applied sciences

MDPI

Article

A Novel RBFNN-CNN Model for Speaker Identification in Stressful Talking Environments

Ali Bou Nassif [1,*], Noha Alnazzawi [2], Ismail Shahin [3], Said A. Salloum [4], Noor Hindawi [3], Mohammed Lataifeh [5] and Ashraf Elnagar [5]

[1] Computer Engineering Department, University of Sharjah, Sharjah P.O. Box 27272, United Arab Emirates
[2] Computer Science and Engineering Department, Yanbu University College, Royal Commission for Jubail and Yanbu, Yanbu Industrial City, Yanbu 46435, Saudi Arabia; alnazzawin@rcyci.edu.sa
[3] Electrical Engineering Department, University of Sharjah, Sharjah P.O. Box 27272, United Arab Emirates; ismail@sharjah.ac.ae (I.S.); u21103472@sharjah.ac.ae (N.H.)
[4] School of Science, Engineering, and Environment, University of Salford, Salford M5 4WT, UK; s.a.s.salloum@edu.salford.ac.uk
[5] Computer Science Department, University of Sharjah, Sharjah P.O. Box 27272, United Arab Emirates; mlataifeh@sharjah.ac.ae (M.L.); ashraf@sharjah.ac.ae (A.E.)
* Correspondence: anassif@sharjah.ac.ae

Abstract: Speaker identification systems perform almost ideally in neutral talking environments. However, these systems perform poorly in stressful talking environments. In this paper, we present an effective approach for enhancing the performance of speaker identification in stressful talking environments based on a novel radial basis function neural network-convolutional neural network (RBFNN-CNN) model. In this research, we applied our approach to two distinct speech databases: a local Arabic Emirati-accent dataset and a global English Speech Under Simulated and Actual Stress (SUSAS) corpus. To the best of our knowledge, this is the first work that addresses the use of an RBFNN-CNN model in speaker identification under stressful talking environments. Our speech identification models select the finest speech signal representation through the use of Mel-frequency cepstral coefficients (MFCCs) as a feature extraction method. A comparison among traditional classifiers such as support vector machine (SVM), multilayer perceptron (MLP), k-nearest neighbors algorithm (KNN) and deep learning models, such as convolutional neural network (CNN) and recurrent neural network (RNN), was conducted. The results of our experiments show that speaker identification performance in stressful environments based on the RBFNN-CNN model is higher than that with the classical and deep machine learning models.

Keywords: Mel-frequency cepstral coefficients; shallow and deep learning models; speaker identification; stressful talking environments

Citation: Nassif, A.B.; Alnazzawi, N.; Shahin, I.; Salloum, S.A.; Hindawi, N.; Lataifeh, M.; Elnagar, A. A Novel RBFNN-CNN Model for Speaker Identification in Stressful Talking Environments. *Appl. Sci.* **2022**, *12*, 4841. https://doi.org/10.3390/app12104841

Academic Editors: Teen-Hang Meen and Chun-Yen Chang

Received: 15 March 2022
Accepted: 9 May 2022
Published: 11 May 2022

1. Introduction

Human beings are able to recognize known voices in a very short time [1]. The voice of an individual is unique, which is due mainly to their developed and physical features. There are physical variations amongst human beings, due to the unique voice-producing organs and the sizes and shapes of their articulators. The larynx (most importantly the vocal folds) is voice-producing, and the rest, including the nose, are mostly responsible for resonances, i.e., articulation [2]. In addition to anatomical features, speech rate, vocabulary, accent, and other personal behaviors taking shape over time are also responsible for the different speech patterns found in individuals These features are exploited by state-of-the-art speaker recognition systems in a way that makes it possible to attain significant recognition accuracy [3]. There are two broad categories of speaker recognition (SR): speaker identification (SI), and speaker verification (SV) [4]. The process of speaker identification involves automatic detection of the speaking person from a set of specific speakers. Conversely,

speaker verification involves automatic detection of whether the speaker is a specific person or not. Speaker identification can be used for various purposes. For instance, applications such as credit card security, confidential data safety, client identity and verification are some examples of broad speaker recognition applications. The functioning of speaker identification is either text-dependent [5–7] or text-independent [8–12].

Many circumstances can cause stress, including noisy backgrounds, emergencies such as aircraft pilot communications, high workloads, physical environmental factors, and multitasking [13]. There are several applications of speech recognition under stressful conditions in real life. This includes emergency call centers, telephone banking and military voice communications applications.

An emotional talking environment occurs when the speaker speaks his or her language under the influence of emotional conditions such as anger, happiness, and sadness. Emotion recognition applications are found in telecommunications, human–robot interfaces, and intelligent call centers. Emotion recognition can be used in intelligent language education systems to sense and adapt to student emotions when they reach a state of boredom during a tutoring session [14,15].

Speaker identification in emotional and stressful conditions is challenging because of the mismatch between training and testing. In the training stage, the models are trained using neutral conditions; however, in the testing phase, models are tested using neutral, as well as emotional and stressful, conditions. This why models perform better when they are tested under neutral conditions.

Speaker identification also has various applications in civil cases and the media. This includes applications in recorded discussions and phone calls made to radio stations or other state and local departments and insurance companies [16]. It is important that the speaker be identified in such cases, since some people tend to disguise their voices to imitate famous people in the media, which may affect their reputation.

The main aim of this research was to further enhance speaker identification performance in stressful talking environments using a novel RBFNN-CNN model. In this research, we evaluated our approach with two distinct speech databases: a local Arabic Emirati-accent dataset and a global English Speech Under Simulated and Actual Stress (SUSAS) corpus. Our speech identification models selected the finest speech signal representation by feature extraction such as Mel-frequency cepstral coefficients (MFCCs), after having been processed. We evaluated our proposed model against different classical classifiers such as support vector machine (SVM), multilayer perceptron (MLP), k-nearest neighbors algorithm (KNN), and different deep learning models such as convolutional neural network (CNN) and recurrent neural network (RNN). Results show that the proposed model outperforms all other models.

