information

Article

Empirical Comparison between Deep and Classical Classifiers for Speaker Verification in Emotional Talking Environments

Ali Bou Nassif [1,*], Ismail Shahin [2], Mohammed Lataifeh [3], Ashraf Elnagar [3] and Nawel Nemmour [1]

[1] Computer Engineering Department, University of Sharjah, Sharjah 27272, United Arab Emirates
[2] Electrical Engineering Department, University of Sharjah, Sharjah 27272, United Arab Emirates
[3] Computer Science Department, University of Sharjah, Sharjah 27272, United Arab Emirates
* Correspondence: anassif@sharjah.ac.ae

Abstract: Speech signals carry various bits of information relevant to the speaker such as age, gender, accent, language, health, and emotions. Emotions are conveyed through modulations of facial and vocal expressions. This paper conducts an empirical comparison of performances between the classical classifiers: Gaussian Mixture Model (GMM), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Artificial neural networks (ANN); and the deep learning classifiers, i.e., Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Gated Recurrent Unit (GRU) in addition to the *i*vector approach for a text-independent speaker verification task in neutral and emotional talking environments. The deep models undergo hyperparameter tuning using the Grid Search optimization algorithm. The models are trained and tested using a private Arabic Emirati Speech Database, Ryerson Audio–Visual Database of Emotional Speech and Song dataset (RAVDESS) database, and a public Crowd-Sourced Emotional Multimodal Actors (CREMA) database. Experimental results illustrate that deep architectures do not necessarily outperform classical classifiers. In fact, evaluation was carried out through Equal Error Rate (EER) along with Area Under the Curve (AUC) scores. The findings reveal that the GMM model yields the lowest EER values and the best AUC scores across all datasets, amongst classical classifiers. In addition, the *i*vector model surpasses all the fine-tuned deep models (CNN, LSTM, and GRU) based on both evaluation metrics in the neutral, as well as the emotional speech. In addition, the GMM outperforms the *i*vector using the Emirati and RAVDESS databases.

Keywords: classical classifiers; deep neural network; emotional speech; feature extraction; speaker verification

check for updates

Citation: Nassif, A.B.; Shahin, I.; Lataifeh, M.; Elnagar, A.; Nemmour, N. Empirical Comparison between Deep and Classical Classifiers for Speaker Verification in Emotional Talking Environments. *Information* **2022**, *13*, 456. https://doi.org/10.3390/info13100456

Academic Editors: Francesco Beritelli and Claude Frasson

Received: 23 June 2022
Accepted: 23 September 2022
Published: 27 September 2022

1. Introduction

Speaker recognition is broadly classified into two distinct areas, i.e., speaker identification and speaker verification or authentication [1]. The former technology describes the process of identifying the identity claim from a pool of already known users using specific acoustic features embedded and retained in speech signals; whereas the latter details a process used to check and inspect whether the claimed user identity is genuine (claimant) or not (imposter/background) based on acoustic models stored in a database.

Speaker recognition systems come in two forms concerning spoken text: text-dependent and text-independent. Text-dependent systems necessitate that the utterances under evaluation utter the same text at training and testing levels, whereas systems with text-independency have no constraint in terms of samples of speech being uttered in the training and testing stages.

Presently, such verification systems are considered prevalent and have a wide range of usages from biometric user verification, surveillance, and forensics, to security applications including credit card payments, access control to computer networks, call monitoring, and online banking access [2].

In general, the speaker verification protocol involves three main phases: development/training, enrollment, and evaluation [3]. The task of each stage is briefly stated below.

- Development/Training: internal representations are learned from the corresponding speaker's acoustic frames.
- Enrollment: voiceprints are derived from voice samples.
- Evaluation: verification is achieved by comparing the test utterance speaker representation against the speaker models [3].

This work carries out an empirical comparison of performance results among fine-tuned Deep Neural Networks (DNNs) models, the *i*vector approach, and several classical classifiers for speaker verification in emotional milieus using Mel Frequency Cepstrum Coefficients (MFCCs) as the extracted features. For this task, the results demonstrate that the traditional GMM classifier outperforms the *i*vector, as well as the fine-tuned deep models using the private Arabic Emirati Speech database and the RAVDESS dataset. Using the CREMA database, the *i*vector surpasses all other classifiers. This emphasizes the fact that deep models are not always the best choice when dealing with machine learning, as per Saez et al. [4]. Zappone et al. [5] stated that the system demands relatively large datasets in order to achieve high performance in deep learning networks. For smaller datasets, classical algorithms could surpass deep learning. In general, results show that GMM is the best model amongst traditional models and that *i*vector surpasses the finetuned CNN, LSTM, and GRU models.

The rest of this paper is structured as follows. In Section 2, a thorough literature review is conducted followed by the contribution of our work. In Section 3, the emotional speech corpora and feature extraction techniques are introduced. In Section 4, the fundamentals of classical classifiers are presented along with their corresponding configurations and verification systems. In Section 5, the DNN-based speaker verification setup is explained. The decision threshold and the verification process are introduced in Section 6, while the experimental results are discussed thoroughly in Section 7. Lastly, in Section 8, the conclusions, limitations, and future work directions are provided.

2. Literature Review

2.1. Speaker Verification Using Classical Classifiers

There have been several efforts to study speaker verification in the neutral acoustic environment using classical models [6–13].

Wan and Campbell [6] studied the speaker verification and speaker identification performance of the SVM model on the YOHO database. The authors developed a novel normalization technique of the polynomial kernel and the experimental results showed that the achieved performance is comparable to that of the GMM model.

Vivaracho et al. [7] conducted a comparative study between GMM-based and ANN-based models in text-independent speaker verification systems using the AHUMADA/GAUDI Spanish dataset. The results using a GMM-based system and microphonic speech demonstrated better verification performance. However, the ANN surpassed the GMM results when testing in specific circumstances and with real telephone speech.

In [10], Alarifi et al. suggested and analyzed the performance of their novel Arabic text-dependent speaker verification system using ANNs. The proposed system can be used as an application for access control in cellular devices. Test results showed that the verification performance obtained using the ANN model is better than that of the SVM.

On the other hand, limited studies have tackled speaker verification problems on emotional speech using the traditional approaches [14–16]. Wu et al. [14] studied the influence of emotional features on the verification performance in a GMM-UBM model. The authors suggested an emotion-dependent score normalization approach for speaker verification tasks under emotional data conditions. The results attained an average speaker verification performance equivalent to 88.5%. Mittal and Dua [17] listed different approaches to the design backend of Automatic Speaker Verification system. These approaches are based on classical, as well as deep learning. The authors concluded that the latest artificial intelligence techniques do not properly target Mimicry and Twins attacks. Ferrer et al. [18] raised the issues of speaker verification in unknown conditions during development. The authors proposed a new model using an adaptive calibrator that modifies the standard backend which yielded better results. Liu et al. [19] proposed a neural acoustic-phonetic approach that assigns differentiated weights dynamically to spectral features for speaker verification. The proposed model surpasses baseline models.

The standard approach involves modeling text-independent speaker recognition applications as (GMM) via Model (UBM) [20]. The GMM-UBM paradigm [21] was followed sequentially by joint factor analysis (JFA) and the widely known *i*vector speaker representations [22,23]. The downside of these approaches is their unsupervised nature which is not consistent with speaker verification tasks. Hence, they were further enhanced by proposing certain techniques which supervise and handle the training of the abovementioned models such as the SVM-based GMM-UBMs [14] and the Probabilistic Linear Discriminant Analysis (PLDA)-based *i*vectors [24].

Current studies are focused on improving some classical classifiers by enhancing their performances or making them hybrid, such as the study by Kumar and Bharathi [25]. With both generative and discriminative classifiers, the presented work aims to evaluate the behavior of a spoof detection countermeasure utilizing linear frequency cepstral coefficients. The analysis is conducted on non-emphasized statements. Bidirectional long short-term memory (discriminative) and Gaussian mixture model (generative) classifiers are employed. The empirical findings reveal that the generative classifier significantly outperformed the discriminative classifier in detecting spoof attacks under logical access conditions, and that the discriminative classifier significantly outperformed the generative model in detecting spoof attacks under physical access conditions.

2.2. Speaker Verification Using Deep Learning

Alam et al. in [26] studied the use of the low-variance multi-taper MFCC along with perceptual linear prediction (PLP) features in an *i*vector speaker verification task in neutral condition. The achieved results demonstrated that both MFCC and PLP features, calculated through multi-tapers, yielded systematic improvements concerning recognition accuracy.

Chen et al. [27] suggested extracting local session variability vectors on diverse phonetic categories from the utterances rather than estimating the session variability throughout the entire utterance as the *i*vector-based architecture does. The local vectors depicted the session variability retained in specific phonetic content employing the posteriors driven using a deep neural network which was modeled to classify phone states. Experimental results indicated that the content-aware local vectors outperformed the DNN *i*vectors with respect to test utterances of short durations.

Zhu et al. in [28] introduced a novel approach to identify speaker embeddings using deep learning models. Their results demonstrated that the suggested self-attentive embeddings achieved superior performance compared to the classical *i*vector approach for both long and short testing utterances.

Mobiny and Najarian [29] proposed a text-independent scenario for the speaker verification problem using LSTM neural networks through MFCC speech features in an end-to-end manner. The performance results demonstrated its superiority compared to other traditional methods. EER results gave 22.9% using the proposed LSTM and 27.1%, 24.7%, 23.5% in GMM-UBM, *i*vector and *i*vector + PLDA, respectively.

In a different work, Hourri et al. [30] proposed the use of two-dimensional CNN filters in order to extract speaker-specific information in speaker verification tasks. Moreover, the authors proposed novel vectors called conVectors (i.e., a convolutional neural network vector) to recognize speakers. The evaluation was conducted on the THUYG-20 SRE gender-dependent database under three noise conditions: clean, 9 db, and 0 db. The results showed that the use of the proposed vectors enhanced the verification performance compared to the state-of-the-art methods for speaker verification.

A recent study by Shahin et al. [31] discussed an empirical comparison examination into the performance of text-independent speaker verification in emotional and stressful talking conditions. The study used a combination of DNN deep models along with classical models, producing a novel hybrid classifier as a result. Three datasets were used to assist their experiments, namely an Arabic Emirati dataset, Speech Under Simulated and Actual Stress (SUSAS) and Ryerson Audio–Visual Database of Emotional Speech and Song (RAVDESS). Based on HMM-DNN, DNN-HMM, DNN-GMM, and GMM-DNN, respectively, the average verification system based on the three databases yielded EERs of 7.19, 16.85, 11.51, and 11.90%. In addition, in both talking conditions, the DNN-GMM model showed the least computational complexity compared to the other hybrid models.

Another recent study by Mohammed et al. [32] focused on analyzing speaker identification, recognition, and verification methods and technique. This study includes a summary of speaker verification literature as well as statistical studies to depict the publications and their categories.

