Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 8279856, 11 pages
https://doi.org/10.1155/2022/8279856

Research Article

An Effective Artificial Intelligence-Enabled Error Detection and Accuracy Estimation Technique for English Speech Recognition System

Lu Han, Xueqin Du, Li Yan, and Jing Yu

School of Humanities, Jiangxi University of Chinese Medicine, Nanchang, 330004 Jiangxi, China

Correspondence should be addressed to Jing Yu; 20081017@jxutcm.edu.cn

Received 31 December 2021; Revised 24 January 2022; Accepted 18 February 2022; Published 4 April 2022

Academic Editor: Deepak Kumar Jain

Error detection and accuracy estimation in automated speech recognition (ASR) systems act a vital part in the design of human-computer spoken dialogue systems, as recognition error can hamper accurate systems in understanding the end user intentions. The major aim is to identify the errors in an utterance, and therefore, the dialogue manager can provide proper clarifications to the user. Therefore, the design of accurate error detection and accuracy determination techniques becomes essential in the ASR system. With this motivation, this paper presents a novel artificial intelligence-enabled accuracy estimation and error detection technique for the English speech recognition system (AIEDAE-ESRS). The goal of the AIEDAE-ESRS technique is to perform three actions such as confidence estimation, out-of-vocabulary (OOV) word identification, and error type categorization. In addition, the AIEDAE-ESRS technique performs different levels of preprocessing including sampling of input speech signal, bandpass filtering, and noise removal. Besides, a new deep neural network with hidden Markov model- (DNN-HMM-) based speech recognition technology is designed, which also aims to estimate the accuracy and error. Finally, the hyperparameters of the DNN-HMM model can be optimally chosen by the use of flower pollination algorithm (FPA) and thereby accomplished improved recognition performance. In order to demonstrate the better performance of the AIEDAE-ESRS technique, a series of simulations were conducted and the results are examined under varying aspects. English voice recognition system's accuracy estimation and error detection were made possible using artificial intelligence (AIEDAE-ESRS). There are three steps in the AIEDAE-ESRS method: confidence estimation; identifying out-of-vocabulary words (OOV); and categorizing mistake types. The simulation results reported the enhanced performance of the AIEDAE-ESRS methodology over current advanced approaches. Our AIEDAE-ESRS methodology outperforms existing methodologies by a factor of ten. The simulation results demonstrated that the AIEDAE-ESRS methodology outperformed previous approaches in terms of efficiency. The improved experimental results indicated that the AIEDAE-ESRS technique produced superior results across a variety of measures.

1. Introduction

The speech signal is one of the essential and common ways of communicating between people. In these communications, the speaker's emotion performs an important role in the transfer of concept in such a way that a change in the emotions may result in distinct translations of speech [1]. Therefore, to make effective communication between man and machine, speech emotion recognition (SER) is becoming a hot research topic. In the selection of important features, together with accurate SER system, an effective way to decrease the data dimension is needed [2]. With the continued growth of science and technology, the global village is shrinking, and the usage of English has become increasingly widespread. The development of artificial intelligence computers that could understand English speech will significantly encourage the new experience and complete intelligence of human life and work eventually [3]. The speech emotion recognition (SER) system is built on CNNs and RNNs that have been trained on a database of emotional speech. Our primary objective is to offer a SER approach that is based on concatenated CNNs and RNNs and does

not rely on any typical hand-crafted features. The literature on speech emotion recognition (SER) has employed a variety of approaches to extract emotions from signals, including numerous well-known speech analysis and classification techniques. Recently, deep learning approaches have been presented as a possible replacement for classic SER techniques. Language interaction and intelligent English speech recognition systems (SRS) affect their study and work life, as well as have promotion significance and extensive application in areas like language promotion, military, and education. Now, there are multiple implementation methods and system designs for SRS. There are different kinds of classification, primarily separated into specific-and nonspecific-persons SRS, continuous and isolated word SRS, embedded/server SRS, small vocabulary, and large vocabulary SRS. In everyday life, people's natural speech is depending on the speaker's need to break at the end of a sentence or add punctuation, and other parts could be continuously pronounced [4].

In the earlier SRS, the isolated word phonetic systems were based primarily on single words or characters [5]. Depending on the way the acoustic method is developed, we could separate SRS as specific- and nonspecific-person recognition. Specific-person recognition implies that the user needs to input a massive number of pronunciations and train recognition in advance. The nonspecific-person is that afterward the scheme is developed, the user does not need to input the trained information before and could recognize directly [6]. The deep learning (DL) method has different areas of application, and several achievements have been found. Another area where DL is effectively used is automated SRS. In automated SRS, better language and acoustic methods are integrated [7]. The SRS problems involve time-series data. In several fields, such as read continuous speech where usually the speech is recorded under clean conditions, the outcomes are satisfied with an error rate under 5%. Since in another field that has high speech differences, like distant conversational speech (meeting) or video speech, the outcomes are still not satisfactory exhibiting 50% of an error rate [8].

To handle these problems and improve the performances of inaccurate ASR systems, the automated correction and detection of the transcript error could be the only choice in some cases [9], especially while tuning the ASR systems by itself is impossible (for example, the system is purchased as a black box) or the manual correction is inconvenient or even impractical as in the case where the transcriptions are not the ultimate objective of the systems (for example, question answering, machine translation, and information retrieval systems). In that respect, ASR classification and error detection are also called confidence estimation [10]. The more commonly studied method is feature-based, where classification is constructed by the feature generated from distinct sources (that is, decoder and nondecoder characteristics) to differentiate the accurately from the inaccurately identified word.

