electronics

MDPI

Article

Improving Mispronunciation Detection of Arabic Words for Non-Native Learners Using Deep Convolutional Neural Network Features

Shamila Akhtar [1], Fawad Hussain [1], Fawad Riasat Raja [2], Muhammad Ehatisham-ul-haq [1], Naveed Khan Baloch [1], Farruh Ishmanov [3,*] and Yousaf Bin Zikria [4,*]

[1] Department of Computer Engineering, University of Engineering and Technology, Taxila 47050, Pakistan; chaudaryshamila10@gmail.com (S.A.); fawad.hussain@uettaxila.edu.pk (F.H.); ehatishamuet@gmail.com (M.E.-u.-h.); naveed.khan@uettaxila.edu.pk (N.K.B.)
[2] Machine Intelligence and Pattern Analysis Laboratory, Institute of Integrated and Intelligent Systems, School of Information and Communication Technology, Griffith University, Nathan, QLD 4111, Australia; faraja352@gmail.com
[3] Department of Electronics and Communication Engineering, Kwangwoon University, Seoul 447-1, Korea
[4] Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Korea
[*] Correspondence: farruh@kw.ac.kr (F.I.); yousafbinzikria@ynu.ac.kr or yousafbinzikria@gmail.com (Y.B.Z.)

Received: 16 March 2020; Accepted: 1 June 2020; Published: 9 June 2020

Abstract: Computer-Aided Language Learning (CALL) is growing nowadays because learning new languages is essential for communication with people of different linguistic backgrounds. Mispronunciation detection is an integral part of CALL, which is used for automatic pointing of errors for the non-native speaker. In this paper, we investigated the mispronunciation detection of Arabic words using deep Convolution Neural Network (CNN). For automated pronunciation error detection, we proposed CNN features-based model and extracted features from different layers of Alex Net (layers 6, 7, and 8) to train three machine learning classifiers; K-nearest neighbor (KNN), Support Vector Machine (SVM) and Random Forest (RF). We also used a transfer learning-based model in which feature extraction and classification are performed automatically. To evaluate the performance of the proposed method, a comprehensive evaluation is provided on these methods with a traditional machine learning-based method using Mel Frequency Cepstral Coefficients (MFCC) features. We used the same three classifiers KNN, SVM, and RF in the baseline method for mispronunciation detection. Experimental results show that with handcrafted features, transfer learning-based method and classification based on deep features extracted from Alex Net achieved an average accuracy of 73.67, 85 and 93.20 on Arabic words, respectively. Moreover, these results reveal that the proposed method with feature selection achieved the best average accuracy of 93.20% than all other methods.

Keywords: computer-aided language learning; deep convolutional neural network; mispronunciation detection; mel frequency cepstral coefficients (MFCC); transfer learning

1. Introduction

Speech is the semantic element of human communication through which human convey their message to each other. Therefore, in the last decade, Computer-Aided Language Learning (CALL) has received considerable attention due to its adaptability, which enables students at their own pace to improve their language abilities. Computer-Aided Pronunciation Training (CAPT) is a more precise sub-region of CALL, which emphasis on areas such as automatic pointing errors in non-native learners’ speech. CALL accomplished many tasks, such as pronunciation error detection, speech recognition, and pronunciation scoring. Moreover, much research has been done to apply speech processing techniques to different languages [1–3]. For a specific language, speakers of different languages tend to form a distinctive muscle of the mouth. Sometimes our muscles may not be able to speak a non-native language because they create pronunciation errors in their utterances. There is no clear meaning of right or wrong pronunciation because it is a challenging task to measure the word "correct pronunciation". Similarly, the variation that produced in accent is recognized and measured, either it is correct or otherwise, through a speaker's prior knowledge about the people having a different accent. For perceiving and exploring foreign accents of speakers, authors in [4] performed many linguistic experiments. There are many reasons for the mispronunciation of a speaker, such as a change in the mother tongue and accent. Consequently, pronunciation mistakes are categorized into prosodic and phonemic errors [5]. Phonemic errors exist due to phones that can differentiate one word from another and create a difference in the meaning of speech. In this case, speakers mostly interchange the complete phonemes with another similar phoneme that creates a difference in sound. On the other side, prosodic error exists due to phonemic variations. The prosodic error includes errors based on annotation, rhythm, and stress. Usually, researchers used acoustic-phonetic features to detect mispronunciation. Many researchers have investigated to detect mispronunciation for different languages (English, Mandarin, Japanese, and Dutch), but little work is done in Arabic. Researchers have used different techniques to detect mispronunciation. These techniques include posterior probability-based methods classifier-based methods, and deep learning-based methods. Mostly authors worked on some confusing phonemes of the Arabic language, but still little work is done on Arabic words.

The Holy book for all Muslims is, "Al-Quran" and it is written in the Arabic language; therefore, it is in high demand to develop a framework to detect the pronunciation errors in the Arabic language. For the recitation of the Holy book, there are some rules known as "Tajweed" that must be followed for proper recitations. These rules define the correct pronunciation of every word of the holy book. By following these rules, one can recite the Holy Book correctly. However, if the recitation is not according to these rules, the meaning of a word is changed. The Arabic words consist of some confusing phonemes that may sound similar, thus making it difficult for non-native learners to pronounce the words correctly. Therefore, extracting appropriate features to differentiate these phonemes is still a very challenging task. Hence, in this paper, we used the Alex net framework to extract high-level features from raw pixel values to differentiate the Arabic words efficiently. In this aspect, we propose a deep Convolution Neural Network (CNN) based feature extraction model for detecting pronunciation errors of the Arabic words. In the first step, we collected the audio dataset of Quranic verses from different speakers. Then, we computed the spectrogram of the speakers' audio and used it as an input to the CNN model for extracting discriminating features from three fully connected layers of Alex Net. These features are the most appropriate features to detect the mispronunciation of the Arabic words. Moreover, we also used the transfer learning-based model. In this model, we passed the spectrogram of audio signals (dataset) to the pre-trained model of CNN (Alex Net) to classify the correct words. We found that the features extracted from the Alex Net perform better than transfer learning-based models and handcrafted features in detecting mispronunciation of Arabic words. Hence, the proposed CNN-based feature extraction method improves the accuracy of mispronunciation detection of Arabic words as compared to the state-of-the-art speech processing features. Major contributions of this paper are as follows:

- To date, there is no standard dataset available for the words of the Holy Quran, which can be used to detect the mispronunciation. Therefore, we created our own dataset of Arabic words from native and non-native speakers, which is also available for future research.
- We developed an algorithm for the detection of mispronunciation of Arabic words by extracting deep features through different layers of CNN.

- A feature selection technique is presented to select the most discriminative set of features for improving the detection of pronunciation errors.
- The performance of the proposed feature-based model is compared with the transfer learning-based method and traditional handcrafted features, where the proposed feature-based model achieves significant accuracy to detect the mispronunciation of Arabic words.

