JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 26 Issue 01, 2023 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

RECOGNITION OF EMOTIONS THROUGH SPEECH USING MACHINE LEARNING TECHNIQUES

[1]Meesala Sudhir Kumar, [2]M.Chitra, [3]Anubhav Sharma, [4]S D Prabu Ragavendiran, [5]Sudharani B Banappagoudar, [6]A Nirmal Kumar

[1] Associate Professor, Department of Computer Sc. & Engineering, MSEIT, MATS University, Raipur(CG) India. mskcec@gmail.com

[2]Rajalakshmi Institute of technology, Chennai, Tamilnadu
chitra.m@ritchennai.edu.in

[3]Assistant Professor, Computer Science and Engineering, IMS Engineering College-Ghaziabad, Ghaziabad, Uttar Pradesh- 201015
mail2anubhav@gmail.com

[4]Professor, Department of Computer science and Engineering, RVS Technical campus-Coimbatore, Coimbatore. sdpgobi@gmail.com

[5]Professor,School of Nursing Science,ITM University,Gwalior(MP).
sudharani.sons@itmuniversity.ac.in

[6]Associate Professor, Department of Computer Science and Engineering, CMR Institute of Technology, Hyderabad. sa.nirmalkumar@gmail.com

Abstract

One of the areas where AI can be used is the recognition of emotions through speech, ensuring the real use of these systems to access and democratize this type of technology. Customer service will be personalized, with bots determining the customer's mood while performing a service and the ability to redirect to human service if slurred speech is noticed. Call centers for emergency and insurance services, in particular, can be positively impacted by emotional recognition. This work presents a Recurrent Neural Network Unit (RNN)-gate recurrent (GRU) and a Convolutional Neural Network (CNN) for speech emotion classification with excellent performance in experimental conditions. The Ryerson Audiovisual Emotional Speech and Song (RAVDESS) dataset was used to train these models, which allowed for the creation of an evaluation and testing environment. Evaluation of a model trained in English provided an accuracy of approximately 42%, which was considered unsatisfactory for the classifier. The main characteristics identified as responsible for performance are sample group characteristics, classification bias without cross-validation, and lack of noise processing. The neural network that gave the best accuracy was RNN-GRU, which achieved 79.69% using a technique that increases the size of the dataset through a stretching process.

Keywords: neural networks, artificial intelligence, emotion recognition, speech feature extraction, machine learning, emotions, speech.

1. Introduction

Methods for extracting speech features have long been sought, and important research in the early 1950s allowed us to divide history into before and after. AT&T Bell Labs researchers were able to extract information from the voice signal, creating the first speech recognition

Submitted: 13/01/2023 **Accepted**: 30/01/2023

system capable of recognizing 10 words in the English language, and these 10 words are numbers from 0 to 9 [1]. The Audrey system, which was 90% accurate when its creator spoke, became 70%/80% accurate when someone else spoke. This aspect of the Audrey system has already demonstrated the level of challenge of creating a system capable of recognizing multiple dialects between each culture, at different speeds and dialects. Since then, much research has been done to understand how the process of speech production affects feature extraction. Other studies have sought to understand how environmental noise and sound quality interfere with speech recognition. This interest is mainly due to the use of technology as a way to improve Human-Machine Interaction (HMI). Currently, we observe that the interaction of robots with humans is limited to the context of the task that the robot performs [2,3].

The interaction between human beings is modulated by emotional contexts: either in perception or in the execution of actions. For a natural communication modulation it is important to consider emotional factors in human-robot interactions for a natural and efficient interface [4]. Therefore, the key elements of natural communication are the skills to perceive and express emotions [5]. In a robot with emotional recognition capability, an application example would be: children, the elderly, sick or disabled people may not be able to interact with social robots that will be ubiquitous in everyday life in the future. They may not be able to pronounce a command correctly or press the appropriate button to stop the robot's current action. Perceiving and identifying emotional states of human beings would help to better understand the physical and emotional needs of human beings and thus could improve the robot's reaction. Especially in the context of social robots, it can be important to recognize feelings of pain or anger. A social robot could, for example, react to perceived pain when it touches a human, or identify fear if the robot gets too close. This could increase human trust and allow for a more natural and safer interaction with humans [6].

Recent advances in deep learning have provided an increase in the popularity and robustness of the voice emotion recognition task [7, 8, 9]. These models usually use a large number of labeled samples to learn general representations for emotion recognition, providing state-of-the-art results in different speech-related scenarios [10, 11, 12].

The research objective is that recognition of emotions through speech is necessary for machines to recognize human needs. It can, for example, speed up calls in emergency systems, improve referral systems and, of course, voice will play a fundamental role in autonomous cars, therapeutic treatments, depression diagnosis and the IoT devices [13].

The development of artificial intelligence (AI) has allowed us to find new ways to recognize emotions through speech, and based on this, this work aims to compare two neural networks, convolutional networks and recurrent networks.

2 Related Works

In this chapter, works using speech recognition techniques will be presented. Works have been selected to allow diversity in the concepts and methods used, allowing a larger view of what is being studied in this field of knowledge.