This paper presents a novel artificial intelligence-enabled accuracy estimation and error detection technique for the English speech recognition system (AIEDAE-ESRS). The AIEDAE-ESRS technique intends to accomplish three actions such as OOV word identification, confidence estimation, and error type categorization. Furthermore, the AIEDA-ESRS technique's architecture incorporates a deep neural network with hidden Markov model- (DNN-HMM-) based speech recognition model. Furthermore, the flower pollination algorithm (FPA) is used to fine-tune the DNN-HMM model's hyperparameters. Flower pollination algorithm (FPA) is a nature-inspired metaheuristic algorithm that replicates the pollination activities of blossoming plants. The implementation of several FPA variants based on tweaks, parameter adjustment, and hybridization with other algorithms is addressed in this article. The design of FPA for hyperparameter optimization of the DNN-HMM model shows the novelty of the work. The experimental result analysis of the AIEDAE-ESRS technique takes place using benchmark dataset and investigated the results under several aspects.

2. Literature Review

In Alhamada et al. [11], the usage of DL in SRS was examined and an appropriate DL framework has been was recognized. A technique using CNN is employed to improve the efficiency of SRS. Han et al. [12] examined the efficacy of different DL-based acoustic models for conversation telephone speech, especially CNN-bLSTM, bLSTM, and TDNN systems. They estimated this model on research test sets, like recordings, Switchboard, and CallHome from a real-time call center applications. In Blaise. O. Yenke et al., due to the large variety of applications and interfaces or computing equipment that can enable speech processing, automatic speech recognition (ASR) is a very active research subject. It is true that well-resourced languages outnumber underresourced languages in most applications. It is evident that ASR may be used to enhance illiterate people's languages. Starting with a small vocabulary is one way to construct an ASR system for underresourced languages. Assertive speech recognition (ASR) with a limited vocabulary recognizes words or sentences in small groups.

Grozdić et al. [13] extended a method for whispered SRS that is the most difficult challenge in ASR. Specifically, because of the profound variances among acoustic features of whispered and neutral speech, the efficiency of conventional ASR system trained on neutral speech greatly reduces once whisper is used. Misbullah et al. [14] investigated the efficiency of SRS for dysarthric speakers using time delay DNN. Furthermore, examine the system performances by integrating dysarthria and normal speech corpus. Lastly, well-tuned hyperparameter of DNN structure gives potential outcomes on English dysarthria and Mandarin speech.

Ogawa and Hori [15] explored three kinds of ASR error detection processes, that is, OOV word recognition, error type classification (ETC), and confidence estimation, and also evaluated the detection rate from the ETC result. The simulation result shows that the DBRNN considerably outperforms conditional random field (CRF). Ogawa et al. [16] presented detection accuracy estimation method based on ETC. The ETC is an extension of confidence estimate.

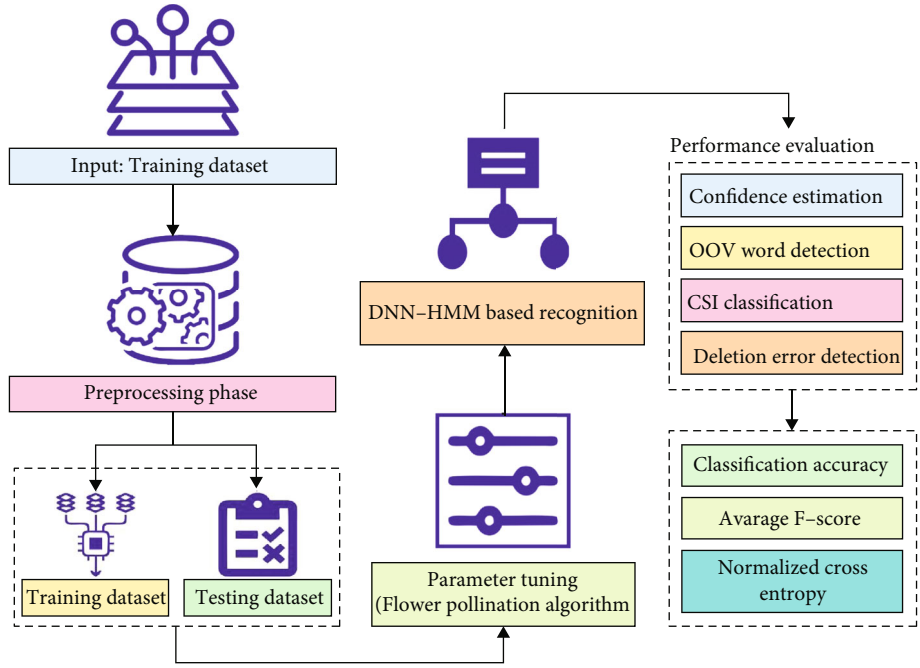

Figure 1: Overall process of AIEDAE-ESRS technique.

In ETC, all the words in the detection outcomes (detected word sequence) for the targeted speech information are categorized into three classes: insertion error (I), substitution error (S), and correct recognition (C).

3. Materials and Methods

In this study, an effective AIEDAE-ESRS technique has been developed for the error detection and accuracy evaluation in SRS. The AIEDAE-ESRS technique involves three major processes, namely, preprocessing, DNN-HMM-based speech signal recognition, and FPA-based hyperparameter tuning. The utilization of the FPA helps to properly alter the hyperparameter of the DNN-HMM model which assists in significantly boosting the detection performances. Figure 1 demonstrates the overall working procedure of the suggested AIEDAE-ESRS technique.

3.1. Level I: Speech Signal Preprocessing. The speech input is the original voice signal gathered by the voice tool; the preprocessing method chiefly consists of three factors: antialiasing bandpass filtering, eliminating the noise effect, and sampling the input original voice signal; the feature extraction method extracts the reflection in the voice. The acoustic parameter of the speaker's key features primarily includes short-term average zero-crossing rate, cepstrum, short-term energy, and linear prediction coefficient. In the recognition phase, the speech feature parameter is attained, and the test template is made. In the test, the template is matched with the reference template as per some discriminative rules (i.e., semantic and grammar rules), later in the training phase, the feature parameter is processed for establishing a reference model, and the better reference template is attained as the detection outcome. Better matching results are closely associated with the matching template, quality of speech feature parameter, and speech technique.

