"Application of Convolutional Neural Networks (CNNs) in Emotional Speech Recognition for the Spanish Language"

Josué Arriaga[*]
josue.arriaga@utec.edu.pe

Nicolás Arroyo[†]
nicolas.arroyo.c@utec.edu.pe

Camila Rodriguez[‡]
camila.rodriguez@utec.edu.pe
Universidad de Ingeniería y Tecnología
Lima, Perú

Abstract

This study delves into the efficacy of Convolutional Neural Networks (CNNs) in identifying emotional speech in the Spanish language, a domain filled with unique linguistic characteristics and challenges. Our research primarily revolves around evaluating the adaptability and accuracy of CNNs in discerning and categorizing emotional nuances within Spanish speech patterns. Notably, we achieved a significant milestone with a CNN model, attaining a low noise training with a loss rate. However, some instances of misclassification were observed. In contrast, another CNN model with the same architecture exhibited a higher misclassification peak, coupled with a loss rate of . Additionally, our exploration extended to experiments involving 1D CNN-LSTM and 2D CNN-LSTM architectures. These experiments were pivotal in contrasting the performance of these models across different linguistic contexts, thereby providing insights into the model's versatility and limitations in emotion recognition tasks across languages.

1 Introduction

Emotion is a fundamental aspect of human existence, intricately woven into our daily interactions and decision-making processes. It manifests in diverse forms and can fluctuate rapidly, reflecting the complex interplay of circumstances, mood, and relationships Qayyum et al. (2019). In the realm of human-computer interaction, understanding and accurately interpreting these emotional states is paramount. It enhances the flexibility and effectiveness of interactions, as systems can respond adaptively to the user's emotional cues Ullah et al. (2023). This importance is underscored by the relatively low requirements for heavy-duty hardware setups and complex computational algorithms in emotion recognition research, making it a highly attractive field for scholars.

The surge in machine learning and its application in speech processing has revolutionized emotion recognition. A plethora of tools are now employed to detect emotions from speech, with many studies conducted under real-life scenarios to affirm their practical applicability and realism Venkataramanan & Rajamohan (2019). Consequently, the accuracy of these research endeavors has seen significant fluctuations, primarily due to the adoption of varied methodologies.

[*]Universidad de Ingeniería y Tecnología (UTEC).
[†]Universidad de Ingeniería y Tecnología (UTEC).
[‡]Universidad de Ingeniería y Tecnología (UTEC).

Final Paper for the ML3061-2023.2 - Machine Learning Class at UTEC. Lima, Peru.

Numerous studies have been dedicated to identifying human emotions instantaneously through speech analysis. These have predominantly employed machine learning approaches, leading to varied results and discrepancies in effectiveness Qayyum et al. (2019). This paper introduces a novel approach using a Convolutional Neural Network (CNN) for emotion classification.

In the ensuing sections, we will delve into the specifics of the database used for this study, the methodology followed, and the results obtained. Additionally, a comparative analysis focusing on the accuracy of our CNN model relative to other methods will be presented, emphasizing its efficacy in recognizing emotional speech, particularly in the context of the Spanish language Ullah et al. (2023). Our exploration also extends to the performance of 1D CNN-LSTM and 2D CNN-LSTM architectures, providing a comprehensive view of the model's versatility Venkataramanan & Rajamohan (2019).

2 Problems and Motivations

The burgeoning field of Artificial Intelligence (AI) has significantly enhanced the intricacies of human-computer interactions, particularly in the realm of emotion recognition in speech Venkataramanan & Rajamohan (2019). However, this technological advancement encounters unique challenges when applied to the Spanish language, characterized by its rich phonetic subtleties and expressive variations Ullah et al. (2023). Recognizing the lack of extensive research in emotion detection using Convolutional Neural Networks (CNNs) in Spanish, this study aims to fill this gap. The scarcity of studies utilizing CNNs for emotion recognition in Spanish provides a compelling opportunity to explore and demonstrate the capabilities of these advanced neural network models in a new linguistic context.Issa et al. (2020)

Furthermore, this research is driven by the necessity to understand and address the impact of environmental noise on the accuracy of emotion recognition systems. Real-world applications are often subject to varying degrees of acoustic disturbances, which can significantly affect the performance of AI systems Qayyum et al. (2019). By focusing on the implementation of CNNs, our study not only aims to contribute to the advancement of emotion recognition in Spanish speech but also seeks to develop robust models capable of maintaining high accuracy in challenging and noise-prone environments. This dual focus on language specificity and noise resilience underscores the innovative and practical significance of our research in the evolving landscape of human-computer interaction. Gupta et al. (2020)

3 Methodology

3.1 Dataset

The primary dataset used in this study is the Mexican Emotional Speech Database (MESD), sourced from Kaggle. This database comprises a total of 862 datapoints. It includes a diverse array of audio recordings, encapsulating six distinct emotional states: Neutral, Anger, Disgust, Fear, Happiness, and Sadness. Uniquely, the dataset encompasses three different vocal types: infant voices, and adult voices classified by gender - female and male. This variety within the dataset provides a comprehensive foundation for training and evaluating our CNN model in recognizing and classifying a wide spectrum of emotional expressions in speech.