In the work presented by [14], demonstrated how CNN can be used for spectrogram image analysis. This method was introduced with the aim of breaking the standard pipeline of

identifying emotions through speech. Taking into account the mean of the 7 trained emotion categories of anger, boredom, disgust, sadness, fear, happiness and neutrality, the results obtained by the proposed model are 84.3%. The author also mentions the problems of network confusion, such as the persistent confusion of fear, anger, disgust, and happiness. To reduce this error, one of the methods mentioned by the author is to augment the data set to provide the characteristic elements of these classes to the neural network.

[13] compares and integrates two neural network architectures for emotion recognition. Both data sets contain the same seven categories, although only anger, happiness, sadness, and neutral categories were used. A final comparison is high-level feature extraction using CNN and information modeling over time using LSTM. In this scenario, the best performance was obtained for the Emo-DB cluster with 82.35% MFCC. For the IEMOCAP group, the spectral logarithmic honeycomb method obtained 50.05%, which is the best result in this data set. The author cites the high normality of IEMOCAP to justify the difference between the two data sets. In the work done by [15], K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) are recommended as the best method to compare classifiers. The Emo-DB data set was used in all seven categories. To perform emotion recognition, the teacher also uses a set of sounds he created. Analyzing the performance of KNN on 7 segments and K = 3, he obtained 66.7% accuracy. When the same algorithm was used with four categories: neutral, angry, happy and sad and the same k = 3, the accuracy increased to 82.2%. The accuracy obtained in the validation phase was 82.1%, using a technique that equates the number of samples in the language set and the number of samples in the SVM algorithm. When brute force was used, the accuracy result in the validation phase was 84.4%. When the test samples are used, the accuracy is 70.83%. The work of [16] uses IEMOCAP, Emo-DB and RAVDESS datasets. The author uses five methods to extract the features provided as input to the proposed convolutional neural network, those methods are MFCC, Honeycomb Gradient Spectral Plot, Histogram, Spectral Contrast Feature and Donets Plot. The panel compared the author model output with the author's validation score [17] and obtained high human validation, 71.61% author and 67% human validation.

3 Methodology

The research methodology used in the development of this research is based on books, theses, theses and publications in journals. The basics are outlined in the books, but what we do has its entire structure based on recent publications, with the aim of providing up-to-date content for proposed research. Early research was done to understand the process by which the vocal spectrum is formed when a human speaks. After this first step, speech features important for emotion recognition are identified, which is a key step in selecting which emotions the machine learning model recognizes. For a good feature extraction process, attention was paid to MFCC, which generates a vector for audio components and aims to extract only relevant features in the speech recognition process. After extraction, various deep learning frameworks were used for emotion recognition. A set of sounds was used to test and determine the impact of one language on the recognition phase of another language.

4 Development

4.1 Data Set

The dataset used for training is RAVDESS and was produced as part of an investigation to develop a set of audio-visual profiles [18]. In this work, only the sound set will be used.

The distribution and characteristics of these phonemes are as follows:

- 24 professional representatives, 12 men and 12 women.
- 60 votes for each representative, for a total of 1,440 files.
- In American English.
- The feelings are: calm, joy, sadness, anger, fear, surprise and disgust.
- Each emotion is reproduced in two levels of intensity, normal and strong.

The phonemes are marked with a unique identifier and are divided into seven parts, as follows:

- Method: 01 = Audio - Video, 02 = Video and 03 = Audio
- Channel Type: 01 = Voice, 02 = Music
- Emotion: 01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = afraid, 07 = upset, 08 = surprised
- Intensity of emotion: 01 = normal, 02 = strong

Sentence: 01 = “Children talking at the door”, 02 = “Dogs are sitting by the door”

- Iteration: 01 = the first iteration, 02 = the second iteration
- Representative: from 01 to 24

4.3 Data Exploration

Before starting the feature extraction step, it is necessary to explore the data set, allowing the identification of possible opportunities.

The first analysis to be done is to count the number of occurrences of each emotion.

The graph below shows the distribution of the number of emotions in the set:

A first observation to be made is about the number of audios classified as neutral, it is lower than the others. The set was created considering two recordings for the same emotion with different intensity, however, only the neutral emotion does not apply. Since neutral, happy, sad, and angry emotions will be used for training and classification, it is important to note that the amount of audio with the neutral emotion can affect performance.

The next analysis, figure 1, is to compare the waveform of each emotion, subdividing between male and female.

The audios have an average duration of 3.5 seconds, where the first second is practically empty, that is, it will not help in the training, as well as the end. Another important observation is the difference between male and female, which may at the moment of sorting the test set to cause false positives.

