Detection and classification of Emotion Recognition System for TESS and Crema-d Audio Datasets Using Hybrid Deep Learning Architecture*

Hammou Djalal Rafik[1,*,†]

[1]Djillali Liabes University, BP 89 22000 Sidi Bel Abbes, Algeria, Faculty of Exact Science, Department of computer sciences.

Abstract

Humans communicate their desires through spoken language, which expresses various emotions. This process has led to the development of speech recognition systems, where machine learning enables computers to recognize and analyze vocal cues to interpret emotions, resulting in application creation focused on human-machine interaction. Advancements in technology, the evolution of artificial intelligence, and the influence of deep learning via CNN architectures have propelled research in emotion recognition systems forward. In this paper, we evaluated our method for detecting and classifying emotions in two architectural models (Model-A and Model-B) that utilized Mel-frequency cepstral coefficients to extract features from audio files. The experiments were conducted using the TESS and Crema-d audio file databases. The outcomes are promising, showing an accuracy of 54,07% for Model-B with the Crema-d dataset and 98,92% for Model-A with the TESS dataset.

Keywords

Speech, Architecture, Emotion, Accuracy, Recognition

1. Introduction

Speech is a means of communication with the outside world. Each human being has his speech, and it is thanks to natural language that individuals can discuss and understand each other. Speech is unique and expressed through a well-defined language addressed to an interlocutor (oneself). It allows us to express needs such as feelings, suffering, aspirations, observations, and the formulation of requests. It also allows us to constitute different natural languages and dialects.

Speech is the most popular tool of expression because it is easier to speak than to write or make a diagram. Nonetheless, The process of producing speech, from the brain to the articulation of the mouth using the vocal cords, is intricate. This difficulty makes the automation of speech by machine complicated [1].

The system of the phonatory apparatus of speech is an acoustic mechanism that differs from other sensory devices. It comprises elements such as the vocal cords, the oral and nasal cavities, the air, the nervous system, and the tongue and lips [3].

These elements are described in detail (see Fig.1):

- **The throat (pharynx)**, which is the largest surface of the neck and head, is made up of three primary parts: the hypopharynx, the nasopharynx, and the oropharynx.

Proceedings of the International IAM'24: International Conference on Informatics And Applied Mathematics, December 4-5, 2024, Guelma, Algeria

*Corresponding author.

†These authors contributed equally.

r_hammou@esi.dz (H. D. Rafik)

0000-0002-0038-0424 (H. D. Rafik)

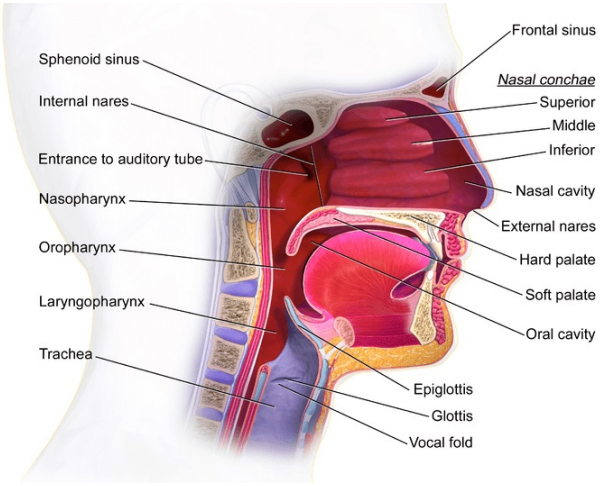

Figure 1: The phonatory anatomy system of the human speech mechanism [2].

- **The hypopharynx**, found behind the pharynx, is a section of the throat. It has a flap-like structure that acts as a lid for the larynx and shuts when swallowing to prevent food and liquid from entering the trachea.

- **The nasopharynx** is located behind the nose, and its purpose is to transfer the air to breathe down through the voice box of the throat and into the lungs.

- **The oropharynx** is situated posterior to the mouth and is responsible for conveying food and liquid to the digestive tract and stomach.

- **The larynx** is the voice box, and it is efficient because it protects the lungs from food, drink, and foreign bodies and acts as a corridor for air from the nasopharynx.

- **The epiglotis** protects the lungs and only lets air pass through.

- **The vocal tract** extends from the vocal cords to the lips and measures approximately 17.5 cm in length. It consists of two types of ducts: a pharyngeal cavity and an oral cavity.

- **The vocal cords** open when breathing and close when swallowing or producing voice sounds. Understanding the vocal process is straightforward. It starts with the movement of air passing through the two vocal cords (see Fig.2), which are soft and vibrate as air flows through them, producing 100 to 1 000 vibrations per second.

- **Articulators** are the tongue, lips, jaws, mouth, etc. These articulators allow for the modification of the shape of the vocal tract.

- **The nasal cavity** forms a part of the vocal tract and is situated beneath the velum.