3.2 Architecture

The Convolutional Neural Network (CNN) architecture utilized in this study is identical for both the experiments involving the original dataset and the dataset augmented with noise. This architecture is specifically designed to efficiently process and classify emotional content in speech signals. The CNN comprises multiple layers, each contributing to the extraction and interpretation of features from the input speech data. Below is a detailed breakdown of the CNN architecture:

Layer Type	Output Shape	Number of Parameters	Description
Conv2D	(None, 457, 255, 16)	160	This layer applies 16 filters of size 3x3 with 'relu' activation. It is the first layer and takes an input shape of (459, 257, 1).
Conv2D	(None, 455, 253, 16)	2,320	Another convolutional layer with the same number of filters and size, continuing the feature extraction process.
Flatten	(None, 1841840)	0	This layer flattens the 3D output of the previous layer into a 1D array to be used in the dense layers.
Dense	(None, 128)	235,755,648	A dense layer with 128 neurons and 'relu' activation, designed to interpret the features extracted by the convolutional layers.
Dense	(None, 6)	774	The final layer with 6 neurons corresponding to the emotion classes, using 'softmax' activation for classification.

Table 1: CNN Architecture for Emotion Recognition in Speech

Total trainable parameters: 235,758,902
Total non-trainable parameters: 0

3.3 Implementation

In the implementation phase, TensorFlow is employed for both preprocessing and model development. Audio files are first standardized to a uniform sampling rate and converted to mono-channel signals to maintain consistency across the dataset. Emotional states within the audio are categorized and one-hot encoded, a necessary step for classification by the CNN. The audio is then preprocessed to uniform length, either by trimming or padding, and transformed into spectrograms suitable for CNN input.

To bolster the model's robustness, we apply noise augmentation techniques using white, pink, and brown noise, which simulate various real-world audio conditions. This step is crucial for enhancing the model's performance in noisy environments. Following this, the dataset undergoes shuffling and batching, before being split into training and testing sets in a 70/30 ratio.

3.4 Experimentation

The experimental framework consisted of four distinct tests to evaluate the performance of emotion recognition models under various conditions. The first experiment was conducted using the original dataset without any added noise to establish a baseline for the CNN model's performance in ideal acoustic conditions.

In the second experiment, the dataset was augmented with a mixture of white, pink, and brown noise to test the CNN model's robustness in noisy environments. This aimed to reflect more challenging and realistic scenarios where ambient noise is present.

The third experiment compared a CNN model with no added noise against a 1D CNN-LSTM architecture. The 1D CNN-LSTM model used Mel-Frequency Cepstral Coefficients (MFCC) as input, and consisted of one-dimensional convolutional layers, dropout layers for regularization, LSTM layers for temporal sequence learning, a flattening layer, and a dense layer for output, predicting the estimated emotion.

In the fourth experiment, a 2D CNN-LSTM model was examined, also without noise interference. This architecture took spectrogram inputs, using two-dimensional convolutional layers followed by dropout layers, LSTM layers, a flattening layer, and a final dense layer for emotion classification. Both models were configured with a set number of filters in the CNN layers, a predefined dropout rate, and were optimized using the Adam optimizer with cross-entropy loss.

The obtained results are summarized in the following confusion matrices and performance metrics:

