Performance Evaluation of Emotion Recognition Algorithms in Brazilian Portuguese Language Audios

Omar Rodrigues da Silva, Luisa Medina Fermino Carlos, Felipe Corchs, Fátima L. S. Nunes, and Ariane Machado-Lima

***Abstract*—Emotion recognition in humans is a multidisciplinary field that involves the analysis of several types of data. Computational techniques in pattern recognition and machine learning have been applied to emotion analysis using various modalities, including gestures and facial expressions (visual signals), the lexical content of spoken or written language (textual signals), and the sound of speech (acoustic signals). Acoustic analysis leverages characteristics of speech, such as frequency, tone, intensity, and harmonics, which are strongly linked to emotional states. This type of acoustic analysis has numerous applications, such as examining relationships through dialogue, improving human-machine interaction, and detecting psychiatric disorders, among others. While the performance of audio-based emotion recognition algorithms is well explored in several languages, there is a notable gap in the literature regarding emotion recognition in audio dialogues in Portuguese. This article aims to address this gap by evaluating the performance of three algorithms that use different models to recognize discrete emotions – happiness, anger, fear, disgust, sadness, surprise, and neutral – in Brazilian Portuguese audios. The results indicate that significant advances are still needed for effective emotion recognition in this language. Among the algorithms studied, the maximum accuracy and F1-score achieved were 0.53, and no peer-reviewed publications were found specifically on emotion recognition in Portuguese involving multiple datasets.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/9577

***Index Terms*—Emotion recognition. Portuguese language audios. Emotional data sets. Acoustic features. Affective computing. Algorithm performance comparison.**

I. Introduction

THE recognition of discrete emotions, such as happiness, anger, fear, disgust, sadness, and surprise, has been extensively studied across various domains. Applications range from therapeutic contexts, such as couples therapy, drug addiction treatment assessment, and psychotherapeutic process evaluation [1].

The use of acoustic characteristics of speech (frequencies, tone, intensity, harmonics, etc.) in emotion recognition has attracted the attention of many researchers, because these characteristics have a strong connection with people's emotional state and they are deeply studied [2]. To be processed by emotion recognition algorithms, audio files first undergo a feature extraction process that involves capturing the acoustic signals present in the recordings, applying mathematical algorithms, and transforming them into numerical quantities. Hundreds of audio features can be extracted, commonly grouped in cepstral, time and frequency features. The extraction can be local (from periods of miliseconds) or global (from longer periods, such as seconds or minutes). Most of the studies that analyzed emotions between 2016 and 2021 used characteristics of these three groups [3]. The main features applied in overall studies in multiple languages are Mel-Frequency Cepstrum Coefficients (cepstral), intensity (time), fundamental frequency (frequency) and pitch (time).

Emotion recognition through acoustic processing using machine learning has been widely explored in numerous studies employing a variety of technologies, using a diverse set of features and algorithms, being the most commonly used the neural networks, SVM-Support Vector Machines, FIS-Fuzzy Inference System and GMM-Gaussian Mixture Model, among others.

The effectiveness of emotion recognition is known to be language-dependent [4]. Studies have shown that the more languages included in the training phase, the better the model's ability to recognize emotions in languages not included in the training set [4]. Most studies focusing on a single language have predominantly targeted English [5] or German [6]. Recently deep learning approaches have emerged to improve the scenario, such as a new method called Improved and a Faster Region-based Convolutional Neural Network (IFR-CNN) using the German EMODB and Serbian GEES datasets [6] and Multi-task learning enhanced speech emotion recognition (MTLSER) [7] applying a pre-trained model wav2vec2.0 [8] using the English IEMOCAP. Other innovation applied Convolutional Neural Networks (CNNs) with pre-trained model HuBERT (Hidden-Unit BERT) [9] using English datasets RAVDESS. Although these new approaches have been successful in some cases, they require huge data volume for training in the target language. No approach nor available dataset were found considering Portuguese idiom.

Although the majority of articles declares the dependency on the language, there are studies mentioning common features able to turn the emotion recognition independent from the

The associate editor coordinating the review of this manuscript and approving it for publication was Carlos Thomaz (*Corresponding author: Omar Silva*).

Omar Silva, F. Nunes, F. Corchs, and A. Machado-Lima are with Universidade de São Paulo, Sao Paulo, Brazil (e-mails: omarrodriguesdasilva@gmail.com, fatima.nunes@usp.br, felipe.corchs@usp.br, and ariane.machado@usp.br).

L. M. F. Carlos is with Paradigma Centro de Ciências e Tecnologia do Comportamento, Sao Paulo, Brazil (e-mail: luisamedinafc@gmail.com).

language. This kind of attempt is showed in a study that, despite using English, analyze different accents (audios of people from Australia, Singapore, Kenya, India and United States) [10]. However, there are no studies published in peer-reviewed journals that focus on emotion recognition from audio in Portuguese or multilanguage studies applied to Portuguese. Only three relevant studies have been found on Brazilian university websites, specifically in Portuguese [11] [12] [13]. However, these studies trained and tested their approaches with specific datasets. No additional study was conducted to prove the generalization capacity of the proposals. Notably, a comparison of different algorithms for emotion recognition in Portuguese, together to an evaluation of their generalization capacity, remains a gap in the literature.

The aim of this study was to compare the emotion recognition performance of three tools [11] [12] [13], proposed elsewhere for Portuguese audios using the same dataset and a standardized approach for testing. The main contribution is to define an experimental standardized protocol to evaluate the cited approaches, consequently evaluating the reproducibility of the results reported in the mentioned studies. For accomplish this protocol, we established two adequate datasets – one of them generated by the research group under controlled situations [14]. This dataset is a second contribution, since actors reproduced some planned speeches, allowing us to correctly compare the results with the ground truth emotion present in each speech. We conducted a statistical testing of the features from both datasets, which is close to a real world scenario. This is an innovative testing methodology.

Each of these algorithms incorporates its own feature extraction methods. Therefore, the comparison encompasses not only the differences in classification algorithms but also the distinct feature sets employed by each solution. The evaluation process involves two simulated audio datasets and uses macro-average recall, macro-average precision, and macro-average F1-score as performance metrics.

II. Related Work

Studies comparing different algorithms for emotion recognition based on audio in languages other than Portuguese are found in the literature, such as those mentioned in this section.