Speech production system:

The human speech production system involves multiple stages before the voice is produced. It operates rapidly and is complex, relying on the respiratory system, the phonatory system, and the articulatory system, all working together. These systems collectively create speech and sound. The respiratory system manages the intake and release of air in the lungs. Inhalation involves the diaphragm

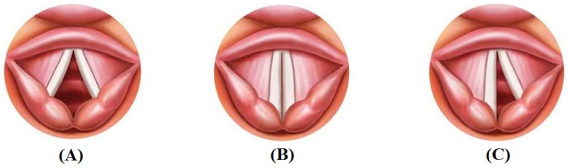

Figure 2: The different positions of the vocal cords in humans: A) vocal cords opening position, B) vocal cords closing position, c) vocal cords in closed position (semi-paralysis) [4].

lowering and the intercostal muscles facilitating vacuum creation in the lungs for air intake. Exhalation entails the diaphragm relaxing, allowing air to escape and produce sounds. Air then passes through the larynx, containing muscles and cartilages called the vocal cords, with the space between them known as the glottis. The vocal cords, capable of rapid opening and closing, can reach up to 400 movements per second in children. The articulators, situated between the arytenoids, enable the vocal cords to move. The articulation process starts with air leaving the larynx, passing through the pharynx before being modulated by resonators like the lips, tongue, mandible, and velum, which impart characteristics to the sound (air flowing freely results in a vowel, while encountering an obstacle yields a consonant).

Analysis of a person's emotions is an important area of research in Natural Language Processing (NLP). For example, it allows us to recognize the individual's anxiety. In certain cases, we use the analysis of emotions to diagnose diseases. NLP is a branch of deep learning research. Because deep learning has revolutionized the world of artificial intelligence, such as in medical research for early diagnosis of diseases such as COVID-19 [5] or in the field of biometric recognition [6].

The strategy of our approach consists of:

1. Making a consistent bibliographic study in the speech emotion recognition field.
2. Extracting adequate scientific knowledge to apply them in our approach.
3. Using quality and available datasets that are made available to the scientific community.
4. Applying our approach to deep neural networks of the LSTM type using the Mel-frequency cepstral coefficients (MFCC).
5. Testing our experiments on two datasets, which are TESS and CREMA-D.
6. Evaluating the results obtained with the evaluation parameters: number of trained and untrained parameters, max accuracy val, accuracy test, score, loss, Precision, Recall, F1-score, and Support.
7. Establishing a Comparative Table between our Results and those of the state of the art.

The rest of our paper is structured as follows: Section 2 is devoted to a literature review of speech recognition systems with the individual analysis of emotions and the datasets used. Section 3 is dedicated to the implementation and methodology of the application of our approach with the use of the architecture of the deep neural network of the LSTM type by focusing on the use of Mel-frequency cepstral coefficients (MFCC). Section 4 concerns the experimental results obtained on the datasets with the evaluation metrics used to validate our approach. In conclusion, we will wrap up with a section discussing the primary obstacles and future research directions about speech recognition systems (SER).

2. Related word

In May 2020, De Pinto G. et al. [7] developed a deep neural network for speech recognition with eight emotions. The developed model is based on convolutional neural networks (CNN). The tests were carried out on the RAVDESS database, and the results obtained are of the order of the F1 score of 91,00%, the emotion class Angry with a score of 95,00%, and the class Sad with 87,00%. In February