The remaining part of the study is summarized as follows. Section 2 describes previous methods that are used for mispronunciation detection. In Section 3, we discussed in detail the dataset we used and the proposed methodology. In Section 4, we analyzed the results, and at last, the conclusion and future work are given in Section 5.

2. Related Work

Traditionally, for mispronunciation-detection, researchers have used the posterior probability and log-likelihood scores. Strik et al. [6] proposed acoustic–phonetic features, for example, zero-crossing rate and log root mean square, energy for identifying velar plosive, and velar fricative. To improve the quality of pronunciation error detection for the Mandarin language, Feng Zhang et al. [7] proposed Scaled Log Posterior Probability (SLPP) and weighted phone SLPP. To determine the effectiveness of these techniques, experiments are performed by reducing Fault Acceptance Rate (FAR) from 41.1 to 31.4 at 90 False Rejection Rate (FRR) and 36.0 to 16.3 at 95 FRR. The issues in the Likelihood Ratio Test (LRT) and posterior probability-based methods are acoustic observations and precision of an acoustic model. These techniques fail to identify the exact location and type of error and only suitable for pronunciation proficiency.

George Georgoulas et al. [8], proposed Discrete Wavelet Transform (DWT) technique that extracts handcrafted features. After extracting the features, they used the SVM classifier for the classification of a task. Wei et al. [9], proposed a method that was different from the traditional posterior probability-based approach. The authors used log-likelihood ratios between all acoustic models as features and used SVM for the classification to detect the text-dependent mispronunciation. They also proposed Pronunciation Space Models (PSM) for pronunciation variations to enhance the competence of acoustic models. Therefore, the proposed framework outperforms all traditional probability-based techniques, which are only suitable for pronunciation proficiency.

Wenping Hu et al. [10] proposed a technique in which they used logistic regression and a Deep Neural Network (DNN) for enhancement in the mispronunciation detection system. Witt et al. [11], proposed a method that used the likelihood-based goodness of pronunciation score by the span of each phone part. This method sets an individual threshold for each phone to find whether each phone is pronounced correctly or not. The value of the threshold is based on both the human judge score and averaged native confidence score. Gu L et al. [12], worked on mispronunciation detection of English words by using the posterior probability-based method. The goodness of pronunciation is used to detect mispronounce words.

Dahan H et al. [13], develops a system that helps the students to improve their pronunciation skills of the Arabic language. The proposed log-likelihood probability for giving a mark on each pronunciation and achieved an average accuracy of 89.63. Abdou et al. [14] used a confidence scoring scheme, but the proposed method fails to differentiate confusable phonemes. Khaled Necibi et al. [15] used a method that detects the pronunciation errors with high accuracy but the proposed method only works with three Arabic words. Muazzam et al. [16] used acoustic–phonetic features which is only suitable to detect individual phone errors, therefore, a separate classifier is required for each pronunciation error. Hanani et al. [17] proposed MFCC with GMM-UBM technique but this technique cannot locate the exact position and type of errors. Table 1 summarizes some of the studies which have addressed for mispronunciation detection.

Table 1. Literature summary.

Author	Techniques	Pros	Cons
Strik et al. [6]	Acoustic–phonetic features (APF)	The proposed method used APF that represent best features	These features are computationally intensive due to high-dimensionality
Zhang et al. [7]	Scaled log-posterior probability (SLPP)	Improve the performance of mispronunciation detection at syllable level	These methods produce acoustic observations and precision of an acoustic model
Georgoulas et al. [8]	DWT + SVM	Used different sets of features that achieved high performance	Similar sounds difficult to classify
Wei et al. [9]	log-likelihood ratio +SVM	Proposed method outperforms all traditional probability -based techniques	Each phone is modeled with several parallel acoustic models
Hu et al. [10]	DNN based on extended recognition networks (ERN)	Proposed ERNs leverage existing automatic speech recognition	acoustic models are trained independently and hence, contextual information is lost
Witt et al. [11]	likelihood-based method	This method detects phone-level pronunciation errors	The proposed technique fails to identify the exact location of error
Gu et al. [12]	Posterior probability-based method.	Proposed method achieved excellent results	These techniques are not sufficient to describe speech characteristics
Dahan et al. [13]	HMM-based method	Help teachers to learn the Arabic language	The proposed method fails to differentiate some errors
Abdou et al. [14]	Confidence scoring	Helps students to learn holy Quran	The proposed method fails to differentiate confusable phonemes
Khaled Necibi et al. [15]	Global average log likelihood	Proposed system is able to detect words errors pronunciation with high accuracy	Only three words used in dataset (Chamsoun, Kataba, Kourssi)
Muazzam et al. [16]	Acoustic phonetic features	These features are more specific to detect an individual phone's error	A separate classifier is needed for every pronunciation error
Abual soud Hanani et al. [17]	MFCC with GMM-UBM	Only suitable for pronunciation proficiency	The proposed technique fails to identify the exact location and type of error

As mentioned earlier, many researchers have investigated the use of non-statistical feature selection methods and machine learning-based techniques for mispronunciation detection of Arabic phonemes, but still, there is lack of attention towards detecting mispronunciation of Arabic words which is a challenging task owing to the fact that one word consists of many confusing phonemes. Therefore, in this paper, we focus developing a system that can solve many of the problems of Arabic word pronunciation. Therefore, we extract deep features from Alex Net and use an automated feature selection method that can be proved useful and achieved better results. Extracting features is a major important stage in speech recognition. Currently, much attention is given to deep learning algorithms because of their ability to learn high-level features from the data.

3. Proposed Methodology

In this paper, we developed two models to detect mispronunciation of Arabic words that are shown in Figure 1. In CNN features-based model, we extracted deep features using pre-trained model of Alex Net to systematize the procedure of feature extraction. For extracting features from Alex net, we converted audio signals into spectrograms and then passed these spectrograms to CNN. We applied three different classifiers KNN, SVM, and RF, to estimate the performance of these features. In the transfer learning-based model, we fed the spectrogram dataset to Alex net through which we extracted the features automatically and then performed classification on these features. The complete description of each step is specified in the following subsections.

