Optimizing Speech Emotion Recognition: Evaluating Combinations of Databases, Data Augmentation, and Feature Extraction Methods

Lara Toledo Cordeiro Ottoni
Graduate Program in Electrical Engineering
Universidade Federal da Bahia
Salvador, Bahia
lara.toledo@ufba.br

Jés de Jesus Fiais Cerqueira
Department of Electrical and Computer Engineering
Universidade Federal da Bahia
Salvador, Bahia
jes@ufba.br

***Abstract*—Speech emotion recognition is a challenging and essential task with numerous applications in human-computer interaction, healthcare, and entertainment. However, achieving high accuracy in this task is complicated by the need to select the best combination of machine learning algorithms, databases, data augmentation techniques, and feature extraction methods. This paper discusses the difficulty of choosing appropriate combinations of these factors and proposes a methodology to address this challenge. The proposed method evaluates the performance of various combinations of databases, data augmentation techniques, and feature extraction methods to determine the most effective approach for speech emotion recognition. The paper also presents a convolutional neural network to classify the emotions of happiness, sadness, fear, anger, surprise, disgust, and neutral. The results showed that the optimal combination proposed, with 94% accuracy, uses the combined RAVDESS and TESS databases, using data augmentation with noise, stretch, and pitch, and using MFCC to extract the characteristics of the audios.**

***Index Terms*—convolutional neural network, data augmentation, MFCC, RAVDESS, speech emotion recognition.**

I. Introduction

Speech Emotion Recognition, or SER, is a rapidly growing field in computational intelligence that focuses on developing systems to detect and interpret emotions from speech signals. With the increasing popularity of voice assistants and the widespread use of smartphones and other communication devices, recognizing emotions from speech has become increasingly important [1]. Speech emotion recognition has many applications, from improving customer service to enhancing human-computer interactions and even in medical and psychological diagnosis. This technology uses machine learning and signal processing techniques to identify patterns in speech signals indicative of emotions such as happiness, sadness, anger, fear, disgust, neutral and surprise. As the demand for more personalized and intuitive human-computer interfaces continues to grow, the development of speech emotion recognition systems will undoubtedly play an increasingly important role in shaping our interactions with technology [2].

The performance of speech emotion recognition is influenced by various factors, such as the choice of the algorithm used to classify emotions, the choice of database, and the feature extraction technique [3]. The choice of algorithm is crucial for successful recognition, as different algorithms have varying degrees of precision and efficiency. The choice of database is also essential, as the data quality used for training and testing can directly affect the system's ability to recognize emotions [4]. Additionally, the feature extraction technique is a critical factor, as selecting relevant features for the recognition task is essential for obtaining good results. It is important to use techniques that allow for extracting relevant and discriminative features to maximize the system's ability to recognize emotions present in speech [5].

This paper aims to propose a methodology to select the best combination of databases, data augmentation and feature extraction. This work also proposes a convolutional neural network to be used as a classifier for the emotions of happiness, sadness, fear, anger, surprise, disgust, and neutral. By selecting the best combination of these factors, we aim to improve the accuracy and robustness of speech emotion recognition systems. The proposed methodology and neural network architecture will be evaluated using performance metrics and compared against existing state-of-the-art methods to demonstrate their effectiveness.

The structure of this article is as follows: The Section II reviews the literature on the subject. The Section III presents the proposed method to identify the optimal combination of database, data augmentation, and feature extraction along with the proposed CNN model. Section IV presents the results obtained, and Section V gives the study's conclusion.

II. Related Work

Given that SER is a highly active research field across the technology industry, numerous research works and innovations have been developed in this domain. Over time, advancements and improvements have emerged. Previously, it was typical for research to employ techniques like Hidden Markov Model (HMM), Support Vector Machine (SVM), and Multi-Layer

This research is funded by FAPESB (Foundation for Research Support in the State of Bahia).

Perceptron (MLP) [6], [7]. However, with the advent of Deep Learning (DL) algorithms and the promising results they have demonstrated, the majority of recent works in the last five years have utilized techniques such as Convolutional Neural Networks (CNN) [3], [5], [8] and a hybrid version incorporating Long Short Term Memory (LSTM) [9]–[11].

As Deep Learning techniques progress, it becomes increasingly important to have access to sufficient data to support new advancements and enable comparison between predictive models. Consequently, selecting a suitable database can significantly impact the effectiveness and accuracy of the chosen technique. Within the field of SER, many datasets are relatively small. Therefore, there is a growing need to merge various databases to construct more generalized classifiers that can recognize emotions across a broader range of individual characteristics in the datasets [6].

In addition to database analysis, data augmentation is another approach to expanding the available data. This technique involves generating new data from existing databases through slight modifications. Standard data augmentation techniques within the SER field include adding noise, stretching, altering pitch, and shifting [10], [12], [13].

Also, a crucial aspect of speech emotion recognition involves converting speech inputs into digital signals and processing them to extract suitable features that can be used to train the model [14]. There are various methods for extracting audio characteristics from databases. The literature has employed techniques such as Mel-Frequency Cepstral Coefficients (MFCC) [13], [15], Zero Crossing Rate (ZCR) [3], Chromagram [9], Mel Spectrogram [10], and Root Mean Square (RMS) values [12]. The choice of feature extraction technique directly influences the classifier's performance and the resulting outcomes.