2022, Puri, T. et al. [8] built a hybrid architecture (LSTM + CNN) using the Hidden Markov Model and Deep Neural Networks (DNN) for a speech recognition system based on eight emotions. They applied their approach to the RAVDESS dataset with a three-branch division strategy (for males and females); the first branch concerns emotions in two positive classes (male and female); the second branch is divided into three emotion classes (positive, negative, and neutral); and finally, the last branch is divided into eight different emotion classes. They obtained an accuracy of 98,00%. In September 2022, Gupta, M. V. et al. [9] developed a computer system to detect stress, which has behavioral, emotional, and physical effects. The authors proposed a cascaded RNN-LSTM architectural system and applied their approach to the RAVDESS dataset and obtained an accuracy of 91,00%. In October 2022, Ullah, S. et al. [10] developed an architecture model based on speech recognition of emotions with human-machine interaction. The researchers proposed a one-dimensional CNN (convolutional neural network). They tested it on a combined emotional dataset (Crema-D, Ravdess, Savee, and Tess) with a feature set and classifier ZCR+energy+entropy of energy+RMS+MFCC. The proposed model obtained an accuracy of 92,62%. In November 2022, Vijayan, D. M. et al. [11] proposed two architecture models for speech recognition based on deep learning. The aim is to analyze the emotions of speech and classify them while extracting the spatial and temporal technical characteristics. The first model is a combined CNN-LSTM architecture, and the second is a CNN-Transform encoder architecture. The RAVDESS database was used for the experimentation.The first model obtained an accuracy of 74,00% while the second achieved a better high precision with 82,00%. In January 2023, Ahmed, R. et al. [12] contributed to the implementation of a speech recognition system based on emotion analysis and feature extraction motivation employed on Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). The authors deployed three different architectures: the first architecture uses 1D CNN followed by FCN networks (Model-A), the second architecture uses 1D CNN followed by LSTM-FCN networks (Model-B), and the third architecture uses 1D CNN followed by GRU-FCN networks (Model-C). They also used data augmentation by adding Gaussian noise. The experiments were carried out on five databases, and they obtained very accuracy with 95,62% for RAVDESS, 99,46% for TESS, 90,47% for CREMA-D, 95,42% for EMO-DB, and 93,22% for SAVEE. In March 2023, Shah, N. et al. [13] contributed to creating a powerful computational model based on the Mel frequency cepstral coefficients by combining three datasets: RAVDESS, TESS, and SAVEE. The model uses two classifiers, Random Forest and Boosting Ensemble, and the prediction accuracy results are 86,30% for the first classifier and 85,80% for the second classifier. Another learning model was experimented with on the dataset and obtained an accuracy of 75,00%. In July 2023, Bhawesh, K.et al. [14] developed four neural networks for emotion speech recognition using features such as MFCC, chrominance, spectral attenuation, etc. The architecture models used are LSTM, CNN, MLP, and Random Forest models. The experiments were carried out on a combined data set (SAVEE, RAVDESS, CREMA-D, and TESS), and the results obtained with accuracies are of the order of 57,50% for the 1st model, 75,80% for the 2nd model, 59,60% for the 3rd model, 67,80% for the last model. In December 2023, Tyagi, S. et al. [15] developed a computer application based on vocal emotions that allows to understand and identify human emotions from speech. The proposed prototype uses the LSTM architecture with a GWO optimization on a combined dataset with the following databases SAVEE, TESS, EMO-DB, and RAVDESS, and obtained a good accuracy of 65,47% (SAVES), 99,93% (TESS), 78,00% (EMO-DB), and 87,00% (RAVDESS) respectively. In February 2024, Lata, S. et al. [16] experimented with two neural network architectures. The first model is a hybrid architecture of a convolutional neural network (CNN) and long short-term memory (LSTM). The second model consists of an architecture composed of MFCC+LSTM. The aim is to exploit a stack of depth layers in linear form to improve the accuracy of the speech sentiment recognition system. The tests were carried out on the TESS database, and the results obtained are encouraging, with an accuracy of 98,00% for the first model and 96,00% for the second model. In March 2024, Yuan, Z. et al. [17] developed a computer module that relies on a speech recognition system, and more particularly on identity information (this method harms the model generalization). The authors proposed a DTNet-type neural network to dissociate acoustic features from emotional features. The experiments were tested on two databases. They obtained an accuracy of 74,80% for IEMOCAP and 95,00% for Emo-DB. In April 2024, Islam A. et al. [18]

built a consistent computing system for speech emotion detection and enhancement. In this context, the authors experimented with their approaches by merging three databases: RAVDESS, TESS, and CREMA-D. They also proposed a hybrid architectural model using CNN and BiLSTM for the eight emotions. The proposed model is based on root mean square energy (RMSE), zero crossing rate (ZCR), and Mel frequency cepstral coefficient (MFCC). The proposed model achieved an accuracy of 97,80%. In June 2024, Akinpelu, S. et al. [19] proposed a computer application of speech recognition based on machine learning to detect emotions. This system is based on the principle of the Vision Transformer (ViT) model. It allows the capture of the characteristics in the images that are adequate indicators of emotional states from the input data of the mel spectrogram introduced into the model. The TESS (Toronto English Speech Set) and EMODB (Berlin Emotional Database) were used for the experiments. The results are satisfactory, as they obtained an accuracy of 98,00% (TESS), 91,00% (EMO-DB), and 93,00% (TESS+EMO-DB). In August 2024, Hossain, I. et al. [20] implemented a hybrid model to extract information and improve prediction accuracy with probability calculated. The model uses the convolutional neural network (CNN) architecture to extract features from the speech spectrogram. After that, the long short-term memory processes the features (LSTM). The authors used a KNN classifier to classify emotions and make predictions. They conducted experiments using the TESS database and achieved an accuracy of 98,21%.

The proposed approach is based on two architectural models, and one of them has given excellent results and is even better than some methods in the literature. The contributions of our model are defined as follows:

- The development of a new neural network architecture based on the long short-term memory (LSTM) network, which is dedicated to the classification and analysis of human emotions.
- The proposed LSTM structure features nine layers, beginning with an input layer for data and ending with an output layer that uses the softmax activation function to classify seven distinct emotions. It is composed of seven hidden layers that implement optimization methods, such as dropout and dense layers.
- The suggested LSTM model achieved an accuracy of 98,92% for classifying emotions using the TESS database, including pleasant (ps), anger, happiness, disgust, fear, sadness, and neutral.
- The LSTM model is lightweight and contains about 307 655 parameters. It allows for accelerated learning and reduces the computation time for emotion class prediction.

3. Methodology and Implementation

3.1. Data collection

The basis of a speech-emotion recognition system is the quality of the audio file database because a good-quality dataset implies an efficient and robust system. If the audio file database contains noise, it must go through refining preprocessing to clean it, and this is done through the process of filtering, encapsulation, and integration [21]. A SER goes through the data collection phase, preprocessing, feature extraction, feature selection, classification, and recognition [22]. The data collection stage is the most sensitive phase of the system (Collect the data thoroughly and ensure that it meets high-quality standards). The University of Marburg in Slovenia was the birthplace of the first database for an SER system, as per the literature [23]. It consists of six types of emotions, and the audio files are in MPEG-4 format [24]. The number of utterances in the dataset is 186 for each emotional category.