2.3. Contribution

To the best of our knowledge, this work is the first of its kind where an empirical study is conducted by executing a diligent assessment and a pragmatic comparison of speaker verification performance between four classical classifiers (GMM, SVM, KNN, and ANN) with three distinct deep learning models (CNN, LSTM, and GRU), using the d-vector approach, in emotional talking environments, in addition to the *i*vector approach. The tuning of hyperparameters using grid search is applied to deep models. Evaluation is assessed on three different speech datasets: the private Arabic Emirati speech corpus, CREMA database, and RAVDESS dataset.

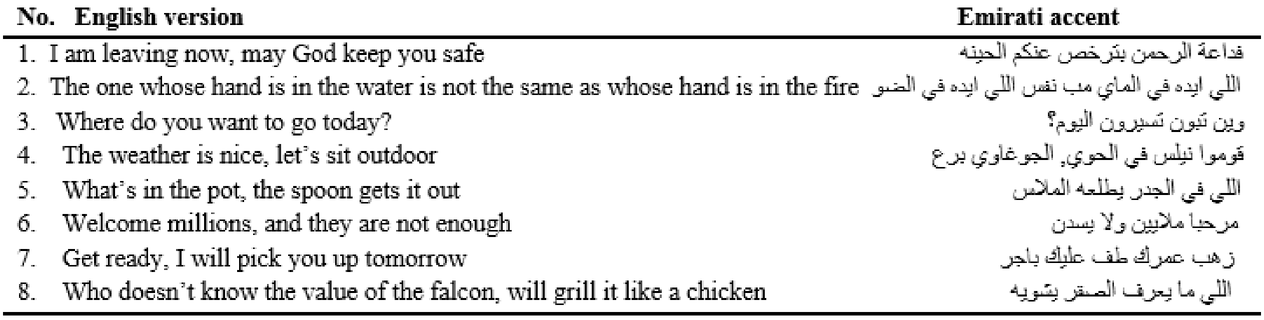
- Unlike previous studies, the d-vector approach implemented in this work uses CNN, as well as recurrent neural networks (LSTM and GRU) layers in order to extract speaker intrinsic voice characteristics from unbiased utterances rather than using CNNs and locally connected networks (LCNs) as in [33], or fully connected maxout layers as in [34], or LSTM layers only as in [35].
- Optimum values of CNN, LSTM, and GRU model hyperparameters are computed using the Grid Search (GS) tuning approach.
- In addition, all state-of-the-art studies examined the verification performance using the d-vector as well as the *i*vector method on neutrally uttered speech only. However, this paper focuses on neutral speech in addition to speech expressed as a function of emotions, namely, anger, sadness, happiness, disgust, and fear.

3. Datasets

A total of three datasets were used for our experiments: A private Arabic Emirati speech database (ESD), the Crowd-Sourced Emotional Multimodal Actors dataset (CREMA), and the Ryerson Audio–Visual Database of Emotional Speech and Song dataset (RAVDESS).

3.1. Arabic Emirati Speech Dataset

This speech dataset consists of 31 inexpert native Emirati speakers (22 females, 9 males), with ages ranging from 15 to 50 years old. Each selected speaker was asked to utter eight different local Arabic phrases commonly used in the United Arab Emirates. Every single phrase was replicated nine times with a duration range between 2–5 s in both neutral and emotional acoustic atmospheres. The acted emotions are: anger, fear, disgust, happiness, sadness, and neutral. The utterances and their corresponding English translation are presented in Figure 1. The training stage is composed of 24 speakers out of 31 speakers (7 males and 17 females) articulating 5 out of the 8 sentences. During this phase, each speaker replicates each sentence 9 times in the neutral state. Therefore, the overall number of utterances utilized in the training stage is equivalent to 1080 utterances (24 speakers × 5 sentences × 9 replicates/sentence × neutral state). The enrollment and evaluation phases comprise each one of the remaining 7 speakers (2 males and 5 females). In enrollment, the corresponding speakers utter the first 5 statements neutrally resulting in a total of 315 utterances (7 speakers × 5 sentences × 9 replicates/sentence × neutral state). In the test phase, the corresponding speakers expressed the remaining 3 sentences (out of 8) with 9 replications per sentence under each of the neutral, angry, happy, sad, fear, and disgust emotions. Hence, the test phase involves a total of 1134 utterances (7 speakers × 3 sentences × 9 repetitions × 6 emotions). In other words, 189 utterances are the total number of speech samples for neutral and each of the emotional states. This privately collected database was recorded at the College of Communication at the University of Sharjah, United Arab Emirates. It was obtained using a speech acquisition board through a 16-bit linear coding analog-to-digital converter and sampled at a rate of 44.6 kHz.



No.	English version	Emirati accent
1.	I am leaving now, may God keep you safe	فداعة الرحمن بترخص عنكم الحينه
2.	The one whose hand is in the water is not the same as whose hand is in the fire	اللي ايده في الماي مب نفس اللي ايده في الضو
3.	Where do you want to go today?	وين تبون تسيرون اليوم؟
4.	The weather is nice, let's sit outdoor	قوموا نيلس في الحوي, الجوغاوي برع
5.	What's in the pot, the spoon gets it out	اللي في الجدر يطلعه الملاس
6.	Welcome millions, and they are not enough	مرحبا ملايين ولا يسدن
7.	Get ready, I will pick you up tomorrow	زهب عمرك طف عليك باجر
8.	Who doesn't know the value of the falcon, will grill it like a chicken	اللي ما يعرف الصقر يشويه

Figure 1. The Emirati dataset and its English version.

3.2. Crowd-Sourced Emotional Multimodal Actors Dataset

The emotional CREMA dataset is a multimodal audio–visual English dataset composed of 7442 clips issued from a total of 91 speakers (48 male and 43 female) of varied ages and ethnicity [36]. Each speaker recorded 12 distinct utterances in six different emotional categories: neutral, anger, sad, happy, fear, and disgust. Examples of utterances are: "Don't forget a jacket (DFA)", "I think I've seen this before (ITS)", "The surface is slick (TSI)" and "We'll stop in a couple of minutes (WSI)". In the training phase, 70 speakers out of 91 speakers are allocated. Each speaker expresses 8 out of 12 sentences with one replication per sentence under the neutral emotion. Therefore, the training phase involves a total of 560 sentences (70 speakers × 8 sentences × 1 time/sentence × neutral state). For enrollment, the remaining 21 speakers uttering the first 8 sentences in the neutral state are used. In enrollment, the total number of sentences utilized is equal to 168 (21 speakers × 8 sentences × 1 time/sentence × neutral state). The evaluation stage includes the remaining 21 speakers speaking the last remaining 4 sentences under each of the emotional states. Therefore, the size of the dataset in the evaluation phase is equivalent to 504 (21 speakers × 4 sentences × 1 time/sentence × 6 emotional classes). Consequently, the number of utterances under each emotional state is 84.

3.3. Ryerson Audio–Visual Database of Emotional Speech and Song Dataset

RAVDESS is an English database with a total of 7356 recorded files involving a total of 24 professional actors and actresses (12 male and 12 female) [37]. It consists of two different utterances expressed in a neutral North American English accent. The corpus contains 1440 audio files (60 trials/speaker × 24 speakers), 1012 song files (44 trials/speaker × 23 speakers), and 4904 video and audio–visual files. In this work, only audio and song files are used. The RAVDESS dataset comprises eight different emotional classes. Only six of them are considered which are neutral, anger, sad, happy, fear, and disgust. The training stage involves 20 speakers (10 male and 10 female) out of 24 speakers expressing the first utterance (out of 2) with a replication of 2 times per utterance expressed in the neutral emotion. Therefore, the total number of utterances in the training phase is equal to 78 utterances originating from both audio and song files: 40 utterances from the audio files (20 speakers × 1 sentence × 2 trials/sentence × neutral state) + 38 utterances from the song files (19 speakers ×

1 sentence × 2 trials/sentence × neutral state). The remaining 4 speakers (2 male and 2 female) are utilized for enrollment and test stages. For enrollment, the corresponding speakers utter the first statement neutrally resulting in a total of 8 utterances from audio files (4 speakers × 1 sentence × 2 trials/sentence × Neutral state) + 8 utterances from the song files (4 speakers × 1 sentence × 2 trials/sentence × Neutral state). For the test phase, the corresponding speakers speak the second utterance neutrally and under each of the emotional conditions. Therefore, a total of 176 utterances from both audio and song files are designated for the test phase. Under the neutral state, a total of 16 utterances (8 from audio and 8 from song files) (4 speakers × 1 sentence × 2 trials/sentence × Neutral state). Under each of Disgust, Angry, Sad, Happy and Fear emotions: a total of 32 utterances (16 from audio and 16 from song files) are utilized (4 speakers × 1 sentence × 4 trials/sentence × 5 emotional states).

3.4. Feature Extraction

Various aspects are involved in uniquely characterizing one's voice such as the structure of the vocal tract, regional accents, intonation, mood, and speaking style. Similarly, audio signals are affected by several parameters which vary with the different recording circumstances such as background noise, microphone distortion, etc. Consequently, the need for a robust feature extraction process is vital in automatic speaker recognition systems to grasp only informative and non-redundant linguistic content [14,38].

In this study, MFCC features are the coefficients utilized to characterize the phonemic content and to learn a compact representation of the spectrogram of audio signals in both speech corpora [39].

The feature extraction uses a concatenation of MFCCs, MFCCs-delta and delta-delta. In this work, 40-dimensional audio features were extracted using the libROSA package designated for music and audio analysis in Python [40]. Features were resampled to 16 kHz, then segmented into frames of 32 ms with a Fast Fourier Transform (FFT) window length of 2048 and frame length of 512 which yielded 75% overlap between successive frames.

4. Classical Classifiers

4.1. Gaussian Mixture Models

A Gaussian mixture model is considered as a probabilistic clustering model which represents the existence of normally distributed subpopulations within an overall population [21,41].

The mixture density utilized for the likelihood function can be numerically expressed as follows [41],

$$P(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{1}$$

where x is the D-dimensional feature vector, $p_i(x)$ is defined as a weighted linear combination of M unimodal Gaussian densities, each parameterized by a mean $D \times 1$ vector, u_i and a D × D covariance matrix, $\sum i$, [41]:

$$p_i(x) = \frac{1}{(2\,\pi)^{\frac{D}{2}}\,|\sum i|^{\frac{1}{2}}}\, exp\left\{-\frac{1}{2}(x-u_i)'\,\left(\sum i\right)^{-1}(x-u_i)\right\} \tag{2}$$

4.2. Support Vector Machines

SVM is considered a binary discriminative model defined by a linear decision boundary or a hyperplane that optimally separates given points into two predefined classes [9].

The SVM classifier is constructed from sums of a known kernel function $K(.,.)$ in order to outline a hyperplane

$$f(x) = \sum_{i=1}^{N} \propto_i\, y_i K(x, x_i) + b \tag{3}$$

where $y_i \epsilon \{-1, 1\}$ denotes the target values, $\sum_{i=1}^{N} \propto_i y_i = 0$ and $\propto_i > 0$.