A search was conducted on the SCOPUS platform in the past five years using the keywords ”speech emotion recognition” and ”combined databases” to investigate the literature requirements. The search yielded 23 articles, and a thorough analysis of these works revealed that only eight papers conducted investigations on the optimal combinations for SER. Table I presents the eight studies that compared databases, data augmentation, or feature extraction.

TABLE I
ARTICLES ON SER THAT PERFORM DATABASE COMPARISON, DATA AUGMENTATION, AND FEATURE EXTRACTION.

Papers	Dataset	Data Augmentation	Feature Extraction
[3]		✓	
[16]	✓		
[10]		✓	✓
[11]	✓		
[12]	✓	✓	
[14]	✓		✓
[13]		✓	
[2]	✓		✓
Proposed	✓	✓	✓

In [3] compares the use and non-use of data augmentation techniques, specifically the addition of noise, shifting, pitching, and stretching. On the other hand, [16] conduct a comparison study to determine the optimal combination of data sets for speech emotion recognition. The study evaluates the performance of RAVDESS, TESS, SAVEE, CREMA-D, and a combination of all datasets.

In [10] compares the impact of data augmentation and feature extraction techniques on speech emotion recognition. The authors apply data augmentation techniques, including noise, stretching, and pitch shifting. They also evaluate the performance of different feature extraction techniques, such as MFCC, Mel, chroma, ZCR, RMS, and Roll off. In contrast, [11] investigate the impact of dataset selection on speech emotion recognition. The authors compare the performance of several datasets, including RAVDESS, TESS, SAVEE, RAVDESS+TESS, TESS+SAVEE, and RAVDESS+TESS+SAVEE.

In [12], a comparison is made between different databases and data augmentation. The databases compared include RAVDESS, CREMA-D, SAVEE, TESS, and a combination of all of them. Additionally, the effectiveness of using data augmentation techniques (noise and pitch) is evaluated. On the other hand, [14] presents a comparison based on databases and feature extraction techniques. The databases compared are RAVDESS, TESS, SAVEE, and a combination of all of them, while the feature extraction techniques evaluated are MFCC and chroma.

In [13], the authors solely compared data augmentation with noise and pitch to not using it at all. On the other hand, in [15], the focus was on comparing the use of different databases and feature extraction techniques. The databases reached were RAVDESS, TESS, SAVEE, and a combination of all three. As for feature extraction techniques, the authors compared MFCC, Mel, chroma, ZCR, and RMS.

The conducted search reveals a gap in the literature concerning the research field of SER. Most papers present their experiments and results without exploring the optimal combination of database, data augmentation, and feature extraction. Therefore, the main contributions of this work are:

1) Propose a methodology that combines databases with data augmentation and several feature extraction methods to investigate an optimal combination.
2) Propose a convolutional neural network that works as a classifier for happy, sad, fear, anger, surprise, disgust, and neutral emotions.

III. PROPOSED METHOD

Speech Emotion Recognition (SER) is a thriving research field in which human emotions are estimated based on speech. However, there are several essential aspects to consider in this classification process, such as selecting the appropriate database, utilizing data augmentation techniques, determining the feature extraction method to be used, choosing the proper machine learning algorithm as a classifier, and finally, understanding how all these factors impact the model's accuracy value.

Therefore, this article proposes a methodology to search for the best combination of these factors: database, data augmentation, feature extraction, and machine learning algorithm. As described in Section 2, none of the works mentioned above performs a search covering all these criteria. A classic combination was established to compare the accuracies, which consists of using the RAVDESS database without data augmentation, extracting the characteristics of the audios with the MFCC technique, and using the convolutional neural network as a classifier.

Then, three experiments are conducted following the methodology depicted in Figure 1. In the first experiment, we examine whether the model's accuracy value can be significantly improved by combining different databases (RAVDESS+TESS, RAVDESS+TESS+SAVEE, and RAVDESS+TESS+SAVEE+CREMA-D). The second experiment investigates the impact of data augmentation by introducing various modifications, such as noise, stretch, and pitch, to the audio files in the database. The objective is to determine if these modifications influence the accuracy value.

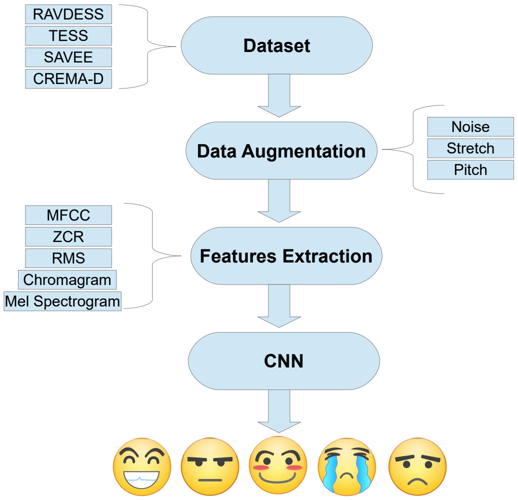


Fig. 1. Flowchart of the proposed method.

The third stage of the investigation focuses on determining the most effective method for feature extraction from databases to optimize the classifier's performance. Various techniques, including Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Root Mean Square (RMS) values, Chromagram, and Mel Spectrogram, will be tested. Throughout all stages of the investigation, a convolutional neural network will be utilized to accurately classify emotions such as happy, sad, fear, anger, surprise, disgust, and neutral. The following will present each of the steps in more detail.