The problem of emotion recognition in German audios is addressed by Dogdu et al. [15]. Using the EMODB dataset [16], the authors compared the performance of seven classification algorithms: Multilayer Perceptron Neural Network (MLP), J48 Decision Tree (DT), Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO), Random Forest (RF), k-Nearest Neighbor (KNN), Simple Logistic Regression (LOG) and Multinomial Logistic Regression (MLR) with 10-fold cross-validation using four predefined acoustic feature sets in the openSMILE library [17] (IS-09, emobase, GeMAPS and eGeMAPS). The results indicated that the emobase feature set and the SMO, MLP and LOG algorithms performed best (respective accuracies of 87.85%, 84.00% and 83.74%) than RF, DT, MLR and KNN (respective accuracies of 73.46%, 53.08%, 70.65% and 58.69%).

Focusing on the English language, Atmaja and Akagi [18] evaluated whether emotions expressed through singing are more noticeable and intense than those expressed in dialogues. To conduct this evaluation, the authors used the RAVDESS dataset (*Ryerson áudio-Visual Database of Emotional Speech and Song*) [19], three feature sets (GeMAPS from the openSMILE library, pyAudioAnalysis [20], and LIBROSA [21]), two categories of features (low level and high level) and compared four algorithms (MLP-Multilayer Perceptron, LSTM-Long Short-Term Memory, GRU-Neural Network Gated Recurrent Unit, and CNN-Convolutional Neural Networks). The results demonstrated that emotional expression in music is more noticeable and intense than in spoken dialogue. In emotion recognition, the high-level feature sets, as well as the LIBROSA HSA feature set and the LSTM algorithm performed best (respective accuracies for LSTM, GRU, MLP and CNN of 82.0%, 81.2%, 79.4% and 74.3%).

Kannadaguli and Bhat [22] conducted a study focused on the Indian language Kannada, where they developed an emotion recognition system using an Artificial Neural Network (ANN) and compared its performance with a system based on a Hidden Markov Model (HMM). The acoustic features employed in their study were Mel-Frequency Cepstral Coefficients (MFCC), Delta/Double Delta MFCC (DMFCC/DDMFCC). The ANN-based model outperformed the HMM-based model, achieving an Emotional Error Rate (EER), the ratio of wrongly classified emotions to the total number of emotions, ranging from 0% to 0.15%, while the HMM-based model had an EER between 0.10% and 0.25%.

Gondohanindijo et al. [23] conducted a study using 639 Indonesian audio samples to compare the accuracy of four algorithms, Support Vector Machine (SVM), Neural Network (NN), Random Forest (RF), and Naive Bayes (NB), in recognizing emotions based on Mel-Frequency Cepstral Coefficients (MFCC) acoustic features. The algorithms achieved accuracy rates of 100%, 99.7%, 97%, and 94.1%, respectively.

No studies were found that compare emotion recognition algorithms specifically for Portuguese language audio data.

III. Materials and Methods

This section outlines the materials and methods used in this performance evaluation.

A. Datasets used

This study selected two datasets in Brazilian Portuguese for the performance evaluation: BADEM and VERBO, which summarizes the main features in Table I. The duration of the audios was calculated using the FFMPEG software [24]. Both datasets were recorded in a controlled environment.

VERBO (Voice Emotion Recognition dataBase in Portuguese language) [25] is an actor-based dataset: six men and six women of different ages, from different regions and accents of Brazil. It contains 14 sentences for each discrete emotion and neutral which were validated by specialized annotators, achieving 76% agreement between the emotion intended by the actor/actress and that perceived by the annotators. The annotators were selected based on the following criteria: health professionals, professionals qualified in an area related to emotion or professionals with clinical experience. Each

TABLE I
INFORMATION FROM THE BADEM AND VERBO DATASETS. AGREEM: PERCENTAGE OF AUDIOS WITH AGREEMENT BETWEEN THE EMOTION INTENDED BY THE ACTOR/ACTRESS AND THAT PERCEIVED BY THE ANNOTATORS

Dataset	No. audios	No. actors + actress	Agreem. (%)	Length (sec)		
				min	max	average
BADEM	1008	6 + 6	80%	2.39	7.68	4.26
VERBO	1167	6 + 6	76%	1.19	5.48	2.40

annotator classified the 1167 audios without prior knowledge of the emotion represented but with knowledge about the categories involved. The sentences vary in length and type, ranging from short to long sentences, including questions and some nonsensical phrases, with durations lasting from one to five seconds

BADEM [14] (Brazilian Audio Dataset of EMotions) consists of videos featuring 12 sentences for each discrete emotion and neutral, performed by 12 actors (six male and six female), totaling 1008 recordings (84 videos of each actor/actress). The videos were evaluated by eight psychologists being each video classified by an evaluator. The evaluators' agreement with the intended emotion simulated by the actors ranged from 70.63% to 98.41%. The emotions with the highest agreement percentages were fear and disgust at 92.36%, while fear had the lowest agreement at 63%. In terms of expressiveness, the actors' ability to convey emotions in line with the evaluators' classifications ranged from 72.6% (the lowest agreement) to 92.8% (the highest agreement).

B. Algorithm Evaluation Process

According to Figure 1, the performance evaluation involved the two data sets described in the previous section (BADEM and VERBO).

Three tools that perform emotion recognition in Portuguese-language audio were compared: DEEP [12], Souza [13] and RosaJr [11]. The Souza and RosaJr tools directly receive the audios to be processed in Waveform Audio File Format with the feature extraction process carried out within the tool itself (Section III-D). The DEEP tool uses as input vectors of certain features pre-extracted by the openSMILE software [17]. Therefore, a pre-processing step of the input was necessary (Section III-C). Some adaptations were also made to the tools (Section III-D).

Five experiments were performed for these three tools:

- **VV**: 10-fold cross-validation using only the VERBO dataset. Both training and test audios were audios from the VERBO dataset;
- **BB**: 10-fold cross-validation using only the BADEM dataset. Both training and test audios were audios from the BADEM dataset;
- **AA**: 10-fold cross-validation was performed on a dataset formed by combining the VERBO and BADEM datasets (referred to as ALL). Both the training and test audio samples were sourced from this combined dataset;
- **VB**: training using the entire VERBO dataset and testing the entire BADEM dataset;
- **BV**: training the entire BADEM dataset and testing the entire VERBO dataset.