3.2. Level II: Design of DNN-HMM-Based Speech Signal Recognitions. In traditional DNN-HMM-based recognition, the probability is modelled by GMM under the maximal probability condition. Such potential models are constrained because GMM is statistically ineffective to model information that lies on or near a nonlinear in the data space. To conquer this limitation, we proposed a DNN-HMM method for recognizing speech, in which the outputs of the DNN are given to the HMM as substitute for the GMM. GMM simulates the observed probability distribution of a feature vector in the presence of a phone. It establishes a sound foundation for determining the "distance" between a phone and the audio frame being observed. The GMM is a probabilistic model capable of simulating normally distributed subpopulations. GMM's components are all Gaussian distributions. A statistical Markov model (HMM) is a sort of hidden Markov model. When the data is continuous, a Gaussian distribution is used to represent each hidden state.

3.2.1. Overview of Hidden Markov Models. The HMM is a statistical Markov method where the algorithm that has been modelled is presumed as a Markov model using unobserved (hidden) state. An HMM, denoted by $= (A, B, \pi)$, contains the subsequent element:

(1) The amount of states in the system represented as Q, the number of states represented by $S = \{s_1, s_2, \cdots, s_Q\}$, and q_t the states at time t

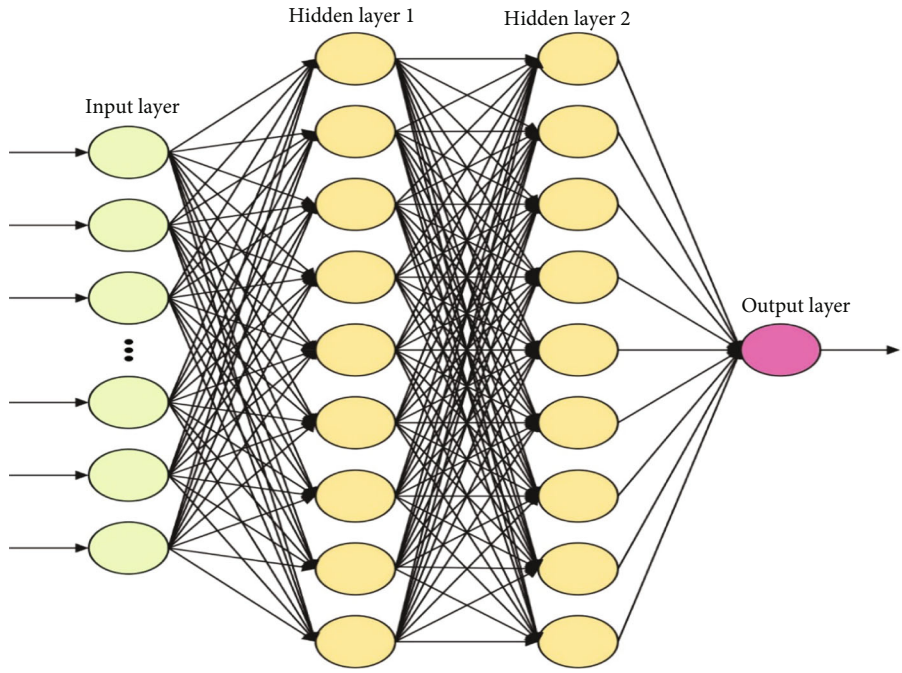

Figure 2: DNN structure.

(2) $A = \{a_{ij}\}$, the transition state likelihood distribution

$$a_{ij} = P\left(q_{t+1} = s_j \mid q_t = s_i\right), 1 \le i, j, \le Q \tag{1}$$

(3) $B = \{b_i(O_t)\}$, the observation probability, in which $b_i(O_t)$ signifies the likelihood of observing O_t at state s_i.B is denoted as a finite mixture:

$$b_i(\mathrm{O}_t) = \sum_{m=1}^{M} c_{im} \aleph(O_t, \mu_{im}, U_{im}), 1 \le i \le Q \tag{2}$$

Let c_{im} be the mixture coefficients for mth mixture in state s_i, as well as $\aleph$ elliptically symmetric density or log-concave, with covariance matrix U_{im} and mean vector μ_{im} for mth mixture element in i state

(4) $\pi = \{\pi_i\}$, the first state distribution, in which $\pi_i = P(q_1 = s_i), 1 \le i \le Q$

In order to apply HMM, two issues must be resolved:

(i) Learning issue: assuming a collection of ground truth X (represent as trained set), the learning process detects the group of variables $\lambda^* = \{A^*, B^*, \pi^*\}$; therefore, $\lambda^* = \arg\ \max_\lambda P(X \mid \lambda)$ that detects the model parameter that well fits the trained set. The forward-backward method is utilized for calculating $(X \mid \lambda)$ [17]. It finds the model parameter that best fits the training data. In order to compute $(X \mid \lambda)$, the forward-backward approach is used

(ii) Decoding issue: assuming a λ parameter and a series of new observation $O = (O_1, O_2, \cdots, O_T)$ (represent as testing set), the decoder process is determined as

$$P(O \mid \lambda) = \max_{q_1, \cdots, q_T} \pi_{q_1} \prod_{t=2}^{T} p(q_t \mid q_{t-1}) b_{q_t}(O_t) \tag{3}$$

In the event of speech recognition, train C HMM $\{\lambda_c, (c = 1, \cdots, C)\}$ for C discrete class. For novel speech input O,

$$c^* = \mathrm{argmax}_{1 \le c \le C} P(O \mid \lambda_c), \tag{4}$$

with $P(O \mid \lambda_c)$ estimated from the Viterbi model.

3.2.2. Structure of DNN-HMM Model. The main variation among GMM-HMM and DNN-HMM is the utilizing of GMM (rather than DNN) to evaluate the observation probability. We employ the DNN for modelling $p(q_t \mid 0_t)$; the following probabilities of the parameters provided the vector 0_t, i.e., feasible, while $p(q_t)$ is easier for estimating from a first state-level position of the trained set. Figure 2 depicts the framework of DNN technique.