The remaining sections are organized as follows: Section 2 presents related work. Section 3 describes the datasets used in this research, while the model architecture and methodology are explained in Section 4. The results of the models are presented in Section 5. Section 6 concludes the paper.

2. Related Work

The performance of speaker identification is exceptionally high in a neutral talking environment compared to other talking environments such as emotional and stressful [16–18]. Conversely, its performance is lower in the case of a stressful talking environment [19–21]. A neutral talking environment is the kind of talking environment whereby the speaker utters the speech without any stressful or emotional talking condition. On the other hand, a stressful talking environment is different from neutral talking conditions in the sense that the speakers deliver their speech under stressful talking conditions, such as shouting or speaking loudly, and quickly.

The authors in [19] investigated “talker-stress-induced intra-word variability”, and the algorithm meant to counter these systematic changes, based on “hidden Markov models (HMMs)” as classifiers that had been trained with the help of speech indications

under different types of talking conditions. By using the hypothesis-driven compensation technique, the error rate was reduced from 13.9% to 6.2%.

Raja and Dandapat [20] focused on studying speaker recognition in stressed conditions with the aim of improving the decline in performance usually observed in these conditions. They made use of four types of stressed conditions of the SUSAS database [22,23], including neutral, angry, Lombard, and question conditions. The study revealed that speech uttered by speakers under angry conditions exhibited the lowest speaker identification performance [20]. The average speaker identification rate using the SUSAS database (stressed condition) was about 57%.

Zhang and Hansen [24] examined five different vocal modes, including whispered, soft, neutral, loud, and shouted, with the aim of studying various aspects of speech. They intended to identify distinguishing features of speech modes. The average accuracy rate was about 96%. Chatzis [25] tried to learn through the use of data with sequential dynamics and put forward infinite-order HMM models that were based on the assumption that first-order Markovian dependencies existed among the successive label values denoted by y. The models that were designed were better than other techniques, for a couple of reasons. First, extended and complex temporal dependencies can be captured by these models. Second, margin maximization paradigms are employed for performing model training in these models, ultimately leading to a convex optimization design [25]. The highest average accuracy obtained was 71.68% for the iMMS model.

In another work, Prasetio et al. (2020) [26] proposed a method using the deep time-delay Markov network (DTMN) to predict emotions by studying earlier emotional states. The novel approach has been evaluated on the English SUSAS database. They concluded in that study that the proposed DTMN outperformed the baseline systems. TDNN-4 was the optimal temporal context for predicting the emotional state of 8.31% PER.

There are very few research studies that focus on the spoken Arabic language as speech [27,28]. This is mainly because Arabic speech databases are quite limited with reference to the speaker recognition area. There are four main regional dialects of the Arabic language. They include "Egyptian, Levantine (e.g., Jordan), North African (e.g., Tunisian), and Gulf Arabic (e.g., Emirati)" [29].

There is currently no published work dealing with speaker identification in stressful talking environments using the RBFNN-CNN model. In this paper, a significant contribution has been made for enhancing speaker identification within stressful talking environments. For this purpose, two speech databases were applied for model testing of speaker identification within stressful talking environments. "Speech Under Simulated and Actual Stress (SUSAS)" is the first database that has been recorded using stressful and neutral talking [22]. The "Emirati speech database" [30] is the second database, in which 30 Emirati speakers (15 males and 15 females) were used as respondents and subjected to neutral, shouted, slow, loud, soft and fast talking conditions.

Results based on different classifiers and compensators as reported in our previous results, Refs. [31–33] reported speaker identification performance under shouted/stressful talking conditions using the Emirati accent dataset. Their reported speaker identification performance in shouted/stressful talking conditions was 58.6%, 61.1%, 65.0%, 68%, 74.6%, 75%, 78.4%, 81.7%, 78.7%, 83.4%, and 85.8% based, respectively, on "First-Order Hidden Markov Models (HMM1s), Second-Order Hidden Markov Models (HMM2s), Third-Order Hidden Markov Models (HMM3s), Second-Order Circular Hidden Markov Models (CHMM2s), First-Order Left-to-Right Suprasegmental Hidden Markov Models (LTRSPHMM1s), Suprasegmental Hidden Markov Models (SPHMMs), Second-Order Left-to-Right Suprasegmental Hidden Markov Models (LTRSPHMM2s), Third-Order Left-to-Right Suprasegmental Hidden Markov Models (LTRSPHMM3s), First-Order Circular Suprasegmental Hidden Markov Models (CSPHMM1s), Second-Order Circular Suprasegmental Hidden Markov Models (CSPHMM2s), and third-order circular suprasegmental hidden Markov models (CSPHMM3s)". Table 1 shows more comparison with previous studies.

Table 1. Comparison with previous studies.

Prior Work	Classifier	Accuracy
[31–33]	CSPHMM2s	85.8%
[20]	Speaker and Stress Information based on Compensation (SSIC)	56.74%
[24]	Gaussian Mixture Model (GMM)	97.62%
[26]	Deep Time-delay Markov Network (DTMN)	8.55 PER *

* PER refers to prediction error rate.

The main contributions of this study are as follows:

- To the best of our knowledge, this is the first work that uses and evaluates an RBFNN-CNN model for speaker identification under stressful/emotional conditions.
- We conduct extensive comparisons between traditional classifiers and deep learning models to predict the model that yields the highest performance.
- We showed that the proposed RBFNN-CNN model outperforms other models.

