Speech Emotion Recognition in Portuguese for SofiaFala: SER SofiaFala

Alexander Scaranti[1], Douglas Antonio Rodrigues Silva[1], Prof. Fernando Meloni[1], D.Sc. and Prof. Alessandra Alaniz Macedo[1], D.Sc.

[1]*University of São Paulo (USP)*

Abstract

Emotion recognition through speech processing has been increasingly demanded as a response to scientific advances and improvement in information technologies. However, a gap exists when the demand concerns projects in the Portuguese language. Here, we propose a method for extracting and recognizing emotion in the Portuguese language. We have evaluated response time, length, silence ratio, long silence ratio, and silence rate. According to the SER 2022 evaluation, our strategy can reach a macro-averaged F1 score of 55% on a very imbalanced dataset. We have aligned our results with the SofiaFala project, which supports speech training in children with Down syndrome.

Keywords

Speech Processing, Emotion Recognition, Portuguese Language, Natural Language Processing, Artificial Intelligence, SofiaFala

1. Introduction

In the last two years, the COVID-19 pandemic has swept the world, leading to new demands for different approaches to communication and interaction. In turn, the 5G technology, which emerged in the second decade of the 21st century, supports new possibilities. In this context, modern algorithm-aligned voice processing tools have paved new ground for improving people's quality of life, assisting people with incapacity, or even assisting long-distance interaction. These algorithms, created with researchers' hard work, have allowed new opportunities such as the Speech Emotion Recognition task to be envisioned.

Portuguese-speaking countries suffer from a scarcity of tools to support speech and emotion recognition. For instance, speech sound and language vary in the many regions of Brazil, a country with continental dimensions. This situation demands research into speech manipulation by considering utterances that sound prosodically distinct. Speaking manner or speech disorders can interfere with speech emotion recognition.

The SofiaFala software[1], developed in the LIS laboratory at USP-Ribeirão Preto-SP, recognizes sounds and images produced during exercises and provides reports on assistive speech training for speech disorders of children with Down syndrome [2].

International Conference on the Computational Processing of Portuguese, March 21, 2022, Fortaleza, Brazil

alexander.scaranti@gmail.com (A. Scaranti); douglasarsilva@gmail.com (D. A. R. Silva); fernandomeloni@alumni.usp.br (F. Meloni); ale.alaniz@usp.br (A. A. Macedo)

Expressing emotions through speech is a part of oral communication through the voice. For voice analysis and knowledge to be generated, different data types (texts, images, and types of speech) must be manipulated through a coordinated analysis that considers connections and particularities of sound. This manipulation is challenging and desirable. For instance, SofiaFala can take advantage of emotion recognition during speech training.

Here, we propose a speech emotion recognition method that uses the corpus provided by the SER committee, namely CORAA version 1.1, which is composed of approximately 50 minutes of audio segments. Our work focuses on the clipping of emotions in speech. We intend to incorporate SER as a module of the SofiaFala app.

2. Our Proposal: SER System

Considering the dataset CORAA available for the shared task and aiming at recognizing emotion, we have developed a computer system called SER to carry out natural language processing and other steps.

SER was built in Python, and it executed the experiments presented in Section 3. Figure 1 illustrates the process and the computational modules.

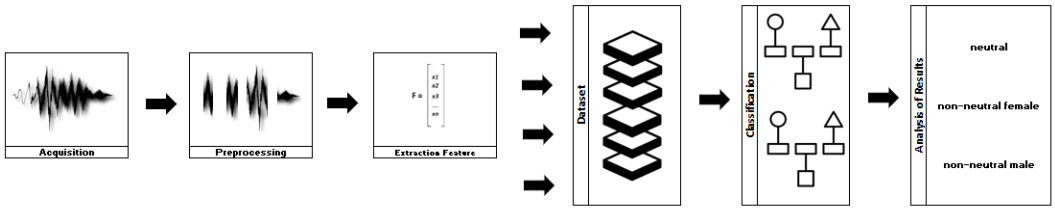


Figure 1 - The SER System: Process and Computational Modules.

SER is composed of the following stages:

- *Acquisition.* All information acquired from the dataset CORAA-v1.1 has three classes: neutral, male neutral, and female neutral, amounting to 625 audio fragments that total 50 minutes of speech. The neutral class comprises audio segments without a well-defined emotional state. The non-neutral class represents segments associated with one of the primary emotional states in the speaker's speech. This non-neutral dataset, called the C-ORAL-BRASIL I corpus, has informal spontaneous speech of Brazilian Portuguese (Raso and Mello, 2012).
- *Preprocessing.* We processed all the acquired audios to clean and to try to improve the performance of the next step, feature extraction. We also applied filters to remove noise from the audios [3]. Moreover, we converted all the audios from stereo to mono and distributed them into three classes: neutral, non-neutral female, and non-neutral male.
- *Prosody and Feature Extractions.* Extraction is the method that analyzes and brings out information from the audio so that the learning model can be developed. Next, we will detail it. In terms of feature extraction, our system carried out some steps by considering:

- *Prosody Extraction.* Prosody or speech elements are properties of linguistic functions with features. We extracted the following features from all the audios in the base: response time, response length, silence ratio, long silence ratio, silence rate, frequency, and intensity.
- *Feature extraction with MFCC.* MFCC is a feature extraction method for audio that uses the Fourier transform [4]. MFCC is the most used method in speech processing because it is the most suitable for representing audio and signal characteristics. This method captures sound exactly as humans recognize it.
- *Transformation with Spectrogram (MEL).* Logarithmic Transformation of an audio signal frequency is said to be a MEL scale whose its central idea is sounds of equal distances (MEL scale) that mimic our perception of sound[5]. Transformation from the Hertz scale to the Mel scale is as follows:

$$m = 1127.log(1 + f/700)$$

- *Aggregation of Chromagram.* We used this strategy to increase the robustness of our logarithmic frequency spectrogram to variations in timbre and instrumentation. The main idea of chroma features is to aggregate all spectral information related to a given pitch class into a single coefficient.