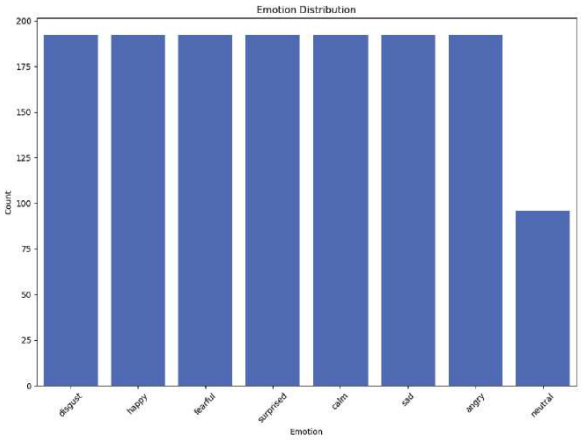

Figure 1. Distribution of emotions.

4.4 Feature Extraction

The step to extract features is very important for any development that aims to make a classification, especially those that use neural networks, which is why the MFCC was presented in section two of this project, a way of analyzing the features that best represent the auditory perceptions of the humans, in addition to capturing better the low frequencies present in the voice and the more abrupt spectral changes.

The figure below presents Waveforms of various emotions in gender wise.

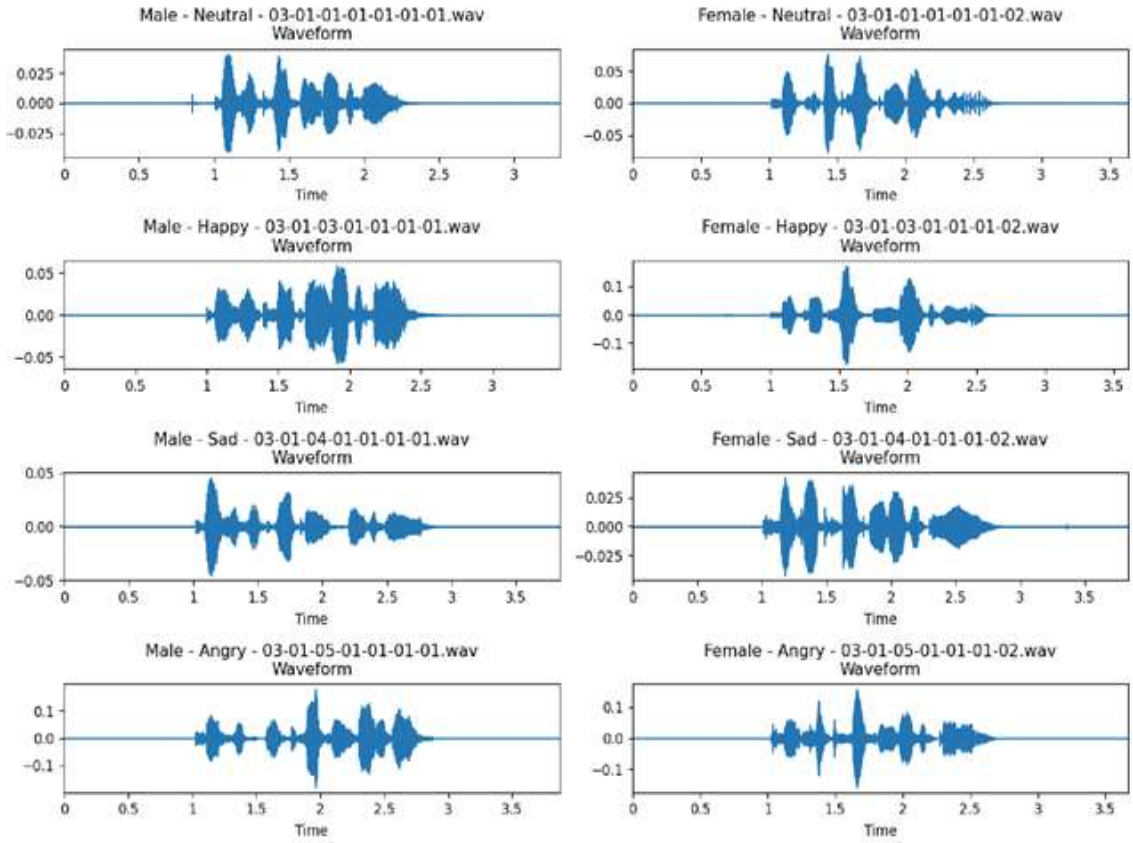

Figure 2 - Waveform.

4.5 Data Strategy

There are some techniques that can improve the performance of a supervised learning neural network, and one of them is to increase the size of the data set based on the data you have, that is, new data is generated by making small transformations in the existing data. This technique, known as data augmentation, can improve accuracy, since it is thought that the larger the data set, the better the performance. Another technique is balancing the data set, ensuring that the number of occurrences of each emotion is equal, ensuring that the least present classes have the same probability of being learned.