3. Speech Databases and the Extraction of Features

3.1. Captured Emirati-Accent Speech Corpus

This task involved the communication of the Emirati-emphasized speech database (Arabic database) by 30 local Emirati speakers from each gender, between the ages of 14 and 55 years. Eight Emirati expressions, which are widely spoken by the UAE public, were uttered by speakers. Eight sentences that would be spoken in six stressful talking conditions were used. These talking conditions were "neutral, angry, slow, loud, soft, and fast". The tone of each talking condition was expressed nine times, inserting 2 to 5 s gaps between utterances. The speakers were asked to speak these sentences on the spot, not giving them the chance to practice the sentences, in order to prevent fake outcomes. The overall number of recorded utterances was 7560 ((15 speakers × 2 genders × first 4 sentences × 9 duplicate sentences in the neutral environment for the training session) + (15 speakers × 2 genders × last 4 sentences × 9 replications/sentence × 6 talking conditions for the testing session)). For more details, please check Section 4.1 [30].

3.2. Speech under Simulated and Actual Stress (SUSAS) Database

A wide variety of emotions and stresses are part of the five domains present within the SUSAS database. Actual domain (actual speech when stressed) and simulated domain (simulated speech when stressed) are part of the database. There were close to 16,000 utterances by 32 speakers (from 22–76 years of age), of which 19 were men and 13 were women [22]. To conduct fair comparisons with other studies, the Lombard condition was removed. This research took into consideration 13,890 records uttered by eight speakers (training considered five, and testing was conducted on the others) two times (each word repeated twice), who talked in 6 stressful talking conditions, which included fast, soft, low, slow, angry, and neutral.

3.3. Extraction of Features

Our speech identification models selected the finest speech signal representation by feature extraction, such as Mel-frequency cepstral coefficients (MFCCs), after having been processed. The most extensively used feature of the speech was the Mel frequency scale, which involved easy calculation, suitable potential for the distinction, anti-noise, as well as other benefits [34,35]. Sound processing also made use of MFCC as a feature extraction technique. It has been found that MFCCs function better than other coefficients in the two areas and award a high-level approximation of auditory perception of individuals [36,37].

We used the Python library (python_speech_features), and both (mfcc) and (logfbank) were imported. The used parameters were: n_mfcc = 96, n_fft = 1024, win_length = int(0.025 ∗ rate), hop_length = int(0.01 ∗ rate).

The used filter bank was the Mel filter bank that was introduced by the library.

4. Model Architecture

Generally, the speaker identification process may be categorized into two crucial parts: feature extraction and classification. Figure 1 shows the speaker identification process. Speakers may be distinguished on the basis of some exceptional characteristics that are typical to a particular speaker by using the extracted features. Feature extraction is an important part of speaker identification; hence, features call for additional fine-tuning and use of appropriate approaches to come up with the ultimate identity of a speaker. Individual speaker models are formulated for each and every speaker by making use of these feature sets. All of the developed speaker models are stored. The features of the speech uttered by the unidentified speaker are extracted and compared with the developed speaker models using a speaker identification classifier, which checks if the features of the unknown speaker match with those in the developed speaker models. This leads to the identification of the unknown speaker. The speaker identification process explained in the current study uses a machine learning approach, whereby speakers are identified by considering the features extracted from the speaker's recorded speech. The features out of Mel-frequency cepstrum coefficients (MFCC) were used for classifier training. A 96-dimension feature of MFCCs was used to determine the observation vectors in all deep learning and classical techniques. A "continuous mixture observation density" was selected in each classifier with six states. MFCC determines the change in the straight cosine in the range of log control related to the direct Mel recurrence size. The human voice can be represented with high precision in Mel recurrence, due to uniform or similar dispersion of recurrence groups therein. The models are fed by an array of MFCCs of each time frame. Closed-set speaker identification is presented as follows: The words spoken by the given speaker are recorded, and the recording is compared with the developed speaker models (a finite set). The developed model that shows the most resemblance with the recording is considered. The recorded speech signals of the given speaker are used for extraction of MFCC. Overall, the MFCC technique will generate features from the inserted audio signal samples that are used as input for the speech recognition model. This is followed by classifier training with these features. The feature extraction process is also followed for the new speech signals we want to classify. The speaker with the closest resemblance is predicted by the trained classifiers. Figure 1 depicts the speaker identification approach employed in the current study.

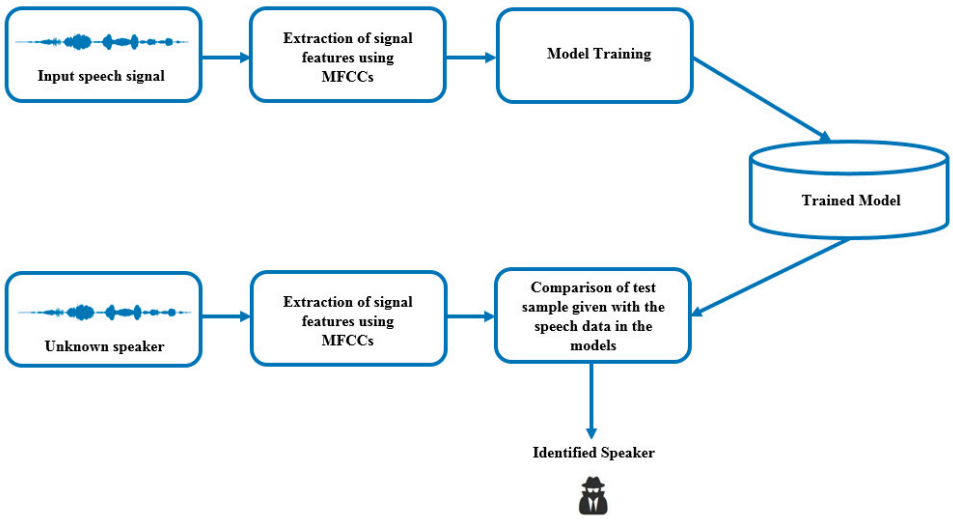

Figure 1. Speaker identification process.