3.3. DNN-HMM Training Process. The thoroughly trained procedure for recognition is given below:

```
Input: Objective function f(x)
Population initialization: Flower/Pollen gamete with arbitrary solutions;
Determine the optimal solution g_* in the primary population;
Represent a switching possibility p ∈ [0, 1];
while (t<Maximum_iterations) do
   for i =1:n (every n flower in the population) do
      if rand < p then
Draw a (d dimension) step vector L from a Levy delivery;
Apply global pollination;
      Else
         Draw ϵ from a similar distribution in [0, 1];
         Apply local pollination;
      End
      Determine new solutions;
      When better solutions are obtained, updating it in population;
   End
   Compute the present optimum solution g_*;
End
```

Algorithm 1: Pseudocode of FPA

(a) For all the classes $c(c = 1, \cdots, C)$, a GMM-HMM λ_c with Q state is instructed by the training sentence of parameter c

(b) For all the sentences $O = (O_1, O_2, \cdots, O_T)$ in trained set c, the Viterbi model of GMM-HMM as per Equation (3) is executed on λ_c to attain an optimum state sequence $(q_1^c, \cdots, q_T^c)$, and all the states q_t^c are allotted a label $L_i(i \in (1, \cdots, C \times Q))$ [18]

(c) Each training sentence and labelled state sequence is utilized as input to train a DNN, where output is the previous probability of the $C \times Q$ output unit. The DNN is executed by BP model using (i) unsupervised pretraining or (ii) the discrimination pretraining

3.4. DNN-HMM Recognition Process. In the detection procedure, for input sentence $O=(0_1, 0_2, \cdots, 0_T)$, one must evaluate the likelihood $p(O \mid \lambda_c)$ for all the classes c and attain the last recognition results as per Equation (4). In GMM-HMM, this likelihood is attained by the Viterbi model using Equation (3).

In DNN-HMM, adapt the subsequent process to estimate the likelihood $(O \mid \lambda_c)$.

(a) The input structure O is initially inputted to DNN, obtaining the previous probability $\{p(L_i \mid o_t)\}_{i=1,\cdots,C\times Q}$ as output. Next, previous probabilities $p(q_t = S_k^c \mid O_t)$ are attained from $(L_i \mid o_t)$, through representing the label L_i to the k and c, as follows

(b) As per the Bayesian principles, estimate the possibility $p(o_t \mid q_t)$ as

$$p(O_t \mid q_t) = \frac{p(q_t \mid O_t)p(O_t)}{p(q_t)} \quad (5)$$

Table 1: Result analysis of AIEDAE-ESRS technique with existing approaches.

Classifier	Accuracy	NEC	Average F-score
Confidence estimation			
CRF	0.8433	0.3630	0.7847
DNN	0.8445	0.3660	0.7854
DURNN	0.8496	0.3820	0.7923
DULSTM	0.8503	0.3820	0.7925
DBRNN	0.8552	0.4040	0.8009
AIEDAE-ESRS	0.8819	0.4410	0.8302
OOV word detection			
CRF	0.9429	0.3050	0.7046
DNN	0.9402	0.2700	0.6789
DURNN	0.9432	0.3140	0.7034
DULSTM	0.9442	0.3120	0.7014
DBRNN	0.9460	0.3480	0.7212
AIEDAE-ESRS	0.9708	0.3720	0.7497
CSI classification			
CRF	0.8213	0.3400	0.6302
DNN	0.8197	0.3160	0.5964
DURNN	0.8250	0.3400	0.6206
DULSTM	0.8255	0.3380	0.6199
DBRNN	0.8333	0.3810	0.6560
AIEDAE-ESRS	0.8579	0.4120	0.6796
Deletion error detection			
CRF	0.9625	0.1850	0.6411
DNN	0.9645	0.1770	0.6332
DURNN	0.9645	0.1950	0.6397
DULSTM	0.9646	0.1920	0.6392
DBRNN	0.9645	0.2430	0.6626
AIEDAE-ESRS	0.9921	0.2640	0.6909

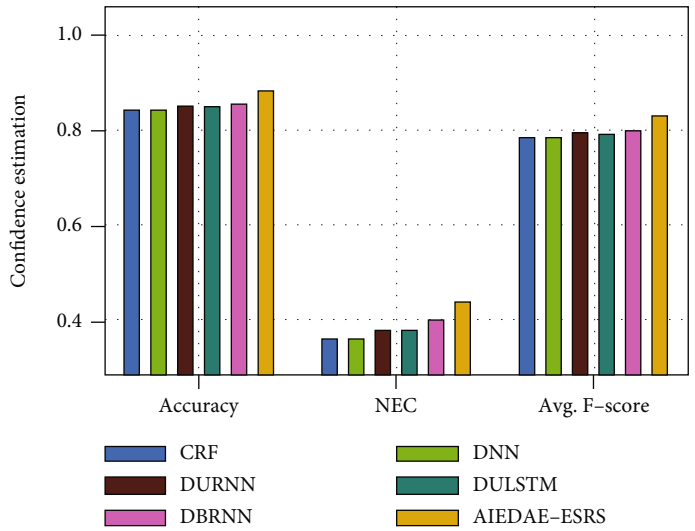

Figure 3: Confidence estimation analysis of AIEDAE-ESRS technique.