The figure below presents emotions comparison of the trait vectors, broken down by emotion and gender:

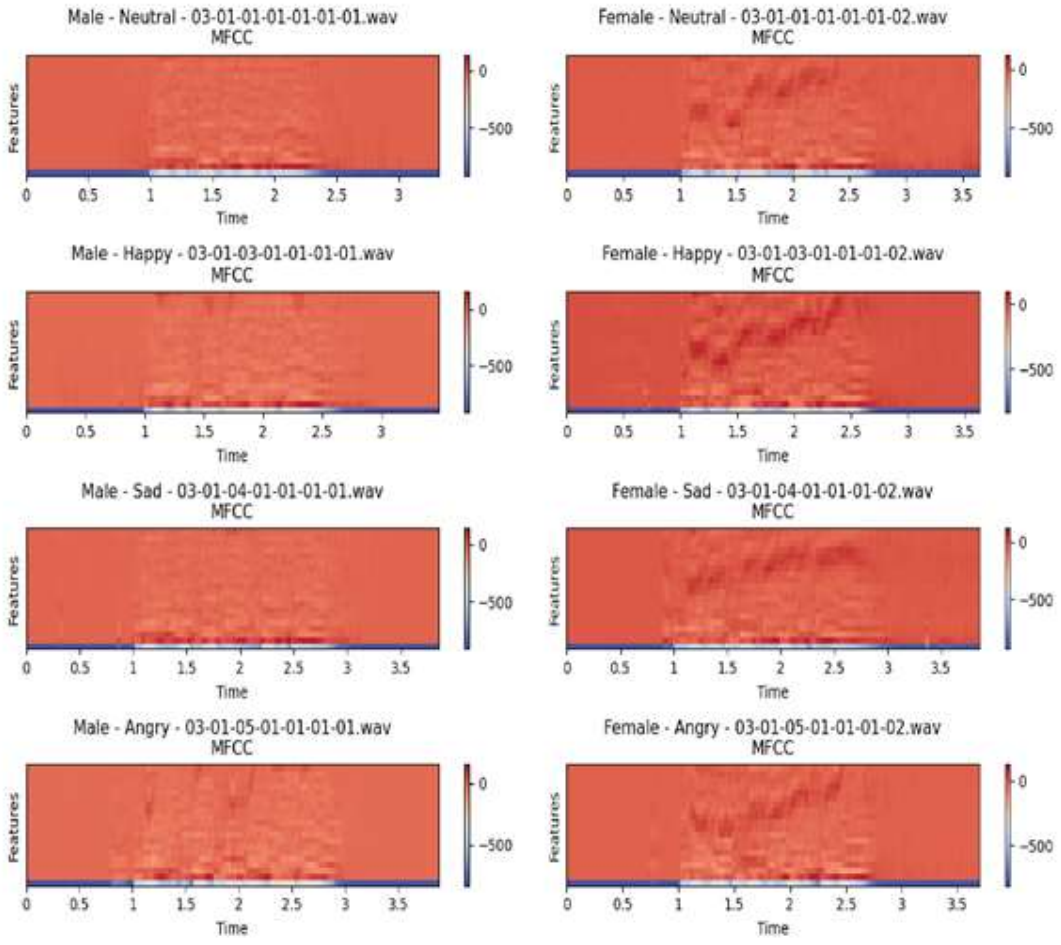

Figure 3 - Vector of features over time.

4.5.1 Data Augmentation

To apply transformations on the dataset that will be used for training the neural network, operations such as noise, shift, stretching and pitch are applied [7]. These are the main operations performed to expand the dataset with new audios.

These operations can be applied separately or together, and some operations can be combined separately. It is not possible to guarantee that just by increasing the training set the result will be better, so each test must be isolated to identify which operation contributes to increasing the accuracy of the model.

These are factors that must be taken into account when strategies such as these are applied.