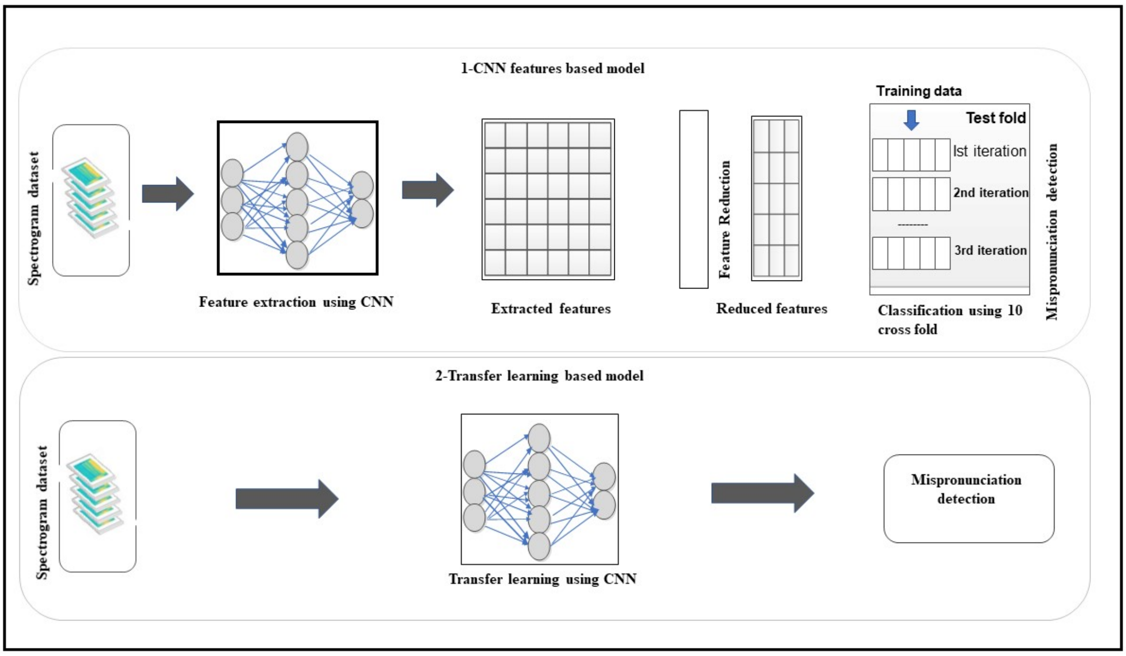
1-CNN features based model
Spectrogram dataset
Feature extraction using CNN
Extracted features
Feature Reduction
Reduced features
Training data
Test fold
Ist iteration
2nd iteration
3rd iteration
Classification using 10 cross fold
Mispronunciation detection
2-Transfer learning based model
Spectrogram dataset
Transfer learning using CNN
Mispronunciation detection

Figure 1. The methodology of our proposed framework.

3.1. CNN Features-Based Model

The detailed description of each step for pronunciation error detection using the CNN feature-based model is shown in Figure 2. First, we collected the audio dataset consisting of Arabic words from different speakers that belong to different regions of Pakistan. Then, we extracted deep features from different layers of Alex net. We performed pre-processing on the data and then passed the preprocessed data to the pre-trained model of Alex Net. We extracted the features through fully connected layers of Alex Net, i.e., layer 6, layer 7, and layer 8. These layers have features with high dimensions and need a lot of time to process these features. Therefore, we applied the correlation-based feature selection technique on these features to reduce the dimensions of the data. The detailed description of each step is described in the following subsection.

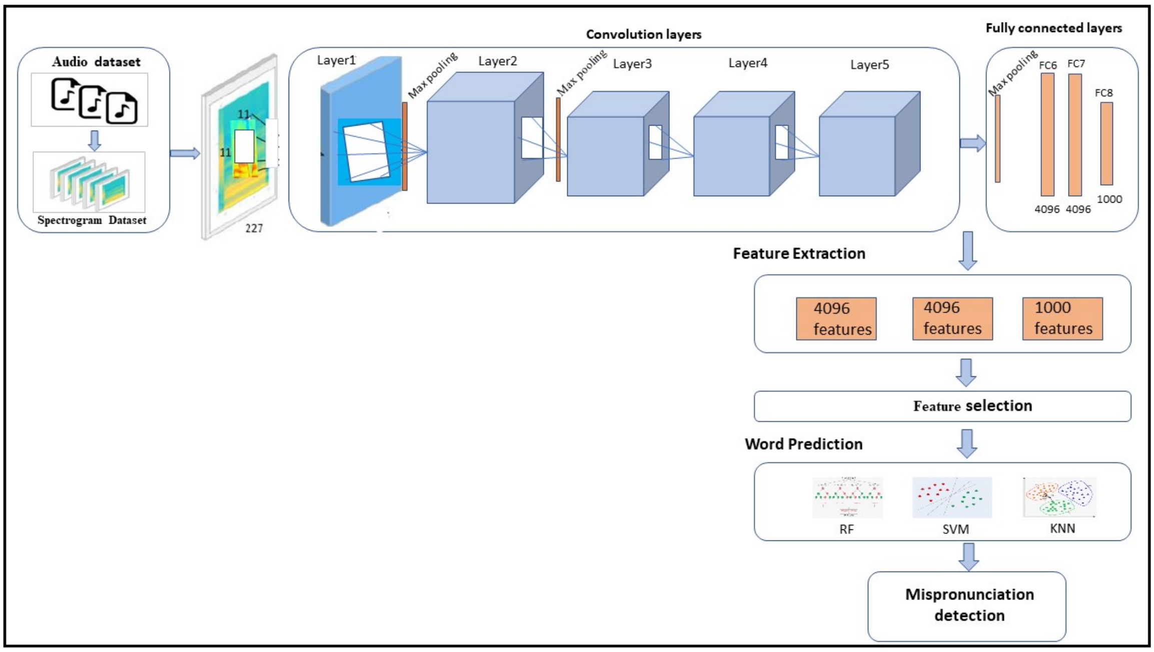

Figure 2. The features extracted from different layers of Convolution Neural Network (CNN) (layer 6, 7, 8) to detect mispronunciation.

3.1.1. Dataset Acquisition

Arabic is the fifth-largest native language all around the world, but still, there is no standard dataset available for this language. Therefore, we created the dataset consisting of Arabic words, and then we choose multiple words that cover all the Arabic letters. We recorded each word 30 times, and each recording is from a different Pakistani speaker who has been learning the Arabic language. The details of each speaker are described in Table 2. We choose people of different ages (8 to 60 years old) and genders. We used a standard microphone that is readily available from the market, and recorded the audio in an office environment with minimum background noise. The recording frequency of sound was 44,100 Hz. After collecting the dataset, the next stage was labeling the data, which has been done by the three different Arabic language experts. These experts have extensive knowledge of Tajweed. We used a voting mechanism to label the data, and only if two of the three experts agree on the same label, then it has been assigned to the data. On the basis of labeling the dataset the ratio of mispronounced words is 45% in our collected dataset while 55% data is correctly pronounced.

Table 2. The detail of speakers used in dataset.

No. of Speakers	Adult Male	Adult Female	Children	Total
Native	8	7	2	17
Non-Native	6	6	1	13
Total	14	13	3	30
No. of Words	**Adult Male**	**Adult Female**	**Children**	**Total**
Native	490	490	210	1190
Non-Native	490	420	70	980
Total	980	910	280	2170

3.1.2. Data Pre-Processing

For feature extraction, spectrograms (2-D) are used by Alex Net. After removing noise from the raw audio signal, we converted all audio signals into spectrograms. Spectrogram re-sampling is performed because deep learning algorithm CNN (Alex net) requires input images of size 227 × 227 × 3. Therefore, for extracting deep features through Alex Net, we used augmented data-store to automatically re-size the spectrograms' dataset according to the input layer of Alex Net before using it for the training of model. After data pre-processing, we forwarded the spectrogram data to CNN model (Alex net) for feature extraction.