3.2. Data preprocessing

Before entering the data into the neural network for learning, it is necessary to go through a preprocessing for feature extraction of the audio files and transform them into mathematical

coefficients for the classification and recognition of emotions. For this, we need the following parameters:

- **Mel scale:** It is a mathematical scale that allows the height to be measured (acoustic scale), so the unit of measurement is the Mel, with a sound characteristic of high or low.
- **Frequency:** represents the number of oscillations (vibrations) of the sound per second, with the unit of measurement Hertz.
- **Chromagram:** represents the intensity of the audio signal at a given moment, it is made up of a chrome vector (size of 12 dimensions with 12 semitones of the chromatic scale).
- **Pitch:** represents the vibration frequency corresponding to the sounds.
- **Fourier transform:** allows the decomposition of the periodic signals and is the link between the temporal signal and the frequency representation.

Subsequently, the Mel spectrogram will be used for individual identification with speech recognition and emotional states. It is represented by an image whose x-axis represents time, and the y-axis represents the frequency with the application of a logarithmic scale in the dot diagram [25].

The Mel-frequency cepstral coefficients (MFCC) are used for speech recognition, and this is done by following the steps above:

1. Calculate the Fourier transform of the signal.
2. Establish a power mapping of the spectrum obtained previously on the Mel scale using interleaved triangular windows.
3. Collect the power recordings for each Mel frequency.
4. Observe the discrete cosine transform (DCT) of the list of mel log powers as if it were a signal.
5. The amplitudes of the resulting spectrum are called MFCC.
6. For each audio recording, 40 MFCC values [26] are used as input data into the neural network.

3.3. Architectural model

The proposed neural network Model-A is of the LSTM (language short-term memory) type [27], with as input the first audio recording vector and these 40 values of the MFCC coefficients. The network comprises an LSTM input layer of 256 neurons and an output layer with the softmax activation function, and these seven neurons correspond to emotions (selection probabilities). The network has seven hidden layers with four dropout layers of 0,2 %, three dense layers with the relu activation function, a dense layer of 128 neurons, a dense layer of 64, and another thick layer of 32 neurons (see Fig.3).

Model-B is a hybrid CNN-LSTM architecture [28], and it is a combined neural network architecture with a total of 18 layers. The input layer consists of 32 neurons with a relu activation function and, as input data, a vector that contains 40 MFCC coefficient values. An output layer with the softmax activation function and the six neurons (correspond to the emotions of speech recognition). The network is composed of 16 hidden layers with five dropout layers (2 layers of 30% and three layers of 50%), two convolution layers of 32 and 64 neurons with a relu activation function, two max-pooling layers of two dimensions, a flattened layer, two LSTM layers with 128 neurons, a dense layer of 128 neurons with the relu activation function, and finally a batch normalization layer (see Fig.4).

3.4. Hardware and Software

During the experimentation phase in both neural network architecture models we used the following hyperparameters (see Table 1).

We also used hardware and software to execute our approach (see Table 2).

Layer (type)	Output Shape	Param
lstm (LSTM)	(None, 256)	264,192
dropout (Dropout)	(None, 256)	0
dense (Dense)	(None, 128)	32,896
dropout_1 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8,256
dropout_2 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 32)	2,080
dropout_3 (Dropout)	(None, 32)	0
dense_3 (Dense)	(None, 7)	231

Total params: 307,655 (1.17 MB)
Trainable params: 307,655 (1.17 MB)
Non-trainable params: 0 (0.00 B)

Figure 3: Architecture of neuronal network LSTM for Dataset TESS Emotions.

Layer (type)	Output Shape	Param #
conv2d_161 (Conv2D)	(None, 98, 11, 32)	320
max_pooling2d_161 (MaxPoolin	(None, 49, 5, 32)	0
dropout_381 (Dropout)	(None, 49, 5, 32)	0
conv2d_162 (Conv2D)	(None, 47, 3, 64)	18496
max_pooling2d_162 (MaxPoolin	(None, 23, 1, 64)	0
dropout_382 (Dropout)	(None, 23, 1, 64)	0
time_distributed_48 (TimeDis	(None, 23, 64)	0
lstm_128 (LSTM)	(None, 23, 128)	98816
dropout_383 (Dropout)	(None, 23, 128)	0
lstm_129 (LSTM)	(None, 128)	131584
dropout_384 (Dropout)	(None, 128)	0
dense_164 (Dense)	(None, 128)	16512
batch_normalization_64 (Batc	(None, 128)	512
dropout_385 (Dropout)	(None, 128)	0
dense_165 (Dense)	(None, 64)	8256
batch_normalization_65 (Batc	(None, 64)	256
dropout_386 (Dropout)	(None, 64)	0
dense_166 (Dense)	(None, 6)	390

Total params: 275,142
Trainable params: 274,758
Non-trainable params: 384

Figure 4: Architecture of neuronal network CNN-LSTM for Dataset Crema-D.

