26] M. T. J. Ansari, D. Pandey and M. Alenezi, "STORE: Security threat oriented requirements engineering methodology," *Journal of King Saud University-Computer and Information Sciences*, vol. 54, no. 5, pp. 1–18, 2018.

[27] N. Mogran, H. Bourlard and H. Hermansky, "Automatic speech recognition: An auditory perspective," *Speech Processing in the Auditory System Journal*, vol. 68, no. 12, pp. 309–338, 2004.

[28] V. Radha and C. Vimala, "A review on speech recognition challenges and approaches," *Encyclopedias of Journals*, vol. 2, no. 1, pp. 1–7, 2012.

[29] C. H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.

[30] K. Sahu, F. A. Alzahrani, R. K. Srivastava and R. Kumar, "Hesitant fuzzy sets based symmetrical model of decision-making for estimating the durability of web application," *Symmetry*, vol. 12, no. 11, pp. 1–20, 2020.

[31] M. T. J. Ansari, A. Baz, H. Alhakami, W. Alhakami, R. Kumar *et al.,* "P-STORE: Extension of store methodology to elicit privacy requirements," *Arabian Journal for Science and Engineering*, vol. 64, no. 3, pp. 1–24, 2021.

[32] R. Kumar, S. A. Khan and R. A. Khan, "Durability challenges in software engineering," *CrossTalk*, vol. 42, no. 4, pp. 29–31, 2016.

[33] K. Sahu, F. A. Alzahrani, R. K. Srivastava and R. Kumar, "Evaluating the impact of prediction techniques: Software reliability perspective," *Computers, Materials & Continua*, vol. 67, no. 2, pp. 1471–1488, 2021.

[34] A. Attaallah, M. Ahmad, M. Tarique, A. K. Pandey, R. Kumar *et al.,* "Device security assessment of internet of healthcare things," *Intelligent Automation & Soft Computing*, vol. 27, no. 2, pp. 593–603, 2021.

[35] K. Sahu and R. K. Srivastava, "Needs and importance of reliability prediction: An industrial perspective," *Information Sciences Letters*, vol. 9, no. 1, pp. 33–37, 2020.

[36] K. Sahu and R. K. Srivastava, "Revisiting software reliability," *Advances in Intelligent Systems and Computing*, vol. 802, pp. 221–235, 2019.

[37] R. Kumar, S. A. Khan and R. A. Khan, "Secure serviceability of software: Durability perspective," *Communications in Computer and Information Science*, vol. 628, pp. 104–110, 2016.

[38] R. Kumar, S. A. Khan and R. A. Khan, "Fuzzy analytic hierarchy process for software durability: Security risks perspective," *Advances in Intelligent Systems and Computing*, vol. 508, pp. 469–478, 2017.

[39] R. Kumar, S. A. Khan and R. A. Khan, "Revisiting software security: Durability perspective," *International Journal of Hybrid Information Technology*, vol. 8, no. 2, pp. 311–322, 2015.

[40] W. Alosaimi, A. Alharbi, H. Alyami, M. Ahmad, A. K. Pandey *et al.,* "Impact of tools and techniques for securing consultancy services," *Computer Systems Science and Engineering*, vol. 37, no. 3, pp. 347–360, 2021.

[41] K. Sahu and R. K. Srivastava, "'Predicting software bugs of newly and large datasets through a unified neuro-fuzzy approach: Reliability perspective," *Advances in Mathematics: Scientific Journal*, vol. 10, no. 1, pp. 543–555, 2021.

[42] R. Kumar, S. A. Khan and R. A. Khan, "Revisiting software security risks," *Journal of Advances in Mathematics and Computer Science*, vol. 11, no. 6, pp. 1–10, 2015.

[43] R. Kumar, S. A. Khan and R. A. Khan, "Analytical network process for software security: A design perspective," *CSI Transactions on ICT*, vol. 4, no. 2, pp. 255–258, 2016.

[44] K. Sahu and R. K. Srivastava, "Soft computing approach for prediction of software reliability," *ICIC Express Letters*, vol. 12, no. 12, pp. 1213–1222, 2018.

[45] R. Kumar, S. A. Khan and R. A. Khan, "Software security testing: A pertinent framework," *Journal of Global Research in Computer Science*, vol. 5, no. 3, pp. 23–27, 2014.

[46] F. A. Alzahrani, M. Ahmad, M. Nadeem, R. Kumar and R. A. Khan, "Integrity assessment of medical devices for improving hospital services," *Computers, Materials & Continua*, vol. 67, no. 3, pp. 3619–3633, 2021.

[47] R. Kumar, S. A. Khan and R. A. Khan, "Durable security in software development: Needs and importance," *CSI Communications*, vol. 10, no. 10, pp. 34–36, 2015.

[48] R. Kumar, S. A. Khan, A. Agrawal and R. A. Khan, "Measuring the security attributes through fuzzy analytic hierarchy process: Durability perspective," *ICIC Express Letters*, vol. 12, no. 6, pp. 615–620, 2018.

[49] H. Alyami, M. Nadeem, W. Alosaimi, A. Alharbi, R. Kumar *et al.,* "Analyzing the data of software security life-span: Quantum computing era," *Intelligent Automation & Soft Computing*, vol. 31, no. 2, pp. 707–716, 2022.

[50] A. H. Seh, J. F. Alamri, A. F. Subahi, A. Agrawal, R. Kumar *et al.,* "Machine learning based framework for maintaining privacy of healthcare data," *Intelligent Automation & Soft Computing*, vol. 29, no. 3, pp. 697–712, 2021.

[51] A. H. Seh, J. F. Alamri, A. F. Subahi, M. T. J. Ansari, R. Kumar *et al.,* "Hybrid computational modeling for web application security assessment," *Computers, Materials & Continua*, vol. 70, no. 1, pp. 469–489, 2022.