3.1.3. Feature Extraction Using CNN

The Alex Net model was trained on the Image-Net database that consists of millions of images [18]. This model can classify 1.2 million images into 1000 object categories. This model is also applied to other databases because it can learn more vibrant feature representation for the wide range of images. The deep learning algorithm has been extensively recognized to learn predictive features directly from the original images. This model has been developed by the Super Vision group. The architecture of Alex-Net with the complete procedure of feature extraction using this CNN model is shown in Figure 2. It consists of 11 layers in which five layers are convolutional layers, three layers are fully connected, and the remaining three layers are Max Pooling layers. This model is successful due to the dropout regularization system and the Rectified Linear Unit (ReLU). The ReLU is a non-linearity layer that works as a half-wave rectifier and used to speed-up the training process of the network. It also prevents over-fitting problem as described in (1).

$$f(x) = max(x, 0) \tag{1}$$

3.1.4. Feature Selection

After extracting features from different layers of Alex net, a feature selection method, "Correlation-based feature Subset Evaluator", was applied on these features for eliminating the irrelevant and redundant features that do not play a vital role in classification. The dimensions of features extracted from CNN is large. To process the large features, we need a lot of time to train the classifier, which reduced the proficiency of the classification algorithm. Therefore, we cut the feature dimensions using a correlation-based feature selection technique.

3.1.5. Correlation-Based Feature Subset Evaluator (CFS)

Correlation-based feature selection evaluator estimates the worth of a subset of features based on hypothesis, i.e., "Good feature subsets are extremely correlated with the class and are preferred while having low inter-correlation with other features". CFS [19] specify ranks attributes, giving a heuristic evaluation function based on the correlations. The process of the feature selection method consists of four steps, which are explained in [20]. In CFS, correlation measures the resemblance between two features. If the two features correlate with each other, then their correlation-coefficient is 1. However, if the features are not correlated with each other, then their correlation coefficient is 0. Evaluation function [21] of CFS is shown in (2).

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{2}$$

where $Merit_s$ is the heuristic merit with s that has k features, $\overline{r_{ff}}$ is average feature–feature intercorrelation and $\overline{r_{cf}}$ is the mean of feature–class correlation.

3.2. Classification Algorithms

After feature extraction and selection, the next step is to choose a suitable classifier that can correctly classify Arabic words. In this paper, we used KNN, SVM, and RF classifiers for the detection of mispronunciation. These classification algorithms are explained in the next subsection.

3.2.1. K-Nearest Neighbor

The KNN [22] is a simple classifier that classifies the new instance and stores all instances by measuring similarity. For mispronunciation detection, We used k = 1 that finds most nearest neighbor based using Euclidean distance to predict a word/mispronunciation. If we have two words *a* and *b*, then Euclidean distance [23] can be measured by using (3).

$$|a - b| = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{3}$$

where, *n* signifies features that exist in the feature set of correct words *a* and incorrect words *b*.

3.2.2. Support Vector Machine

SVM [24] is the most powerful supervised machine learning algorithm that is used for binary classification. In this research work, we used one versus one approach. We standardize the data before applying the SVM algorithm. The following parameters are used to classify the data by using polynomial kernel [25,26], as shown in (4).

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \tag{4}$$

where x_j is the support vector, x_i represents the input data, K represents the kernel function, r represents the constant term, and d is the degree of polynomial. We can adjust these parameters according to our data.

3.2.3. Random Forest

The RF [27] is a meta estimator and easy to use supervised machine learning algorithm. It generates multiple decision trees and added randomness to achieve maximum performance. By using the voting technique of RF, the proximity function $prox(i, j)$ [28] between two elements *i* and *j* can be defined as:

$$prox(i, j) = \frac{\sum_{t=1}^{ntree} ht(i) = ht(j))}{ntree} \tag{5}$$

In (5), *ht* is the *t*th tree in the forest and *prox(i,j)* = 1 if all trees classify *i* and *j* in the same classes.

3.3. Transfer Learning-Based Model

The detail description of each step for pronunciation error detection using transfer learning is shown in Figure 3. In this model, we don't need to use a separate classification algorithm to detect mispronunciation. CNN automatically learns discriminative features that correctly classify the words. We performed pre-processing that is the same as we used in features-based model. The description of each step by using this model is described below.

3.3.1. Load Pre-Trained Network of CNN (AlexNet)

In the transfer learning-based model, first, we performed pre-processing and then import the pre-trained model of AlexNet. After pre-processing on an audio dataset, we load this model, which is trained on the ImageNet database, consisting of millions of images and classify 1.2 million images into 1000 object groups. Then, we passed the spectrogram of the dataset as an input to this model (AlexNet) and split the data into two parts, i.e., training and testing. We input 70 percent data for training while 30 percent data for testing purposes to the CNN model.

learning1.jpg

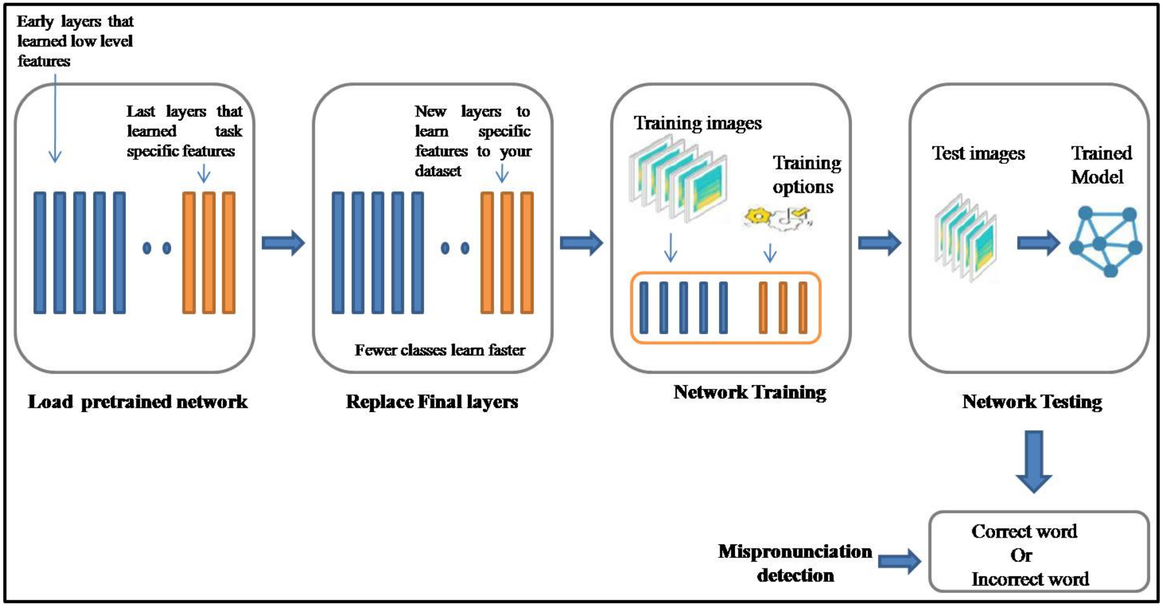
Early layers that learned low level features
Last layers that learned task specific features
Load pretrained network
New layers to learn specific features to your dataset
Fewer classes learn faster
Replace Final layers
Training images
Training options
Network Training
Test images
Trained Model
Network Testing
Correct word
Or
Incorrect word
Mispronunciation detection

Figure 3. The Methodology of the proposed transfer learning-based model for mispronunciation detection.