A. Datasets

Data availability to support new developments and enable comparison between predictive models is currently one of the pillars of science. In the field of SER, there are many small datasets, and selecting the appropriate ones is crucial for the performance of the machine learning algorithm [6].

In this study, an investigation was conducted to determine the most efficient combination of databases that achieves the highest accuracy. To this end, four of the most popular databases in the SER research field, RAVDESS, TESS, SAVEE, and CREMA-D, were combined. More information about each database will be presented below.

1) RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [17] is frequently used for tasks related to speech emotion recognition. It consists of recordings from 24 professional actors, 12 women and 12 men, who recite two statements in both singing and speaking modes. The dataset contains expressions of calm, happy, sad, angry, fearful, surprised, neutral, and disgusted emotions, each produced at two levels of emotional intensity - normal and strong. This dataset is available in three formats: audio-only (16-bit, 48kHz .wav), audio-video (720p H.264, AAC 48kHz, .mp4), and video-only (no sound). However, it is important to note that the data set is unbalanced, as shown in Fig. 2.

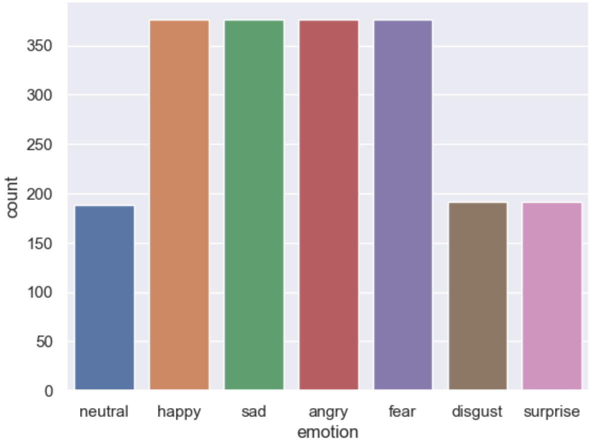


Fig. 2. RAVDESS dataset.

2) TESS: The Toronto Emotional Speech Set (TESS) [18] is a database developed by researchers at the University of Toronto's Department of Psychology. The database features recordings of two English actresses, one aged 26 and the other aged 64, portraying various emotions. The TESS dataset includes anger, disgust, fear, happy, neutral, surprise, and sad expressions. The database contains 2800 audio files in .wav format, with 400 audio recordings per emotion, see Fig. 3. It is important to note that this database is balanced, meaning each emotion category has an equal number of audio files.

3) SAVEE: The SAVEE (Surrey Audio-Visual Expressed Emotion) dataset [19] includes 480 speech utterances spoken by four English actors aged between 27 to 31 years, expressing seven diverse emotions: angry, happy, neutral, disgust, sad, fear, and surprise, in a phonetically stable manner. Nonetheless, the dataset suffers from a class imbalance issue, where the neutral class contains almost double the amount of samples compared to all other classes, as seen in Fig. 4. This study only utilized the speech audio samples from the dataset.

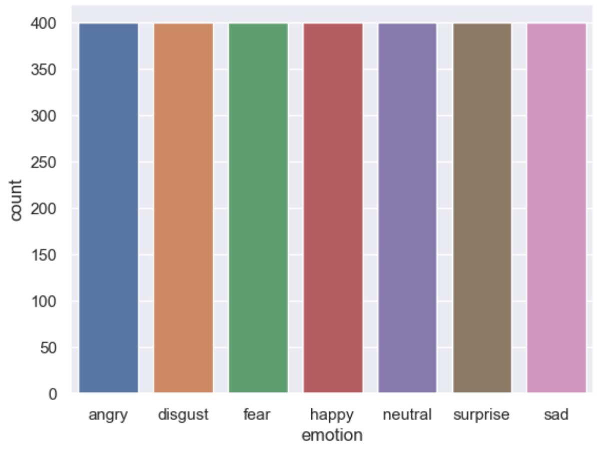


Fig. 3. TESS dataset.

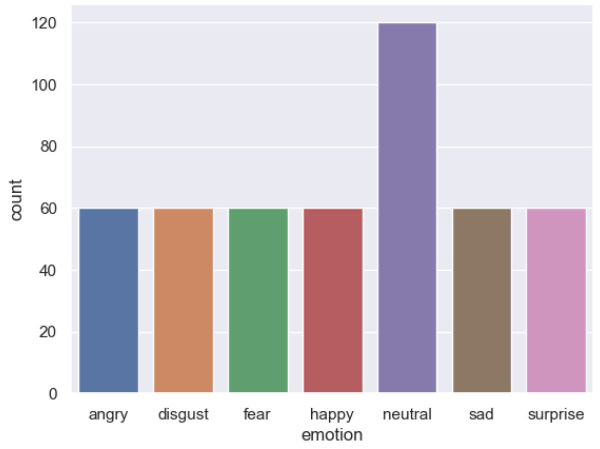


Fig. 4. SAVEE dataset.