4. Results and discussion

Applying our approach requires the use of two datasets.

Dataset 1: The TESS dataset is a database that contains audio files for seven emotions (pleasant (ps), anger, happiness, disgust, fear, sadness, and neutral) [29]. These recordings were made by two actresses aged 26 and 64 recruited in the Toronto area as English-speaking actresses with university studies and musical training. The dataset contains 2 800 audio files (These sound recordings were based on the Northwestern University Hearing Test Number 6) with a set of 200 words relating to a sentence that says "Say the word .....".

Dataset 2: The CREMA-D data set consists of 7 442 original recordings featuring 91 actors, with 48

Table 1
The values and properties of hyper-parameters.

N°	Property	Values
1	Number class	6 and 7
2	Batch-size	64
3	Epochs	100
4	Learning rate	0.0001
6	Optimizer	Adam
7	Beta 1	0,9
8	Beta 2	0,999
9	Epsilon	1e-06
10	Loss	Categorical crossentropy

Table 2
Hardware and software characteristics.

Hardware and Software Used	Characteristics detail
Programming language	Python 3.7.10
Mathematical Library	Numpy 1.19.5
Memory (RAM)	32 GB
Operating Système	x86-64 GNU/Linux
Deep Learning Framework	Keras 2.4.3
Graphics Card (GPU)	NVIDIA Tesla T4, 16 GB GDDR6, NVIDIA CUDA Cores 2560, PCIe Gen 3.0 x16.
Architecture	Tensorflow 2.4.1
Numerical software Library	Panda 1.1.5
Notebook	Jupyter
Processor (CPU)	Intel(R) Xeon(R) CPU @ 2.00GHz
Visualization Library	Matplotlib 3.4.3

male and 43 female participants aged between 20 and 74 from diverse racial and ethnic backgrounds such as Hispanic, African, Caucasian, Asian, American, and Unspecified [30]. The actors delivered 12 specific sentences while expressing six different emotions (Happy, Neutral, Sad, Fear, Anger, and Disgust) at four varying levels (Low, Medium, High, and Unspecified).

To evaluate the results obtained from the experiments of our approach on the two datasets, we used the following criteria: number of trained parameters (train param), number of untrained parameters (untrain param), total number of parameters (total param), top loss (validation), top accuracy (validation), score (test), and accuracy (test).

On the other hand, we have taken specific evaluation criteria to measure the performance of our approach on the image corpus. F1-score, Recall, Precision, and Support parameters are calculated based on the TP (true positive), TN (true negative), FP (false positive), and FN (false negative) [31]. These parameters are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

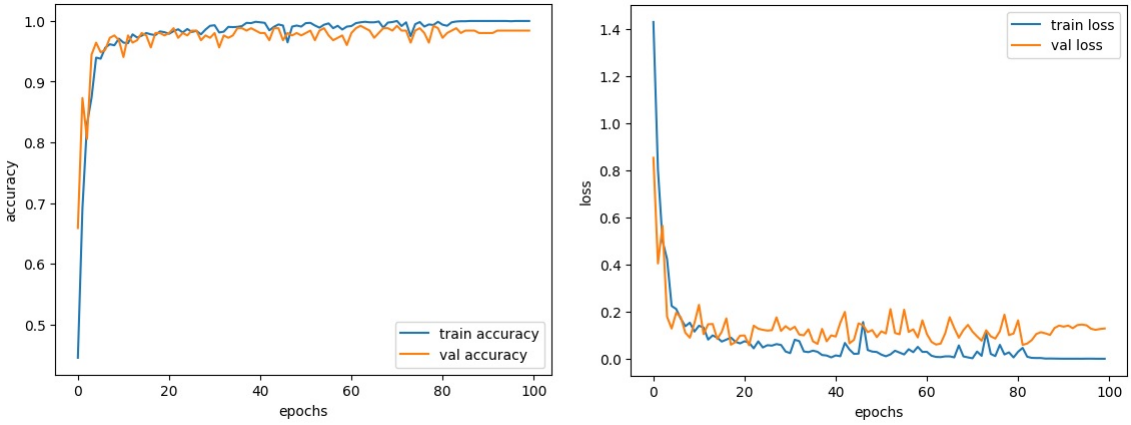

Figure 5: Result of experimentation of accuracy and loss with LSTM architecture (dataset TESS)

.

$$Support \quad : \sum \quad instances \quad in \quad each \quad class. \tag{4}$$

Table 3
Results of experiments with the two architectural models.

Architecture	Total param	Train param	Non-train param	Accuracy (test) (%)	Score (test)	Top Accuracy (val) (%)	Top loss (val)
LSTM	307 655	307 655	0	98,92	0,0561	99,21	0,0573
CNN+LSTM	274 758	275 142	384	54,07	1,4708	55,02	1,2211

Figure 5 displays the outcomes of the accuracy and loss experiments conducted on the TESS database using the LSTM architectural model. The diagram shows that the training and validation accuracy curves closely align, and the same pattern is observed for the loss, indicating that the neural network has effectively learned with minimal risk of overfitting.