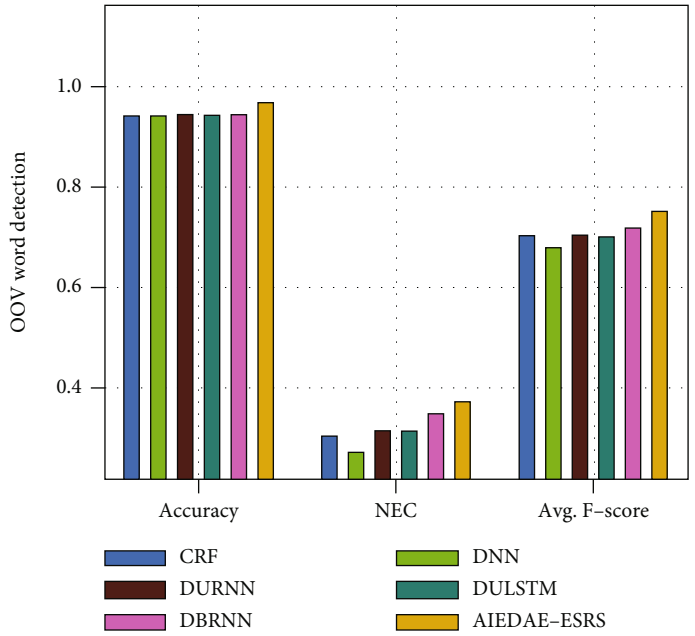

Figure 4: OOV word detection analysis of AIEDAE-ESRS technique.

In this process, the previous probabilities of each state (q_t) is estimated from (occurrence of) the trained set, and $p(O_t)$ is allocated a constant because the feature vector is considered independent of one another [19].

(c) For all the models λ_c, the Viterbi model is executed to estimate the prospect $p(O \mid \lambda_c)$. But, the likelihood $b_{q_t}(O_t)$ is substituted with $p(o_t \mid q_t)$ estimated by Equation (5)

3.5. Level III: Design of FPA-Based Hyperparameter Tuning. The FPA is used to effectively adjust the DNN-HMM model's hyperparameter settings. The abiotic pollination is discussed as well as induced in a flower pollination approach in the optimization. Pollination challenges encompass a difficult process in plant generation theory. The pollen gamete and bloom are more likely to provide a consistent solution to the optimization challenge. The advantages of FPA are listed below. FPA, unlike GA, HS, and PSO, provides a simple floral analogy with lightweight computationally based control parameters (that is, switch conditions, p). It also provides a balanced diversity and intensity of solution through the adaptation of levy flight (random walks punctuated by larger leaps) and switch conditions, which are utilized to transition between intensive local search and global search.

Flower constancy was identified as a precise solution that might be differentiated. In the case of global pollination, the pollinator transports pollen from a great distance

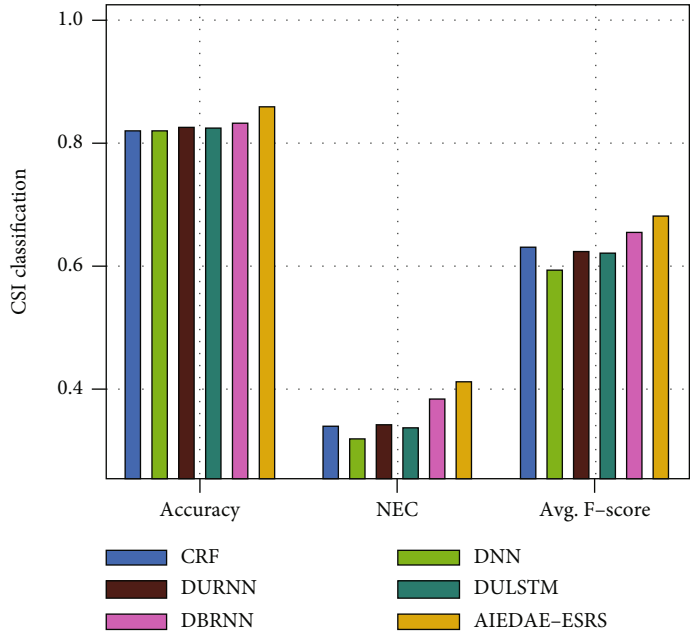

Figure 5: CSI classification analysis of AIEDAE-ESRS technique.

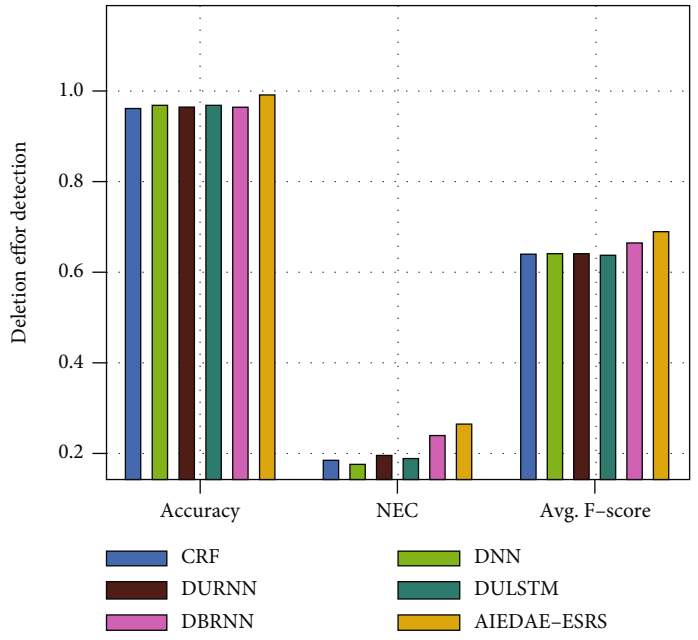

Figure 6: Deletion error detection analysis of AIEDAE-ESRS technique.

to a more suitable location. In another example, local pollination was carried out inside a smaller region of a unique bloom in shade water [20]. Global pollination is carried out through a possibility known as switch probability. Pollination occurs all across the world when a pollinator transports pollen from vast distances to higher fitting. Global pollination is accomplished by the use of a probability known as switch probability. In a tiny area of a unique bloom, local pollination is carried out in water shade. Pollinators like bees are vital to the sexual reproduction of about ninety percent of wildflowers. Ecosystems depend on these plants to function. They provide food, shelter, and other resources for many animal species, including humans. Once the stage was removed, local pollination is substituted. In FPA technique, the following 4 rules are used (also shown in Algorithm 1):