3.3.2. Replacing Final Layers

The last three fully connected layers of the pre-trained model are constructed for 1000 classes. These layers learn high-level features that are used to classify the data while other layers of Alex Net learn low-level features. Therefore, we fine-tuned these layers for performing classification on our dataset. In this model, we replaced the last fully connected layers with new layers to learn discriminative features for our unique dataset. The layers C1–C5 of the pre-trained network are trained on a large ImageNet dataset and transfer these layers for new tasks by replacing the last three layers of Alex Net with the SoftMax layer, which is a fully connected layer and a classification output layer. After transferring the last layers with new layers, we specify the parameters for a fully connected layer with new data. We adjusted the output size of a fully connected layer to the same number of classes of our dataset. The other parameters, such as weight and bias, learn the rate factor, and are used in a fully connected layer. We used cross-entropy as a loss function, and the output size is the same as the number of classes in the new spectrogram dataset.

3.3.3. Network Training and Testing

The pre-trained model of AlexNet has eight convolutional layers. The last three layers must be fine-tuned for new spectrogram dataset and replaced with a fully connected layer, a softmax layer, and a classification output layer. These newly replaced layers are trained on the dataset of the spectrogram to classify the words of the Arabic language only correctly. In the transfer learning-based model, we also used k-fold cross-validation for results evaluation. In the k-fold cross-validation method, the whole dataset is divided into k equal sized subsamples, where a single subsample is used as testing data and remaining k-1 subsamples are used for training data. This process is repeated for k times, where we used k = 10. The hyperparameters for the proposed model are tuned based on the training data, where we trained the CNN model with a learning rate of 10^{-4} and batch size of 10.

4. Experimental Results

As the proposed framework focuses on the mispronunciation detection of the Arabic words; therefore, in this work, we used three classifiers RF, SVM, KNN to detect mispronunciation that classifies the words correctly. For these classifiers, the hyper-parameters were tuned based on the training set as follows. For the SVM classifier, a pairwise classification, i.e., 1-vs.-1 and polynomial kernel with sequential minimal optimization (SMO) algorithm [29] was used. A random tree was used with a base-learner for the RF classifier and the number of iterations was set to 100. For K-NN classifier, K was set equal to 1. We used a 10-fold cross-validation, where we used 90% data for model training with 10 percent test data during each round of the validation. The selected features are used for training the classifier, which classifies the features to predict whether the words are pronounced correctly or incorrectly.

4.1. Performance Evaluation

In this paper, the performance of the proposed framework is evaluated based on accuracy. For the classification task, accuracy can be calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

In (6) *TP*, *FN*, *FP* and *TN* represent the number of true positives, false negatives, false positives and true negatives, respectively.

In general, the accuracy of a prediction model is computed based on the error rate, where the error rate represents the difference between a measurement and the true value of the measurand (i.e., the quantity being measured). Hence, the error rate is a better estimate for continuous value prediction, where accuracy can be obtained as 1-error. However, this is not valid in case of classification with discrete classes, where we only evaluate if the correct class is predicted or not, and do not evaluate the closeness of the prediction. Hence, the above-discussed definition of accuracy is more meaningful.

In addition to accuracy, we also define another performance evaluation parameter, i.e., negative predictive value (NPV), for the proposed model. NPV is the measure of probability that an algorithm classifies the instances as negative while actual instances also belong to the negative class. The formula for NPV is provided in (7).

$$NPV = \frac{TN}{TN + FN}, \tag{7}$$

where *TN* represents the number of true negatives and *FN* represent the number of false negatives.

4.2. Results of CNN Features-Based Model

In our proposed features-based model, we extracted the features from different layers, i.e., layer 6, layer 7, and layer 8 of Alex net. Table 3 shows the results which are achieved on individual words by three different classifiers SVM, RF, and KNN by using the features' set of fully connected layer, i.e., FC6. The features extracted from the FC6 layer have 4096 dimensions. The performance of the selected classifiers was evaluated on the FC6 layer, where SVM achieved the best results.

Table 3. The comparison of different classifiers using features set of layer 6 on individual words.

Words	RF	SVM	KNN	Words	RF	SVM	KNN
Bis'mi	80	83.33	86.66	Yawmi	86.66	83.33	86.66
Allahi	63.33	73.33	60	I-dini	70	76.66	66.67
I-rahmani	63.33	63.33	70	Iyyaka	80	73.33	70
I-rahimi	80	86.66	73.33	Nabadu	63.33	80	60
Al-hamdu	90	90	90	Wa-iyaka	70	66.66	56.66
Lillahi	83.33	76.66	70	Nastainu	56.66	46.66	50
Rabbi	80	66.66	80	Ihdina	56.66	53.33	56.66
I-alamina	66.66	66.66	66.66	I-sirata	66.67	83.33	40
Maliki	73.33	73.33	80	mustaqima	70	76.66	73.33
Sirata	63.33	76.66	73.33	magdubi	73.33	73.33	70
Alladhina	63.15	57.89	57.89	Alayhim	90	86.66	86.66
Anamta	60	60	50	Wala	80	80	80
Alayhim	90	86.66	86.66	I-dalina	76.66	73.33	76.66
Ghayri	90	90	90	-	-	-	-

In the next step, we evaluated the performance of the same three classifiers KNN, SVM, and RF, on the feature set that is attained from a fully connected layer (FC7) of Alex Net. The features extracted from layer 7 have 4096 dimensions. Firstly, we experimented on individual words by using a feature set of the FC7 layer, as shown in Table 4.

Table 4. The comparison of different classifiers using features set of layer 7 on individual words.