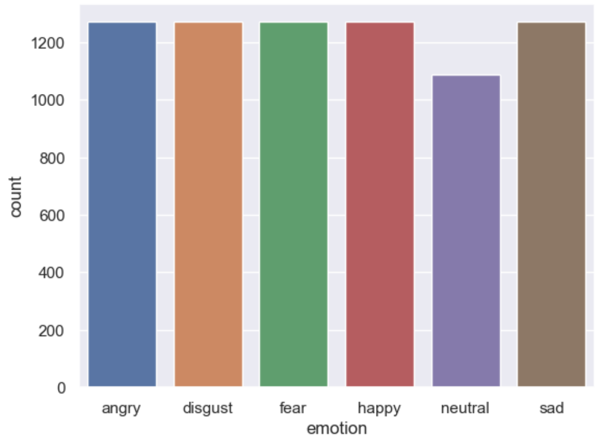


Fig. 5. CREMA-D dataset.

4) CREMA-D: The authors [20] presented the Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) as a suitable dataset for investigating multimodal emotion expression and perception. This audio-visual dataset contains a wide range of data that researchers can use for model training and later apply to new datasets. The dataset includes 7442 unique audio samples that 91 actors of various races and ethnicities recorded. Additionally, 48 male and 43 female actors recited 12 sentences in six different emotions, as can be seen in Fig. 5.

B. Data Augmentation

Insufficient dataset size and class imbalance are common issues in the SER task. With the expansion of complexity and scale of DNNs, a significant dataset is essential for achieving optimal performance. One potential solution is to augment the dataset using diverse data augmentation (DA) techniques [3].

Data Augmentation involves applying minor modifications to the original training dataset to generate new artificial training samples. Given the relatively low number of speech utterance records in each class, this study employs three types of audio DA techniques: additive white Gaussian noise injection, time-stretching, and pitch shifting of the audio files. The impact of these techniques is visually demonstrated in Fig. 6.

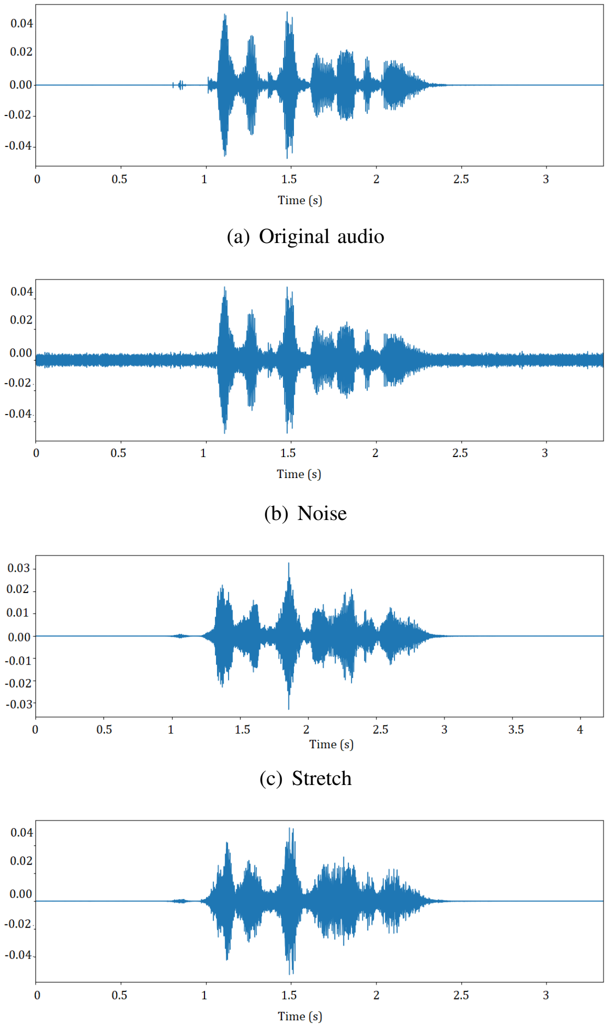
This study employed the noise injection technique to add random values to the data using NumPy's normal and uniform methods with a rate of 0.035. It also utilized the stretching technique to stretch time series by a fixed rate of 0.8, which was implemented using the time_stretch method of the Python Librosa library. Finally, it randomly altered the pitch with a pitch factor of 0.7 by utilizing the pitch_shift method of the Librosa library.

C. Feature extraction

Extracting features from speech audio signals is one of the essential measures in SER-related activities. Precise extraction of crucial features improves the performance in terms of the SER accuracy of the model [7]. Specifically, this study uses five different spectral features: Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Chromagram, Mel Spectrogram, and Root Mean Square (RMS) values of the speech audio files as the input for the deep learning algorithms.

1) Mel-frequency cepstral coefficients (MFCC): To extract MFCC features, the first step is to divide the speech signal into short frames of 20-30 ms each, which are advanced every 10 ms to capture temporal features of individual speech signals. Discrete Fourier Transform (DFT) is then applied to each windowed frame, and they are transformed into magnitude spectrum. Next, 26 filters are applied to the signal obtained in the previous step to calculate the Mel-Scaled Filter-bank (MSFB). MSFB is a unit of measurement based on the frequency perception of the human ear, resulting in 26 values that describe the energy of each frame. Log energies are then calculated to obtain log filter-bank energies. In Eq. 1 quantifies the estimation of Mel from the physical frequency [9] [10] [2].

$$f_{Mel} = 2590 \log_{10} 1 + \frac{f}{700} \tag{1}$$



(a) Original audio

(b) Noise

(c) Stretch

(d) Pitch

Fig. 6. Example of modifications used for Data augmentation.