C. Preprocessing

The audio files from the BADEM dataset, originally in MPEG-4 Part 14 (MP4) format, were converted to the Waveform Audio File (WAVE) format using the FFMPEG package (community 2024). The VERBO dataset, already in WAVE format, did not require any conversion.

Feature extraction for the DEEP tool, applied to both the BADEM and VERBO datasets, was carried out using the openSMILE software [17] with specific configuration files: chroma-fft.conf, MFCC_CSV_OUT.conf, and prosody-Acf.conf. For all three groups of features, the parameters were set to a frame size of 0.050 seconds and a frame step of 0.010 seconds, meaning that features of 0.050 seconds in length were extracted every 0.010 seconds.

D. Execution of the Evaluated Tools

Table II shows the number of features produced by audio and the classifier-inducing algorithm used by each tool. All of them were used to recognize the six basic emotions (surprise, disgust, fear, happiness, sadness and anger) and the neutral expression. DEEP tool uses 32 types of features extracted every 0.01 sec of audio. The total number of features depends on the audio size and the set of features is complemented using the *padding* technique, described below, to reach the number of features of the largest audio in the training sample.

TABLE II
TOOLS EVALUATED. NO. FEAT.: NUMBER OF FEATURES EXTRACTED PER AUDIO. CLASSIF. ALGOR.: CLASSIFICATION ALGORITHM

Tool	No. feat.	Classif. algor.	Ref.
RosaJr	136	SVM	[11]
DEEP	32	CNN	[12]
Souza	20	RNN	[13]

*1) **Souza Tool**:* The Souza tool [13] is based on a recurrent neural network (RNN) trained from MFCC features of audios from the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset [19].

The MFCC technique is used to represent audio features by transforming the signal into a frequency spectrum using the Fourier transform, which is then mapped to the Mel scale. This non-linear scale is designed to better capture how the human ear perceives different frequencies, and then the signal is analyzed in terms of its frequency components. These features are extracted using the $feature.mfcc$ function from the LIBROSA library [21].

The parameters used are: (i) $sample\_rate$: 44100 Hz - representing the number of audio samples captured per second; (ii) $offset$: 0.5 seconds - means that the first half second will be disregarded; (iii) n_mfcc: 20 - number of MFCC coefficients that make up the calculated feature vector; (iv)

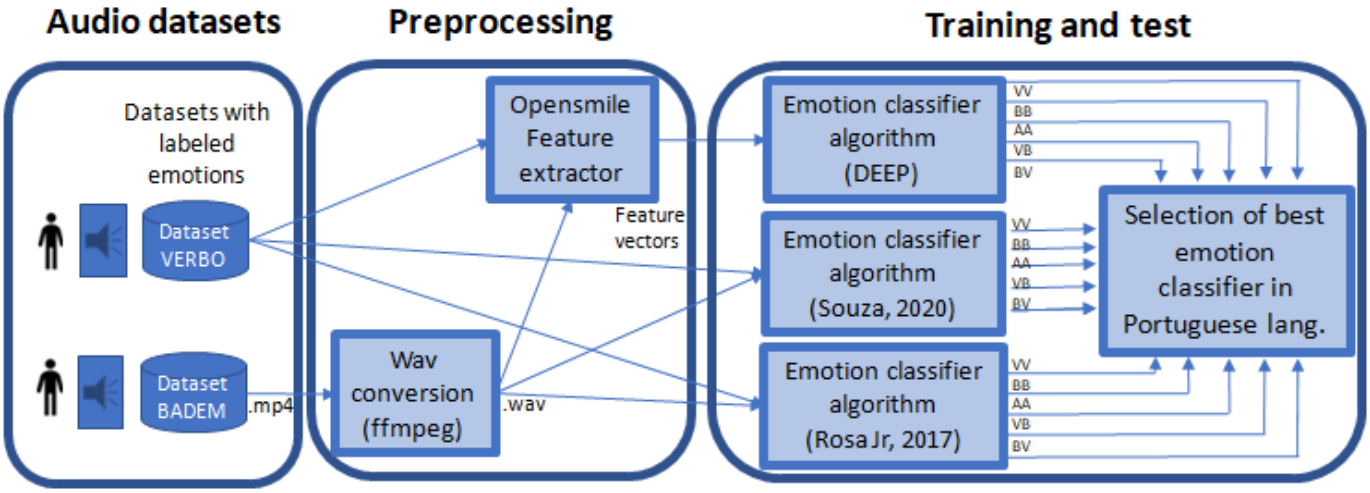

Fig. 1. Overview of performance evaluation process of emotion recognition tools. Two datasets were used (VERBO e BADEM) for training and tests, having three cross validation experiments (VV, BB e AA) and two experiments training on one dataset and testing on another (VB e BV).

duration: 3 seconds - audio interval considered for feature extraction. 20 features are extracted for each audio, each corresponding to the average value of an MFCC coefficient for every three seconds of the audio. Finally, each feature is normalized by the z-score technique [26].

Some audios from the VERBO dataset presented errors during feature extraction, as they were recorded with a coder-decoder parameterization (codec, a program used to compress or decompress digital media files, particularly audio and video) not supported by LIBROSA (codec pcm_f32le, 32 bits). To solve this problem, these audio files were converted to a compatible codec (pcm_s16le, 16 bits).

The RNN-GRU (Recurrent Neural Network - Gated Recurrent Units) classifier induction algorithm was executed with the same hyperparameters used by the author of the tool as described in the original project [13].

*2) **RosaJr Tool**:* The RosaJr tool [11] uses the pyAudioAnalysis library [20] to extract features and recognize emotions through the SVM classification algorithm using the *one-against-one* strategy for multiclass classification. The pyAudioAnalysis library is used for feature extraction, audio conversion, training, and testing, taking as input the directories containing the audio files for each emotion. Table III describes the parameters used by the RosaJr tool to extract the 136 acoustic features from each audio, by the pyAudioanalysis library. The parameters *short_term_window* and *short_term_step* define the window and the step used to extract features from short audio intervals, while *mid_term_window* and *mid_term_step* are used to calculate the averages over the short-interval features extracted from the audios. This combination of short- and medium-term features is a strategy that enhances emotion recognition in audio by generating averages, deltas (the difference between two consecutive measurements), and standard deviations.