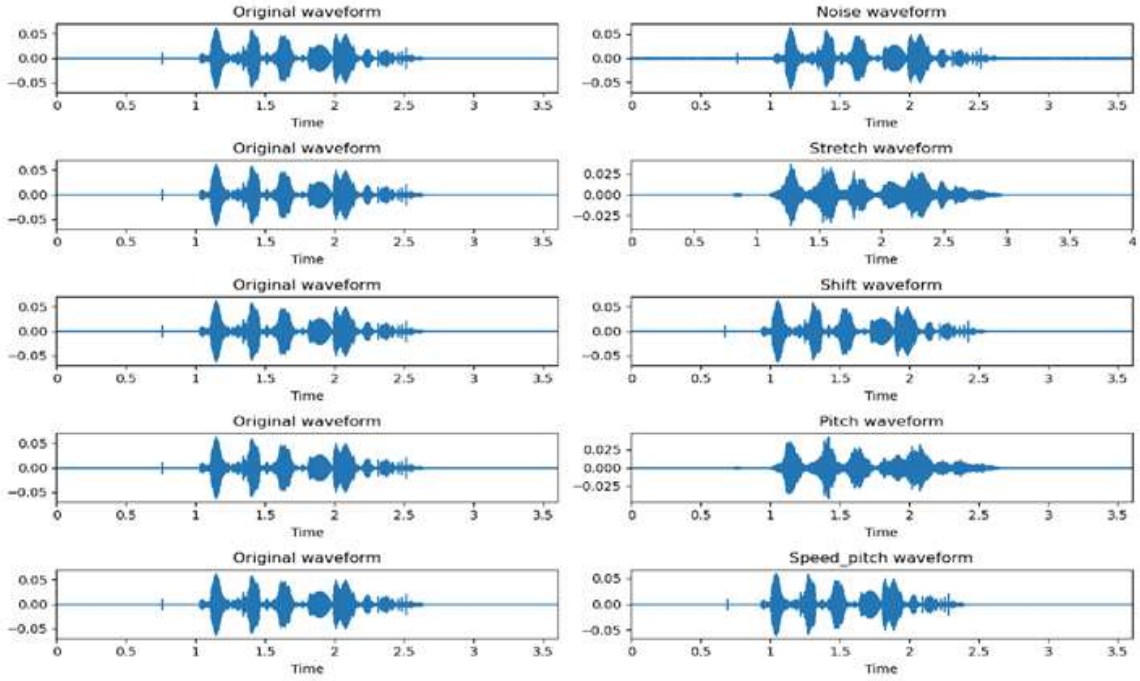

Figure 4- Comparison of transformation operations.

4.5.2 Balancing

The dataset is not balanced, the neutral emotion does not have the same amount of audios as the other emotions. To ensure that the number is equal and there is no impact on training, twice as many audios of the neutral emotion were added with the noise transformation operation, allowing the set to be matched, since the set is not large and removing audios could affect performance of model.

4.6 Proposed Architectures

The proposal is to compare convolutional neural networks and GRU recurrent neural networks to identify which architecture achieves better performance using the accuracy metric. Based on this, the four models will be presented below, two of each architecture, which will be trained and evaluated.

Table 3 presents the layers of the two CNN networks that were compared. The choice of each one was based on tests performed on the original and expanded data set, which obtained the best accuracy.

The models were built using 1D convolution, best suited for time series analysis. The input of the convolutional layer has the format (batch_size, input_dim), where the first parameter is the number of vectors (training set size) and the second is the dimension of the feature vector.

Table 1 – CNN models comparison table.

CNN 1	CNN 2
Input (batch_size, input_dim)	
Conv1d 20-64	Conv1d 20-64
Activation ReLu	
Conv1d 20-32	Conv1d 20-32
Activation ReLu	
MaxPool	
Conv1d 20-16	Conv1d 20-4
Activation ReLu	
MaxPool	Flatten
Conv 20-8	Dense
Activation ReLu	Activation Softmax
MaxPool	
Conv 20-4	
Activation ReLu	
Flatten	
Dense	
Activation Softmax	

CNN 1 has two convolution layers, two ReLu and two MaxPool more than CNN 2, in addition to the extra layers, the third convolution layer of CNN 2 applies only 4 filters, while the third layer of CNN 1 applies 16 filters. The kernel size is the same for both networks, twenty.
Below, in table 2, the two RNN-GRU that presented the best precision in the test stage are shown. A characteristic that can be highlighted here is that in the training stage, the accuracy was generally below the value when compared to applying the model to the test set.
Two different filters (units) were applied, 256 in the RNN-GRU 1 network and 128 in the RNN-GRU 2 network. In the RNN-GRU 1 network, the units size was doubled, followed by a completely connected layer and three GRU layers of 256 units, ending up with a fully connected layer the size of classes to be sorted and a softmax activation function. In RNN-GRU 2, the network is much smaller, with a fully connected layer of size four, representing the number of classes for classification, and a softmax layer.

Table 2 – Comparison table of the RNN-GRU models.

RNN-GRU 1	RNN-GRU 2
Input (*dim*, *input_dim*)	
GRU 256	GRU 128
GRU 512	Dense
Dense	Activation Softmax
GRU 256	
GRU 256	
GRU 256	

Dense	
Activation Softmax	

The training set was divided between training and validation, where 75% was used for training and 25% for validation. For the test set, actors one and two were separated, the first is male and the second is female, ensuring that the model had not previously had contact with these audios. The data augmentation method presented was also measured, so there is no standard amount in the training and validation set, what remains is the test set with two actors. In the tests it was possible to identify that the architectures present different performance as the applied operations are changed.

5 Implementation

The implementation carried out during the development of the project is presented in this section, as well as relevant information about the parameters used and characteristics of each architecture that perhaps have not yet been demonstrated. The code is presented in the following order: set for prediction, feature extractor, build model, train model, and test model. The Pandas framework was used to index the audio set information within a data frame.

To test the trained model, that is, to predict a set that the neural network does not know, two actors were separated according to the algorithm below:

The feature extraction algorithm uses the librosa.feature.mfcc() function. The offset parameter means that the first half second will be disregarded and the duration considers the next 3 seconds. This standardization helps in training the network and discards the beginning and end of the audios that are silent, as shown in figure 4. The output of the algorithm is a dataframe with the feature vector with size 20 (parameter n_mfcc) in the feature column.

The next step is to implement the models that will be trained by the neural network, be it CNN or RNN-GRU. All models presented were implemented using python, which facilitated the definition of parameters for each layer and streamlined the process of testing the best network conversion. So much so that, below, only four networks are demonstrated, however, tests were carried out with other layers until these were chosen.