For this process, f represents the physical frequency (in Hz), and f_{Mel} represents the frequency perceived by the human ear. After obtaining the log filter-bank energies, the Discrete Cosine Transform (DCT) is applied to generate the MFCCs [9] [6]. This study utilized the Librosa library [21] to extract the MFCC values from the datasets.

2) Zero crossing rate (ZCR): It is commonly utilized in SER. It quantifies the number of times the amplitude of a speech signal crosses the zero-value mark during a given time frame. ZCR is an effective method for distinguishing between voiced and unvoiced expressions. Low-frequency fluctuations, with numerous zero crossings, do not have a prominent effect. Mathematically, ZCR can be defined by Eq. 2, where s represents the signal of length T and $1_{\mathbb{R}<0}$ is an indicator function. This study utilized the Librosa library to extract the ZCR values from the datasets [9].

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{\mathbb{R}<0}(S_t S_{t-1}) \tag{2}$$

3) Chromagram: The Chromagram (or Chroma) feature represents the tonal content of an audio signal and is closely related to the 12 classes of the pitch. One of the main attributes of Chroma features is that they capture audio's harmonic and melodic characteristics. The Chromagram features are obtained by applying Short-Time Fourier Transforms (STFT) to the waveform created from the audio files in the dataset [22]. This study utilized the Librosa library to extract the Chroma values from the datasets.

4) Mel Spectrogram: A Spectrogram visualizes the frequency spectrum of a signal over time using a signal analysis method. It calculates the spectrogram for each window by transforming the signal from the time domain to the frequency domain using Fast Fourier Transform (FFT). It divides the frequency spectrum into evenly spaced Mel scale frequencies, producing a Mel spectrogram for each window. Then, it decomposes the signal's magnitude into components corresponding to the Mel frequencies [22] [10]. The study extracted the values from the datasets using the Librosa library.

5) Root Mean Square (RMS) Value: The RMS calculates the value for each frame from the speech audio samples. It analyzes the overall amplitude of the signal, providing an average signal amplitude. The RMS method measures the magnitude of a signal as a signal strength, regardless of the amplitude's positive or negative level. For a given signal, $x = x_1, x_2, x_3, ..., x_n$, the RMS value, x_{RMS}, can be calculated using Eq. 3 [9]. This study utilized the Librosa library to extract the RMS values from the datasets.

$$x_{RMS} = \sqrt{\frac{x^2}{n}} = \sqrt{\frac{1}{n}(x_1^2, x_2^2, x_3^2, ..., x_n^2)} \tag{3}$$

D. Deep Learning

For the classifier, we opted to use a Convolutional Neural Network (CNN) as a deep learning model, as it has shown good performance in audio classification and speech recognition tasks [23]. Some papers in the literature that use CNN for the Speech Emotion Recognition (SER) problem can be seen in [3], [4], [8], [24], [25].

We used the Keras library for Deep Learning and Neural Networks [23] in conjunction with Google's machine learning framework TensorFlow to develop the classifier. These libraries collectively offer the essential tools required for the development, training, and validation of various machine learning algorithm. In this work we used the architecture presented by Tab II

The architecture consists of three Conv1D layers and two MaxPooling1D layers. Additionally, there are two Dropout layers, two BatchNormalization layers, one Flatten layer, and two Dense layers, one utilizing ReLU activation and the other utilizing Softmax activation. The Dense final layer, which uses Softmax activation, generates a probability distribution for each audio emotion class and serves as the output layer for the CNN.

Once the architecture is defined, the next step is to train the model. In this phase, 80% of the input data is used for

TABLE II
CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE FOR SPEECH EMOTION RECOGNITION.

Layer	Size	Number of Parameters
Conv1D (+ relu)	(36, 64)	384
Conv1D (+ relu)	(36, 128)	41088
BatchNormalization	(36, 128)	512
MaxPooling1D	(7, 128)	0
Dropout	(7, 258)	0
Conv1D (+ relu)	(7, 256)	164096
BatchNormalization	(7, 256)	1024
MaxPooling1D	(1, 256)	0
Dropout	(1, 256)	0
Flatten	(256)	0
Dense (+ relu)	(64)	16448
Dense (+ softmax)	(7)	455

the training process. The remaining 20% is reserved for the stages of validation (10%) and testing (10%).

IV. RESULTS AND DISCUSSION

This session will present the results of the three investigation stages proposed in the methodology described in Section 3. We conducted all experiments on a notebook with a Windows 11 operating system, an Intel i5 1135G7 2.40GHz processor, 8 GB RAM, and an Nvidia GeForce MX350 GPU with 2GB VRAM. The development environment for this work consists of the Jupyter IDE associated with the Python language.

The first experiment investigated the influence of the combination of databases on the accuracy value. These experiments were carried out without data augmentation, with the MFCC feature extraction technique, and using the convolutional neural network described in the previous section. The results of this first step are described in Tab. III, in which it is possible to see that the combination of databases improves the accuracy value.