Table IV describes the feature groups used by this tool. Four features are extracted for each group: mean of the feature values fextracted from each audio segment ($\overline{f}$), standard deviation of the values of that feature (σ_f), mean of the differences between the value of that feature in one segment and in the subsequent segment ($\overline{\Delta f}$), and standard deviation of the differences ($\sigma_{\Delta f}$).

TABLE III
PARAMETERS USED FOR FEATURE EXTRACTION IN THE ROSAJR TOOL

Parameter	Value
SHORT_TERM_WINDOW	0.1 sec
SHORT_TERM_STEP	0.033 sec
MID_TERM_WINDOW	1.2 sec
MID_TERM_STEP	0.6 sec

*3) **DEEP Tool**:* The DEEP tool [12] approaches emotion recognition differently from other methods that treat it as a multiclass problem. It recognizes the six basic emotions plus the neutral expression, generating a binary classifier based on a convolutional neural network (CNN) for each emotion. The training of the classifier for each target emotion is carried out in such a way that audios of the six emotions different from the target emotion (positive class) are part of the negative training sample, which must contain the same number of examples of the positive class. Therefore, to use this tool in the multiclass context in this study, we used these binary classifiers in a *one-vs-rest* approach, in which the probabilities of the audio being of each emotion (given by the seven binary classifiers) were used to classify the audio according to the emotion that presented the highest probability. When an instance presented two or more equal maximum probabilities, the instance was excluded from classification. Each binary classifier uses MFCC, chromatic and prosodic acoustic features extracted using the openSMILE package (Table V). The feature vector of each audio consists of the concatenation of 32 acoustic features from several overlapping audio segments. More specifically, every 10 ms of the audio, a set of 32 features corresponding

TABLE IV
GROUPS OF FEATURES EXTRACTED FROM AUDIO BY THE PYAUDIOANALYSIS LIBRARY USED IN THE CLASSIFICATION OF AUDIO EMOTIONS BY THE ROSAJR TOOL

Feature groups	No. feat.	Description
Zero *Crossing* rate	4	The rate at which a signal changes over the duration of a specific frame.
Energy	4	The sum of the squares of the signal values, normalized by the respective frame length.
Entropy of Energy	4	The entropy of the normalized energies of the subframes, which can be interpreted as a measure of rapid changes.
Spectral Centroid	4	The center of gravity of the spectrum.
Spectrum Spread	4	The second central moment of the spectrum.
Spectral Entropy	4	Entropy of normalized spectral energies for a set of subframes.
Spectral Flux	4	The squared difference between the normalized magnitudes of the spectrum of two successive frames.
Spectral shift	4	The frequency below which 90% of the spectrum's magnitude distribution is concentrated.
MFCC	13×4 = 52	13 Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the Mel scale.
Chroma Vector	13×4 = 52	A vector of 13 components representing the amount of energy present in each pitch class over a given period.

TABLE V
FEATURES USED BY THE DEEP TOOL, EXTRACTED BY THE OPENSMILE LIBRARY

Feature(s)	Description
chroma	14 tonal and spectral characteristics of an audio signal (semitone spectrum converted to octaves).
pcm_fftMag_mfcc	12 features that include the PCM *(Pulse Code Modulation)* representation of the signal, the magnitude of the spectral components obtained by FFT *(Fast Fourier Transform)* and the mel frequency cepstral coefficients.
voiceProb_sma	Probability of voice presence in the audio signal, calculated using (*simple moving average*).
F0_sma	Fundamental frequency of the audio signal, calculated using the simple moving average.
pcm_loudness_sma	Perception of sound intensity or volume of the audio signal, calculated using the simple moving average.
jitterLocal_sma	Local temporal variation or fluctuations in the periodicity of an audio signal calculated using the simple moving average.
jitterDDP_sma	Temporal variation in the Double Derivative of Jitter calculated using simple the moving average.
shimmerLocal_sma	Local variation or fluctuations in audio signal intensity calculated using the simple moving average.

to 50 ms of audio is extracted (sliding window with step 10 ms and size 50 ms). The other parameters used the default values of the openSMILE tool. Since different audio samples can have varying durations, their initial feature vectors may also differ in length. To standardize the number of features, the length of the longest feature vector in the training set is identified (max), and each audio has its feature vector completed with zeros until it reaches length max, a process called *padding*. Thus, in each stage of the cross-validation (Section III-E), the maximum number of features from the audios in the training sample (max) was identified. Feature vectors with sizes smaller than max, whether from the training, validation or test sample, went through the *padding* process. Feature vectors from the validation or test sample with sizes greater than max were truncated to this length.

E. Performance Evaluation

Section III-B shows the experiments performed to evaluate the three emotion recognition tools: three cross-validation experiments using only the VERBO dataset, only the BADEM dataset, and the two combined dataset (VV, BB, and AA, respectively), and two experiments training with one set and testing on the other (VB and BV). In all five experiments, the performance measures macro precision, macro recall and macro f1-score were calculated. The training set was redivided into training and validation samples within each tool, respecting the fraction defined by each of them: 70%/30% in the DEEP tool and 75%/25% in the Souza and RosaJr tools.

K = 10 was used in the k-fold cross-validation experiments, and the division into folds was stratified using the *StratifiedKFold* function of the *scikit learn* library with the parameterization *n_splits*=10, *shuffle*=True and *random_state*=42.

Since the DEEP tool addresses the problem from a binary rather than a multiclass perspective, additional procedures were needed to perform the fold splitting for the DEEP evaluation. For the tests based on training with one dataset and testing with the other dataset, the training dataset was used to create a training sample for each emotion E_i that contained all the audios of that emotion, in addition to the same amount of audios of the other emotions, in a balanced way, as illustrated in Table VI. This Table contains a row describes the number of training audios of each classifier *Clsf* E_i., with the number of audios of the target emotion E_i, totaling 50% of the audios of the positive class and 50% of the audios of the negative class. Thus, if N_{E_i} is the number of audios of the emotion E_i,$N_{E_i}/6$ audios of each of the other six emotions were randomly selected to compose the negative sample. For the tests based on cross-validation (BB, VV and AA), the sets presented in Table VI created for each emotion (N_{E_i} audios of the emotion E_i and $N_{E_i}/6$ of each of the other six emotions), were used to perform the stratified division into 10 folds.

IV. RESULTS AND DISCUSSION

A. Comparison of the Three Tools

Table VII has the values of the performance measures obtained for each of the tools in the five experiments.