4.3. K-Nearest Neighbors

The KNN is a supervised learning algorithm and is a nonparametric method; meaning no prior underlying assumptions are needed about data distribution. It can be used in both classification and regression problems. When KNN is used for the classification problem, the output can be predicted as the class with the highest frequency from the K-most identical instances [42].

In this paper, KNN has been selected to perform classification tasks with K = 5. The distance between the different data points is computed using the Euclidean distance calculated as the square root of the sum of the squared differences between a new point y_i and an existing point x_i across all input attributes k.

$$Euclidean = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{4}$$

4.4. Artificial Neural Networks

Similar to neurons in a biological brain, an artificial neural network is a computational network that is based on a collection of partially linked units or nodes, arranged in layers. In addition, they contain interconnections between the nodes of successive layers [43]. Figure 2 demonstrates a schematic diagram of the basic configuration of a single neuron or node within a neural network model which consists of inputs, an activation function, and a single output. The nodes are linked through weights. The weights W characterize the relative importance of the connection between neurons. Y_i represents the output [43].

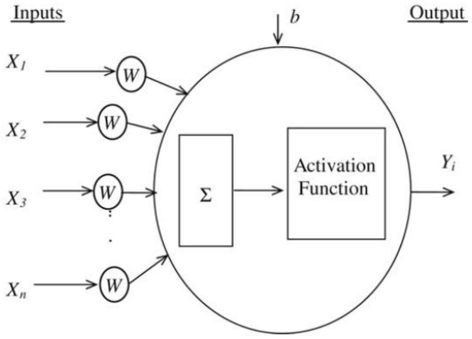


Figure 2. Schematic configuration of the basic structure of a node in ANN [43].

4.5. Model Configuration and Verification

The configuration of each of the GMM, SVM, KNN, and ANN models for the Emirati database are explained in the subsequent paragraphs. The Scikit-learn machine learning library in Python was utilized.

4.5.1. The GMM Model

In the training phase, every speaker is characterized by a unique GMM model which is parameterized by 16 Gaussian mixture components, each having its diagonal covariance matrix with 200 iterations of the EM algorithm. For the Emirati database, each model is obtained using the first 5 sentences. Each speaker repeats each sentence 9 times with neutral emotion. This results in a total of 45 utterances (5 sentences × 9 repetitions/sentence) for each speaker model. In the evaluation stage, the remaining 3 sentences were involved. A total of 837 utterances (31 speakers × 3 sentences × 9 repetitions/sentence) were used in each emotional category.

Upon the arrival of a test utterance, MFCC audio features are first extracted from the raw signal, then log-likelihood scores are computed using Equations 10 and 11 to obtain true (claimants) speakers and false speakers' scores (imposters), respectively. Figure 3 shows the histograms of scores for true and false speakers of neutral and emotional speech for the GMM model using the Emirati database.

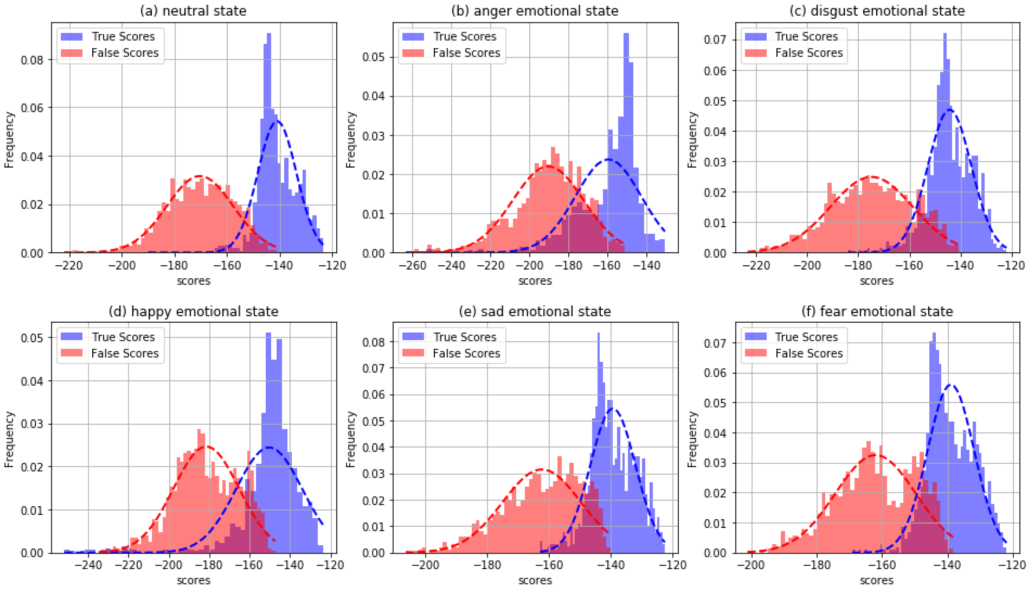


Figure 3. Histograms of scores for true and false speakers of neutral and emotional speech for the GMM model in the Emirati database.

The conditional probability of the observation sequence O given that it comes from the true speaker $P(O|\lambda_{model,\,C})$ is defined by:

$$\log\ P(O|\lambda_{model,\,C}) = \frac{1}{T}\sum_{T=1}^{T} log(o_t|\lambda_{model,\,C}) \tag{5}$$

where, $O = o_1, o_2, \ldots, o_t, \ldots, o_T$ and T is the duration of an utterance.

The probability of the observation sequence O given that it comes from a false speaker $P\left(O\middle|\lambda_{model,\overline{C}}\right)$ is calculated using a set of B imposter speaker models $\left\{\lambda_{model,\,\overline{C_1}}, \lambda_{model,\,\overline{C_2}}, \ldots, \lambda_{model,\,\overline{C_B}}\right\}$ as follows:

$$\log P\left(O\middle|\lambda_{model,\overline{C}}\right) = \frac{1}{B}\sum_{b=1}^{B} log\left[\left(O\middle|\lambda_{model,\,\overline{C_b}}\right)\right] \tag{6}$$

4.5.2. SVM, KNN and ANN Models

For the SVM model, this work utilizes a C-support vector classification with radial basis function kernel or RBF kernel. For the KNN model configuration, the number of neighbors is set to 5 with the Euclidean distance metric. The ANN model uses one hidden layer with 100 neurons with a stochastic gradient descent solver for weight optimization.

Training or fitting of each of the SVM, KNN, and ANN models is accomplished by providing the set of scaled training data as well as their corresponding target values (class labels). The training data are the 40-order vertically stacked MFCC vectors + 40-order MFCC-deltas + 40-order MFCC-delta-delta, whereas the class labels represent the number of speakers available in the dataset.

5. Deep Neural Networks

5.1. System Overview

Figure 4 depicts an overview of the model's DNN-based topology used in this verification problem. In general, the verification task is composed of the following three stages:

development/training, enrollment and testing. Each phase comprises input and output. During training, each input of the DNN model is the set of training utterances 1, . . . , N, while the output is the predicted probabilities relevant to every speaker.

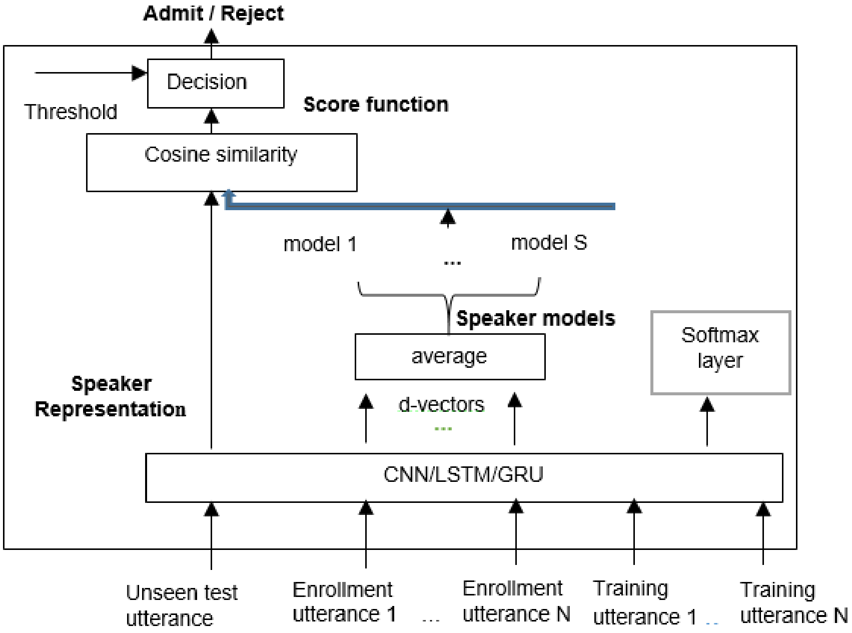


Figure 4. Architecture used for DNN models in speaker verification.

During the development stage, three distinct and independent supervised DNN-based models (CNN, LSTM, and GRU) operating at the frame level, are designed and eventually trained in order to learn speaker specific features. The utterances expressed neutrally form the input of the DNNs. As DNNs assume a fixed size window of input vectors, the utterances are framed at a sufficiently large window size (frame length), and then vertically stacked and used as the input of the DNN. In this work, the input feature map size is equivalent to 120 × 120 which corresponds to frame length × number of coefficients. The number of coefficients is the result of the concatenation of 40 MFCCs + 40 deltas + 40 delta-deltas. Using the Emirati database, 24 speakers out of 31 speakers (17 females, 7 males) are designated for the training phase.

Using the Emirati database, 7 unseen and new speakers (5 females and 2 males) are designated for both enrollment and evaluation phases. The enrollment utterances are composed of the first 5 sentences, out of 8 sentences, expressed by each enrollment speaker in the neutral state. Hence, there is a total of 315 utterances from the enrollment phase (7 speakers × 5 sentences × 9 replicates/sentence × neutral state). During this phase, each input to the DNN network is the enrollment utterances 1, . . . , N issued from corresponding speakers, whereas the output is the speaker models.

In the testing phase, the input is the test utterance to be verified and the output is a single node indicating admission or rejection of the speaker identity claim.

5.2. CNN Model

5.2.1. Development Phase

The design of the CNN model during the development phase consists of one convolutional hidden layer made of 128 units and a Rectified Linear Unit (ReLU) activation function followed by one pooling layer. The output from the latter layer is flattened and forwarded to a Dense layer with 128 units. Eventually, the resultant vector is fed to the output layer. This layer is a dense layer with SoftMax activation function which correlates to the number of speakers, 24, involved in this phase. The target labels are encoded and represented as a binary class matrix ranging from 0 to N where the only component not equivalent to zero is allocated to the corresponding speaker identity S. The model is trained using the categorical cross-entropy criterion and stochastic gradient descent optimizer with a learning rate of 0.02. Figure 5 shows the background DNN model's topology for learning speaker audio features during the training phase. The input of the CNN model is the stacked MFCC features of the development utterances. The output is the classification probabilities relevant to each speaker.