Figure 6 illustrates the outcomes of the confusion matrix from experiments conducted on the TESS database for identifying the emotions: anger, happiness, pleasant, disgust, fear, sadness, and neutral. From this matrix, we observe that the accuracy for the emotion categories fear, sad, angry, disgust, and happy is notably higher than that for the other categories; conversely, the neutral and pleasant emotions exhibit a lower accuracy. This indicates that the model maintains an almost uniform distribution across all emotion categories without clear distinctions. Furthermore, the overall accuracy of the emotion recognition system stands at 98,92 %.

Table 3 shows the results of the experiments of the two proposed architectural models. The LSTM model gave excellent results, with an accuracy of 98,48% and a score of 0,0561, while the CNN-LSTM model gave average results, with an accuracy of 54,07% and a score of 1,4708.

Table 4 displays the results from the tests conducted on the TESS emotion database, including the evaluation metrics of recall, precision, F1 score, and support. It is noteworthy that the emotion categories happy, disgust, neutral, pleasant, and sad exhibit high precision, as illustrated in Figure 6, while the emotion classes angry and fear show slightly lower precision. Additionally, both the micro average and macro average precision are 0,99.

Table 5 showcases a comparison of our method's outcomes on both databases against those found in the existing literature.

Table 4
Results and predictions emotion classes from the database TESS with the model LSTM.

Table Emotion	Evaluation parameters			
	Precision	Recall	F1-score	Support
Fear	0,95	0,98	0,96	41
Angry	0,98	1,00	0,99	42
Disgust	1,00	1,00	1,00	43
Neutral	1,00	1,00	1,00	42
Sad	1,00	1,00	1,00	36
Pleasant (ps)	1,00	0,94	0,97	32
Happy	1,00	1,00	1,00	44
Micri avg	0,99	0,99	0,99	280
Macro avg	0,99	0,99	0,99	280
Weighted avg	0,99	0;99	0,99	280
Sample avg	0,99	0,99	0,99	280

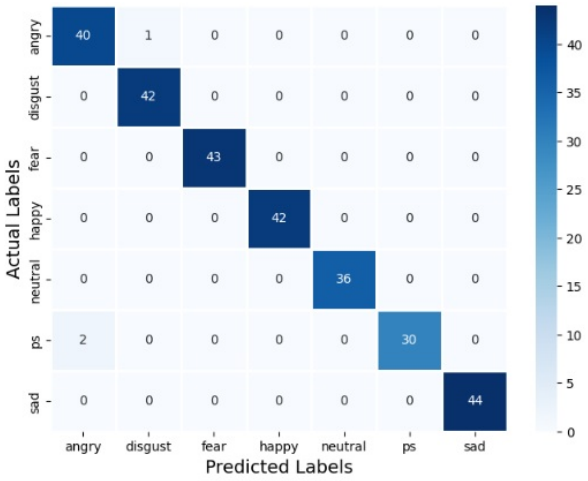

Figure 6: Confusion Matrix for Dataset Tess Emotions (Model-A).

5. Conclusion

The application of our approach and the experiments and tests carried out on the two emotion databases Crema-d and TESS with the two architectural models LSTM and CNN + LSTM demonstrate the results obtained with the CNN + LSTM model on the dataset Crema-d with an accuracy of 54,07% and the LSTM model on the TESS dataset with an accuracy of 98,92%. This also implies that the proposed model is so effective with the LSTM architecture with all the classes of emotions that there is an almost uniform distribution in the accuracy, which is very adaptable for an emotion recognition system.

The introduction of deep learning has significantly enhanced research in areas like medicine [32], biometrics [33], and even more speech recognition systems, especially through the use of convolutional neural network (CNN) architectures. We aim to enhance our model by utilizing a more extensive dataset of emotions, and additionally, we will analyze speech to determine whether the patient is experiencing a psychological disorder.

Table 5
A comparison table between our results and those of the state of the art.