Words	RF	SVM	KNN	Words	RF	SVM	KNN
Bis'mi	80	80	73.33	Yawmi	86.66	80	80
Allahi	56.66	60	60	I-dini	70	70	73.33
I-rahmani	80	90	90	Iyyaka	80	70	66.66
I-rahimi	80	90	90	Nabadu	60	86.66	50
Al-hamdu	90	83.33	90	Wa-iyaka	63.33	66.66	63.33
Lillahi	83.33	73.33	76.66	Nastainu	53.33	40	63.33
Rabbi	80	73.33	56.66	Ihdina	56.66	56.66	46.66
I-alamina	73.33	76.66	66.66	I-sirata	70	70	76.66
Maliki	70	70	73.33	mustaqima	70	70	76.86
Sirata	73.33	76.66	73.33	magdubi	70	73.33	66.66
Alladhina	52.63	63.15	57.89	Alayhim	90	86.66	83.33
Anamta	60	66.66	46.66	Wala	86.66	76.66	86.66
Alayhim	90	86.66	83.33	I-dalina	80	76.66	73.33
Ghayri	90	90	93.33	-	-	-	-

Table 5 shows the performance of selected classifiers on individual words by using the feature set of layer 8, i.e., FC8. The features extracted from layer 8 have 1000 dimensions.

Table 5. The comparison of different classifiers using features set of layer 8 on individual words.

Words	RF	SVM	KNN	Words	RF	SVM	KNN
Bis'mi	73.33	86.66	80	Yawmi	86.66	76.66	80
Allahi	56.55	56.66	53.33	I-dini	83.33	80	76.66
I-rahmani	70	53.33	56.66	Iyyaka	80	73.33	66.66
I-rahimi	70	66.66	56.66	Nabadu	60	63.33	60
Al-hamdu	90	76.66	86.66	Wa-iyaka	66.66	60	70
Lillahi	83.33	66.66	70	Nastainu	50	50	46.66
Rabbi	80	73.33	56.66	Ihdina	56.66	43.33	43.33
I-alamina	66.66	63.33	66.66	I-sirata	70	66.66	73.33
Maliki	70	66.66	80	I-mustaqima	70	66.66	73.33
Sirata	60	66.66	60	I-maghdubi	72.41	68.96	58.06
Alladhina	52.63	68.42	47.36	Alayhim	90	83.33	86.66
Anamta	53.33	66.66	56.66	Wala	83.33	86.66	83.33
Alayhim	90	83.33	86.66	I-dalina	96.66	83.33	90
Ghayri	90	93.33	93.33	-	-	-	-

A comparison of selected classifiers, i.e., KNN, SVM, and RF on features sets extracted from FC6, FC7, and FC8 of CNN without feature selection, is shown in Table 6.The SVM classifier achieved an average accuracy of 75.19 on layer 6 while NPV 0.65 due to its stability property and its ability for automatic feature selection. The RF and KNN classifiers achieved an average accuracy of 73.33 and 70.53 and NPV 0.51, 0.48 respectively.The experimental results achieved from layer 7 shown that RF achieved best results, i.e., an average accuracy of 74.81, NPV 0.58 as compared to SVM and KNN, which have an accuracy of 74 and 71.86, where NPV is 0.55 and 0.5 respectively. RF added the randomness to the model, as shown in Table 6 therefore, it achieves better results. By analyzing Table 6, we can see that RF achieved the best results on layer8 as compared to KNN and SVM with an average accuracy of 72 and NPV 0.52. The classifiers SVM and KNN achieved the lowest average accuracy of 69.55, 69 and NPV 0.52, 0.45, respectively. From the results obtained on layer 7, we can conclude that accuracy is increased by 2%. On FC8 feature set, RF achieves an accuracy of 72 while 74.81 on FC7 features set.

Table 6. The average results extracted from different layers without feature selection.

Average Accuracy			
Classifier	Layer 6	Layer 7	Layer 8
SVM	75.19	74	65.55
RF	73.33	74.81	72
KNN	70.53	71.86	69
Average NPV			
Classifier	Layer 6	Layer 7	Layer 8
SVM	0.65	0.55	0.52
RF	0.51	0.58	0.4
KNN	0.48	0.5	0.45

The comparison results on these three-classification algorithms show that SVM and RF perform better than KNN, and the best accuracy is achieved at FC6. When processing a very large set of features, the efficiency of the classification algorithm may decrease. Therefore, to remove redundant features from feature set, we applied the correlation-based features selection technique on these features to select the relevant features that can correctly classify the data. To check the performance of feature selection technique, we run the algorithm on the features set that is extracted from FC6, FC7, and FC8 layers. Experimental results show that our feature selection technique on features extracted from CNN layers (6, 7 and 8) improves the accuracy, as shown in Table 7.

For results evaluation, we performed experiments on each word individually as our work is mispronunciation detection of individual words. In this regards, the whole network took an average training time of approximately 3 min per word using 10-fold cross validation. After applying feature selection technique on the features extracted from different layers of Alex Net (layer 6, layer 7, and layer 8), this time is reduced further with the increase in overall accuracy. This signifies the benefit of entailing feature-selection in the proposed scheme.

Another comparison of selected classifiers was made on a feature set of layer 6, layer 7, and layer 8 with a feature selection technique that shows SVM achieved best classification accuracy, i.e., 93.02, NPV = 0.9 on Arabic words. RF obtained an accuracy of 91.86, NPV = 0.8 while KNN achieved the lowest average accuracy of 90, NPV = 0.78 on a feature set of FC6 layer.

Table 7. The average results extracted from different layers using feature selection.

Average Accuracy			
Classifier	Layer 6	Layer 7	Layer 8
SVM	93.2	89.86	83.02
RF	91.86	89.99	88.51
KNN	90	89	83.17
Average NPV			
Classifier	Layer 6	Layer 7	Layer 8
SVM	0.9	0.75	0.7
RF	0.8	0.77	0.73
KNN	0.78	0.73	0.7

4.3. Results Achieved Using Transfer Learning-Based Model

Transfer learning is a popular method that learns new tasks by transferring knowledge from previous tasks. In transfer learning, we train the network on the base dataset and extract the learned features. These features are transferred to a second task to train the target dataset. The feature extraction and classification tasks are performed automatically in transfer learning. In this work, we input the data to the model in the form of a spectrogram, and as an output, we got a mispronunciation score. In the transfer learning-based model, the process of fine-tuning is performed through different parameters such as learning rate and bias. In our proposed work, we applied this model to each word and evaluated the result by changing these parameters. We recorded the results with the following configurations of different parameters,bias learn factor from 10 to 100 and learning rate from 10^{-1} to 10^{-10}. The optimal results have been achieved at Mini Batch Size of 10, learning rate 10^{-4}, and bias learns factor 90. We applied the transfer learning model with different numbers of epochs, i.e., 10, 15, 20, and 25 on each individual word. Table 8 summarises the performance of the transfer learning-based model in terms of accuracy and NPV with a different number of epochs on all words that we used in this paper. The classification results of the transfer learning-based model on individual words with a different number of epochs are shown in Table 9.

Table 8. The average accuracy and NPV of transfer learning model on 10, 15, 25 and 30 epochs on all words.