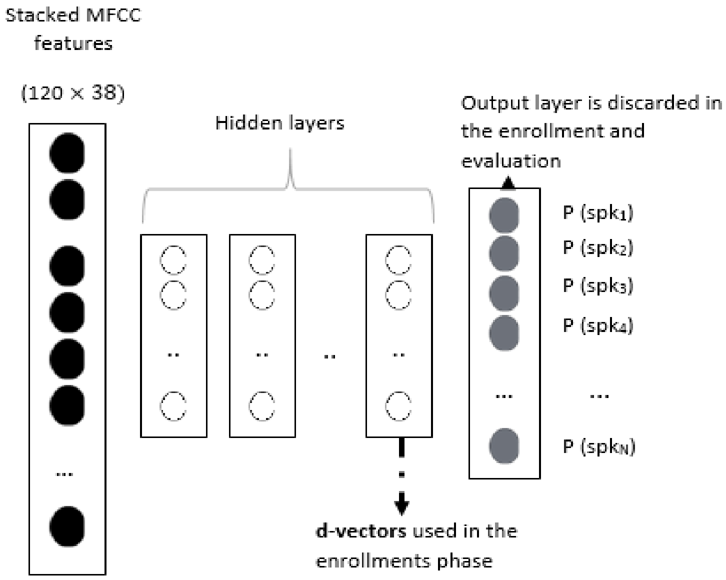

Stacked MFCC features
(120 × 38)
Hidden layers
Output layer is discarded in the enrollment and evaluation
P (spk1)
P (spk2)
P (spk3)
P (spk4)
...
P (spkN)
d-vectors used in the enrollments phase

Figure 5. Background DNN topology for learning speaker-specific features [33].

5.2.2. Enrollment Phase

First, the MFCC features are extracted from the enrollment utterances, framed, and then stacked. Second, they are fed to the pretrained CNN model then forward propagated through the network whose weights are kept static upon completion of model training during the development phase. Then, the d-vector speaker discriminative information is obtained from output activations of the last hidden layer for all frames of an utterance (before the output layer). Each sentence engenders one d-vector. Voiceprints are given by averaging the d-vectors of the corresponding speaker enrollment utterances.

5.2.3. Evaluation Phase

In this phase, test utterances under each of the emotional categories are framed, stacked, and fed to the pretrained CNN model obtaining the test d-vectors. The evaluation is performed for each utterance where the cosine distance is computed, first, between the test d-vector and the real speaker identity model and, second, against all other speaker models forming the true speakers and false speakers' scores, respectively. Ultimately, the distance is compared to a pre-defined threshold allowing for a verification decision to be made. The value of a scoring function of an utterance X and a test speaker spk S(X, spk) is given by:

$$S(X, spk) = \left[f(X)^T m_{spk} \right] / \left[\|f(X)\| \|m_{spk}\| \right] \tag{7}$$

where $f(X)$ is the speaker representation and, m_{spk} is the speaker model.

5.3. LSTM Model

5.3.1. Development Phase

The input of the LSTM model is the stacked MFCC features of the training utterances. The LSTM model comprises one LSTM layer with 64 units followed by a Dense layer whose units are equivalent to 64. The output vector is ultimately fed to the output layer and predicted probabilities of each speaker are generated.

5.3.2. Enrollment Phase

The MFCC coefficients are extracted from the enrolment utterances and fed to the pretrained LSTM model. The d-vector speaker information is obtained from the last hidden layer for all frames of an utterance. At the end of this stage, speaker models are produced by averaging the d-vectors of the corresponding speaker enrollment utterances.

5.3.3. Evaluation Phase

Test utterances under each emotion are fed to the pretrained LSTM model. The corresponding test d-vectors are obtained from the output of the last hidden layer and speaker verification is performed as described in Section 5.2.3.

5.4. GRU Model

5.4.1. Development Phase

The GRU model is designed as follows: one GRU layer with 64 hidden units followed by a Dense layer then an output layer. The input of the GRU model is the set of MFCCs relevant to the enrolment utterances. The output is the predicted probability for each speaker.

5.4.2. Enrollment Phase

The MFCC coefficients are extracted from the enrolment utterances and fed to the pretrained GRU model. The d-vector speaker information is retrieved from the last hidden layer for all frames of an utterance. The output from this phase is the corresponding speaker models.

5.4.3. Evaluation Phase

During this phase, the test utterances are fed to the pretrained GRU model. The test d-vectors are obtained from the last hidden layer and the verification of speakers is achieved as explained in Section 5.2.3. Table 1 shows the configuration of the DNN models.

Table 1. Configuration of DNN-based models using the Emirati database.

DNN Model	Layers	#Layers	Units	Other Params.
CNN	Conv2d	1	128	Relu [1], kernel = 7, strides = 2
	MaxPool2D	1	-	pool_size = 2, strides = 2
	Dense	1	128	-
	Dense (Output layer)		24	SoftMax
LSTM	LSTM	1	64	Relu
	Dense	1	64	-
	Dense (Output layer)		24	SoftMax
GRU	GRU	1	64	-
	Dense	1	64	-
	Dense (Output layer)		24	SoftMax

[1] Rectified Linear Unit activation function.

5.5. Enrollment Phase

Using the ESD database, 7 unseen and new speakers (5 females and 2 males) are designated for both enrollment and evaluation phases. The enrollment utterances are composed of the first 5 sentences, out of 8 sentences, expressed by each enrollment speaker in the neutral state. Hence, a total of 315 utterances from the enrollment phase (7 speakers × 5 sentences × 9 replicates/sentence × neutral state). First, the enrollment utterances of a given speaker S are forward propagated through the pre-trained supervised DNN model whose weights are kept static upon completion of model training during the development phase. Then, the d-vector speaker discriminative information is obtained from output activations of the last hidden layer for all frames of an utterance (before the

SoftMax layer). Each sentence engenders one d-vector. Voiceprints are given by averaging the d-vectors of the corresponding speaker enrollment utterances.

5.6. Evaluation Phase

Likewise, test utterances under each of the emotional categories are stacked and fed to the DNN models, at the evaluation stage, obtaining the test d-vectors. The evaluation is performed for each particular phrase where the cosine distance is computed, first, between the test d-vector and the real speaker identity model and, second, against all other speaker models forming the true speakers and false speakers' scores, respectively. Ultimately, the distance is compared to a pre-defined threshold allowing for a verification decision to be made.

The value of a scoring function of an utterance X and a test speaker spk $S(X, spk)$ is given by [35],

$$S(X, spk) = \left[f(X)^T m_{spk}\right] / \left[\|f(X)\| \|m_{spk}\|\right] \tag{8}$$

where $f(X)$ is the speaker representation and, m_{spk} is the speaker model.

6. Decision Threshold and Verification Process

Speaker authentication systems commonly use a threshold to determine whether a claimed identity counterpart is a formerly enrolled voiceprint or not. It is a paramount parameter and a critical factor in verification and binary decision tasks. In this setup, two potential forms of errors may arise: False Rejection (FR) and False Acceptance (FA). A false rejection error occurs when a genuine speaker identity claim is declined; on the other hand, a false acceptance error occurs when an imposter speaker identity claim is admitted. The value where the False Rejection Rate (FRR) is equivalent to the False Acceptance Rate (FAR) is called EER and it is broadly utilized as one of the key performance metrics in authentication systems. Values where FRR is not equivalent to FAR are commonly evaluated and assessed with detection error trade-off (DET) curves which include FRR at the y-axis and FAR at the x-axis.

The equal error rate (*EER*) is calculated at different threshold values using *FAR* and *FRR*. The EER is equivalent to when both rates are equal.

$$EER\ is\ where:\ FAR == FRR \tag{9}$$

The last step in the authentication procedure is to compete for the "log-likelihood ratio" with the "threshold" θ to admit or decline the requested speaker, i.e.,

$$\begin{array}{c}\text{Accept the claimed speaker if the log} - \text{likelihood ratio} \geq \theta \\ \text{Reject the claimed speaker if the log} - \text{likelihood ratio} < \theta\end{array}$$

The log-likelihood ratio is given by the following formula:

$$\text{log} - \text{likelihood ratio} = log\left[P(O|\lambda_{\text{model},\ C})\right] - log\left[P\left(O\middle|\lambda_{\text{model},\ \overline{C}}\right)\right] \tag{10}$$

7. Results and Discussion

In order to evaluate and assess the verification performance, EER values along with AUC scores are computed for each of the private Emirati dataset, CREMA dataset, and RAVDESS. Table 2 demonstrates that utterances expressed neutrally have the lowest EER percentage value in comparison with other phrases spoken in the different emotional categories based on classical classifiers as well as on deep learning classifiers and *i*vector. On the other hand, results report that phrases that are portrayed with Anger have the highest error rates based on all classifiers.

Table 2. Percentage EER values of classical and deep models using Emirati dataset.

Equal Error Rate (EER) (%) Collected Emirati Dataset								
	GMM EER AUC	KNN EER AUC	SVM EER AUC	ANN EER AUC	*i*vector EER AUC	CNN EER AUC	LSTM EER AUC	GRU EER AUC
Neutral	1.43 0.99	19.00 0.16	9.00 0.09	10.00 0.09	8.55 0.97	12.83 0.93	9.13 0.95	8.91 0.96
Anger	12.49 0.94	42.00 0.24	29.00 0.21	37.00 0.23	12.83 0.94	13.89 0.91	12.70 0.94	14.77 0.92
Happy	5.32 0.98	35.00 0.23	21.00 0.17	23.00 0.18	10.1 0.95	14.86 0.92	11.64 0.94	12.79 0.94
Sad	2.63 0.98	45.00 0.25	25.00 0.19	25.00 0.19	9.08 0.97	15.34 0.91	12.74 0.95	10.54 0.94
Fear	3.70 0.99	45.00 0.25	24.00 0.18	23.00 0.18	9.18 0.97	13.89 0.91	12.26 0.94	11.77 0.95
Disgust	2.27 0.99	29.00 0.20	15.00 0.13	16.00 0.13	10.1 0.96	16.58 0.91	10.11 0.95	11.66 0.94
Average	4.64 0.97	35.83 0.22	20.5 0.16	22.33 0.16	9.97 0.96	14.56 0.92	11.43 0.94	11.74 0.94

Based on EER results using the Emirati database shown in Table 2, it is apparent that the *i*vector surpasses the fine-tuned deep models CNN, LSTM, and GRU under neutral and emotional conditions with average error rates equivalent to 9.97% compared to 14.56%, 11.43%, and 11.74%, respectively. For this dataset, the *i*vector consists of 512 UBM components and 64 total variability space (TVS) rank. The total number of eigenvectors in the projection matrix and the number of dimensions in the probabilistic linear discriminant analysis (PLDA) are both set to 16.

The GMM model performs the best amongst the classical classifiers when phrases are uttered neutrally and emotionally; followed by the SVM, then the ANN, and eventually the KNN models. The GMM model yields the lowest percentage EER value equivalent to EER = 4.64% compared to 20.5%, 22.33%, and 35.83% based on SVM, ANN, and KNN, respectively.