Using only the RAVDESS database, the accuracy value is 77%. However, when combined with the TESS database, the accuracy goes to 87%, with an increase of 10%. The combination RAVDESS+TESS+SAVEE increases accuracy by 9%. However, the combination RAVDESS+TESS+SAVEE+CREMA-D achieved an accuracy of 60%, i.e., 17% below the reference value. Combining the four databases is believed to generate a problem of greater complexity.

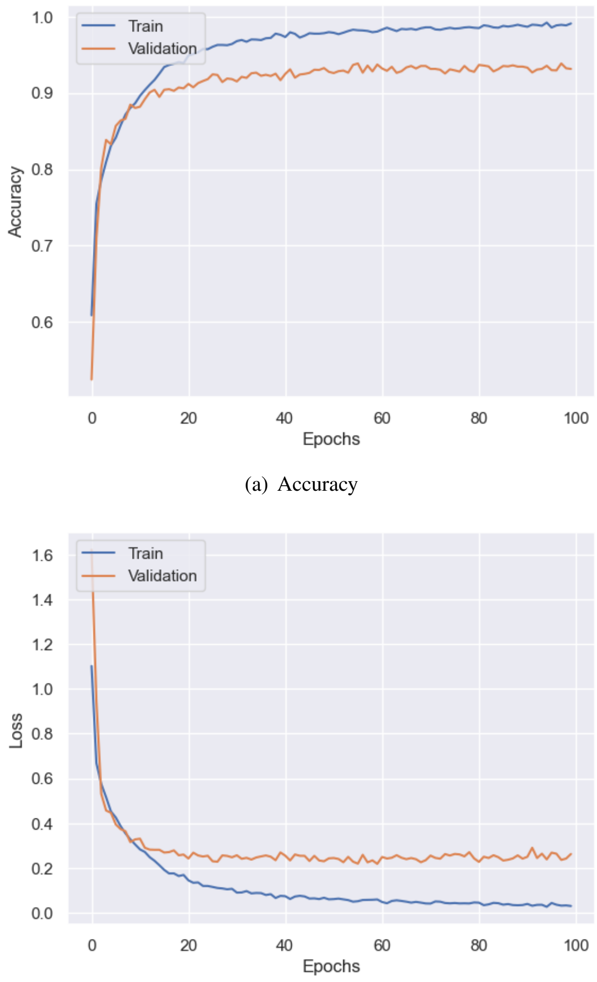
TABLE III
ACCURACY VALUES (%) FOR EXPERIMENTS WITH AND WITHOUT DATA AUGMENTATION, USING MFCC.

Dataset	Without D.A.	With D.A.
RAVDESS	77.0	**85.0**
RAVDESS+TESS	87.0	**94.0**
RAVDESS+TESS+SAVEE	86.0	**92.0**
RAVDESS+TESS+SAVEE+CREMA-D	60.0	**66.0**

Also, in Tab.III, it is possible to observe the second investigation proposed by this paper, which consists of data augmentation. With the application of noise, stretch, and pitch in the databases, the accuracy value increased by 6% in all analyzed combinations.

In the third stage, the investigation focuses on determining the most effective audio feature extraction technique for optimal classification by CNN. Six feature extraction methods were selected, including Mel Spectrogram, Chroma, ZCR, RMS, MFCC, and a combination of all (All). Table IV shows the results obtained for each feature extraction method with data augmentation for each database. It is worth noting that MFCC achieved the best accuracy values for the RAVDESS, RAVDESS+TESS, and RAVDESS+TESS+SAVEE databases, while Mel Spectrogram performed better for the RAVDESS+TESS+SAVEE+CREMA-D dataset.

With this, it is possible to observe that the combination that generates the best model (with an accuracy value of 94%) uses the RAVDESS+TESS database, with data augmentation and the MFCC, to extract the features. In Fig. 7, it is possible to view the training and validation history of the best accuracy combination.



(a) Accuracy

(b) Loss

Fig. 7. Graph of accuracy and loss values during training and testing.

TABLE IV
ACCURACY VALUES (%) FOR DIFFERENT FEATURE EXTRACTION TECHNIQUES USING DATA AUGMENTATION.

Dataset	Mel Spectogram	Chroma	ZCR	RMS	MFCC	All
RAVDESS	74.0	39.0	19.0	27.0	**85.0**	81.0
RAVDESS+TESS	87.0	64.0	20.0	27.0	**94.0**	92.0
RAVDESS+TESS+SAVEE	84.0	60.0	20.0	26.0	**92.0**	91.0
RAVDESS+TESS+SAVEE+CREMA-D	**71.0**	39.0	22.4	29.0	66.0	70.0

It is possible to observe an increase in the accuracy value over the epochs for both the training and validation sets. It is also possible to watch a drop in the value of the loss. This behavior indicates that the model could recognize each class's characteristics during training without losing its generalization capacity.

Figure 8 presents the confusion matrix resulting from the model evaluation process. It is observed that, in general, the network could differentiate the seven emotions. It is noted that there was more incredible difficulty in determining the emotions sad and fear.