The first interpretation of the results focuses on the generalization capacity of the tools. The highest values across all measures were observed in experiments that considered the training and testing sets with the same data set in cross-validation (only with VERBO, BADEM or both together), suggesting that the generalization capacity (training with one set and testing with another) of all of them is limited. This may be due to both the specific features each tool takes into account and the composition of the datasets themselves,

TABLE VI
NUMBER OF AUDIOS IN THE TRAINING SAMPLE OF THE CLASSIFIERS OF EACH EMOTION IN THE EVALUATION PROCESS OF THE DEEP TOOL, CONSIDERING THE TRAINING USING THE ENTIRE BADEM DATASET (BV) AND THE TRAINING USING THE ENTIRE VERBO DATASET (VB). THE NUMBER OF AUDIOS OF THE TARGET EMOTION E_i (POSITIVE SAMPLE) IS IN BOLD AND THE NUMBER OF AUDIOS OF THE OTHER EMOTIONS (NEGATIVE SAMPLE) IN REGULAR FORMAT

BADEM	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**Total**
Clsf E1	**129**	22	22	22	21	21	21	257
Clsf E2	22	**130**	22	22	22	22	22	259
Clsf E3	22	22	**130**	22	22	22	22	259
Clsf E4	22	22	22	**130**	22	22	22	259
Clsf E5	21	22	22	22	**129**	21	21	257
Clsf E6	21	22	22	22	21	**129**	21	257
Clsf E7	21	22	22	22	21	21	**129**	257
VERBO	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**Total**
Clsf E1	**149**	25	25	25	25	25	25	299
Clsf E2	25	**150**	25	25	25	25	25	300
Clsf E3	25	25	**149**	25	25	25	25	299
Clsf E4	25	25	25	**150**	25	25	25	301
Clsf E5	25	25	25	25	**150**	25	25	301
Clsf E6	25	25	25	25	25	**150**	25	301
Clsf E7	25	25	25	25	25	25	**150**	301

TABLE VII
PERFORMANCE MEASURES OBTAINED BY EMOTION RECOGNITION TOOLS. THE HIGHEST VALUE OF EACH MEASURE, IN EACH EXPERIMENT, IS SHOWN IN BOLD. THE BV AND VB EXPERIMENTS DO NOT HAVE A STANDARD DEVIATION BECAUSE THEY ARE SINGLE EXECUTIONS, WHILE FOR THE OTHERS THE STANDARD DEVIATION IS OBTAINED FROM THE TEN CROSS-VALIDATION EXECUTIONS

F1-score macro			
Experiment	**Algorithm ROSAJR**	**Algorithm SOUZA**	**Algorithm DEEP**
BB	**0.40 (0.03)**	0.34 (0.05)	0.28 (0.06)
VV	**0.53 (0.03)**	0.43 (0.04)	0.28 (0.03)
AA	**0.53 (0.04)**	0.35 (0.04)	0.28 (0.03)
BV	0.24	**0.30**	0.23
VB	0.22	**0.25**	0.17
Precision macro			
Experiment	**Algorithm ROSAJR**	**Algorithm SOUZA**	**Algorithm DEEP**
BB	**0.41 (0.03)**	0.34 (0.06)	0.27 (0.07)
VV	**0.53 (0.03)**	0.42 (0.04)	0.27 (0.05)
AA	**0.53 (0.04)**	0.35 (0.04)	0.28 (0.03)
BV	0.26	**0.29**	0.26
VB	**0.34**	0.26	0.18
Recall macro			
Experiment	**Algorithm ROSAJR**	**Algorithm SOUZA**	**Algorithm DEEP**
BB	**0.40 (0.03)**	0.33 (0.05)	0.30 (0.05)
VV	**0.53 (0.03)**	0.43 (0.04)	0.29 (0.02)
AA	**0.53 (0.04)**	0.35 (0.04)	0.28 (0.02)
BV	0.23	**0.31**	0.20
VB	0.16	**0.25**	0.16

which includes variables such as the selection of actors, the evaluators, the duration and content of the sentences used, among other factors.

The tool with the highest values in all measures (0.53) was RosaJr, a value derived from the VV and AA cross-validation experiments. Several factors may have contributed to this result. Firstly, the three tools evaluated, RosaJr is the only one that uses a traditional machine learning algorithm (SVM) instead of deep learning. Perhaps the number of audios in the datasets was insufficient to adequately train the deep neural networks used in the other tools, especially in terms of the classification into seven classes. Secondly, the set of features used by RosaJr tool may have positively influenced its performance. While the Souza tool uses only MFCC coefficients, reported as non-optimal combination of features by other studies [27], RosaJr uses features based on both MFCC and other acoustic information, and combines short-term and medium-term features such as means, standard deviations and variations of these measures (Table IV).

Considering only the BV and VB experiments, the tool with the highest performance values was Souza, except for precision_M in the VB experiment, where the RosaJr tool achieved the highest value (Table VII). A possible cause could be the difference in features used by the tools. A hypothesis to be investigated is whether these features are more generalized than the others for different data sets.

The DEEP tool showed the poorest performance across all measures in every experiment (highest F1-score of 0.28) in the cross-validation tests. To provide further insight into these results, the Supplementary Tables S-I and S-II present the precision and recall values for each emotion separately in the BV and VB experiments, which were the two experiments where the DEEP tool performed the worst. It is observed that for the emotions *anger* and *sadness*, no true positives were identified in either of the two experiments. For the remaining emotions, none demonstrate a strong combination of precision and recall. One possible explanation for the DEEP tool's poor performance could be the set of features used. First, it extracts 32 features every 0.01 seconds of audio, resulting in approximately 12,800 features for a 4-second audio clip. Additionally, DEEP applies *padding* to equalize the number of features across audio samples, which adds even more features with values set to zero for most audios. As the size of the audios varies from 1 to 7 seconds, the shortest audios have more zeros than the longest ones. Thus, due to the high dimensionality of the data and the comparatively low size of the training sample (2175 audios adding the two bases), the estimation errors of the parameters of the trained neural network drastically impacted the performance of the classifier.