Figure 6 depicts the plots of the ROC curves which compare the verification performance based on each of the GMM, CNN, LSTM, GRU, and *i*vector models at different classification thresholds using the Emirati speech database in the neutral and emotional conditions. The performance of each classifier is measured by considering the area under the ROC curve or the AUC score. The area covered by the curve is the entire area underneath the ROC curve. A larger area indicates better performance. The ROC plots demonstrate that the performance of the GMM model outperforms each of the fine-tuned CNN, LSTM, GRU, *i*vector as well as all classical models under both neutral and emotional conditions with AUC = 0.99 in the neutral condition compared to 0.09, 0.09, 0.16, 0.97, 0.94, 0.96 and 0.93 based on SVM, ANN, KNN, *i*vector, LSTM, GRU, and CNN, respectively.

Based on the plots in Figure 6, we deduce the superiority of the *i*vector approach over the fine-tuned CNN, LSTM, and GRU models. The *i*vector succeeds in achieving a larger area underneath the ROC curves corresponding to each of the emotional categories as well as the neutral speech.

In order to validate our results, we conducted the non-parametric Wilcoxon test to observe whether the winning model among the conventional classifiers is statistically different from other models based on 95% confidence interval. Based on the results in Table 3, we find that the GMM model is statistically different from the ANN, KNN, and SVM models in all emotions based on a 95% confidence interval, where all the *p*-values obtained are less than alpha = 0.05. In addition, we conclude that a significant difference does exist between the *i*vector model and each of the LSTM and GRU deep models for neutral as well as emotional speech. In addition, we notice that *i*vector is different from the CNN model in Neutral, Anger, Happy, and Sad. However, the *i*vector fails to be different from CNN in Fear and Disgust emotional states.

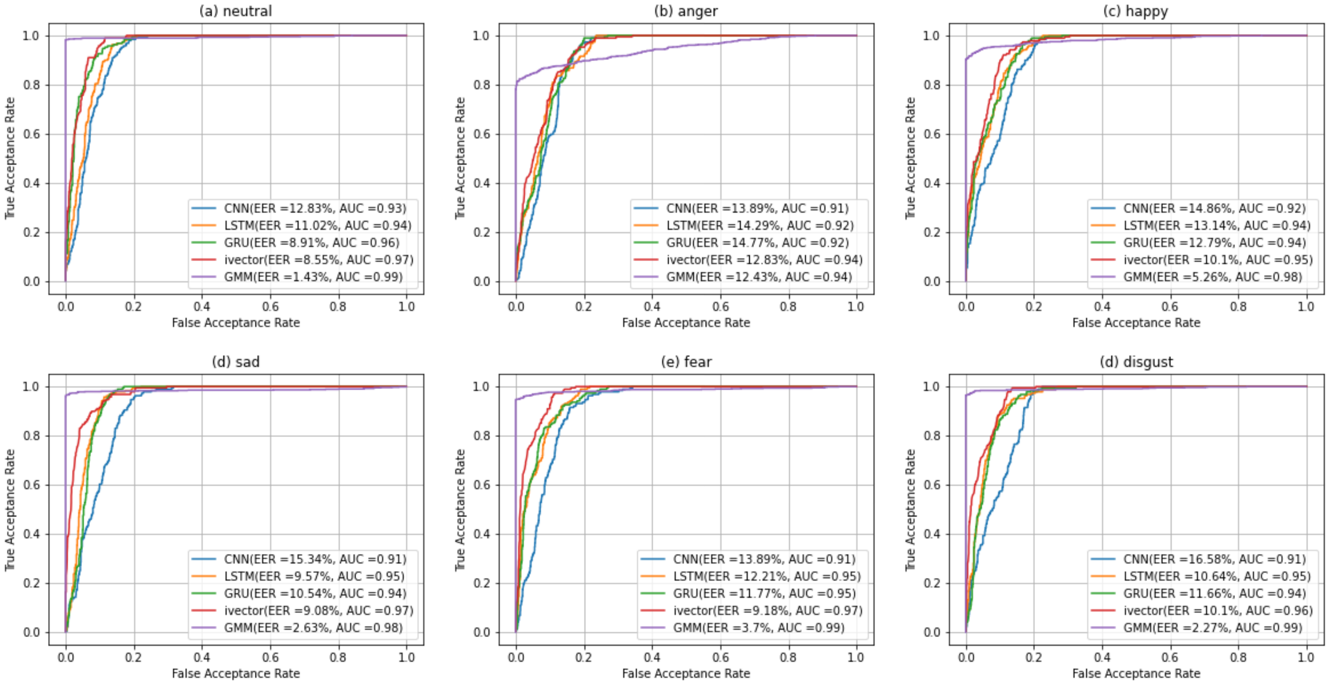


Figure 6. ROC curves with each based on GMM, KNN, SVM, ANN, CNN, LSTM, and GRU models for the different emotional states using the Emirati database.

Table 3. *p*-value using Wilcoxon test for classical models and deep models using the Emirati database.

				Wilcoxon Test				
	KNN	SVM	ANN		CNN	GRU	LSTM	
Neutral	**0.003**	**0.000**	**0.000**	GMM	**0.000**	**0.000**	**0.000**	*i*vector
Anger	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	
Happy	**0.000**	**0.000**	**0.000**		**0.009**	**0.000**	**0.000**	
Sad	**0.000**	**0.000**	**0.000**		**0.000**	**0.005**	**0.000**	
Fear	**0.000**	**0.000**	**0.000**		0.588	**0.000**	**0.000**	
Disgust	**0.000**	**0.000**	**0.000**		0.059	**0.000**	**0.000**	

When shedding light upon deep models, we deduce that in the neutral state, the calculated AUC value of the GRU model is higher than that of the LSTM and the CNN models; 0.96, 0.93, and 0.95 for GRU, CNN, and LSTM, respectively. In all emotional states, the AUC scores obtained through the LSTM model are higher than that of the CNN model. Likewise, the AUC scores of the LSTM model are better than the GRU model except for the neutral and fear emotions. The average EER result attained by the LSTM model is lower than that achieved by GRU (11.43% compared to 11.74%). Figure 7 portrays the graphical plots of the DET curves based on each of the aforementioned classifiers.

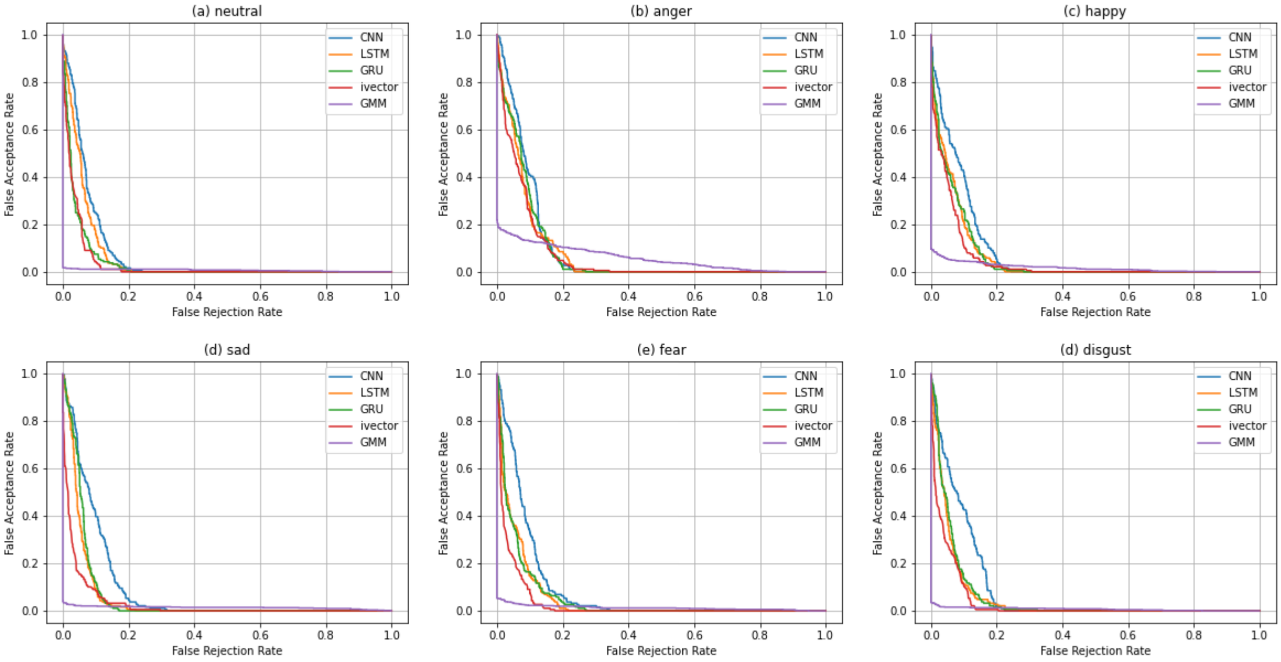


Figure 7. DET curves with each based on GMM, KNN, SVM, ANN, CNN, LSTM, and GRU models for the different emotional states using the Emirati database.

7.1. CREMA Database

The CREMA database is employed in order to assess the verification performance for each GMM, SVM, KNN, ANN, CNN, LSTM, GRU, and *i*vector. Based on the results in Table 4, it is evident that the *i*vector approach yields the best verification performance amongst the fine-tuned CNN, LSTM, and GRU, as well as the classical classifiers with an average percentage EER equivalent to 20.41%, compared to 33.00%, 26.91%, 29.65%, 30.5, 51%, 34% and 44.5% based on CNN, LSTM, GRU, GMM, KNN, SVM, and ANN, respectively. For this dataset, the *i*vector consists of 512 UBM components and 64 TVS rank. The total number of eigenvectors in the projection matrix and the number of dimensions in PLDA are both set to 16.

Table 4. Percentage EER and AUC scores using the CREMA database.

Equal Error Rate (EER) (%) CREMA								
	GMM EER AUC	KNN EER AUC	SVM EER AUC	ANN EER AUC	*i*vector EER AUC	CNN EER AUC	LSTM EER AUC	GRU EER AUC
Neutral	21.00 0.84	44.00 0.07	22.00 0.11	30.00 0.15	12.54 0.94	21.94 0.86	18.75 0.9	17.19 0.91
Anger	35.00 0.70	50.00 0.02	40.00 0.08	47.00 0.05	25.28 0.80	38.54 0.66	36.25 0.69	40.62 0.65
Happy	33.00 0.74	53.00 0.05	35.00 0.10	47.00 0.09	23.61 0.84	32.29 0.73	32.66 0.72	32.81 0.74
Sad	29.00 0.76	54.00 0.10	32.00 0.16	47.00 0.17	16.66 0.89	32.33 0.74	21.88 0.86	22.03 0.86
Fear	37.00 0.69	53.00 0.09	38.00 0.14	50.00 0.14	20.87 0.85	38.54 0.69	25.31 0.8	37.19 0.72
Disgust	28.00 0.78	52.00 0.09	37.00 0.14	46.00 0.14	23.52 0.85	34.38 0.72	26.61 0.8	28.07 0.81
Average	30.5 0.75	51 0.07	34 0.12	44.5 0.12	20.41 0.86	33.00 0.73	26.91 0.8	29.65 0.78

Among classical models, GMM yields the lowest EER values under neutral and emotional environments with a percentage EER equivalent to 21.00% compared to 44.00%, 22%, and 30% based on KNN, SVM, and ANN, respectively, using the neutral speech. When utterances are expressed with Anger, the percentage EER recorded is equivalent to 35.00%, 50.00%, 40.00%, and 47.0% based on GMM, KNN, SVM, and ANN, respectively.