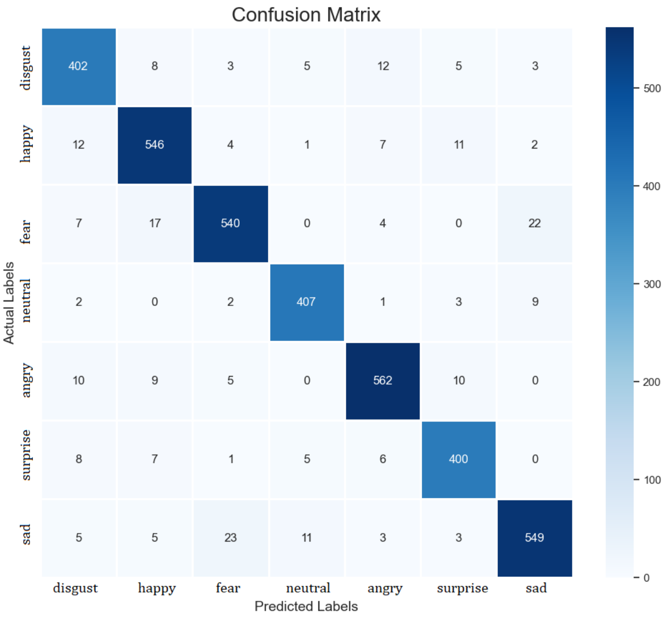


Fig. 8. Confusion Matrix.

Table V presents the values of precision, recall, F1 and support for each class, obtained from the confusion matrix.

TABLE V
PRECISION, RECALL, AND F1 VALUES (%) WERE OBTAINED FROM THE BEST COMBINATION FOUND.

Emotion	Precision	Recall	F1
Disgust	94.0	92.0	94.0
Happy	89.0	94.0	92.0
Fear	97.0	87.0	92.0
Neutral	93.0	98.0	95.0
Angry	94.0	95.0	94.0
Surprise	94.0	93.0	94.0
Sad	92.0	93.0	93.0
Accuracy			94.0

A comparison with the accuracy values reported in the literature is presented next. Tab. VI displays the relevant information for each work, including the dataset used, whether data augmentation was applied, the feature extraction method, the emotion classification algorithm utilized, and the corresponding test accuracy values.

Table VI presents articles that utilized combinations of databases used in this study: RAVDESS+TESS, RAVDESS+TESS+SAVEE, and RAVDESS+TESS+SAVEE+CREMA-D. Interestingly, most of the papers reviewed did not apply data augmentation, which is a relatively recent approach and still uncommon in research articles focused on the speech recognition of emotions. The majority of the authors used the MFCC method, which, similar to our work, achieved better results for most combinations of databases. Additionally, some papers utilized a variety of multiple-feature extraction methods.

Regarding the emotion classification algorithms employed by the studies in Table VI, the majority relied on convolutional neural networks. Others achieved promising results using hybrid methods that combine CNNs with LSTMs. In summary, combining the proposed analysis methodology that considers the databases, data augmentation, and the investigation of the best feature extraction techniques, along with the proposed CNN for speech emotion recognition, resulted in an accuracy of 94%. This accuracy is notable compared to the results presented by the authors in Table VI.

V. CONCLUSION AND FUTURE WORK

The main objective of this study was to propose a methodology for selecting the optimal combination of database, data augmentation, and feature extraction. Another contribution of this research was to propose a convolutional neural network for classifying emotions, including happy, sad, fear, anger, surprise, disgust, and neutral, during human-computer interaction.

It can be observed from Tables III and IV that combining the RAVDESS, TESS, and SAVEE databases yields better results compared to using only the RAVDESS database. Furthermore, Table III demonstrates that incorporating data augmentation techniques (such as noise, pitch, and stretch) significantly improves the accuracy value. Table IV compares various feature extraction techniques used as input to the CNN, with the MFCC technique achieving the best performance, reaching 94% accuracy for the RAVDESS+TESS database. Comparing the results with existing literature (see Table VI) confirms the effectiveness of analyzing optimal database combinations, data augmentation, and feature extraction and the proposed CNN's satisfactory performance.

TABLE VI
COMPARISON OF ACCURACY VALUES (%) FOR SPEECH EMOTION RECOGNITION USING THE COMBINATION OF DATABASES.

Papers	Datasets	Data Augmentation	Feature Extraction	Algoritm	Acuraccy
Proposed	RAVDESS+TESS	Yes	MFCC	CNN	94.00
[11]	RAVDESS+TESS	No	MFCC	CNN+LSTM	84.35
[7]	RAVDESS+TESS	No	MFCC+ZCR+LSF	SVM	88.40
[5]	RAVDESS+TESS	No	MFCC	CNN	90.09
[3]	RAVDESS+TESS	Yes	MFCC+Mel+Chroma+ZCR	CNN	89.00
[22]	RAVDESS+TESS+SAVEE	No	MFCC+Mel+Chroma	MLP	84,96
[14]	RAVDESS+TESS+SAVEE	No	MFCC+Croma	CNN	82.25
[2]	RAVDESS+TESS+SAVEE	Yes	MFCC	CNN+LSTM	90.00
[16]	RAVDESS+TESS+SAVEE	No	Spectrogram images	CNN	89.93
[10]	RAVDESS+TESS+SAVEE+CREMA-D	Yes	MFCC+Mel+Chroma+ZCR+RMS	CNN+LSTM	92.73
[12]	RAVDESS+TESS+SAVEE+CREMA-D	Yes	MFCC+ZCR+RMS	CNN+LSTM	94.50
[26]	RAVDESS+TESS+SAVEE+CREMA-D	Yes	MFCC	CNN+GAP	92.28
[13]	RAVDESS+TESS+SAVEE+CREMA-D	Yes	MFCC	CNN	70.00

The proposed CNN can be further improved by exploring different variations in future work. One possible direction is to investigate hybrid algorithms such as CNN and LSTM, which have shown promising results in the literature. Moreover, different CNN architectures can be proposed and compared. Overall, this is a research field with many avenues for further exploration.