(1) Cross and live pollination is called global pollination as well as the carrier of pollen pollinator apples the LF

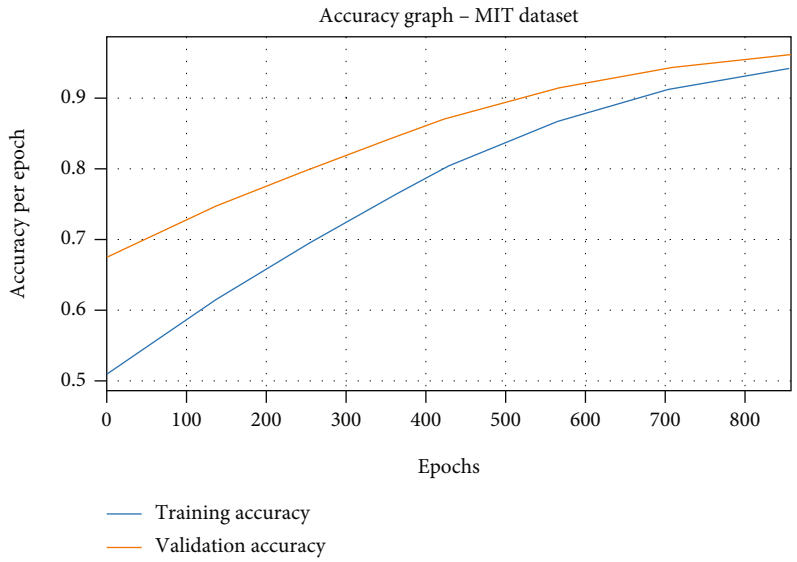

Figure 7: Accuracy graph analysis of AIEDAE-ESRS technique.

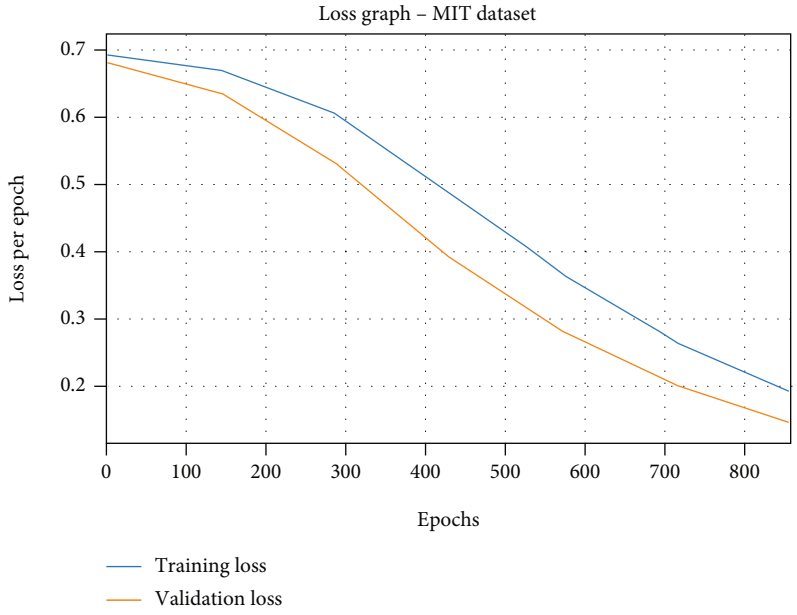

Figure 8: Loss graph analysis of AIEDAE-ESRS technique.

(2) Abiotic and self-pollination are represented as local pollination

(3) Pollinators are insects, i.e., able to develop flower constancy. It is determined as the possibility of two employed flowers

(4) The transmission of local and international pollination is handled by switch possibility

Therefore, the first and second rules are given by

$$x_q^{t+1} = x_p^t + \gamma \times L(\lambda) \times \left(g_* - x_p^t\right), \tag{6}$$

in which x_p^t is the pollen vector at iteration t; g_* indicates a present solution from present generated outcomes; $\gamma = a$ indicates the level factor to control phase size; and L denotes pollination power, which is related with a step-size of levy allocation. The LF is calculated as a collection of random computations that have the duration of all the leaps and use the levy likelihood distribution function with infinite variation. Following that, L represents a levy distribution:

$$L \sim \frac{\lambda \times \Gamma(\lambda) \times \sin\ (\pi\lambda/2)}{\pi} \times \frac{1}{S^1 + \lambda} S \gg S_0 0, \tag{7}$$

in which $\Gamma(\lambda)$ is the basic gamma function.

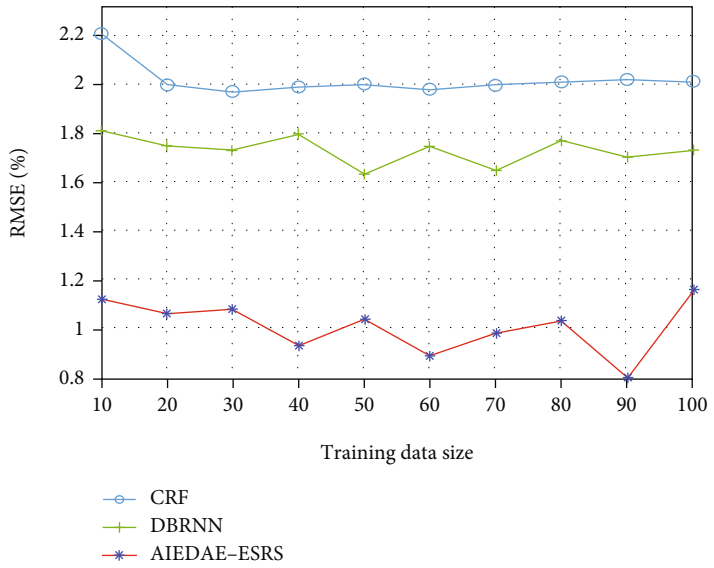

Figure 9: RMSE analysis of AIEDAE-ESRS technique.