B. Comparison of the Five Experiments

The DEEP tool presented very similar values in the three experiments (with the highest standard deviation in the BB experiment), but with the lowest performances, as already mentioned. In the cross-validation experiments (BB, VV, and AA - Table VII), the RosaJr tool showed lower performance in the BB experiment (using only the BADEM dataset, with an f1-score_M of 0.40). Meanwhile, the Souza tool demonstrated lower performance in both the BB and AA experiments (with f1-score_M of 0.34 and 0.35, respectively). The result

of AA is unusual as it would be expected to have similar or higher measures than other cases like VV. This is another result which requires additional investigations. The DEEP tool yielded similar values across all three experiments, with the highest standard deviation in the BB experiment, but consistently showed the lowest performance, as previously noted. Therefore, some characteristic of the BADEM dataset may be negatively influencing the results of the RosaJr and Souza tools, as discussed below.

The performances in the experiments that used different datasets for training and testing (BV and VB) were significantly lower than those obtained in the experiments based on cross-validation on the same dataset (BB, VV and AA). As shown in Table VII, the measures are better when the BADEM dataset is used for training (BV) for the three tools, except for precisão$_M$ for the RosaJr tool. Although we assumed that some characteristic of this dataset may have negatively influenced the performance of the RosaJr and Souza tools, it is possible that the longer duration of the audios (in relation to VERBO) may have provided more diverse characteristics, favoring training with greater generalization capacity. The BADEM set has longer audios as shown in Table I. This factor may have allowed to better train the models, especially through the RosaJr and Souza tools that use global characteristics of the audios extracted from the combined values obtained throughout the audios, such as through averages. In addition, there are other aspects that differentiate the two datasets (evaluators, actors, phrases used, recording format), which may also have contributed to the result obtained.

Interestingly, for the DEEP tool, whose results were already poor in all experiments, the results were even worse in the VB experiment. Section III-D shows that DEEP uses the number of features of the longest training audio to define the number of input neurons of the neural network to be trained, and the *padding* process completes the features of shorter duration audios with zeros. Since the average duration of the audios in the BADEM dataset is almost twice as long as that of VERBO, this causes the feature vectors of the test audios (from the VERBO dataset) to have approximately half of the values equal to zero. This excess of zeros in the test instances should lead to worse results than the opposite (training using the dataset with a shorter average duration, as in the VB experiment), but the results did not show that.

It is evident that factors beyond the differences in audio duration between the two datasets influence the results, such as variations in the sets of evaluators, actors, sentences used, and recording formats (Table I). To assess the extent of these differences, the Mann-Whitney statistical test from the SciPy library was applied to each feature extracted by the tools. This test was used to determine whether the feature values differ significantly between BADEM and VERBO. The difference was considered significant if the p-value calculated by the test was less than 0.05.

Table VIII shows the percentage of features extracted from VERBO and BADEM by each tool that are significantly different. These results show that, although the two datasets have audios with the same six emotions (plus the neutral expression) in balanced quantities, most of the features used by the tools (87.5% to 90%) presented significantly different values between the two datasets. Since most of the features are duration-independent, the different audio durations of the two datasets (Table I) is unlikely the reason. In addition, both datasets have the same number of actors and actress. Although they are equals on the genre distribution, the actors/actress from VERBO are different from those that participated on BADEM construction. It may indicates that the use of 12 speakers may be not enough to generalize personal acoustic features. In fact, it is consistent with the low performance observed when training with one dataset and testing with the other set (BV and VB), which is significantly lower than those obtained in experiments based on cross-validation that use the same dataset for training and testing (BB, VV and AA). These differences also indicate that acoustic differences in the two datasets are associated with better or worse performance when using them as training or testing.

TABLE VIII
PERCENTAGE OF FEATURES EXTRACTED BY THE THREE TOOLS WITH STATISTICALLY DIFFERENT VALUES IN THE BADEM AND VERBO DATASETS ACCORDING TO THE MANN-WHITNEY TEST. (*): 32 FEATURES EVERY 0.01 SEC

Tool	No. feat.	% of significantly different features
ROSAJR	136	88%
SOUZA	20	90%
DEEP	32(*)	87.5%

It is important to acknowledge a limitation of this study: the inherent subjectivity involved in an actor's portrayal of a specific emotion and in a listener's recognition of that emotion. This subjectivity is reflected in the agreement rates between actors and evaluators, which ranged from 76% to 80% (Table I). The discordant audio samples were not excluded from the study since information about which samples were discordant is not available. While the composition of the datasets may present a limitation, it is important to note that these are controlled datasets, previously analyzed and labeled by experts, which enables a comparative study of the tools. Such a comparison would not be feasible without datasets with well-defined features. These results have practical implications for the application of these tools in real-world scenarios. Typically, the audio samples from which emotions need to be recognized are captured differently from those in the training set and are produced by different individuals. Consequently, it is expected that the performance of the tools will be even lower than what is estimated by cross-validation experiments and closer to the values obtained when using experiments that involve different datasets for training and testing.

C. Comparison with Related Work

Table IX shows a comparison of the results obtained in this study with related work presented in Section II. While the distinct datasets employed make direct comparison challenging, we observe that the outcomes across several algorithms in a same language differ, a trend also evident in Portuguese. Other tools used different algorithms, with different testing methodologies and performance measures, different characteristics,

and the datasets used in other languages were also significantly different: e.g. the smallest dataset, EMODB containing 494 audios, the Indonesian dataset containing 639 audios, the RAVDESS dataset containing 1440 audios and the Kannada dataset, the largest, with 10,000 audios (8000 for training and 2000 for testing). Furthermore, there were no experiments correlated to crossing different datasets (training with one dataset and testing with another dataset from a different source) in related studies.

In fact, in the study by Kannadaguli et al. [22], the training and testing datasets were recorded by the same participant. Based on the results observed in this study, such procedures tend to overestimate performance measures.

Finally, although most of the studies presented their results based on accuracy (Table IX), this metric tends to superestimate the classification performance. While accuracy values obtained in this work are comparable to those obtained in other languages, f1-score unveils difficulties in recognition.

D. Future Work

The results presented in this study open up new avenues for future work, including: investigate how some dataset characteristics, such as audio durations and variability of actors, may impact the performance of specific tools; extend the experiments to datasets composed of audios recorded in real-life situations (uncontrolled conditions); investigate how Portuguese linguistic and cultural characteristics affect emotion recognition; extend the evaluation to other algorithms, including multilanguage tools; improvement of datasets for training (using only audios with 100% of agreement among annotators, including more diversity regarding age and gender, and a higher number of participants); evaluate other audio features. In addition, our research group is developing a multimodal emotion classifier, taking into account not only accoustic but also textual and visual signals.