Author	Year	Nbr of class	Nbr of Files	dataset	Architecture	Accuracy
De Pinto, G. et al. [7]	May 2020	8	7356	RAVDESS	CNN-MFCC	91,00 %
Puri, T., et al. [8]	February 2022	8	7356	RAVDESS	CNN-LSTM+DNN	98,00 %
Gupta, M. V. et al. [9]	September 2022	8	7356	RAVDESS	RNN-LSTM	91,00 %
Ullah, S. et al. [10]	October 2022	7	18078	Crema-D+Ravdess+Savee+Tess	CNN	92,62 %
Vijayan, D. M. et al. [11]	November 2022	8	7356	RAVDESS	CNN-LSTM	74,00 %
					CNN-Transform	82,00 %
Ahmed, R., et al. [12]	January 2023	8	7356	RAVDESS	Model-A+Model-B+Model-C	95,62 %
		7	2800	TESS		99,46 %
		6	7442	CREMA-D		90,47 %
		7	535	EMO-DB		95,42 %
		7	480	SAVEE		93,22 %
Shah, N. et al. [13]	March 2023	7	10636	RAVDESS+TESS+SAVEE	Boosting (KNN+MLP+RF)	86,30 %
					Random Forest	85,80 %
					CNN-LSTM	75,00 %
Bhawesh, K.et al. [14]	July 2023	7	18078	SAVEE+RAVDESS+CREMA-D+TESS	LSTM	57.50 %
					CNN	75,80 %
					MLP	59,60 %
					Random Forest	67,80 %
Tyagi, S. et al. [15]	December 2023	8	7356	RAVDESS	CNN-LSTM-GWO	87,00 %
		7	2800	TESS		99,93 %
		7	480	SAVES		65,47 %
		7	535	EMO-DB		78,00 %
Lata, S. et al. [16]	February 2024	7	2800	TESS	CNN-LSTM	98,00 %
					MFCC-LSTM	96,00 %
Yuan, Z. et al. [17]	Marsh 2024	10		IEMOCAP	DTNet	74.80 %
		7	535	EMO-DB		95.00 %
Islam, A. e al. [18]	April 2024	8	17589	RAVDESS+TESS+CREMA-D	CNN+BiLSTM	97,80 %
Akinpelu, S. et al. [19]	June 2024	7	2800	TESS	ViT	98,00 %
		7	535	EMO-DB		91,00 %
		7	3335	TESS+EMO-DB		93,00 %
Hossain, I. et al. [20]	August 2024	7	2800	TESS	CNN+LSTM+KNN	98,21 %
Proposed approach	October 2024	6	7442	CREMA-D	CNN-LSTM	54,07 %
		7	2800	TESS	LSTM	98,92 %

Declaration on Generative AI

Either:
The author(s) have not employed any Generative AI tools.

References

[1] Children's Health Queensland, Vocal cord palsy, https://www.childrens.health.qld.gov.au/health-a-to-z/vocal-cord-palsy#section__signs-and-symptoms. Accessed: 2024-11-19.

[2] Anatomy Corner, Epiglottis, https://anatomycorner.com/main/2016/09/13/epiglottis/. Accessed: 2024-11-12.

[3] J. A. Seikel, D. J. Hudock, D. G. Drumright, Anatomy Physiology for Speech, Language, and Hearing, Seventh Edition, 2021, 6th. ed., Plural Publishing, SIXTH EDITION, 912 pages, Full Color, Hardcover, 8.5" x 11",ISBN13: 978-1-63550-279-4.

[4] THROAT - Anatomy, respiration, voice swallowing, A brief introduction to the throat's anatomy, https://www.aarontrinidade.com/throat. Accessed: 2024-11-12.

[5] D. R. Hammou, Classification and detection of covid-19 in human respiratory lungs using convolutional neural network architectures, in: 2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP), IEEE, 2021. doi:10.1109/ai-csp52968.2021.9671158.
[6] D. R. Hammou, S. A. Mahmoudi, R. Adjoudj, Multi-Biometric Iris Recognition System Using Consensus Between Convolutional Neural Network Architectures, Int. J. Organ. Collect. Intell. 12.1 (2022) 1–30. doi:10.4018/ijoci.305210.
[7] M. G. de Pinto, M. Polignano, P. Lops, G. Semeraro, Emotions Understanding Model from Spoken Language using Deep Neural Networks and Mel-Frequency Cepstral Coefficients, in: 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), IEEE, 2020. doi:10.1109/eais48028.2020.9122698.
[8] T. Puri, M. Soni, G. Dhiman, O. Ibrahim Khalaf, M. alazzam, I. Raza Khan, Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network, J. Healthc. Eng. (2022) 1–9. doi:10.1155/2022/8472947.
[9] M. V. Gupta, S. Vaikole, A. D. Oza, A. Patel, D. P. Burduhos-Nergis, D. D. Burduhos-Nergis, Audio-Visual Stress Classification Using Cascaded RNN-LSTM Networks, Bioengineering 9.10 (2022) 510. doi:10.3390/bioengineering9100510.
[10] S. Ullah, Q. A. Sahib, Faizullah, S. Ullahh, I. U. Haq, I. Ullah, Speech Emotion Recognition Using Deep Neural Networks, in: 2022 International Conference on IT and Industrial Technologies (ICIT), IEEE, 2022. doi:10.1109/icit56493.2022.9989197.
[11] D. M. Vijayan, A. A. V, G. R, A. N. S. A, R. C. Roy, Development and Analysis of Convolutional Neural Network based Accurate Speech Emotion Recognition Models, in: 2022 IEEE 19th India Council International Conference (INDICON), IEEE, 2022. doi:10.1109/indicon56171.2022.10040174.
[12] M. Rayhan Ahmed, S. Islam, A. K. M. Muzahidul Islam, S. Shatabda, An Ensemble 1D-CNN-LSTM-GRU Model with Data Augmentation for Speech Emotion Recognition, Expert Syst. With Appl. (2023) 119633. doi:10.1016/j.eswa.2023.119633.
[13] N. Shah, K. Sood, J. Arora, Speech emotion recognition for psychotherapy: an analysis of traditional machine learning and deep learning techniques, in: 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2023. doi:10.1109/ccwc57344.2023.10099344.
[14] K. Bhawesh, D. Mustafi, A Comparison of Deep Learning and Machine Learning Models for Speech Emotion Recognition Using Multiple Features, in: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2023. doi:10.1109/icccnt56998.2023.10307514.
[15] S. Tyagi, S. Szénási, Optimizing Speech Emotion Recognition with Deep Learning and Grey Wolf Optimization: A Multi-Dataset Approach, Algorithms 17.3 (2024) 90. doi:10.3390/a17030090.
[16] S. Lata, N. Kishore, P. Sangwan, A Comparative Analysis of CNNLSTM and MFCCLSTM for Sentiment Recognition from Speech Signals," Int J Intell Syst Appl Eng, vol. 12, no. 21s, pp. 4392–4402, Mar. (2024), Accessed: Nov. 22, 2024. [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/6295.
[17] Z. Yuan, C. L. Philip Chen, S. Li, T. Zhang, Disentanglement Network: Disentangle the Emotional Features from Acoustic Features for Speech Emotion Recognition, in: ICASSP 2024 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024. doi:10.1109/icassp48485.2024.10448044.
[18] A. Islam, M. Foysal, M. I. Ahmed, Emotion Recognition from Speech Audio Signals using CNN-BiLSTM Hybrid Model, in: 2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), IEEE, 2024. doi:10.1109/icaeee62219.2024.10561755.
[19] S. Akinpelu, S. Viriri, A. Adegun, An enhanced speech emotion recognition using vision transformer, Sci. Rep. 14.1 (2024). doi:10.1038/s41598-024-63776-4.
[20] I. Hossain, M. Islam, T. Nahrin, M. Rashed, M. Rahman, Improving Speech Emotion Recognition and Classification Accuracy Using Hybrid CNN-LSTM-KNN Model. (2024), International Journal of Research Publication and Reviews Journal Homepage: Www.ijrpr.com, 5. Retrieved from URL:https://ijrpr.com/uploads/V5ISSUE8/IJRPR32597.pdf.