The implementation of a recurrent neural network of the GRU type is very similar to the CNN network, the difference lies in the use of a GRU layer instead of a Conv1D layer. Like CNN using Conv1D, the GRU layer is suitable for data with temporal characteristics, therefore, suitable for audio streams.

RNNs have two new parameters, the first recurrent_dropout applies dropout to units of a recurrent network and return_sequences returns the last output of the sequence. The return_sequences parameter in RNN-GRU 2 is False because it is the only GRU layer present in this model. In the RNN-GRU 2 model, the optimization function was also changed, from Adam to RMSprop.

The next algorithm will train the network itself, considering the configuration parameters. The best model will be saved in the file_weights file that will be used in the testing stage. The model.fit() function starts model training, using 25% of the data for validation of each epoch

considering the list of callbacks. Training will stop when it reaches the maximum number of epochs or monitoring by early_stopping callbacks.

As mentioned above, in the training stage, the best model is saved in order to use the audio samples of the two actors (1 and 2) for inference, samples that were never used during the training stage. In addition to saving the best model during training, the model was serialized in a model.json file, allowing it not to be necessary to rebuild each layer to predict the test samples. With these two files saved, it is only necessary to open the file in json format and load the model saved in the training step inside it.

With the model ready and loaded, it is necessary to extract the feature vector from the test samples, as was done for the training step.

The algorithm below shows the use of the predict() function of the model passing the test set.

6 Experiments and Results

In this section, the experiments carried out in order to seek the best result within the test set will be presented. The strategy for carrying out the experiments was first to make a comparison between each convolutional neural network, for this, the tests were conducted with a first model of CNN until it reached the best accuracy, after obtaining the first model, some more significant changes were made in the first model, for example, reduce and increase the number of fully connected and polling layers until obtaining a model to perform the tests comparing the two CNN networks.

For the experiments with the RNN-GRU, the same strategy as for the CNN was applied, with the construction of a recurrent neural network with the best accuracy and then the change to a new RNN in order to obtain two networks for comparison.

All models were evaluated following the accuracy metric.

According to the objective of this project, after the tests and results among the four proposed neural networks, the one that obtained the best performance was used to infer the set of audio samples, a set created [15] when he carried out his final work of course. This set was used due to the difficulty of obtaining a set, and because it is used only for testing the models, a large volume of samples is not necessarily required.

With these four models, the experiments described in the subsections were applied below.

6.1 CNN Experiments

For CNN 1, tests were performed by increasing the training/validation set. The analysis was done with No data augmentation, With all Training Loss and Accuracy increase operations, **With increased noise, With pitch boost, With speed pitch increase, and With increased stretch.**

Tests were applied on each transformation operation that increased the training set, however, just increasing the set does not mean that the result will be better, according to the tests carried out with CNN 1 the transformation operation that had the best performance was stretching, where the loss and accuracy rate during the training and validation process was the one that had the best performance.

And the confusion matrix of the tests, as expected, the best performance of the stretch operation obtained an accuracy of 67.19%. Also, it would need less epochs.

The same tests were applied to the CNN 2 network, and the performance was better than CNN 1 in almost all operations, just not better in the stretch operation. Another point is the low performance using the noise operation, for this reason, another test was carried out only with the pitch and speed_pitch operations. Accuracy was 73.44%, representing an improvement of approximately 6% over CNN 1.

6.2 RNN-GRU Experiments

The training/validation set augmentation tests were also performed on the two RNN networks. The objective is to show how this factor influences the test stage and if the operations performance is the same as the CNN networks.

The RNN-GRU 1 network in general presented better performance than the CNNs, mainly the performance using only the stretch operation, reaching 79.69%. Another important point that should be highlighted when the confusion matrix, is analyzed, is the 100% success rate in the happy and neutral classes, and the 87.5% rate for the angry class cannot be ignored either. The negative point is the assertiveness when analyzing the sad class, which in all tests were well below the other classes. Another relevant piece of information is the difference between the dataset without augmentation, that is, the original set, which had an accuracy of 68.75% against 79.69% by increasing the dataset using the stretch operation, a gain of approximately 11%. In order to seek a better convergence of the model, an attempt was made to increase the learning rate, but the results in the test samples were not satisfactory, even the loss function presenting a smaller value in the training stage.

The tests using the RNN-GRU 2 network, showed no difference increasing the set of tests/validation, remaining with the same 79.69% of the RNN-GRU 1 network. The difference here was with the better performance of the test that did not use any data augmentation technique, with 71.88% and the worst performance using all techniques at the same time, with 65.62%. As it was done in the RNN-GRU 1 network, in the RNN-GRU 2 network the learning rate was also increasing, and the result was unsatisfactory as in the previous network.

6.3 Tests with Set

The model that presented the best performance was used to test a set of audios. This test aims to understand whether the spoken language is a factor that should be taken into account when emotion recognition applications through speech are developed and whether it is possible to use a model trained in one language to predict audios from another. This is important since the audio sets for training/validation must be of considerable size and the vast majority are in English, therefore, a trained and validated model can be a way of skipping the construction stage of these sets, starting directly for the prediction step.