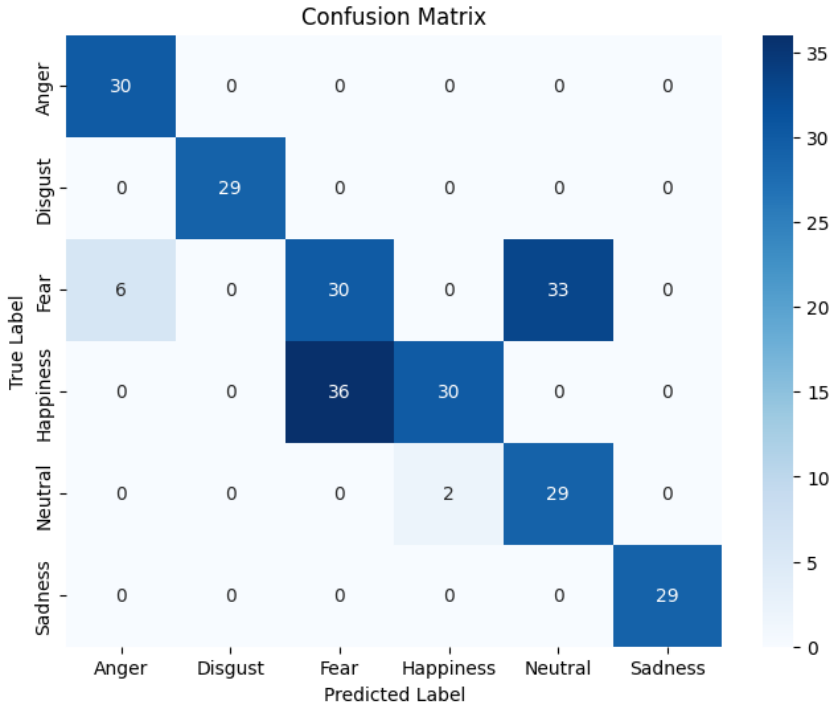

Figure 1: Confusion matrix for CNN model

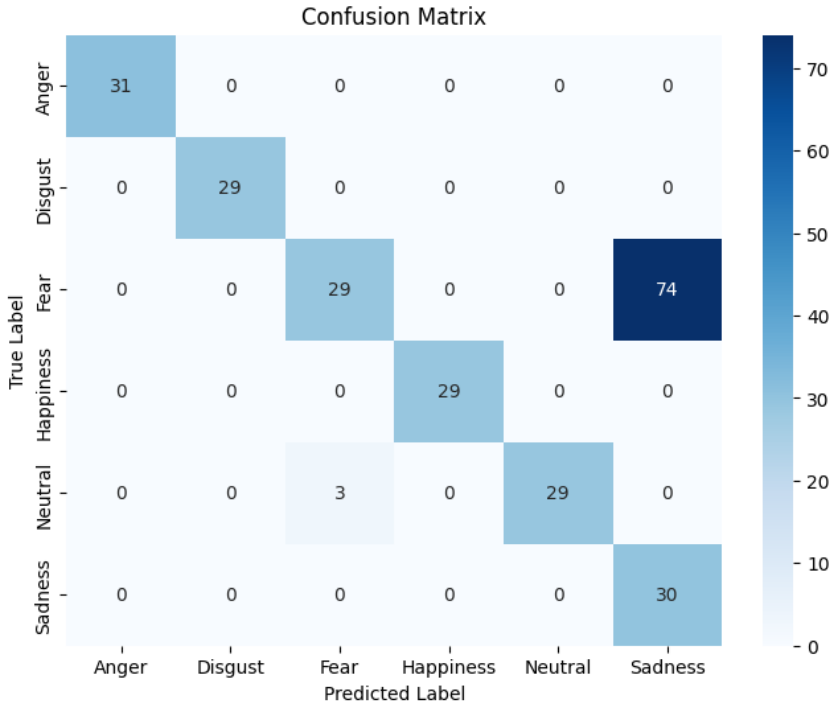

Figure 2: Confusion matrix for CNN model with noise

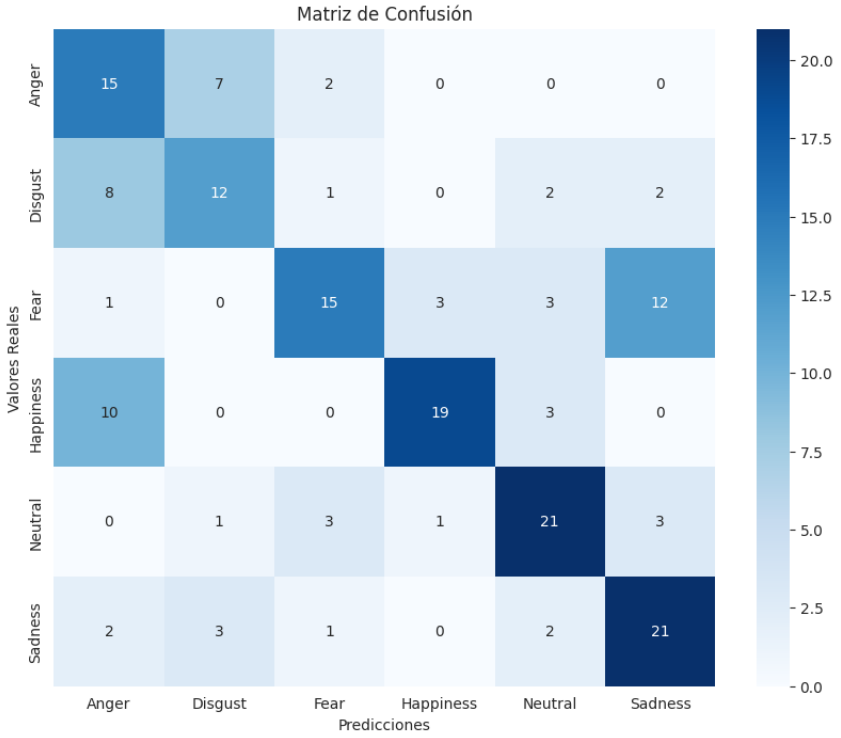

Figure 3: Confusion matrix for 1D-CNN-LSTM model

Figure 4: Training and validation accuracy for 1D-CNN-LSTM model

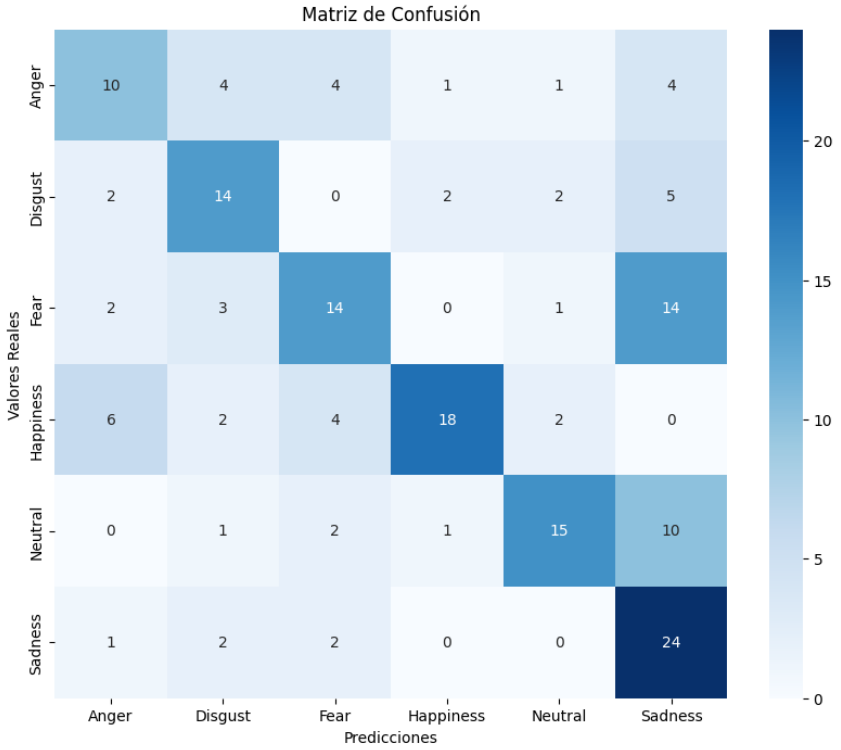

Figure 5: Confusion matrix for 1D-CNN-LSTM model

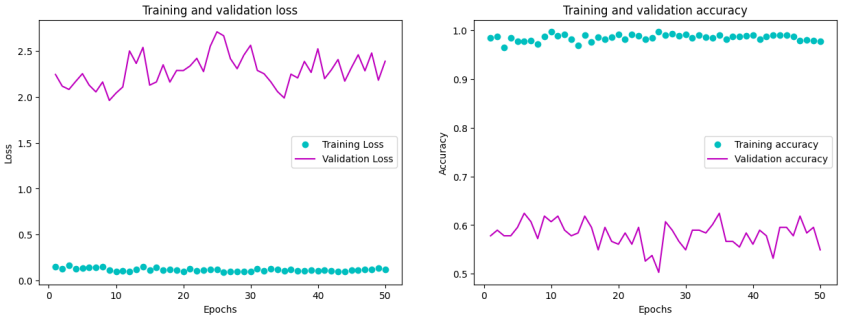

Figure 6: Training and validation accuracy for 2D-CNN-LSTM model

These four experiments collectively aimed to measure the effectiveness of different neural network architectures in recognizing emotions from speech and to determine their efficacy.

4 Results

The results of our study are presented, including the performance metrics such as loss, recall and presicion .