V. CONCLUSION

The literature on emotion recognition in audio predominantly focuses on the English and German languages, with a smaller number of studies dedicated to Chinese and Japanese. Notably, no international publications were found that specifically address emotion recognition in the Portuguese language. This study identified three tools developed for emotion recognition in Brazilian Portuguese, as described in the Portuguese-language literature, and compared their performance using two audio datasets in this language: BADEM and VERBO. These datasets were utilized both in cross-validation experiments and in experiments where training was performed on one dataset and testing on the other.

All three tools work with a set of distinct acoustic features and different classifier induction algorithms. The tool that produced the best overall result was RosaJr [11] with an f1-score$_M$ of 0.53 in the AA and VV experiments. This tool trains an SVM from 136 features based on means and standard deviations of values extracted throughout the audio. The other tools use *deep learning* algorithms, which may have been hampered by the low volume of training data. Reevaluating these tools with a substantially larger volume of audio may clarify how much the size of the training sample or the type of features used may greatly influence the final result.

Another issue observed was the performance difference between the five experiments that used different data sets. The BADEM and VERBO sets presented significantly different mean values related to most of the features. This may be due to several reasons, such as differences in the equipment used and also audio duration. An important aspect is that the actors who participated in the creation of each of the two data sets are different, with a fairly small number. It is expected that

TABLE IX
COMPARISON OF RELATED WORKS. (*1) LB: LIBROSA WAS USED TO OBTAIN THE ACCURACY AMONG THE DIFFERENT ALGORITHMS; (*2) OBTAINED FROM THE AVERAGE OF THE EER (EMOTION ERROR RATE); ACC: ACCURACY; ERR: EMOTION ERROR RATE; F1: F1-SCORE; SVMS: SUPPORT VECTOR MACHINE WITH SEQUENTIAL MINIMAL OPTIMIZATION (SMO); LOG: SIMPLE LOGISTIC REGRESSION; RF: RANDOM FOREST; DT: J48 DECISION TREE; MLR: MULTINOMIAL LOGISTIC REGRESSION; KNN: K-NEAREST NEIGHBOR; LSTM: LONG SHORT-TERM MEMORY; RNNG: RECURRENT NEURAL NETWORK - GATED RECURRENT UNIT; MLP: MULTILAYER PERCEPTRON; CNN: CONVOLUTION NEURAL NETWORK; ANN: ARTIFICIAL NEURAL NETWORK; HMM: HIDDEN MARKOV MODEL; DMFCC/DDMFCC: DELTA/DOUBLE DELTA MFCC.; PAA: PYAUDIOANALYSIS; VB: VOICE PROBABILITY; LD: LOUDNESS; JI: JITTER; SH: SHIMMER; (**) THIS WORK (HIGHEST MACRO F1-SCORE); (*3) SEE TABLE VII

Ref.	Alg.	Language	No. audios	Feature set	Result	Perf. metric
[15]	SVMS, MLP, LOG, RF, DT, MLR, KNN	German	494	IS-09, emobase, GeMAPS, eGeMAPS	87.85% 84.00% 83.74% 73.46% 53.08% 70.65% 58.69%	ACC
[18]	LSTM, RNNG, MLP, CNN	English	1440	GeMAPS, PAA, LB(*1)	82.0% 81.2% 79.4% 74.3% ’	ACC
[22]	ANN HMM	Kannada	10000	MFCC, DMFCC, DDMFCC	93.7% 77.5% (*2)	EER
[23]	SVM, NN, RF, NB	Indonesian	639	MFCC	100% 99.7% 97% 94.1%	ACC
(**)	RNNG	Portuguese VV(*3)	1.167	MFCC	0.43/ 81.4%	F1$_M$/ ACC
	SVM			PAA	0.53/ 74%	F1$_M$/ ACC
	CNN			Chroma, MFCC, F0, VB, LD, JI, SH	0.28/ 86%	F1$_M$/ ACC

the acoustic features present different values among different people. Thus, investigating the minimum adequate number of different speakers participating in the training sample to allow adequate training is a noteworthy investigation question.

In general, better tools for recognizing audio-based emotions in Brazilian Portuguese should be developed, and, to support the training of these tools, as well as larger sets of audio data in this language, they should contain not only more audio but also diverse speakers.

This study contributes for the research of emotion recognition from audios in Portuguese. This topic has multiple applications, such as psychotherapy contexts, behavioral evaluation, evolution of emotions during dialogs such as in call center conversations, human-computer interaction, detection of psychiatric disorders among others.

ACKNOWLEDGMENTS

This work was supported by Brazilian National Council of Scientific and Technological Development (CNPq) (grant 307710/2022-0), and São Paulo Research Foundation (FAPESP) - National Institute of Science and Technology—Medicine Assisted by Scientific Computing (INCT-MACC) - grant 2014/50889-7.

REFERENCES

[1] M. Nasir, B. R. Baucom, P. Georgiou, and S. Narayanan, "Predicting couple therapy outcomes based on speech acoustic features.," *PloS one*, vol. 12, no. 9, p. e0185123, 2017. doi: 10.1371/journal.pone.0185123.

[2] B. Paul, S. Bera, T. Dey, and S. Phadikar, "Machine learning approach of speech emotions recognition using feature fusion technique," *Multimedia Tools and Applications*, vol. 83, pp. 8663–8688, Jan. 2024. doi: 10.1007/s11042-023-16036-y.

[3] O. R. da Silva, "Análise de emoções em diálogos em português brasileiro por meio da análise acústica: uma aplicação em psicoterapia," Master's thesis, EACH - Universidade de São Paulo), 2024. https://doi.org/10.11606/D.100.2024.tde-20012025-123806.

[4] G. Tamulevičius, G. Korvel, A. Yayak, P. Treigys, J. Bernatavičienė, and B. Kostek, "A study of cross-linguistic speech emotion recognition based on 2D feature spaces," *Electronics (Switzerland)*, vol. 9, no. 10, pp. 1–13, 2020. doi: 10.3390/electronics9101725.

[5] H. Li, B. Baucom, and P. Georgiou, "Linking emotions to behaviors through deep transfer learning," *PeerJ Computer Science*, vol. 2020, no. 1, pp. 1–32, 2020. doi: 10.7717/peerj-cs.246.