In this section, the results of the tests carried out with the set of samples will be presented. For this set, no audio transformation operation was applied to increase the set of samples, the tests were performed with the original set.

The RNN-GRU 1 and 2 models showed the same accuracy during the experiments stage, and therefore both were applied to the set of samples.

The RNN-GRU 1 model, figure 5, obtained the best result compared to the RNN-GRU 2 model, figure 6, however the performance of the two models was not satisfactory. The first reached an

accuracy of 42.73% and the second 40.91%, where the class with the highest hit rate was sadness with 54.55%.

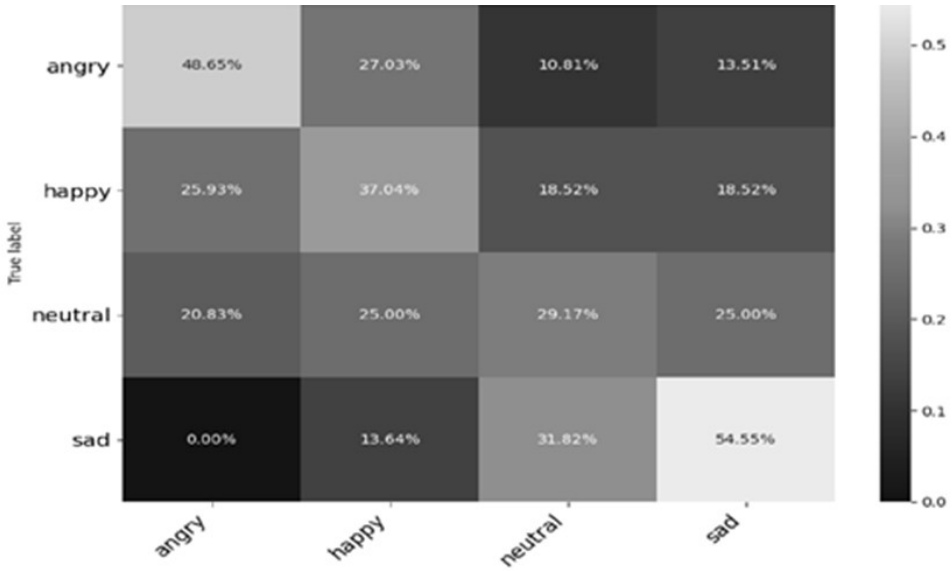

Figure 5 - RNN-GRU 1: Sample confusion matrix.

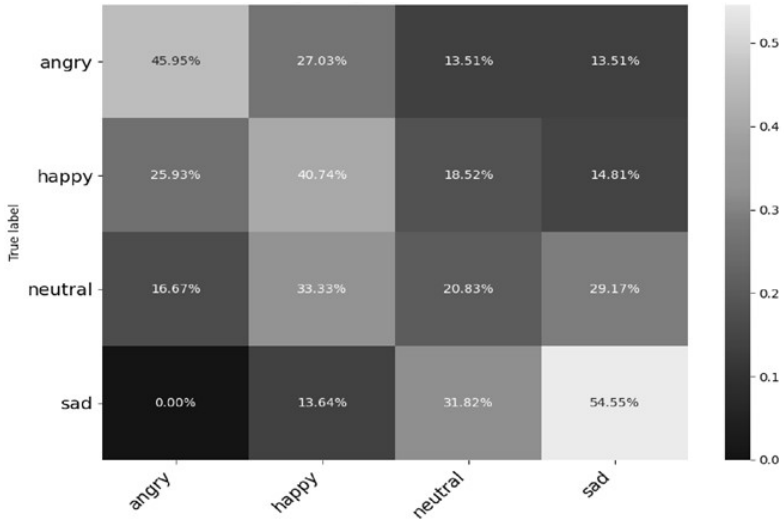

Figure 6 -- RNN-GRU 1: Sample confusion matrix.

6.4 Results

After all the experiments, it was observed that the difference between the models was not so significant, even so, it is important to emphasize the relevance of the RNN-GRU type, which in the tests achieved a better performance when compared to a CNN.

Even not having found references in the literature on predicting emotions in a language using models trained on others, it was believed that better results could be found, but this has not been proven. It is important to highlight some points that also influenced the low performance, being the characteristics of the set of samples, the classification bias without validation and the lack of noise treatment, the main ones.

Just remember that this set of samples represents a real context, where in practice more than one model would be needed to perform the classification, for example, one to analyze the semantics using STT and natural language processing, another to analyze the audio spectrum and finally, a model to consolidate this information and convert it into just one result, being more accurate because it considers physiological and semantic factors.

The table below seeks to show the performance of the models described in the related work stage, even knowing that there are significant differences between these models and what was proposed in this work. To facilitate, the main differences between the models are shown in the table.

Table 3 – Performance comparison table with other models.