- *Classification.* We applied an MLP Neural Network[6] with the following parameters: Hidden Layer = 500, interaction = 600, MLPClassifier.
- *Analysis of Results.* After the procedures described above, we divided the recognized emotions into neutral, neutral-male, and neutral-female.

3. Results and Discussion

The trained model has an F-Score of 84% when 80% (550 audios) of the training base (see Table 1) is used. The other 20% of the training base (125 audios in total) is for the tests. In Table 2, a confusion matrix shows data from the experiments. After we applied the developed model to the available test base and submitted it to the SER, we achieved an accuracy rate measured by the F-Score of 55% in the results.

	%	
Features	**Score performance**	**Official F1 Score**
Extraction of Prosody	82.15	55.00
MFCC	81.53	
Spectrogram	75.80	
Chromagram	78.98	
Extraction of Prosody + MFCC + Spectrogram + Chromagram	84.34	

Table 1 - Distribution of Results

Confusion Matrix			
	neutral	**non-neutral-male**	**non-neutral-female**
neutral	**117**	8	3
non-neutral-male	2	**16**	0
non-neutral-female	2	1	**8**

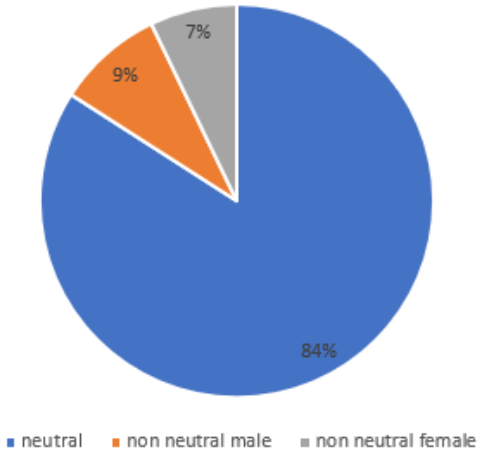
Table 2 - Matrix Confusion

By using the 308 audios, we generated the results from the data available for testing. For classification, we created the MLPClassifier. As a result, 259, 27, and 22 audios were labelled as neutral, non-neutral female, and non-neutral male, respectively as shown in Table 3.

Label	Sum
neutral	259
non-neutral-male	27
non-neutral-female	22

Table 3 - Classification

Graph 1 depicts the classification distribution. Neutral audios (84%) were the majority in the dataset, followed by non-neutral female (9%), and non-neutral male (7%).



Graph 1 - Distribution of Results

4. Final Remarks

We have proposed a method for extracting and recognizing emotion in the Portuguese language. We have carried out a simple process based on preprocessing strategies, prosody extraction, MFCC, MEL, and Chromagram. We have reached our goal by using the dataset CORAA-v1.1,

which has 625 audios classified as neutral, masculine, and feminine language. Our strategy does not take advantage of external models to manipulate the data, and, according to the SER 2022 evaluation, it can reach a macro-averaged F1 score of 55%. Due to simplicity, we have been to generate the results in 18 seconds by considering the whole set of CORAA audios.

By considering the SofiaFala project, we have looked for new possibilities for monitoring, understanding, and even treating speech and emotion. Here, we have developed a SofiaFala module aiming at improving a person's functional capacity of speech, and hence, communication. Moreover, we have contributed to the usability evaluation of SofiaFala[7].

As future work, we will integrate our SER module into the SofiaFala app. Moreover, we will evaluate the use of external models.

Acknowledgments

This research was carried out at the Center for Artificial Intelligence (C4AI- USP), with support by the São Paulo Research Foundation (FAPESP grant 2019/07665-4) and by the IBM Corporation.

The authors would like to thank the SofiaFala group, CNPq, C4AI- USP and SER 2022 organizers for their support.

References

[1] D. S. de Paula, S. R. G. Panico, J. C. Daneluzzi, E. E. S. Ruiz, J. C. Felipe, A. A. Macedo, Sistema de informação de apoio ao programa de educação para pais e famílias, in: Proceedings of XI Congresso Brasileiro de Informática em Saúde, 2008.

[2] P. H. D. G. Rissato, A. A. Macedo, Sofiafala: Software inteligente de apoio à fala, in: Anais Estendidos do XXVII Simpósio Brasileiro de Sistemas Multimídia e Web, SBC, 2021, pp. 91–94.

[3] I. Braga, Avaliação da influência da remoção de stopwords na abordagem estatística de extração automática de termos, in: 7th Brazilian Symposium in Information and Human Language Technology (STIL 2009), So Carlos, SP, Brazil, 2009, p. 18.

[4] C. Ittichaichareon, S. Suksri, T. Yingthawornsuk, Speech recognition using mfcc, in: International conference on computer graphics, simulation and modeling, 2012, pp. 135–138.

[5] K. Venkataramanan, H. R. Rajamohan, Emotion recognition from speech, arXiv preprint arXiv:1912.10458 (2019).

[6] H. Palo, M. N. Mohanty, M. Chandra, Use of different features for emotion recognition using mlp network, in: Computational Vision and Robotics, Springer, 2015, pp. 7–15.

[7] F. Meloni, B. Sicchieri, P. Mandrá, R. Bulcão-Neto, A. A. Macedo, A nonverbal recognition method to assist speech, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2021, pp. 360–365.