[21] T. Vamsikrishna, P. Naga Vyshnavi, (2017). Efficient Speech Emotion Recognition Using SVM and Decision Trees. (2017), In International Research Journal of Engineering and Technology. Retrieved from URL:https://www.irjet.net/archives/V4/i7/IRJET-V4I7663.pdf.

[22] T. Pfister, P. Robinson, Real-Time Recognition of Affective States from Nonverbal Features of Speech and Its Application for Public Speaking Skill Analysis. (2011) IEEE Transactions on Affective Computing, 2(2), 66–78. URL:https://doi.org/10.1109/t-affc.2011.8.

[23] D. C. Ambrus, Collecting and recording of an emotional speech database. (2000), Maribor, Slovenia: University of Maribor.

[24] J. Ostermann, Face Animation in MPEG-4. MPEG-4 Facial Animation: The Standard, Implementation and Applications, 17-55, (2002).

[25] H. Meng, T. Yan, F. Yuan, H. Wei, Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network, IEEE Access 7 (2019) 125868–125881. doi:10.1109/access.2019.2938007.

[26] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy, Kong-Pang Pun, An efficient MFCC extraction method in speech recognition, in: 2006 IEEE International Symposium on Circuits and Systems, IEEE. doi:10.1109/iscas.2006.1692543.

[27] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, B. Schuller, Speech Emotion Classification Using Attention-Based LSTM, IEEE/ACM Trans. Audio, Speech, Lang. Process. 27.11 (2019) 1675–1685. doi:10.1109/taslp.2019.2925934.

[28] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1D 2D CNN LSTM networks, Biomed. Signal Process. Control 47 (2019) 312–323. doi:10.1016/j.bspc.2018.08.035.

[29] P. M. Kathleen, K. Dupuis, Toronto emotional speech set (TESS). (2020) Scholars Portal Dataverse 1. University of Toronto Toronto, ON, Canada: URL:https://doi.org/10.5683/SP2/E8H2MF.

[30] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, R. Verma, CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. IEEE Trans Affect Comput. (2014) Oct-Dec;5(4):377-390. PMID: 25653738; PMCID: PMC4313618. doi:10.1109/TAFFC.2014.2336244.

[31] X. Tang, Y. Lin, T. Dang, Y. Zhang, J. Cheng, Speech Emotion Recognition Via CNN-Transforemr and Multidimensional Attention Mechanism. (2024). ArXiv (Cornell University). URL:https://doi.org/10.48550/arxiv.2403.04743.

[32] D. R. Hammou, S. A. Mahmoudi, R. Adjoudj, B. Mechab, A Model Of A Biometric Recognition System Based On The Hough Transform Of Libor Masek and 1D LogGabor Filter. 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), 1–9. URL:https://doi.org/10.1109/CloudTech49835.2020.9365917.

[33] D. R. Hammou, Y. Z. Feddag, S. Benadane , A New Architecture For Diagnosing Pulmonary Thorax Diseases (Covid19, Pneumonology, Normal) Using Deep Learning Technology. 2023 6th International Conference on Advanced Communication Technologies and Networking (CommNet), 1–10. URL:https://doi.org/10.1109/CommNet60167.2023.10365251.