No of Epochs	**Learning Rate**	**Bias Learn Factor**	**Accuracy**	**NPV**
10	1.00×10^{-4}	50	66.00%	0.48
15	1.00×10^{-5}	90	74.15%	0.61
25	1.00×10^{-5}	90	85.00%	0.75
30	1.00×10^{-5}	90	71.00%	0.5

Table 9. The performance of transfer learning-based model on individual words using 10, 15, 25 and 30 epochs.

Words	10	15	25	30	Words	10	15	25	30
Bis'mi	60.6	70.21	88.89	83.5	Nastainu	55.55	75	66.67	60.5
Allahi	54.67	70.89	75	65.87	Ihdina	55.55	88.88	100	55.5
I-rahmani	74.78	68.89	77.78	54.55	I-sirata	50.89	100	77.78	65.5
I-rahimi	65	60.88	75	70.5	Sirata	75	60.88	87.5	75
Al-hamdu	66.66	88.88	100	83.5	Alladhina	65.55	100	66.67	55.55
Lillahi	77.78	70.67	87.5	80.88	Alayhim	75	85.88	87.5	83.5
Rabbi	85.89	60.88	88.89	83.5	Ghayri	55.55	60.88	88.89	83.5
I-alamina	85.89	60.88	88.89	83.5	magdubi	65.55	88.88	77.78	75
Maliki	55.55	66.67	77.78	60.55	Alayhim	75	85.55	88.89	75
Yawmi	82	60.88	87.5	83.5	Wala	55.55	85.89	88.89	83.5
I-dini	66.66	88.88	87.5	66.66	mustaqima	55.55	75	87.5	65.55
Iyyaka	66.66	88.88	88.89	83.5	Anamta	55.55	66.66	100	65
Nabadu	85.89	60.88	100	83.5	I-dalina	55.55	60.88	88.88	55.5
Wa-iyaka	75	66.66	87.5	55.55	-	-	-	-	-

4.4. Discussion

In this paper, we benchmark our proposed feature-based model with a traditional handcrafted feature-based model and transfer learning-based model. We used MFCC features in the conventional handcrafted feature-based model. We extracted MFCC features from the dataset and forwarded these features as an input to KNN, SVM, and RF classifiers for mispronunciation detection. The proposed CNN feature-based model with feature selection technique obtained the highest accuracy of 93.20% and NPV 0.9 because it automatically extracts enhanced features for the classification. The transfer learning-based model, which used the previously learned knowledge for the classification task to detect the mispronunciation, achieves the second-highest accuracy of 85% and NPV 0.7. In contrast, the traditional handcrafted feature-based model achieves an accuracy of 72.87%, NPV 0.6. The experimental results are shown in Figure 4.

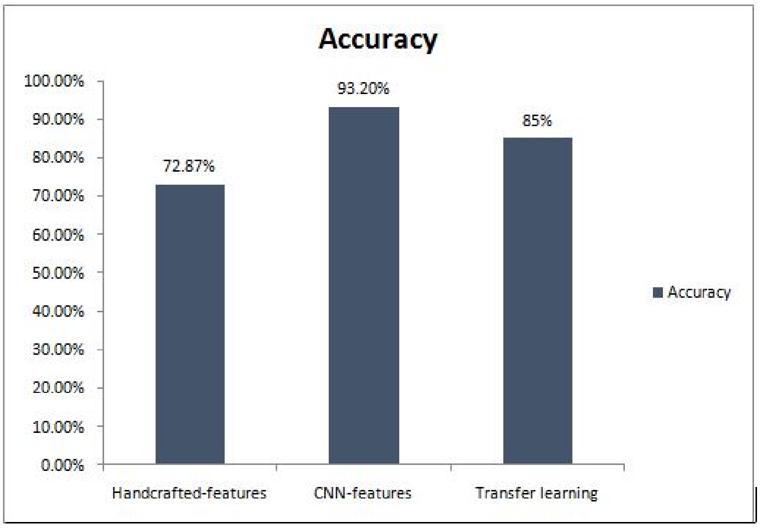

Figure 4. The comparison of the Handcrafted model, deep CNN model and transfer learning model to detect mispronunciation.

4.5. Comparison of the Proposed Framework with Previous Research

To determine the efficiency of our proposed method, we compared the results of our proposed approach with existing approaches, as shown in Table 10. Our proposed CNN-feature based model performs better than the existing approaches those are developed to detect the mispronunciation of Arabic words.

Table 10. The comparison of proposed CNN features with previous approaches.

Sr.no	Paper	Datasets	Techniques	Accuracy
1	Abdou et al. [14]	Database of Holy Qur'an recitation	Confidence scoring	62
2	Khaled Necibi et al. [15]	Manually constructed dataset consists only three Arabic words	Global average log likelihood	76.66
3	Muazzam et al. [16]	Manually constructed Arabic phoneme	ANN	82.7
4	Abual soud Hanani et al. [17]	Manually constructed dataset consists only three Arabic words	MFCC with GMM-UBM	75
5	Wahyuni et al. [30]	Three letters sa, sya and tsa	MFCC + ANN	92.42
6	Proposed method	Manually constructed 27 Arabic words	Convolution Neural Network (CNN)	93.20

5. Conclusions and Future Work

In this paper, we proposed a heterogeneous framework for mispronunciation identification of Arabic words. We compared the results of our proposed CNN feature-based model with traditionally handcrafted feature-based and transferred learning-based models. In the traditional handcrafted feature-based model, we used MFCC features and classified the data by using RF, SVM, and KNN classifiers. The results demonstrate that the SVM classifier achieves best results with an average accuracy of 73.87 while KNN and RF classifiers attained an accuracy of 68.43 and 72.53, respectively. We also performed the experiments by using a transfer learning-based model. The transfer learning-based model achieves an average accuracy of 85 on all words of the Arabic dataset. This method achieves better accuracy than the traditional handcrafted feature-based model because previously learned knowledge has been transferred in the new task. In our proposed framework, we extracted features from different layers of CCN, and a comparison was made on the performance of the selected classifiers, i.e., KNN, SVM, and RF. In the first step, we extracted the features from layers 6–8, and then applied the correlation-based feature selection technique on each of the extracted features. The experimental results have shown that SVM achieves the best accuracy, i.e., 93.20 on layer 6, while RF classifier achieves an accuracy of 89.99 and 88.51 on layers 6 and 7, respectively.

Our proposed approach works well with a clean dataset, i.e., without background noises, and provides satisfactory performance. However, as future research work, exploring new substitutes of the proposed solution can be beneficial in improving the overall accuracy and negative predictive value. A useful direction for future work can be to further improve the mispronunciation detection results by incorporating more speakers, and additionally, more trials per speaker too. Moreover, a gender-specific model can be trained to enhance the results further. We did not test the proposed algorithm in a noisy environment. In the future, we would like to embed the noise in our dataset to evaluate the accuracy of our proposed model on a noisy dataset. Furthermore, we also intend to create a bigger dataset with more number of Arabic words, which will be publically available for other researchers as well.