4.1. RBFNN-CNN Model

RBFNN is a feedforward neural network model that is relatively fast in comparison with other machine learning models such as multilayer perceptron [38]. Figure 2 shows how the RBFNN layers are comprised [39], where the input vector is represented by the first layer. The second layer, which is also called the hidden layer, is where the RBFs of all input data are stored. For example, the node RBF1 is the vector with the length of n where the RBF of X ($[x_1, x_2, \ldots, x_n]$) and C1 (first centroid vector) is described. The RBF1 vector is a measure of how the distance between the first centroid and data X is related to other vectors. Eventually, through utilizing the theory described above, the resulting prediction of the unknown point's class can be made by calculating the RBF of an unidentified data point x in terms of all centroids. Additionally, we calculate the dot product of RBF and W (the weight) and choose the index with the highest value. RBFNN models use a Gaussian function as an activation function in the hidden layer. In this article, the implementation of SUSAS and Arabic Emirati-accent dataset classifications are described.

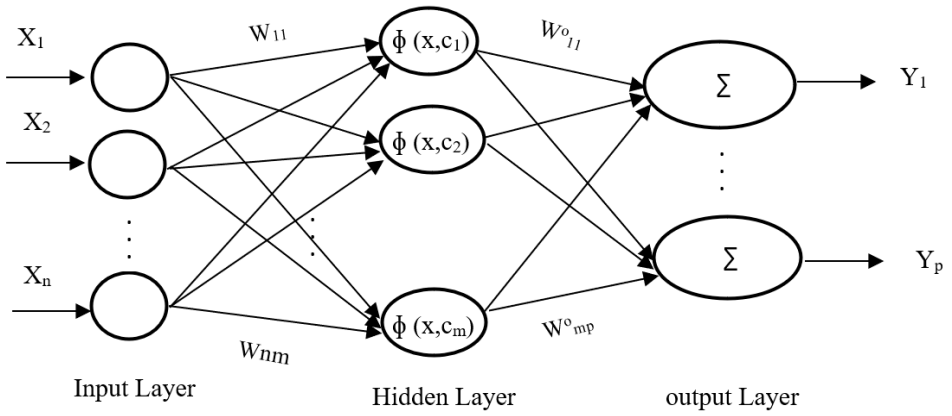

Figure 2. RBF model.

The RBFNN is followed by CNN, in which the final model is the result of cascading RBFNN with CNN (RBFNN-CNN). In the CNN model design, five hidden layers have been used due to the significant outcome in accuracy. The number of input neurons is 957, which represent the features, while the number of output neurons is 50, which represent the classes after categorizing the testing labels, knowing that the output layer is known since the dataset is labeled. Figure 3 illustrates the block diagram of the CNN model that is cascaded with RBFNN, where the output of RBFNN is the input for the CNN model. The block diagram of the RBFNN-CNN model is shown in Figure 4. The activation function that has been applied in this work is "Softmax". The standard (unit) Softmax function $\sigma : R^K \rightarrow R^K$ {\displaystyle \sigma:\mathbb {R} ^{K}\to \mathbb {R} ^{K}} is characterized by the following equation [40].

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \; for\; i = 1, \ldots, K \; and \; z = (z_1, \ldots, z_k) \in R^K \tag{1}$$

The exponential function is applied to all input vector features and then normalizes the values. At first, audio signals are inserted into the model with zero manipulations. The inserted audios are the input for the next step, named "preprocessing step", that is applied to each audio. In addition, this step involves the labeling process for each audio file. The labeling process is where the naming of each sound file is accomplished, e.g., the first sound corresponds to speaker X, etc.

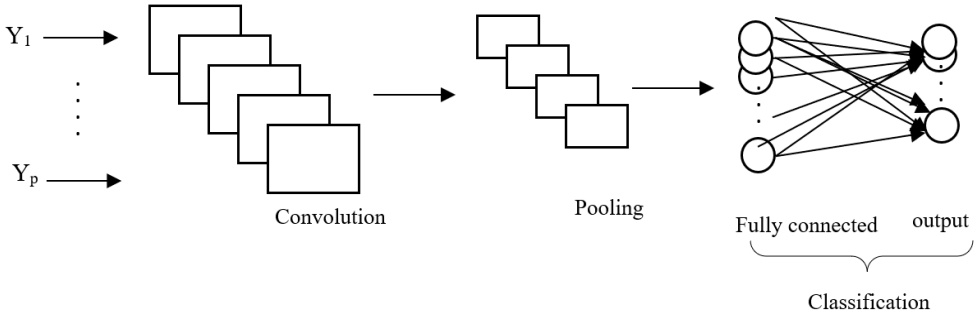

Figure 3. CNN model.

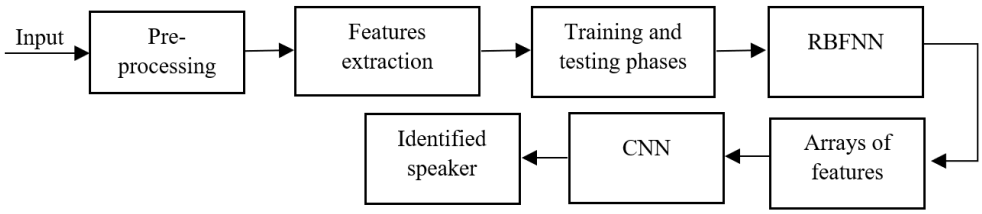

Figure 4. The block diagram of the proposed cascaded model (RBFNN-CNN) under stressful talking conditions.