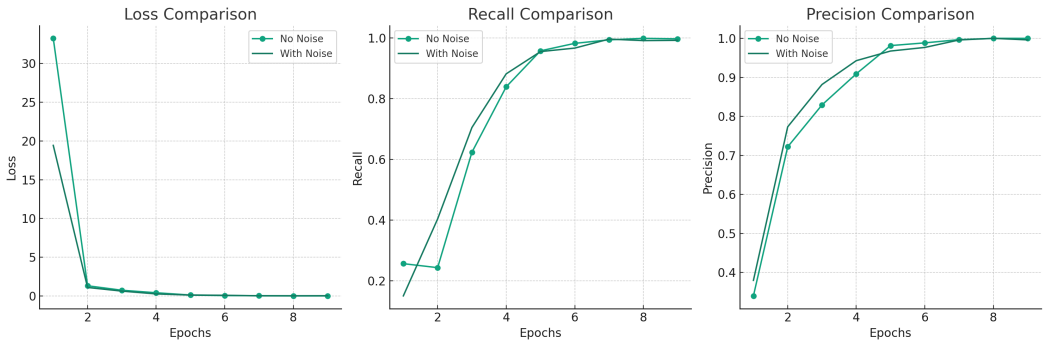

Figure 7: Graphical comparison of Loss, Recall, and Precision for CNN models with and without noise

4.1 Summary of Findings

The results of the experiments are encapsulated as follows:

CNN without noise:

- Achieved the lowest loss (0.0183), indicating superior error minimization.
- Attained the highest recall and precision, both at 98.96%, suggesting excellent model performance.

CNN with noise:

- Exhibited a slightly higher loss (0.0196), yet still performed well.
- Recorded a recall and precision of 98.60%, demonstrating robustness in noisy conditions.

1D CNN-LSTM:

- Reported a significantly higher loss (0.9130398035049438), implying greater error rates.
- Had a decrease in accuracy to 75.72% and recall to 0.7456647157669067, indicating less effectiveness in emotion recognition.

2D CNN-LSTM:

- Showed the highest loss (1.9295319387171362), indicating the largest error among the models.
- Achieved an accuracy of 58.96% and the lowest recall (0.560693621635437), suggesting challenges in capturing relevant features for emotion classification.

These findings indicate that while the CNN models trained with and without noise showcase high efficiency, the CNN without noise demonstrates a marginal advantage in all evaluated metrics. Conversely, the 1D and 2D CNN-LSTM architectures, despite their complexity and potential for capturing temporal and spatial features, fall behind the simpler CNN architectures in performance metrics.

5 Conclusions and Discussion

The deployment of Convolutional Neural Networks (CNN) showed high effectiveness in classifying emotions from clean voice recordings, as evidenced in Experiment 1, where high precision and recall metrics (98.96%) were achieved with a loss of only 0.0183. The confusion matrix from this experiment displays excellent classification among different emotions, with few misclassifications, mainly between fear and sadness.

Introducing various types of noise into the dataset (Experiment 2), the CNN maintained outstanding performance, with a slight decrease in precision and recall (98.60%) and a slightly increased loss (0.0196). The confusion matrix indicates an increase in misclassification between happy and neutral emotions, suggesting that noise primarily affects the distinction between positive and neutral emotions.

In contrast, more complex architectures, such as 1DCNN-LSTM and 2DCNN-LSTM, failed to match the performance of the simple CNN, with a notable drop in precision and recall. These results underline the CNN's capability to capture relevant emotional characteristics in speech, even in the presence of noise, and question the necessity and efficiency of more complex models in this domain.

The findings suggest that CNNs, with proper design and training, can be powerful and efficient tools for emotion recognition in realistic environments, paving the way for future developments in interactive systems sensitive to the user's emotional state.

References

V. Gupta, S. Juyal, G. P. Singh, C. Killa, and N. Gupta. Emotion recognition of audio/speech data using deep learning approaches. *Journal of Information and Optimization Sciences*, 41(6): 1309–1317, 2020.

D. Issa, M. F. Demirci, and A. Yazici. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894, 2020.

A. B. A. Qayyum, A. Arefeen, and C. Shahnaz. Convolutional neural network (cnn) based speech-emotion recognition. In *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, pp. 1–5. IEEE, 2019. Current trends on virtual presence will provide the groundwork for truly immersive communication designed to transcend time and space.

R. Ullah, M. Asif, W. A. Shah, F. Anjam, I. Ullah, T. Khurshaid, L. Wuttisittikulkij, S. Shah, S. M. Ali, and M. Alibakhshikenari. Speech emotion recognition using convolution neural networks and multi-head convolutional transformer. *Sensors*, 23(13):6212, 2023. doi: 10.3390/s23136212.

K. Venkataramanan and H. R. Rajamohan. Emotion recognition from speech. 2019. doi: 10.48550/arXiv.1912.10458.