In the event of local pollination, the second and third rules are formulated as

$$x_p^{t+1} = x_p^t + \varepsilon\left(x_q^t - x_k^t\right), \tag{8}$$

in which x_q^t and x_k^t are 2 pollens from several blooms on the same plant, if x_q^t and x_k^t come from the same species and are chosen from a homogenous population; this is represented as local random walks and is included by a standard distribution in zero and one [21].

FF acts as an important part of the optimization problem. It estimates a positive integer to represent how better the candidate solution is. In the work, classification error rate is considered as a minimalizing FF. The poorer solutions have high fitness scores and the richer solutions have less fitness scores.

$$\begin{aligned}\text{fitness}\,(\chi_i) &= \text{Classifier Error Rate}(\chi_i)\\ &= \frac{\text{number of misclassified instances}}{\text{Toral number of instances}} * 100.\end{aligned} \tag{9}$$

4. Experimental Validation

In this study, the experimental result analysis of the AIEDAE-ESRS technique takes place using the MIT lecture English speech corpus, called MIT dataset [22]. The MIT corpus includes speech information from invited talks and systematic university classes. The length of a lecture exist between 45 and 90 minutes. First, the error detection result analysis of the AIEDAE-ESRS technique takes place under deletion error detection, confidence estimation, OOV word detection, and CSI classification in Table 1.

Figure 3 exhibits the comparative result analysis of the AIEDAE-ESRS technique with existing methods under confidence estimation. The figure reported that the AIEDAE-ESRS technique can accomplish effectual outcomes with the increased values of accuracy, average F-score (AFS), and normalized cross entropy (NCE). It is noticed that the CRF and DNN models have shown least performance with the minimal values of accuracy, NEC, and AFS. In line with this, the DURNN, DULSTM, and DBRNN techniques have resulted in moderately closer accuracy, NEC, and AFS values. However, the AIEDAE-ESRS technique has outperformed the other techniques with the higher accuracy, NCE, and AFS of 0.8819, 0.4410, and 0.8302, respectively.

Figure 4 displays the comparative result analysis of the AIEDAE-ESRS system with present methodologies under OOV word detection. The figure stated that the AIEDAE-ESRS method has the capacity of achieving efficient outcomes with increased value of accuracy, NCE, and AFS. It is noted that the CRF and DNN models have shown minimum performance with minimal values of accuracy, NEC, and AFS. In line with this, the DURNN, DULSTM, and DBRNN systems have resulted in moderately closer accuracy, NEC, and AFS values. But, the AIEDAE-ESRS method has outperformed the other systems with the high accuracy, NCE, and AFS of 0.9708, 0.3720, and 0.7497, respectively.

Figure 5 displays the comparative analysis of the AIEDAE-ESRS procedure with current methods under CSI classification. The figure described that the AIEDAE-ESRS method has the capacity of achieving efficient outcomes with the increased values of accuracy, NCE, and AFS. It is noticed that the CRF and DNN models have shown minimum performance with minimal values of accuracy, NEC, and AFS. In line with this, the DURNN, DULSTM, and DBRNN systems have resulted in moderately closer accuracy, NEC, and AFS values. But, the AIEDAE-ESRS system has outperformed the other techniques with the higher accuracy, NCE, and AFS of 0.8579, 0.4120, and 0.6796 correspondingly.

Table 2: RMSE analysis of AIEDAE-ESRS technique with distinct training size.

Training size (%)	Root mean square error (%) CRF	DBRNN	AIEDAE-ESRS
10	2.21	1.81	1.12
20	2.00	1.75	1.06
30	1.97	1.73	1.08
40	1.99	1.80	0.93
50	2.00	1.63	1.04
60	1.98	1.75	0.89
70	2.00	1.65	0.98
80	2.01	1.77	1.03
90	2.02	1.70	0.80
100	2.01	1.73	1.16

Figure 6 displays the comparative analysis of the AIEDAE-ESRS method with current methods under deletion error detection. The figure described that the AIEDAE-ESRS method has the capacity of achieving efficient outcomes with the increased values of accuracy, NCE, and AFS. It is noticed that the CRF and DNN models have shown smallest performance with the minimal values of accuracy, NEC, and AFS. In line with this, the DURNN, DULSTM, and DBRNN methods have resulted in moderately closer accuracy, NEC, and AFS values.

Figure 7 exhibits the accuracy graph analysis of the AIEDAE-ESRS technique on the test MIT speech recognition dataset. The figure portrayed that the AIEDAE-ESRS technique has reached improved training and validation accuracy with increasing amount of epochs. It is also noticed that the training accuracy is considered to be lower compared to the validation accuracy.

Figure 8 demonstrates the loss graph analysis of the AIEDAE-ESRS technique on the test MIT speech recognition dataset. The figure depicted that the AIEDAE-ESRS technique has attained decreasing training and validation loss with a rise in the number of epochs. It is noticed the training loss is seemed to be higher than the validation loss.

Finally, a brief RMSE analysis of the AIEDAE-ESRS technique takes place under distinct sizes of training data is given in Figure 9 and Table 2 [23]. The result reported that the AIEDAE-ESRS technique has attained improved performance with the minimal values of RMSE compared to CRF and DBRNN techniques. For instance, with TS of 10%, the AIEDAE-ESRS technique has obtained lower RMSE of 1.12%, whereas the CRF and DBRNN techniques have attained higher RMSE of 2.21% and 1.81%, respectively. Meanwhile, with TS of 40%, the AIEDAE-ESRS system has attained lower RMSE of 0.93%, while the CRF and DBRNN techniques have attained higher RMSE of 1.99% and 1.80% correspondingly. Eventually, with TS of 60%, the AIEDAE-ESRS technique has obtained lower RMSE of 0.89%, whereas the CRF and DBRNN techniques have achieved higher RMSE of 1.98% and 1.75% correspondingly. Moreover, with TS of 80%, the AIEDAE-ESRS technique has obtained lower RMSE of 1.03%, while the CRF and DBRNN methods have accomplished higher RMSE of 2.01% and 1.77% correspondingly. Furthermore, with TS of 100%, the AIEDAE-ESRS technique has obtained lower RMSE of 1.16%, while the CRF and DBRNN systems have reached higher RMSE of 2.01% and 1.73% correspondingly.