The output of the preprocessing step includes all of the files along with their labels. This output is the input to the feature extraction process, where each audio file contains special characteristics recognized by the presented MFCC feature extraction method. The output of the feature extraction method entails the essential features as a matrix having feature vectors for each audio. The output encounters all of the matrices that should be used in the classification. Before applying the classification method, we first split the data into training and testing datasets. The next step was to start classifying our data using the proposed model, RBFNN-CNN. RBFNN addresses the theoretical gap in the original RBF and solves the ambiguity regarding the class of the data. Inside the RBFNN, in the hidden layers, each node represents a vector that is stored with its centroid vector. Each node vector represents the distance between first centroid and data X, which are associated with one another. The output of RBFNN is an array of features that contain weights that best estimate the linear relationship among RBFs and the output. This is considered as the input to CNN, where the "Softmax" activation layer was used to normalize it into a probability distribution containing K probabilities, ending up with the identified speaker.

RBFNN is an unconventional machine learning method that is also incredibly fast, effective, and straightforward. RBFNN models the data using smooth transitioning circular shapes rather than harsh cut-off circles, creating a pattern that best approximates the position of the clusters (features of speaker x). It also provides information on the prediction confidence rate, which is the core use of RBFNN as a deep feature match/robust for stressed conditions. The color intensity gradually lowers as you move away from the cluster centroids. An exponential function with a negative power of distance can be utilized to achieve such a smooth transition. We can regulate how fast the function decays by multiplying the distance with a scalar coefficient "Beta." As a result, a larger Beta indicates a more rapid drop. RBFNN is, in general, one of the most powerful models for classification and regression applications. Using a large number of RBF curves, RBF nets can learn to approximate the underlying patterns. In comparison to MLP structured networks, the use of a statistical equation for the optimization process makes the method more conducive and faster.

4.2. Traditional Classifier Experiments

In this section, we discuss the traditional machine learning algorithms that were evaluated for speaker identification using the Emirati speech database and the global

English SUSAS database in stressful talking environments. In this study, we used the KNN, SVM and MLP classifiers. Within each model, there is one classifier, where 80% of utterance for the individual speaker has been used for training. Testing and evaluation include 20% of utterances.

4.2.1. Support Vector Machines (SVM)

A multi-class support vector machine (SVM) classifier was trained by applying the linear function. As the SVM is considered to be simple as well as competent for machine learning algorithm computation, it was applied for pattern recognition and classification issues. Since the training data were limited, the classification performance was quite efficient when compared to other classifiers. Hence, in the current research, the speech data were classified using the support vector machine, and the following parameters were used.

- SVC: C = 1.0
- Kernel = linear
- decision_function_shape = ovr

4.2.2. Multilayer Perceptron (MLP)

With the help of the MLP, it was possible to minimize the system's expected and real output differences. For the current model, the MLP topology design included the following parameters:

- activation = 'tanh'
- hidden_layer_sizes = 100
- activation = 'relu', solver = 'adam'
- validation_fraction = 0.1
- learning_rate_init = 0.001
- max_iter = 200

The hyperbolic tangent activation function was applied for all hidden layer neurons. For the output layer neurons, the linear ones were applied.

4.2.3. The K-Nearest Neighbors Algorithm (KNN)

KNN is the simplest classification algorithm, as it assumes that instances that are close to the instance space will have similar class values. Currently, the KNN classifier is usually adopted by researchers, since it is simple, refined, and direct. If new sample data x arrive, KNN will search for the k neighbors nearest to the unlabeled data initiating from the training space with reference to some distance measure. In the current research, the following parameters were applied:

- KNN: n_neighbors = 25
- weights = 'uniform'
- algorithm = 'auto'
- max_iter = 200

4.3. Deep Learning Experiments

In this research, we implemented CNN and RNN models. There are many researchers [41,42] who have experimented using a combination of RNN and CNN. They have clearly indicated that as compared to the machine learning models, this mixture is much more effective.

4.3.1. Convolutional Neural Network (CNN)

In this paper, the CNN model was developed through the integration of five convolutional layers, each followed by a max-pooling layer and a dropout layer. The input of the neural network was a vector of 96 MFCC features. Each of these five CNN layers had 64 filters, with a kernel size of 3×3, that were applied at a stride setting of three pixels. We used the RELU activation function instead of the typical sigmoid functions, which

improved the efficiency of the training process. The max-pooling layer generates a lower resolution version of the convolution layer; we applied a pool size of 2 × 2. Later, we added a dropout layer, which helped in avoiding overfitting; we set the dropout ratio at 20%. Finally, we added a fully connected layer and a dense layer that had 30 neurons; each represented one speaker in the dataset.

4.3.2. Recurrent Neural Networks (LSTM)

Due to the effectiveness of deep learning techniques, we also adopted an RNN model consisting of three LSTM layers. The input of the neural network was a vector of 96 MFCC features. Each of these three LSTM layers had 64 filters. We added a fully connected layer and a dense layer of 16 nodes that had an activation function, Softmax.

4.3.3. Bidirectional LSTM (BILSTM)

There are three LSTM layers present within the proposed BILSTM model. The neural network input contained a vector of 96 MFCC features. There were 64 filters present within each of the three LSTM layers. A fully connected layer and dense layer were included, which attained an activation function Softmax with 16 nodes.

4.3.4. CNN-BiLSTM

There are four convolutional layers present within the proposed CNN-BILSTM model. The max-pooling layer and a dropout layer follow each of these layers. The neural network input is a vector of 96 MFCC features. There are 64 filters present within the four CNN layers along with a kernel size of 3 × 3 that are included in a 3-pixel stride setting. For the convolution layer, a lower resolution version was generated by the max-pooling layer, and 2 × 2 pool size was applied. A dropout layer, of ratio 20%, was also included later, helping to avoid overfitting. Lastly, an entirely connected layer was included, along with a dense layer that maintained an activation function, Softmax.