Likewise, the AUC scores attained by each classifier indicate that the *i*vector outperforms all classifiers in achieving the largest AUC scores and the best ROC curves under both neutral and emotional environments, as portrayed in Figure 8. Figure 9 shows the DET curves based on each of the CNN, LSTM, GRU, *i*vector and GMM classifiers.

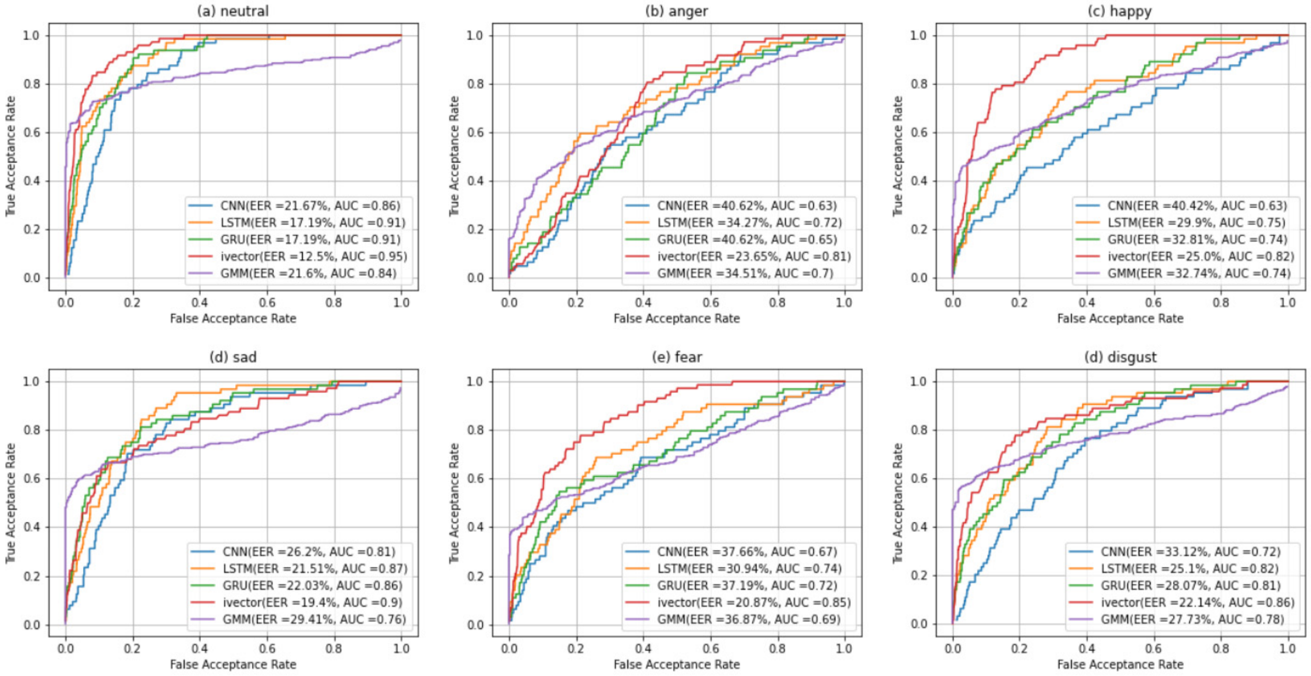


Figure 8. ROC plots using CREMA database based on CNN, LSTM, GRU, *i*vector, and GMM.

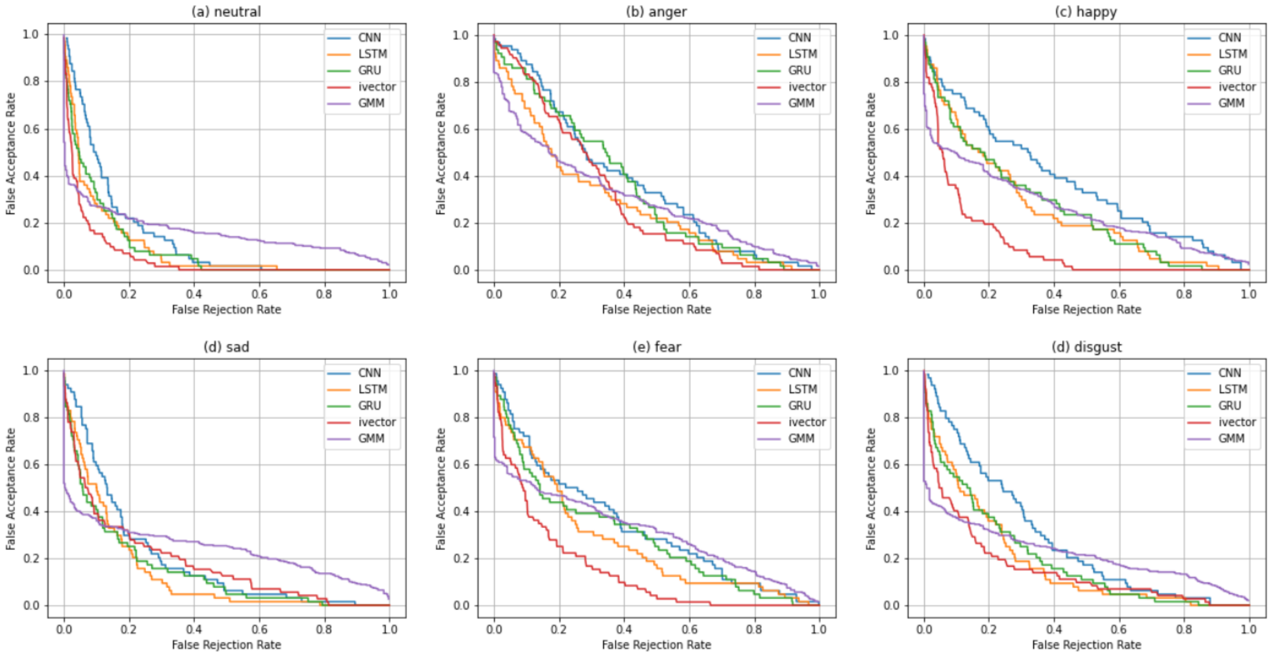


Figure 9. DET curves with each based on CNN, LSTM, GRU, *i*vector, and GMM models for the different emotional states using the CREMA database.

Based on the Wilcoxon results presented in Table 5, it is apparent that GMM is statistically different than KNN, SVM, and ANN when sentences are expressed both neutrally and emotionally. Moreover, the results demonstrate that *i*vector is statistically different than both LSTM and GRU models in neutral and all emotional categories. Nevertheless, the *i*vector model fails to be different from the CNN model in the Sad emotional state with alpha = 0.814.

Table 5. *p*-value using Wilcoxon test for classical and deep models using the CREMA database.

				Wilcoxon Test				
	KNN	SVM	ANN		CNN	GRU	LSTM	
Neutral	**0.000**	**0.000**	**0.000**	GMM	**0.045**	**0.000**	**0.000**	*i*vector
Anger	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	
Happy	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	
Sad	**0.000**	**0.000**	**0.000**		0.814	**0.000**	**0.000**	
Fear	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	
Disgust	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	

7.2. RAVDESS Database

The RAVDESS database is utilized in order to assess the speaker verification performance attained using classical and deep neural models. The percentage EER values of both classical, *i*vector, and the fine-tuned deep models for the RAVDESS speech dataset are tabulated in Table 6. Based on the results in Table 6, the EER values under the neutral speech are the lowest in comparison to each of the emotional categories. The *i*vector approach surpasses each of the fine-tuned CNN, LSTM, and GRU models. For this dataset, the *i*vector is composed of 32 UBM components and 16 total variability space rank. The total number of eigenvectors in the projection matrix and the number of dimensions in the probabilistic linear discriminant analysis (PLDA) are both set to 16.

Table 6. Percentage EER values of classical and deep models for the RAVDESS dataset.

Equal Error Rate (EER) (%) RAVDESS Dataset																
	GMM		KNN		SVM		ANN		*i*vector		CNN		LSTM		GRU	
	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC
Neutral	2.13	0.98	4.00	0.04	7.00	0.07	2.00	0.02	12.5	0.89	25.00	0.85	12.50	0.89	12.50	0.91
Anger	23.40	0.81	62.00	0.24	52.00	0.25	61.00	0.24	28.65	0.72	36.98	0.74	25.00	0.74	43.23	0.63
Happy	27.13	0.83	46.00	0.25	48.00	0.25	47.00	0.25	28.13	0.79	28.12	0.78	24.48	0.80	30.73	0.69
Sad	17.02	0.91	40.00	0.24	40.00	0.24	39.00	0.24	21.88	0.77	21.88	0.85	25.00	0.83	18.75	0.88
Fear	20.48	0.83	63.00	0.23	57.00	0.25	59.00	0.24	28.13	0.75	31.77	0.71	25.52	0.81	31.25	0.68
Disgust	22.40	0.80	64.00	0.23	65.00	0.23	60.00	0.24	19.79	0.89	30.21	0.82	31.25	0.77	37.50	0.70
Average	18.76	0.86	46.50	0.21	44.83	0.21	44.67	0.20	23.18	0.80	28.99	0.79	23.96	0.81	28.99	0.75

Furthermore, we conclude that the fine-tuned CNN, LSTM, and GRU models in addition to *i*vector approach perform poorly compared to the classical GMM model at emotional as well as neutral speech levels using the RAVDESS database. The obtained results are in accordance with our test results achieved using the collected Emirati dataset.

The results in Table 6 reveal that both LSTM and GRU models outperform the CNN in terms of average EER values as well as average AUC scores at the neutral state and all emotional categories except for the Sad and Disgust emotions. The Sad emotion records a percentage EER equal to 25.00% compared to 21.88% and AUC scores equivalent to 0.83 compared to 0.85 based on LSTM and CNN, respectively.

Figure 10 represents the ROC curves of the performance for each of the GMM, CNN, LSTM, GRU and *i*vector models using the RAVDESS database under the neutral and emotional environments. It is evident that the GMM model achieves the lowest percentage

EER values, and the largest AUC scores based on the plots (a), (b), (c), (d) and (e) for the neutral, Anger, Sad, Happy and Fear emotional states. However, *i*vector surpasses the GMM for the Disgust emotion with AUC scores equivalent to 0.89 and 0.8 based on *i*vector and GMM, respectively. Additionally, Figure 11 represents 10 DET curves of RAVDESS database based on CNN, LSTM, GRU, *i*vector, and GMM.

