# A Speech Emotion Recognition Model to Detect Aggressive Behavior in Dialogues

**Gabriel Gonçalves Ferreira**[1], **Johnny Marques**[1]

[1]Postgraduate Program in Electronic and Computer Engineering
Instituto Tecnológico de Aeronáutica (ITA)
Praça Mal. Eduardo Gomes 50 – São José dos Campos – SP – Brasil

gabrifer@gmail.com, johnny@ita.br

***Abstract.*** *Speech Emotion Recognition (SER) is a multidisciplinary field that develops computational models to detect and analyze emotional states automatically conveyed through speech signals. Using signal processing, machine learning, and natural language processing techniques, SER systems extract relevant features from audio data and classify emotions into distinct categories such as happiness, sadness, anger, and more. This work aims to leverage the latest SER techniques to build a robust model that can detect aggressive behavior in dialogues solely based on audio input signals.*

## 1. Context

Audio data can be one of the most valuable sources of information available nowadays. Within audio data, human speech represents the basics of communication and is one of the most complex and valuable when it comes to solving different kinds of problems among different fields of studies [Padhy et al. 2022].

The field of Speech Emotion Recognition (SER) leverages speech audio signals and tries to identify their associated emotions, such as happiness, sadness, anger, and others. SER is expected to be helpful in many different industries, such as marketing, healthcare, customer satisfaction, social media analysis, stress management, and many more [Jolly et al. 2023].

Anger is a negative emotional state that substantially influences human actions and can often give rise to dysfunctional behavior. When left unchecked, it might lead to a plethora of interpersonal problems [Nahar and Ali 2022]. Hence, effectively detecting aggressive behavior in dialogues is an important problem to address, potentially impacting human-computer interactions positively. To address this matter, this work proposes building a SER model specialized in identifying aggressive behavior in dialogues. This model will leverage the latest techniques in the SER field, which are described in the available literature.

## 2. Adopted Process

Predefined steps need to be fulfilled to build the proposed model. This work proposes an SER approach that relies on the latest Machine Learning techniques, which are already used to solve a wide range of problems. Like any other Machine Learning system, it is important that the training and validation dataset present reliable and valuable data. Thus, the first step before building the actual model is selecting the right data.

A literature review was conducted to identify pre-built datasets that would fit the purpose of this work and provide valuable samples to build the proposed model. During this process, two main datasets were identified: the IEMOCAP dataset and the RAVDESS dataset. Both are well-known within the field of SER and are used in several different applications. However, they do not contain audio clips annotated in a way that fits the purpose of the model of identifying aggressive behavior. Plus, both datasets are composed only of samples that are monologues and not dialogues or conversations.

A tailored dataset was curated due to the absence of a specialized dataset for aggressive behavior detection. This dataset was created using podcast audio snippets exhibiting various forms of aggressive behavior during conversations sourced from YouTube clips. The videos were selected in a way that only the English language was selected. Subsequently, these podcast audio snippets were segmented into 10-second units, manually annotated into two classes: **positive**, denoting the presence of aggressive behavior, and **negative**, indicating the absence of detected aggressive behavior. The streamlined process is depicted in the Figure 1.

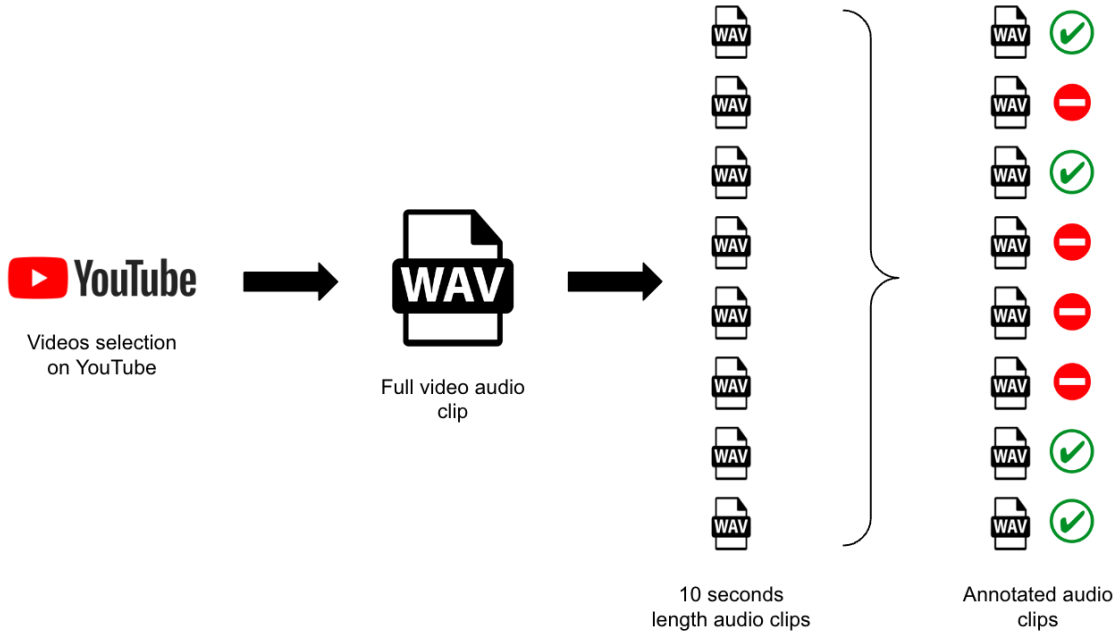


**Figure 1. Audio extraction and annotation process**

The final curated dataset comprises 300 audio clips of 10 seconds in length and is stored in MP3 format. From the samples, 100 are from the positive class, meaning that they contain aggressive behavior, and 200 are from the negative class, meaning the absence of aggressive behavior. Also, the audio clips were stored at a 441 Hertz sample rate and a 16-bit resolution.