4.3.5. Gated Recurrent Units (GRU)

The proposed GRU model consists of three GRU layers. Each of these three GRU layers has 32 filters. We added a fully connected layer and a dense layer that has activation function Softmax with 16 nodes. Finally, we added a dropout layer, which helps in avoiding overfitting; we set the dropout ratio at 20%.

4.3.6. BI-GRU

There are three GRU layers present within the proposed BI-GRU model. Each of these three GRU layers has 32 filters. We added two dropout layers, which help in avoiding overfitting; we set a dropout ratio of 20%. Finally, we added a fully connected layer and two dense layers that have activation functions Softmax and RELU, respectively, with 16 nodes.

4.3.7. Attention-BILSTM BI-GRU

The proposed Attention-BILSTM model consists of three BI-GRU layers. Each of these three GRU layers has 32 filters. We added a dropout layer, which helps in avoiding overfitting; we set a dropout ratio of 30%. Finally, we added a fully connected layer and a dense layer that has activation function Softmax with 16 nodes.

5. Results and Discussion

Experimental Results and Evaluation

The measurement of classifiers in terms of quality involves the use of accuracy, precision, recall, and F1-measure.

One of the most crucial performance measures is accuracy. It is defined as the ratio of observation predicted accurately to all of the observations predicted. It is usually believed that the model with the highest accuracy is best. While the significance of accuracy as an important measure cannot be denied, measuring performance also calls for the availability

of symmetric datasets having close or similar values for false positives and false negatives. Hence, a model's performance evaluation must involve additional parameters. The ratio of the predicted positive observations that are accurate to the overall predicted positive observations is referred to as the precision. High precision is associated with a low false positive rate.

Recall (sensitivity)—Recall refers to the ratio of predicted positive observations that are accurate to the overall observations with respect to actual class-yes. Any value exceeding 0.5 is considered as appropriate. F1 score—Computation of the weighted average of precision and recall yields the F1 score. This implies the inclusion of false positives as well as false negatives in the F1 score. This concept involves more complexity as compared to accuracy; however, F1 outshines accuracy when it comes to usefulness, particularly in the case of irregular class distribution. The higher the similarity of costs of false positives and false negatives, the higher the accuracy will be. We must consider precision as well as recall in the case of high diversity in the cost of false positives and false negatives.

This research employed a novel RBFNN-CNN that was evaluated against multiple machine learning algorithms using the Emirati speech database and global English SUSAS database in stressful talking environments. The stressful talking environments included "neutral, shouted, slow, loud, soft, and fast-talking conditions".

Tables 2 and 3 demonstrate average speaker identification performance in stressful talking conditions using the Emirati and SUSAS datasets, respectively. In Table 2, the standard deviation is added as one of the measurements. A large standard deviation indicates that the data are dispersed, which is unreliable. However, a low standard deviation indicates that the data are tightly grouped around the mean, which is more reliable. The conducted results show that the standard deviation outcomes were relatively small; thus, the data were reliable.

Table 2. Average speaker identification performance using Emirati accent dataset.

	Best Test Accuracy	Standard Deviation	Precision	Recall	F1
			Weighted	Weighted	Weighted
SVM	87%	3.01	97%	87%	0.87
MLP	69%	3.51	74%	69%	0.69
KNN	62%	3.60	63%	63%	0.62
CNN_BILSTM	88%	2.90	89%	88%	0.87
Att-BILSTM	79%	3.24	82%	79%	0.77
BI-LSTM	78%	3.36	81%	78%	0.78
BIGRU	73%	3.44	76%	73%	0.73
GRU	70%	3.49	72%	70%	0.69
LSTM	62%	3.57	64%	62%	0.62
Proposed RBFNN-CNN	**98%**	**1.80**	**98%**	**98%**	**0.98**

Table 3. Average speaker identification performance using SUSAS dataset.

	Best Test Accuracy	Precision	Recall	F1
		Weighted	Weighted	Weighted
SVM	90%	90%	90%	0.90
MLP	91%	92%	92%	0.92
KNN	70%	71%	70%	0.70
CNN_BILSTM	92%	93%	92%	0.92
Att-BILSTM	84%	84%	84%	0.84
BI-LSTM	80%	81%	80%	0.81
BIGRU	77%	78%	77%	0.77
GRU	73%	74%	73%	0.73
LSTM	63%	65%	63%	0.63
Proposed RBFNN-CNN	**98%**	**98%**	**98%**	**0.98**

Results in the above tables show that the proposed RBFNN-CNN outperformed all classical and deep learning models in both Emirati and SUSAS datasets. To prove that the proposed model is significantly better than other models, we conducted a statistical test between the proposed model and other models. We chose the non-parametric Wilcoxon test [43] because data were not normally distributed. The reported p-values were less than 0.05, which indicated that the proposed model was statistically significant. On the other hand, based on classical classifiers only, SVM surpassed other classifiers using the Emirati dataset; while MLP was the winning model with the SUSAS dataset. Furthermore, the CNN_BILSTM model proved to be the winning model in both datasets after the proposed RBFNN-CNN model. Based on these results, MLP had better performance as compared to the other classifiers using the SUSAS speech corpus. However, the results attained based on SVM were better than those achieved based on each of the MLP and KNN classifiers using the Emirati-accent dataset. Therefore, we can conclude that there is no rule that states the superiority of a specific classifier over another in any classification problem. Furthermore, the neutral talking environment was noted for the best accuracy followed by slow, soft, fast, loud, and shout, respectively. Based on Tables 1–3, it is evident that our proposed model outperformed the models in previous studies.