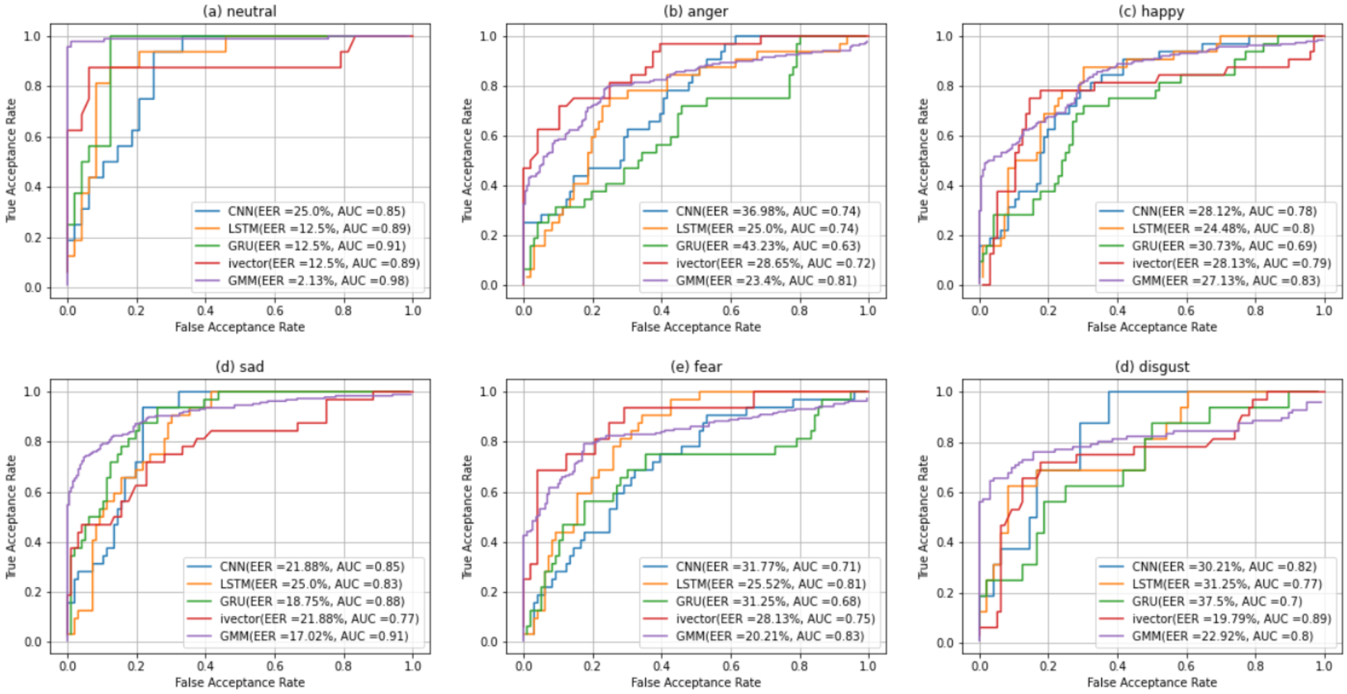


Figure 10. Plots of ROC curves for the RAVDESS database each based on the CNN, LSTM, GRU, and *i*vector.

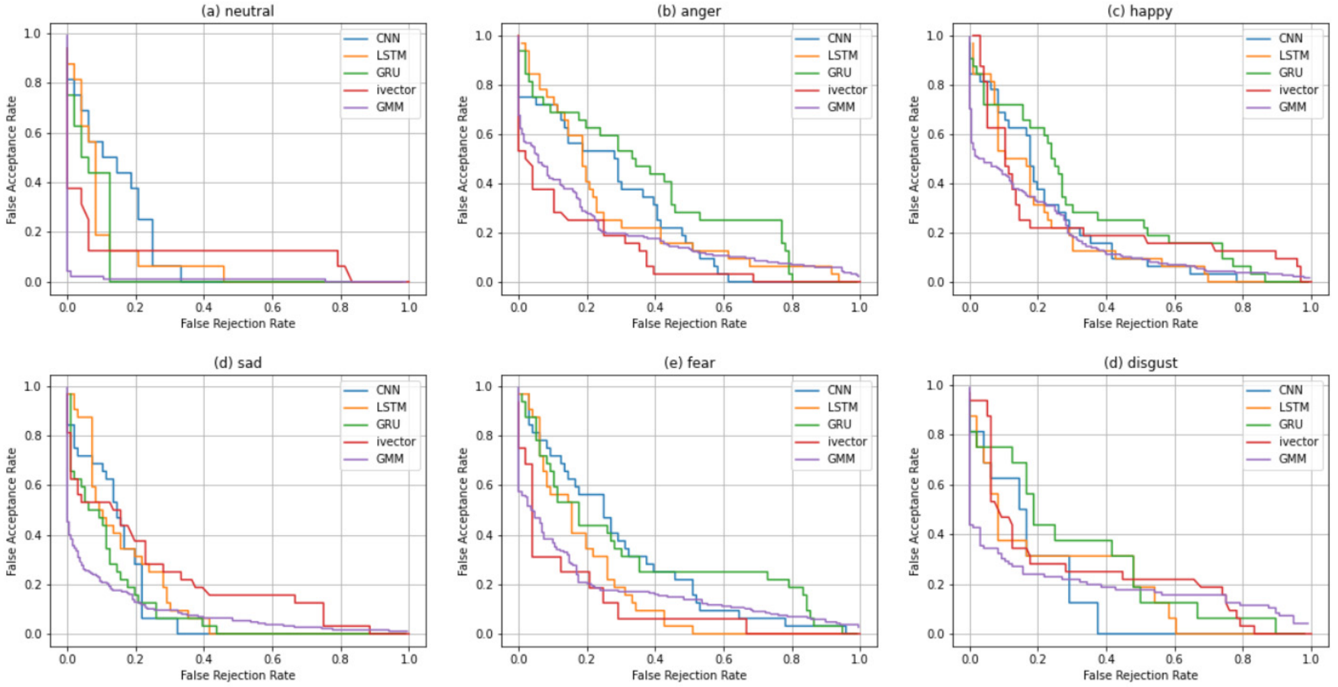


Figure 11. DET curves of RAVDESS database based on CNN, LSTM, GRU, *i*vector, and GMM.

In order to validate our results, we conducted the non-parametric Wilcoxon test to observe whether the winning model among the conventional classifiers is statistically different from other models based on 95% confidence interval. The results in Table 7 show that GMM is, indeed, statistically different from KNN, SVM, and ANN models, according to the Wilcoxon test.

Table 7. *p*-value using Wilcoxon test for the classical and fine-tuned deep models using the RAVDESS database.

				Wilcoxon Test				
	KNN	SVM	ANN		CNN	GRU	LSTM	
Neutral	**0.000**	**0.000**	**0.000**	GMM	**0.000**	**0.000**	**0.000**	*i*vector
Anger	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	
Happy	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	
Sad	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	
Fear	**0.000**	**0.000**	**0.000**		**0.000**	**0.000**	**0.000**	
Disgust	**0.000**	**0.000**	**0.000**		**0.004**	0.586	**0.000**	

Based on EER and AUC evaluation metrics, we deduce that the *i*vector is the winning model amongst deep learning models. To test statistical dissimilarity between the latter model and other models, we conducted the Wilcoxon test between *i*vector-CNN, *i*vector-LSTM, and between *i*vector-GRU, as shown in Table 7. From the achieved results, we observe that there is a significant difference between *i*vector and CNN and between *i*vector and LSTM. The *i*vector model fails to be different from the GRU model when utterances are expressed with Disgust (alpha = 0.586).

In order to validate our results using the Emirati dataset, we conducted statistical tests between the winning models amongst classical classifiers; the GMM and the other models (KNN, SVM, and ANN). Likewise, we perform the test between *i*vector and each of the LSTM, GRU, and CNN models. Based on the Kolmogorov–Smirnov test, we find that the error rates are not normally distributed. For this reason, we use the Wilcoxon test which is a non-parametric test.

In Table 6, it is clear that GMM outperforms the deep learning models as well as the *i*vector at the neutral and the emotional speech levels using Emirati and RAVDESS databases. This emphasizes the fact that deep models are not always the best option when dealing with machine learning, as per Saez et al. [4]. Zappone et al. [5] stated that to achieve high performance, deep learning networks necessitate relatively large datasets. For smaller datasets, classical algorithms could outperform deep learning. Therefore, in our specific task, deep classifiers do not provide statistically better verification results, even with hyperparameters finetuning, compared to the classical GMM, the winning classifier among classical methods. These results are consistent with previous work in [4] and [44].

7.3. Comparison with Other Related Work

Our work demonstrates the preeminence of the GMM model over all other models in terms of verification performance using Emirati and RAVDESS databases. Therefore, and as an additional experimental setup, we compare the error rates achieved by GMM with a variety of classifiers previously deployed in a text-independent speaker verification system [45] using the same set of databases.

As can be observed from Table 8, the performance of the GMM model, for the Emirati database, surpasses that of HMM1, HMM2 as well as HMM3 classifiers in the neutral speech with error rates equivalent to 1.43%, 11.5%, 9.6%, and 4.9%, respectively. The literature work did not address emotional speech.

Table 8. Equal error rates (%) for the GMM and different classifiers for the Emirati database.

Models	Neutral
HMM1, HMM2, HMM3 [45]	11.5, 9.6, 4.9
GMM [our winning model]	**1.43**

7.4. Computation Performance Study

Regarding the computation performance study, Table 9 shows the calculated average testing time, measured in seconds, of each machine learning classifier using ESD, CREMA, and RAVDESS databases over Google Colab. The results demonstrate that the fine-tuned CNN model has the fastest test time among deep classifiers and that the GMM is the slowest among classical classifiers across all databases.

Table 9. Testing time (in seconds) of classical, *i*vector, and deep models.

	Models	Emirati	RAVDESS	CREMA
Classical Classifiers	GMM	94.530	13.149	66.375
	KNN	35.365	3.446	11.898
	SVM	6.949	1.153	5.161
	ANN	19.231	2.455	7.203
Deep Classifiers	CNN	0.963	1.482	0.767
	LSTM	1.054	2.058	2.269
	GRU	0.980	1.526	2.203
	*i*vector	90.850	6.4542	34.6124

The *i*vector approach has the longest test time compared to the fine-tuned CNN, LSTM, GRU as well as SVM, KNN, and ANN yet it yields the best speaker verification results using the CREMA database. The GMM model attains the best verification results using Emirati and RAVDESS databases, yet it has the longest test time in comparison to all classical classifiers (KNN, SVM and ANN), *i*vector and deep models.