ACKNOWLEDGMENT

The authors are grateful to CAPES (Financing Code 001), FAPESB (084.0508.2022.0000201-09), and UFBA.

REFERENCES

[1] L. T. C. Ottoni and J. J. F. Cerqueira, “A review of emotions in human-robot interaction,” in *2021 Latin American Robotics Symposium (LARS)*. IEEE, 2021, pp. 7–12.

[2] J. Singh, L. B. Saheer, and O. Faust, “Speech emotion recognition using attention model,” *International Journal of Environmental Research and Public Health*, vol. 20, no. 6, pp. 1–21, 2023.

[3] U. Asiya and V. Kiran, “Speech emotion recognition-a deep learning approach,” in *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. IEEE, 2021, pp. 867–871.

[4] E. Guizzo, T. Weyde, S. Scardapane, and D. Comminiello, “Learning speech emotion representations in the quaternion domain,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[5] M. Gupta and S. Chandra, “Speech emotion recognition using mfcc and wide residual network,” in *2021 Thirteenth International Conference on Contemporary Computing (IC3-2021)*, 2021, pp. 320–327.

[6] J. de Lope and M. Graña, “An ongoing review of speech emotion recognition,” *Neurocomputing*, pp. 1–11, 2023.

[7] A. Ashok, J. Pawlak, S. Paplu, Z. Zafar, and K. Berns, “Paralinguistic cues in speech to adapt robot behavior in human-robot interaction,” in *2022 9th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, 2022, pp. 01–06.

[8] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, “Real-time speech emotion analysis for smart home assistants,” *IEEE Transactions on Consumer Electronics*, vol. 67, no. 1, pp. 68–76, 2021.

[9] M. R. Ahmed, S. Islam, A. M. Islam, and S. Shatabda, “An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition,” *Expert Systems with Applications*, vol. 218, pp. 1–21, 2023.

[10] M. Gupta, T. Patel, S. H. Mankad, and T. Vyas, “Detecting emotions from human speech: role of gender information,” in *2022 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 2022, pp. 1–6.

[11] S. K. Hazra, R. R. Ema, S. M. Galib, S. Kabir, and N. Adnan, “Emotion recognition of human speech using deep learning method and mfcc features,” *Radioelectronic and Computer Systems*, no. 4, pp. 161–172, 2022.

[12] S. Jothimani and K. Premalatha, “Mff-saug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network,” *Chaos, Solitons & Fractals*, vol. 162, pp. 112–512, 2022.

[13] R. Mittal, S. Vart, P. Shokeen, and M. Kumar, “Speech emotion recognition,” in *2022 2nd International Conference on Intelligent Technologies (CONIT)*. IEEE, 2022, pp. 1–6.

[14] N. Kumar, R. Kaushal, S. Agarwal, and Y. B. Singh, “Cnn based approach for speech emotion recognition using mfcc, croma and stft hand-crafted features,” in *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE, 2021, pp. 981–985.

[15] I. Shahin, O. A. Alomari, A. B. Nassif, I. Afyouni, I. A. Hashem, and A. Elnagar, “An efficient feature selection method for arabic and english speech emotion recognition using grey wolf optimizer,” *Applied Acoustics*, vol. 205, p. 109279, 2023.

[16] N. Chitre, N. Bhorade, P. Topale, J. Ramteke, and C. Gajbhiye, “Speech emotion recognition to assist autistic children,” in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*. IEEE, 2022, pp. 983–990.

[17] S. R. Livingstone and F. A. Russo, “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*, vol. 13, no. 5, p. e0196391, 2018.

[18] M. K. Pichora-Fuller and K. Dupuis, “Toronto emotional speech set (tess),” *Scholars Portal Dataverse*, vol. 1, p. 2020, 2020.

[19] P. Jackson and S. Haq, “Surrey audio-visual expressed emotion (savee) database,” *University of Surrey: Guildford, UK*, 2014.

[20] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “Crema-d: Crowd-sourced emotional multimodal actors dataset,” *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[21] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[22] A. S. Nasim, R. H. Chowdory, A. Dey, and A. Das, “Recognizing speech emotion based on acoustic features using machine learning,” in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2021, pp. 1–7.

[23] F. Chollet, *Deep learning with Python*. Simon and Schuster, 2021.

[24] Z. Zhong, “Speech emotion recognition based on svm and cnn using mfcc feature extraction,” in *International Conference on Statistics, Data Science, and Computational Intelligence (CSDSCI 2022)*, vol. 12510. SPIE, 2023, pp. 445–452.

[25] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, “Recognition of emotions in speech using convolutional neural networks on different datasets,” *Electronics*, vol. 11, no. 22, p. 3831, 2022.

[26] P. Mishra and R. Sharma, “Gender differentiated convolutional neural networks for speech emotion recognition,” in *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. IEEE, 2020, pp. 142–148.