Having prepared the dataset, the next step is building a feature extraction process to extract the most relevant audio features that will feed the Machine Learning model. For this phase, all audio clips were converted to a raw waveform and later translated into their frequency representation using the Short-Term Fourier Transform (STFT). Based on the frequency representation of the audio, it is now possible to extract the so-called

Mel-frequency cepstral coefficients (MFCCs). Those coefficients are currently known to be the most representative features of a speech signal, given their ability to capture the different nuances of human speech, such as pitch and timbre [Seknedy and Fawzi 2023]. This process is illustrated in the Figure 2.

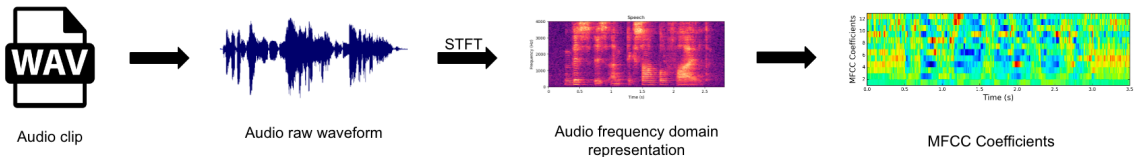


**Figure 2. Feature extraction process**

After extracting the features, the next phase relies on training the model using modern machine-learning techniques. Literature demonstrated that two effective algorithms for SER problems are Support Vector Machines (SVM) and Convolutional Neural Networks (CNN). For training purposes, both algorithms used the 10-fold validation strategy. This strategy divides the dataset into 10 subsets (folds), and the model is trained and evaluated 10 times. Each iteration uses a different fold as the test set, while the remaining nine folds are used for training. This process helps obtain a more reliable estimate of the model's performance by reducing the impact of variability in a single train-test split. Figure 3 represents the simplified training and prediction process.

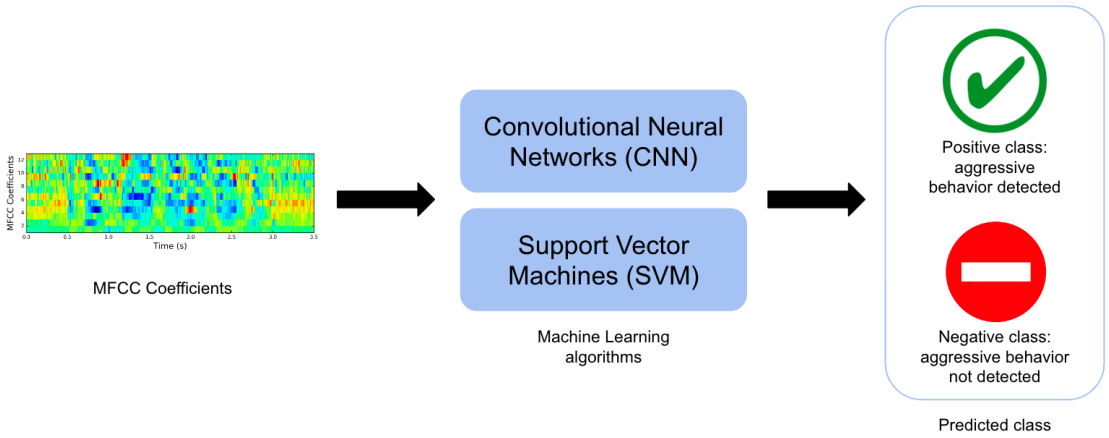


**Figure 3. Training process**

Finally, the trained model can be deployed in real-time systems such as a Web App or any batching process mechanism. For instance, the model could be used to automatically audit batches of call center recordings for quality assurance purposes.

## 3. Solution

Currently, the model is being trained and tuned using several different data points extracted from diverse audio signals. Preliminary results demonstrate that the CNN approach can reach accuracy levels close to 90%. However, it is important to note that the first version of the model is trained using audio clips in the English language and probably would not perform well in different language contexts.

Nevertheless, despite the model's specialization in English, the procedure for constructing diverse models applicable to other languages remains similar, differing primarily in the choice of training datasets. Therefore, the process of developing new models is intricately linked to the task of acquiring suitable data points for training.

Finally, it is reasonable to assume that, given the importance of effective ways of detecting aggressive behavior based on speech, automated Machine Learning models that perform such tasks can be useful. While the proposed model has room for improvement, it already showcases positive and encouraging preliminary results.

## Acknowledge

This work received support from the Financiadora de Estudos e Projetos (FINEP), contract 01.22.0615.00.

## References

Jolly, M., Gupta, P., Bansal, S., Gupta, G., and Goel, R. (2023). Machine learning based speech emotion recognition in hindi audio. In *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE.

Nahar, M. and Ali, M. E. (2022). A deep ensemble approach of anger detection from audio-textual conversations. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE.

Padhy, S., Das, N., Tiwari, S., and Arora, S. (2022). AI based web app and framework for detecting emotions from human speech. In *2022 2nd Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON)*. IEEE.

Seknedy, M. E. and Fawzi, S. (2023). Arabic english speech emotion recognition system. In *2023 20th Learning and Technology Conference (L&ampT)*. IEEE.