Among the deep network models, it is evident from the results reported in Tables 2, 4 and 6 that the LSTM model achieves the best verification performance, followed by the GRU, and then the CNN in both neutral and emotional talking environments across all datasets. Based on our results in Table 9, the GRU model optimizes more rapidly compared to the LSTM model while attaining equal error rates within a narrow margin to it. The CNN optimization is the fastest, yet, the model yields the least verification performance. This is because recurrent neural networks (LSTM and GRU), unlike the feed-forward networks (CNN), can use their internal memory in order to store information for a long period of time [29].

8. Conclusions, Limitations, and Future Work

This study focuses on performance appraisal of text-independent speaker verification tasks in emotional acoustic environments using three different datasets: Arabic Emirati-accented, English CREMA dataset, and RAVDESS database. We compare the following classical classifiers: GMM, SVM, KNN, and ANN with i-vector and three DNN-based systems, namely CNN, LSTM, and GRU. Hyperparameter tuning is applied to each of the CNN, LSTM, and GRU models using Grid Search. From the reported findings, we observe the superiority of the GMM classical classifier over the deep neural network and *i*vector classifiers in neutral environments as well as in emotional milieus using the private Emirati dataset as well as the RAVDESS databases. For the CREMA dataset, the *i*vector model yields the most accurate and the best verification results in terms of EER and AUC compared to all other models although at higher computational complexity in comparison to SVM, KNN, ANN, CNN, LSTM, and GRU models.

In terms of test time, the SVM is the fastest amongst all other classical classifiers, while the CNN attains the lowest computational performance compared to deep learning models.

However, CNN surpasses SVM with respect to the average percentage values of EER and AUC by a large margin.

In the future, we aim to scrutinize the performance of hybrid DNN-based classifiers, such as HMM-DNN, for speaker verification applications in stressful and emotional talking environments, and subsequently compare it with the verification performance of HMM alone, DNN alone, and with many classical classifiers.

Author Contributions: A.B.N.: Conceptualization, Methodology, Supervision, Writing—Original Draft, Writing—Review and Editing. I.S.: Data Curation, Methodology, Writing—Review and Editing. M.L.: Methodology, Writing—Review and Editing. A.E.: Methodology, Writing—Review and Editing, N.N.: Investigation, Writing—Original Draft. 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 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.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. 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]
2. Reynolds, D.A. An Overview of Automatic Speaker Recognition Technology. In Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, 13–17 May 2002; Volume IV, pp. 4072–4075.
3. Salehghaffari, H. Speaker Verification using Convolutional Neural Networks. *arXiv* **2018**, arXiv:abs/1803.0.
4. Baldominos, A.; Cervantes, A.; Saez, Y.; Isasi, P. A Comparison of Machine Learning and Deep Learning Techniques for Activity Recognition using Mobile Devices. *Sensors* **2019**, *19*, 521. [CrossRef] [PubMed]
5. Zappone, A.; Di Renzo, M.; Debbah, M. Wireless Networks Design in the Era of Deep Learning: Model-Based, AI-Based, or Both? *IEEE Trans. Commun.* **2019**, *67*, 7331–7376. [CrossRef]
6. Wan, V.; Campbell, W.M. Support vector machines for speaker verification and identification. In Proceedings of the Neural Networks for Signal Processing X. In Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No.00TH8501), Sydney, NSW, Australia, 11–13 December 2000; Volume 2, pp. 775–784.
7. Vivaracho-Pascual, C.; Ortega-Garcia, J.; Alonso, L.; Moro-Sancho, Q.I. A comparative study of MLP-based artificial neural networks in text-independent speaker verification against GMM-based systems. In Proceedings of the Eurospeech, Aalborg, Denmark, 3–7 September 2001; pp. 1753–1757.
8. Campbell, W.M.; Sturim, D.E.; Reynolds, D.A. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Process. Lett.* **2006**, *13*, 308–311. [CrossRef]
9. Chen, S.-H.; Luo, Y. Speaker Verification Using MFCC and Support Vector Machine. In Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, China, 18–20 March 2009.
10. Alarifi, A. Arabic text-dependent speaker verification for mobile devices using artificial neural networks. *Int. J. Phys. Sci.* **2012**, *7*, 1073–1082. [CrossRef]
11. Mahmood, A.; Alsulaiman, M.; Muhammad, G. Automatic Speaker Recognition Using Multi-Directional Local Features (MDLF). *Arab. J. Sci. Eng.* **2014**, *39*, 3799–3811. [CrossRef]
12. Taylor, S.; Hanani, A.; Basha, H.; Sharaf, Y. Palestinian Arabic regional accent recognition. In Proceedings of the 2015 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 14–17 October 2015; pp. 1–6. [CrossRef]
13. Chauhan, N.; Chandra, M. Speaker recognition and verification using artificial neural network. In Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 22–24 March 2017; pp. 1147–1149.

14. Wu, W.; Zheng, T.F.; Xu, M.-X.; Bao, H.-J. Study on Speaker Verification on Emotional Speech. In Proceedings of the NTERSPEECH, Pittsburgh, PA, USA, 17–21 September 2006.
15. Pillay, S.G.; Ariyaeeinia, A.; Pawlewski, M.; Sivakumaran, P. Speaker verification under mismatched data conditions. *Signal Process. IET* **2009**, *3*, 236–246. [CrossRef]
16. Shahin, I.; Nassif, A.B. Three-stage speaker verification architecture in emotional talking environments. *Int. J. Speech Technol.* **2018**, *21*, 915–930. [CrossRef]
17. Mittal, A.; Dua, M. Automatic speaker verification systems and spoof detection techniques: Review and analysis. *Int. J. Speech Technol.* **2022**, *25*, 105–134. [CrossRef]
18. Ferrer, L.; McLaren, M.; Brümmer, N. A speaker verification backend with robust performance across conditions. *Comput. Speech Lang.* **2022**, *71*, 101258. [CrossRef]
19. Liu, T.; Das, R.K.; Lee, K.A.; Li, H. Neural Acoustic-Phonetic Approach for Speaker Verification with Phonetic Attention Mask. *IEEE Signal Process. Lett.* **2022**, *29*, 782–786. [CrossRef]
20. Bhattacharya, G.; Alam, J.; Kenny, P. Deep speaker embeddings for short-duration speaker verification. In Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, Stockholm, Sweden, 20–24 August 2017; Volume 2017, pp. 1517–1521. [CrossRef]
21. Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process. A Rev. J.* **2000**, *10*, 19–41. [CrossRef]
22. Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 788–798. [CrossRef]
23. Kenny, P.; Ouellet, P.; Dehak, N.; Gupta, V.; Dumouchel, P. A Study of Inter-Speaker Variability in Speaker Verification. *IEEE Trans. Audio Speech Lang. Process.* **2008**, *16*, 980–988. [CrossRef]
24. Garcia-Romero, D.; Espy-Wilson, C. Analysis of i-vector Length Normalization in Speaker Recognition Systems. In Proceedings of the Interspeech, Florence, Italy, 28–31 August 2011; pp. 249–252.
25. Rupesh Kumar, S.; Bharathi, B. Generative and Discriminative Modelling of Linear Energy Sub-bands for Spoof Detection in Speaker Verification Systems. *Circuits Syst. Signal Process.* **2022**, *41*, 3811–3831. [CrossRef]
26. Alam, M.J.; Kinnunen, T.; Kenny, P.; Ouellet, P.; O'Shaughnessy, D. Multi-taper MFCC Features for Speaker Verification using I-vectors. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, Waikoloa, HI, USA, 11–15 December 2011; pp. 547–552.
27. Chen, L.; Lee, K.A.; Chng, E.; Ma, B.; Li, H.; Dai, L.-R. Content-aware local variability vector for speaker verification with short utterance. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5485–5489.
28. Zhu, Y.; Ko, T.; Snyder, D.; Mak, B.; Povey, D. Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification. In Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech, Hyderabad, India, 2–6 September 2018; pp. 3573–3577.
29. Mobiny, A.; Najarian, M. Text-Independent Speaker Verification Using Long Short-Term Memory Networks. *arXiv* **2018**, arXiv:1805.00604.
30. Hourri, S.; Nikolov, N.S.; Kharroubi, J. Convolutional neural network vectors for speaker recognition. *Int. J. Speech Technol.* **2021**, *24*, 389–400. [CrossRef]
31. Shahin, I.; Nassif, A.B.; Nemmour, N.; Elnagar, A.; Alhudhaif, A.; Polat, K. Novel hybrid DNN approaches for speaker verification in emotional and stressful talking environments. *Neural Comput. Appl.* **2021**, *33*, 16033–16055. [CrossRef]
32. Mohammed, T.S.; Aljebory, K.M.; Abdul Rasheed, M.A.; Al-Ani, M.S.; Sagheer, A.M. Analysis of Methods and Techniques Used for Speaker Identification, Recognition, and Verification: A Study on Quarter-Century Research Outcomes. *Iraqi J. Sci.* **2021**, *62*, 3256–3281. [CrossRef]
33. Chen, Y.H.; Lopez-Moreno, I.; Sainath, T.N.; Visontai, M.; Alvarez, R.; Parada, C. Locally-connected and convolutional neural networks for small footprint speaker recognition. In Proceedings of the Interspeech, Dresden, Germany, 6–10 September 2015.
34. 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]
35. Heigold, G.; Moreno, I.; Bengio, S.; Shazeer, N. End-to-end text-dependent speaker verification. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20–25 March 2016.
36. Cao, H.; Cooper, D.G.; Keutmann, M.K.; Gur, R.C.; Nenkova, A.; Verma, R. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans. Affect. Comput.* **2014**, *5*, 377–390. [CrossRef] [PubMed]
37. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef] [PubMed]
38. Kumar, D.S.P. Feature Normalisation for Robust Speech Recognition. *arXiv* **2015**, arXiv:abs/1507.0.
39. Li, L.; Wang, D.; Zhang, Z.; Zheng, T.F. Deep Speaker Vectors for Semi Text-independent Speaker Verification. *arXiv* **2015**, arXiv:1505.06427.
40. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. Librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; pp. 18–25.

41. Reynolds, D.A.; Rose, R.C. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 72–83. [CrossRef]
42. Pulgar, F.J.; Charte, F.; Rivera, A.J.; del Jesus, M.J. AEkNN: An AutoEncoder kNN-Based Classifier With Built-in Dimensionality Reduction. *Int. J. Comput. Intell. Syst.* **2018**, *12*, 436. [CrossRef]
43. Arce-Medina, E.; Paz-Paredes, J.I. Artificial neural network modeling techniques applied to the hydrodesulfurization process. *Math. Comput. Model.* **2009**, *49*, 207–214. [CrossRef]
44. Saez, Y.; Baldominos, A.; Isasi, P. A Comparison Study of Classifier Algorithms for Cross-Person Physical Activity Recognition. *Sensors* **2016**, *17*, 66. [CrossRef]
45. Shahin, I. Emirati speaker verification based on HMMls, HMM2s, and HMM3s. In Proceedings of the 2016 IEEE 13th International Conference on Signal Processing (ICSP), Chengdu, China, 6–10 November 2016; pp. 562–567.