6. Conclusions

In this paper, we focused on improving speaker identification performance in stressful talking environments by introducing a novel RBFNN-CNN model. The proposed model was compared against shallow and deep learning models using a local Arabic Emirati-accent dataset and a global English Speech Under Simulated and Actual Stress (SUSAS) corpus. MFCCs were used as the extracted features of the database based on classical machine learning algorithms SVM, MLP, and KNN as classifiers, as well as deep learning techniques, such as RBFNN-CNN, CNN, and RNN in stressful talking conditions. Some conclusions can be presented. First, the proposed RBFNN-CNN model outperformed all other models based on both datasets. Second, the results attained based on SVM were better than those achieved based on each of the MLP and KNN classifiers using the Emirati-accent dataset. Third, the MLP had better performance as compared to the other classifiers using the SUSAS speech corpus. Fourth, the CNN_BILSTM classifier came second in terms of performance after the proposed model using both datasets. Furthermore, the proposed model surpassed the models used in studies in related work based on the same datasets.

The main limitation in our study is that we used two different datasets for model evaluation. In the future, we plan to extend the Emirati-accent speech dataset to include more speakers from different emirates and genders.

Author Contributions: A.B.N.: Conceptualization, Methodology, Supervision, Writing—Original Draft, Writing—Review and Editing. N.A.: Funding Acquisition, Methodology, Writing—Review and Editing. I.S.: Data Curation, Methodology, Writing—Review and Editing. S.A.S.: Investigation, Writing—Original Draft. N.H.: Methodology, Software. M.L.: Methodology, Writing—Review and Editing. A.E.: Methodology, Writing—Review and Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This project was funded by the "Machine learning and Arabic language processing research group" at the University of Sharjah.

Institutional Review Board Statement: The authors have authorization from the University of Sharjah to gather speech databases from UAE nationals based on the competitive research project titled Emirati-Accented Speaker and Emotion Recognition Based on Deep Neural Network, No. 19020403139.

Informed Consent Statement: This study did not involve any experiments on animals.

Data Availability Statement: Speech datasets are described in Section 3.

Acknowledgments: The authors would like to convey their thanks and appreciation to the University of Sharjah for supporting the work through the research group, Machine Learning and Arabic Language Processing. Additionally, the authors would like to thank engineers Ridhwan Al-Debsi and Hozayfa El-Rifai for assisting the research team.

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

References

1. Alsharhan, E.; Ramsay, A. Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition. *Lang Resour. Eval.* **2020**, *54*, 975–998. [CrossRef]
2. Zhang, Z. Mechanics of human voice production and control. *J. Acoust. Soc. Am.* **2016**, *140*, 2614–2635. [CrossRef] [PubMed]
3. Kinnunen, T.; Li, H. An overview of text-independent speaker recognition: From features to supervectors. *Speech Commun.* **2010**, *52*, 12–40. [CrossRef]
4. Nassif, A.B.; Shahin, I.; Attili, I.; Azzeh, M.; Shaalan, K. Speech recognition using deep neural networks: A systematic review. *IEEE Access* **2019**, *7*, 19143–19165. [CrossRef]
5. Variani, E.; Lei, X.; McDermott, E.; Moreno, I.L.; Gonzalez-Dominguez, J. Deep neural networks for small footprint text-dependent speaker verification. In Proceedings of the 2014 in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 4052–4056. [CrossRef]
6. Heigold, G.; Moreno, I.; Bengio, S.; Shazeer, N. End-to-end text-dependent speaker verification. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; Volume 2016.
7. Larcher, A.; Lee, K.A.; Ma, B.; Li, H. Text-dependent speaker verification: Classifiers, databases and RSR2015. *Speech Commun.* **2014**, *60*, 56–77. [CrossRef]
8. Auckenthaler, R.; Carey, M.; Lloyd-Thomas, H. Score normalization for text-independent speaker verification systems. *Digit. Signal Process.* **2000**, *10*, 42–54. [CrossRef]
9. Bimbot, F.; Bonastre, J.-F.; Fredouille, C.; Gravier, G.; Magrin-Chagnolleau, I.; Meignier, S.; Merlin, T.; Ortega-García, J.; Petrovska-Delacrétaz, D.; Reynolds, D.A. A tutorial on text-independent speaker verification. *EURASIP J. Adv. Signal Process.* **2004**, *2004*, 101962. [CrossRef]
10. Reynolds, D.A. Comparison of background normalization methods for text-independent speaker verification. In Proceedings of the Fifth European Conference on Speech Communication and Technology, Rhodes, Greece, 22–25 September 1997.
11. Snyder, D.; Garcia-Romero, D.; Povey, D.; Khudanpur, S. Deep neural network embeddings for text-independent speaker verification. *Interspeech* **2017**, 999–1003. Available online: https://www.isca-speech.org/archive/interspeech_2017/snyder17_interspeech.html (accessed on 14 March 2022). [CrossRef]
12. Nammous, M.K.; Saeed, K.; Kobojek, P. Using a small amount of text-independent speech data for a BiLSTM large-scale speaker identification approach. *J. King Saud Univ. Inf. Sci.* **2020**, *34*, 764–770. [CrossRef]
13. Bou-Ghazale, S.E.; Hansen, J.H.L. A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Trans. Speech Audio Process.* **2000**, *8*, 429–442. [CrossRef]
14. Petrushin, V.A. Emotion recognition in speech signal: Experimental study, development, and application. In Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP 2000), Beijing, China, 16–20 October 2000.
15. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
16. Furui, S. Speaker-dependent-feature extraction, recognition and processing techniques. *Speech Commun.* **1991**, *10*, 505–520. [CrossRef]
17. Farrell, K.R.; Mammone, R.J.; Assaleh, K.T. Speaker recognition using neural networks and conventional classifiers. *IEEE Trans. Speech Audio Process.* **1994**, *2*, 194–205. [CrossRef]
18. Yu, K.; Mason, J.; Oglesby, J. Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation. *IEE Proc. Vis. Image Signal Process.* **1995**, *142*, 313–318. [CrossRef]
19. Chen, Y. Cepstral domain talker stress compensation for robust speech recognition. *IEEE Trans. Acoust.* **1988**, *36*, 433–439. [CrossRef]
20. Raja, G.S.; Dandapat, S. Speaker recognition under stressed condition. *Int. J. Speech Technol.* **2010**, *13*, 141–161.
21. Nassif, A.B.; Shahin, I.; Hamsa, S.; Nemmour, N.; Hirose, K. CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions. *Appl. Soft Comput.* **2021**, *103*, 107141. [CrossRef]
22. Hansen, J.H.L.; Bou-Ghazale, S.E. Getting started with SUSAS: A speech under simulated and actual stress database. In Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997), Rhodes, Greece, 22–25 September 1997; pp. 1–4.
23. Available online: https://catalog.ldc.upenn.edu/LDC99S78. (accessed on 30 November 2019).
24. Zhang, C.; Hansen, J.H.L. Analysis and classification of speech mode: Whispered through shouted. In Proceedings of the Eighth Annual Conference of the International Speech Communication Association, Antwerp, Belgium, 27–31 August 2007.
25. Chatzis, S.P. Margin-maximizing classification of sequential data with infinitely-long temporal dependencies. *Expert Syst. Appl.* **2013**, *40*, 4519–4527. [CrossRef]