By looking into the abovementioned figures and tables, it is ensured that our AIEDAE-ESRS methodology has gained maximal performances over the existing techniques.

5. Conclusion

In this study, an effective AIEDAE-ESRS technique has been developed for the accurate estimation and error detection in speech recognition model. The AIEDAE-ESRS technique involves three major processes, namely, preprocessing, DNN-HMM-based speech signal recognition, and FPA-based hyperparameter tuning. The utilization of the FPA helps to properly adjust the hyperparameters of the DNN-HMM model which supports to greatly increase the detection performance. The experimental result analysis of the AIEDAE-ESRS technique take place using benchmark dataset and investigated the results under several aspects. The simulation results reported the outstanding efficiency of the AIEDAE-ESRS methodology over the recent approaches. The improvements in experimental results reported the enhanced outcomes of the AIEDAE-ESRS technique based on various measures. With accuracy, NCE, and AFS values of 0.9921, 0.2640, and 0.6909, respectively, the AIEDAE-ESRS system outperformed the other techniques. With a TS of 100%, the AIEDAE-ESRS technique achieved a reduced root mean square error of 1.16 percent, whereas the CRF and DBRNN systems achieved a higher root mean square error of 2.01% and 1.73 percent, respectively. In the future, the performance of the AIEDAE-ESRS technique is additionally improved by the advanced DL models for speech recognition.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] Y. C. Tam, Y. Lei, J. Zheng, and W. Wang, "ASR error detection using recurrent neural network language model and complementary ASR," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2312–2316, Florence, Italy, 2014.

[2] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] T. Mikolov, S. Kombrink, D. Anoop, L. Burget, and J. Cernocky, "RNNLM–recurrent neural network language modeling toolkit," in *Proc. of the 2011 ASRU Workshop*, 2011.

[4] L. L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," in *Computer Speech and Language*no. 4, pp. 373–400, 2000.

[5] H.-A. Chang and J. Glass, "Discriminative training of hierarchical acoustic models for large vocabulary continuous speech recognition," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4481–4484, Taipei, Taiwan, 2009.

[6] G. Evermann and P. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, pp. 1655–1658, Istanbul, Turkey, 2000.

[7] A. Marin, T. Kwiatkowski, M. Ostendorf, and L. Zettlemoyer, "Using syntactic and confusion network structure for out-of-vocabulary word detection," in *2012 IEEE Spoken Language Technology Workshop (SLT).*, Miami, FL, USA, 2012.

[8] S. Kombrink, L. Burget, P. Matějka, M. Karafiát, and H. Hermansky, "Posterior-based out of vocabulary word detection in telephone speech," in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association*, Brighton, United Kingdom, September 6-10, 2009, DBLP.

[9] K. Yao, B. Peng, G. Zweig, D. Yu, X. Li, and F. Gao, "Recurrent conditional random field for language understanding," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4105–4109, Florence, Italy, 2014.

[10] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, 2001.

[11] A. I. Alhamada, O. O. Khalifa, and A. H. Abdalla, "Deep learning for environmentally robust speech recognition," *AIP Conference Proceedings*, vol. 2306, article 020025, 2020.

[12] K. J. Han, S. Hahm, B. H. Kim, J. Kim, and I. R. Lane, "Deep learning-based telephony speech recognition in the wild," in *Interspeech*, Stockholm, Sweden, 2017.

[13] Đ. T. Grozdić, S. T. Jovičić, and M. Subotić, "Whispered speech recognition using deep denoising autoencoder," *Engineering Applications of Artificial Intelligence*, vol. 59, pp. 15–22, 2017.

[14] A. Misbullah, H. H. Lin, C. Y. Chang, H. W. Yeh, and K. C. Weng, "Improving acoustic models for dysarthric speech recognition using time delay neural networks," in *2020 International Conference on Electrical Engineering and Informatics (ICELTICs)*, pp. 1–4, Aceh, Indonesia, 2020.

[15] A. Ogawa and T. Hori, "ASR error detection and recognition rate estimation using deep bidirectional recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4370–4374, South Brisbane, QLD, Australia, 2015.

[16] A. Ogawa, T. Hori, and A. Nakamura, "Estimating speech recognition accuracy based on error type classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2400–2413, 2016.

[17] B. Mor, S. Garhwal, and A. Kumar, "A systematic review of hidden Markov models and their applications," *Archives of Computational Methods in Engineering*, vol. 28, no. 3, pp. 1429–1448, 2021.

[18] L. Li, Y. Zhao, D. Jiang et al., "Hybrid deep neural network–hidden Markov model (dnn-hmm) based speech emotion recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 312–317, Geneva, Switzerland, 2013.

[19] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pretrained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[20] T. T. Nguyen, J. S. Pan, and T. K. Dao, "An improved flower pollination algorithm for optimizing layouts of nodes in wireless sensor network," *Ieee Access*, vol. 7, pp. 75985–75998, 2019.

[21] S. Rahini, "Large scale optimization to minimize network traffic using MapReduce in big data applications," in *International Conference on Computation of Power, Energy Information and Communication (ICCPEIC)*, pp. 193–199, 2016.

[22] S. Kalra and S. Arora, "Firefly algorithm hybridized with flower pollination algorithm for multimodal functions," in *Proceedings of the International Congress on Information and Communication Technology*, pp. 193–199, 2016.

[23] S. Neelakandan, M. A. Berlin, S. Tripathi, V. Brindha Devi, I. Bhardwaj, and N. Arulkumar, "IoT-based traffic prediction and traffic signal control system for smart city," *Soft Computing*, vol. 25, no. 18, pp. 12241–12248, 2021.