	[13]	[15]	[16]	Proposed model
Accuracy	78.16%	94.3%	71.61%	79.69%
Model	CNN	SVM	CNN	RNN-GRU
Dataset	Emo-DB	Emo-DB	RAVDESS	RAVDESS
Classes	4	4	8	4

7 Conclusion

In this work, we sought to recognize emotions through speech using two of the main architectures of neural networks, which are applied to different areas of knowledge. Both convolutional neural networks and recurrent neural networks were presented as an option and their literature was reviewed in order to understand their specificities and their real application within the proposed problem.

Through the bibliographic review, a recent type of recurrent neural network was identified with little research, mainly on the subject of emotion recognition through speech. Unlike the neural network type, the MFCC technique for feature extraction is widely used for automatic voice emotion recognition systems and can better capture the intentions behind speech.

During the development, it was identified that the techniques to increase the size of the data set for training and validation influenced the precision of the model, characterizing a point of attention when starting the construction of this type of application. Considering that related research on emotion recognition is not as advanced as image manipulation and there is still a lack of data sets ready to evolve research in this area, some techniques were exposed to increase the sample set and its effectiveness in improving the performance of each proposed model.

The data augmentation technique improved the performance of the neural network, but not all audio transformation operations improved performance, the one that presented the best result was stretching.

The difference between the male and female voices was not analyzed, but during the experiments it was identified that the sadness emotion class had a low performance. A possible explanation may be related to the same intensity of the voice of the neutral emotion. In the image, it is possible to identify a relationship between the intensity of the male and female voice in the neutral and sadness classes, corroborating for the low performance in the sadness

class. Another point that can be highlighted is the small improvement in the sadness class without using data augmentation of the best result obtained, which should be considered a point of attention.

The tests carried out on the set of audio samples presented a low performance and it is relevant that new researches focus on this objective, since using only the MFCC extraction method was not enough.

References

1. Yu, Y. Research on speech recognition technology and its application. In: *2012 International Conference on Computer Science and Electronics Engineering*. [S.l.: s.n.],2012. v. 1, p. 306–309.
2. Minami, Y. Executive summary: world robotics 2012 industrial robots. Available online on http://www.worldrobotics.org, p. 8–18, 2013.
3. Ferrús, R. M.; Somonte, M. D. Design in robotics based in the voice of the customer of household robots. Robotics and Autonomous Systems, Elsevier, v. 79, p. 99–107, 2016.
4. Schwarz, N.; Bless, H.; Bohner, G. Mood and persuasion: Affective states influence the processing of persuasive communications. Advances in experimental social psychology, Elsevier, v. 24, p. 161–199, 1991.
5. Nummenmaa, L. et al. Emotional speech synchronizes brains across listeners and engages large-scale dynamic brain networks. NeuroImage, Elsevier, v. 102, p. 498–509, 2014.
6. Wieser, I. Towards Understanding Auditory Representations in Emotional Expressions. Tese (Doutorado) — Universität Hamburg, Fachbereich Informatik, 2016.
7. Trigeorgis, G. et al. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: IEEE. Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. [S.l.], 2016. p. 5200–5204.
8. Zheng, W.; Yu, J.; Zou, Y. An experimental study of speech emotion recognition based on deep convolutional neural networks. In: IEEE. Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on. [S.l.], 2015. p. 827–831.
9. Huang, Z. et al. Speech emotion recognition using cnn. In: ACM. Proceedings of the 22nd ACM International Conference on Multimedia. [S.l.], 2014. p. 801–804.
10. Chang, J.; Scherer, S. Learning representations of emotional speech with deep convolutional generative adversarial networks. arXiv preprint arXiv:1705.02394, 2017.
11. Ashwin, T.; Saran, S.; Reddy, G. R. M. Video affective content analysis based on multimodal features using a novel hybrid SVM-RBM classifier. In: IEEE. Electrical, Computer and Electronics Engineering (UPCON), 2016 IEEE Uttar Pradesh Section International Conference on. [S.l.], 2016. p. 416–421.
12. Weisskirchen, N.; Bock, R.; Wendemuth, A. Recognition of emotional speech with convolutional neural networks by means of spectral estimates. In: IEEE. Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2017 Seventh International Conference on. [S.l.], 2017. p. 50–55.

13. Pandey, S. K.; Shekhawat, H. S.; Prasanna, S. R. M. Deep learning techniques for speech emotion recognition: A review. In: *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. [S.l.: s.n.], 2019. p. 1–6.
14. Badshah, A. M. et al. Speech emotion recognition from spectrograms with deep convolutional neural network. In: *2017 International Conference on Platform Technologyand Service (PlatCon)*. [S.l.: s.n.], 2017. p. 1–5.
15. Zhao Kang. Improved emotional speech recognition based on SVM and decision tree [J]. *Information technology*, 2020. (8): 17–22.
16. Issa, D.; Fatih Demirci, M.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, v. 59, p. 101894, 2020.
17. 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*, Public Library of Science, v. 13, n. 5, p. 1–35, 052018.
18. Wu Junqing, Ni Jiancheng, Wei Yuanyuan, et al. Statistical feature oriented RxK ensemble model for speech emotion recognition [J]. *Journal of Qufu Normal University (NATURAL SCIENCE EDITION)*, 2020, V. 46; No.176(02):57–62.