26. Prasetio, B.H.; Tamura, H.; Tanno, K. Deep time-delay Markov network for prediction and modeling the stress and emotions state transition. *Sci. Rep.* **2020**, *10*, 18071. [CrossRef]
27. Al-Dahri, S.S.; Al-Jassar, Y.H.; Alotaibi, Y.A.; Alsulaiman, M.M.; Abdullah-Al-Mamun, K. A word-dependent automatic Arabic speaker identification system. In Proceedings of the 2008 IEEE International Symposium on Signal Processing and Information Technology, Sarajevo, Bosnia and Herzegovina, 16–19 December 2008; pp. 198–202.
28. Krobba, A.; Debyeche, M.; Amrouche, A. Evaluation of speaker identification system using GSMEFR speech data. In Proceedings of the 5th International Conference on Design & Technology of Integrated Systems in Nanoscale Era, Hammamet, Tunisia, 23–25 March 2010; pp. 1–5.
29. Kirchhoff, K.; Bilmes, J.; Das, S.; Duta, N.; Egan, M.; Ji, G.; He, F.; Henderson, J.; Liu, D.; Noamany, M. Novel approaches to Arabic speech recognition: Report from the 2002 Johns-Hopkins summer workshop. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03), Hong Kong, China, 6–10 April 2003; Volume 1, p. I. Available online: https://ieeexplore.ieee.org/document/1198788 (accessed on 14 March 2022).
30. Shahin, I.; Hindawi, N.; Nassif, A.B.; Alhudhaif, A.; Polat, K. Novel dual-channel long short-term memory compressed capsule networks for emotion recognition. *Expert Syst. Appl.* **2022**, *188*, 116080. [CrossRef]
31. Shahin, I. Speaker identification in a shouted talking environment based on novel third-order circular suprasegmental hidden Markov models. *Circuits Syst. Signal Process.* **2016**, *35*, 3770–3792. [CrossRef]
32. Shahin, I. Speaker identification in the shouted environment using suprasegmental hidden Markov models. *Signal Process.* **2008**, *88*, 2700–2708. [CrossRef]
33. Shahin, I.; Nassif, A.B. Emirati-accented speaker identification in stressful talking conditions. In Proceedings of the 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, United Arab Emirates, 19–21 November 2019.
34. Han, Y.; Wang, G.; Yang, Y. Speech emotion recognition based on MFCC. *J. ChongQing Univ. Posts Telecommun. (Natural Sci. Ed.)* **2008**, *20*, 507–602.
35. Pan, Y.; Shen, P.; Shen, L. Speech emotion recognition using support vector machine. *Int. J. Smart Home* **2012**, *6*, 101–108.
36. Falk, T.H.; Chan, W.-Y. Modulation spectral features for robust far-field speaker identification. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *18*, 90–100. [CrossRef]
37. Grozdić, Đ.T.; Jovičić, S.T.; Subotić, M. Whispered speech recognition using deep denoising autoencoder. *Eng. Appl. Artif. Intell.* **2017**, *59*, 15–22. [CrossRef]
38. Vachkov, G. *Multistep Modeling for Approximation and Classification by Use of RBF Network Models BT—Innovative Issues in Intelligent Systems*; Sgurev, V., Yager, R.R., Kacprzyk, J., Jotsov, V., Eds.; Springer International Publishing: Cham, Denmark, 2016; pp. 325–353, ISBN 978-3-319-27267-2.
39. Di, C.; Yang, X.; Wan, X. A four-stage hybrid model for hydrological time series forecasting. *PLoS ONE* **2014**, *9*, e1046639. [CrossRef]
40. Hinton, G.E.; Salakhutdinov, R.R. Replicated softmax: An undirected topic model. *Adv. Neural Inf. Process. Syst.* **2009**, *22*, 1607–1614.
41. Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; Zhu, X. Learning entity and relation embeddings for knowledge graph completion. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
42. Tang, D.; Qin, B.; Liu, T. Document modeling with gated recurrent neural network for sentiment classification. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1422–1432.
43. Gehan, E.A. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **1965**, *52*, 203–224. [CrossRef]