Author Contributions: Conceptualization, S.A.; data curation, Y.B.Z.; formal analysis, S.A.; Investigation, S.A., F.H., M.E.-u.h., N.K.B. and Y.B.Z.; methodology, S.A., F.H., F.R.R., M.E.-u.h., N.K.B. and Y.B.Z.; project administration, F.I. and Y.B.Z.; resources, F.H., M.E.-u.h. and N.K.B.; software, S.A., F.R.R. and N.K.B.; supervision, F.H., F.I. and Y.B.Z.; validation, F.R.R., F.I. and Y.B.Z.; visualization, S.A., F.R.R., M.E.-u.h. and N.K.B.; writing—original draft, S.A., F.H. and Y.B.Z.; writing—review and editing, F.I. and Y.B.Z. All authors have read and agree to the published version of the manuscript.

Funding: No funding was utilized in conduction of this reswarch work.

Acknowledgments: This research was supported in part by Yeungnam University, Korea and in part by Kwangwoon University, Korea and in part by University of Engineering & Technology, Taxila, Pakistan.

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

References

1. Cucchiarini, C.; Wet, F.D.; Strik, H.; Boves, L. Assessment of Dutch pronunciation by means of automatic speech recognition technology. In Proceedings of the ICSLP '98, Sydney, Australia, 30 November–4 December 1998; pp. 751–754.
2. Minematsu, N. Pronunciation assessment based upon the compatibility between a learner's pronunciation structure and the target language's lexical structure. In Proceedings of the 8th International Conference on Spoken Language Processing, Jeju Island, Korea, 4–8 October 2004.
3. Neumeyer, L.; Franco, H.; Digalakis, V.; Weintraub, M. Automatic scoring of pronunciation quality. *Speech Commun.* **2000**, *30*, 83–93. [CrossRef]
4. Witt, S.M. Automatic error detection in pronunciation training: Where we are and where we need to go. In Proceedings of the International Symposium on Automatic Detection on Errors in Pronunciation Training, Stockholm, Sweden, 6–8 June 2012; Volume 1.
5. Flege, J.E.; Fletcher, K.L. Talker and listener effects on degree of perceived foreign accent. *J. Acoust. Soc. Am.* **1992**, *91*, 370–389. [CrossRef] [PubMed]
6. Strik, H.; Truong, K.; De Wet, F.; Cucchiarini, C. Comparing different approaches for automatic pronunciation error detection. *Speech Commun.* **2009**, *51*, 845–852. [CrossRef]
7. Zhang, F.; Huang, C.; Soong, F.K.; Chu, M.; Wang, R. Automatic mispronunciation detection for Mandarin. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 5077–5080.
8. Georgoulas, G.; Georgopoulos, V.C.; Stylios, C.D. Speech sound classification and detection of articulation disorders with support vector machines and wavelets. In Proceedings of the 2006 International Conference on the IEEE Engineering in Medicine and Biology Society, New York, NY, USA, 30 August–3 September 2006; pp. 2199–2202.
9. Wei, S.; Hu, G.; Hu, Y.; Wang, R.H. A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Commun.* **2009**, *51*, 896–905. [CrossRef]
10. Hu, W.; Qian, Y.; Soong, F.K. An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech. In Proceedings of the SLaTE, Leipzig, Germany, 4–5 September 2015; pp. 71–76.
11. Witt, S.M.; Young, S.J. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Commun.* **2000**, *30*, 95–108. [CrossRef]
12. Gu, L.; Harris, J.G. SLAP: A system for the detection and correction of pronunciation for second language acquisition. In Proceedings of the 2003 International Symposium on Circuits and Systems (ISCAS'03), Bangkok, Thailand, 25–28 May 2003; Volume 2, pp. 583–580.
13. Dahan, H.; Hussin, A.; Razak, Z.; Odelha, M. Automatic arabic pronunciation scoring for language instruction. In Proceedings of the EDULEARN, Barcelona, Spain, 4–6 July 2011,; p. 150.
14. Abdou, S.M.; Hamid, S.E.; Rashwan, M.; Samir, A.; Abdel-Hamid, O.; Shahin, M.; Nazih, W. Computer aided pronunciation learning system using speech recognition techniques. In Proceedings of the 9th International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 September 2006.
15. Necibi, K.; Bahi, H. An arabic mispronunciation detection system by means of automatic speech recognition technology. In Proceedings of the 13th International Arab Conference on Information Technoloy, Zarqa, Jordan, 10–13 December 2012; pp. 303–308.
16. Maqsood, M.; Habib, H.A.; Nawaz, T. An efficientmis pronunciation detection system using discriminative acoustic phonetic features for arabic consonants. *Int. Arab J. Inf. Technol.* **2019**, *16*, 242–250.
17. Hanani, A.; Attari, M.; Farakhna, A.; Hussein, M.; Joma'a, A.; Taylor, S. Automatic identification of articulation disorders for arabic children speakers. In Proceedings of the Workshop on Child Computer Interaction, San Francisco, CA, USA, 6–7 September 2016.
18. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 26th Annual Conference on the Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
19. Hall, M.A. Correlation-Based Feature Subset Selection for Machine Learning. Ph.D. Thesis, University of Waikato, Hamilton, New Zealand, 1998. Unpublished work.

20. Dash, M.; Liu, H. Feature selection for classification. *Intell. Data Anal.* **1997**, *1*, 131–156. [CrossRef]
21. Glick, W.H. Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Acad. Manag. Rev.* **1985**, *10*, 601–616. [CrossRef]
22. Cai, Y.L.; Ji, D.; Cai, D. A KNN Research Paper Classification Method Based on Shared Nearest Neighbor. In Proceedings of the 8th NTCIR Workshop, Tokyo, Japan, 15–18 June 2010; pp. 336–340.
23. Danielsson, P.E. Euclidean distance mapping. *Comput. Graph. Image Process.* **1980**, *14*, 227–248. [CrossRef]
24. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
25. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media, Germany, 2013.
26. Hsu, C.W.; Chang, C.C.; Lin, C.J. A practical Guide to Support Vector Classification. 2003. Available online: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (accessed on 10 December 2019).
27. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
28. Resende, P.A.A.; Drummond, A.C. A survey of random forest based methods for intrusion detection systems. *Acm Comput. Surv. (CSUR)* **2018**, *51*, 1–36. [CrossRef]
29. Zeng, Z.Q.; Yu, H.B.; Xu, H.R.; Xie, Y.Q.; Gao, J. Fast training support vector machines using parallel sequential minimal optimization. In Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering, Xiamen, China, 17–19 November 2008; Volume 1, pp. 997–1001.
30. Wahyuni, E.S. Arabic speech recognition using MFCC feature extraction and ANN classification. In Proceedings of the 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 1–3 November 2017; pp. 22–25.