[6] C. Suneetha and R. Anitha, "Speech based emotion recognition by using a faster region-based convolutional neural network," *Multimedia Tools and Applications*, vol. 84, pp. 5205–5237, Mar. 2025. doi: 10.1007/s11042-024-19004-2.

[7] Z. Chen, C. Liu, Z. Wang, C. Zhao, M. Lin, and Q. Zheng, "Mtlser: Multi-task learning enhanced speech emotion recognition with pre-trained acoustic model," *Expert Systems with Applications*, vol. 273, p. 126855, 2025. doi: https://doi.org/10.1016/j.eswa.2025.126855.

[8] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, (Red Hook, NY, USA), Curran Associates Inc., 2020. doi: 10.5555/3495724.3496768.

[9] M. A. Gismelbari, I. I. Vixnin, G. M. Kovalev, and E. E. Gogolev, "Speech emotion recognition using deep learning," in *2024 XXVII International Conference on Soft Computing and Measurements (SCM)*, pp. 380–384, 2024. doi 10.1109/SCM62608.2024.10554077.

[10] P. Laukka, N. Thingujam, F. Iraki, H. Elfenbein, T. Rockstuhl, W. Chui, and J. Althoff, "The expression and recognition of emotions in the voice across five nations: A lens model analysis based on acoustic features," *Journal of Personality and Social Psychology*, vol. 111, no. 5, pp. 686–705, 2016. doi: 10.1037/pspi0000066.

[11] J. da Rosa Jr, "Reconhecimento automático de emoções através da voz," 2017. Undergraduate thesis. Universidade Federal de Santa Catarina. doi not available. https://repositorio.ufsc.br/bitstream/handle/123456789/182186/reconhecimento_automatico_emocoes_voz_final_pdfa.pdf.

[12] G. A. Campos and L. da S. Moutinho, "Deep: Uma arquitetura para reconhecer emoção com base no espectro sonoro da voz de falantes da língua portuguesa," 2020. Undergraduate thesis. Universidade de Brasília. doi not available. https://bdm.unb.br/bitstream/10483/27583/1/2020_GabrielCampos_LucasMoutinho_tcc.pdf.

[13] R. M. de Souza, "Reconhecimento de emoções através da fala utilizando redes neurais," 2020. Undergraduate thesis. Universidade Federal de Santa Catarina. doi not available. https://repositorio.ufsc.br/bitstream/handle/123456789/218146/TCC.pdf.

[14] L. M. F. Carlos, "A criação de um dataset audiovisual brasileiro de expressões emocionais," Master's thesis, Paradigma Centro de Ciências e Tecnologia do Comportamento, 2022. doi not available. https://mestrado.institutopar.org/wp-content/uploads/sites/2/2024/04/Luisa-Medina-2022.pdf.

[15] C. Doğdu, T. Kessler, D. Schneider, M. Shadaydeh, and S. R. Schweinberger, "A comparison of machine learning algorithms and feature sets for automatic vocal emotion recognition in speech," *Sensors*, vol. 22, no. 19, 2022. doi: 10.3390/s22197561.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Interspeech*, pp. 1517–1520, ISCA, 2005. doi: 10.21437/Interspeech.2005-446.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, (New York, NY, USA), p. 1459–1462, Association for Computing Machinery, 2010. doi: 10.1145/1873951.1874246.

[18] B. T. Atmaja and M. Akagi, "On the differences between song and speech emotion recognition: Effect of feature sets, feature types, and classifiers," in *2020 IEEE REGION 10 CONFERENCE (TENCON)*, pp. 968–972, Nov 2020. doi:10.1109/TENCON50793.2020.9293852.

[19] 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, pp. 1–35, 05 2018. doi: 10.1371/journal.pone.0196391.

[20] T. Giannakopoulos, "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis," *PloS one*, vol. 10, no. 12, 2015. doi: 10.1371/journal.pone.0144610.

[21] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," pp. 18–24, 01 2015. doi: 10.25080/Majora-7b98e3ed-003.

[22] P. Kannadaguli and V. Bhat, "Comparison of hidden markov model and artificial neural network based machine learning techniques using DDMFCC vectors for emotion recognition in Kannada," 2019. doi: 10.1109/WIECON-ECE48653.2019.9019936.

[23] J. Gondohanindijo, E. Noersasongko, Pujiono, Muljono, A. Z. Fanani, Affandy, and R. S. Basuki, "Comparison method in indonesian emotion speech classification," p. 230 – 235, 2019. doi: 10.1109/ISEMANTIC.2019.8884298.

[24] ffmpeg.org, "FFmpeg - a complete, cross-platform solution to record, convert and stream audio and video," 2024. Accessed: 2024-09-07 https://ffmpeg.org/.

[25] J. R. Torres Neto, G. P. Filho, L. Y. Mano, and J. Ueyama, "VERBO: Voice Emotion Recognition dataBase in Portuguese Language," *Journal of Computer Science*, vol. 14, pp. 1420–1430, Nov 2018. doi: 10.3844/jcssp.2018.1420.1430.

[26] P. Crewson, "Applied statistics handbook," *AcaStat Software*, vol. 1, pp. 103–123, 2006. doi not available.

[27] W. Zhang, X. Zhang, and Y. Sun, "A new fuzzy cognitive map learning algorithm for speech emotion recognition," *Mathematical Problems in Engineering*, vol. 2017, 2017. doi: 10.1155/2017/4127401.

Omar Rodrigues da Silva master's in computing techniques (pattern recognition) at EACH-USP (School of Arts, Sciences, and Humanities of the University of São Paulo) from the Postgraduate Program in Information Systems. Bachelor's degree in Physics and working in the Information Technology sector since 1985, with experience in application development, reliability engineering, database administration, and data replication across heterogeneous sources, as well as infrastructure monitoring and performance.

Luisa Medina Fermino Carlos clinical psychologist graduated in Psychology from the State University of Londrina (UEL). Postgraduate in Analytical-Behavioral Therapy from UniFil (Centro Universitário Filadélfia). Master's degree in Applied Behavior Analysis from the Paradigma Center for Behavioral Science and Technology.

Felipe Corchs holds a Medical degree (2001), a Medical Residency in Psychiatry (2005), a PhD in Sciences (2008), and completed a postdoctoral fellowship in the Department of Neuroscience and Psychiatry at the Icahn School of Medicine at Mount Sinai in New York. He is currently a Collaborating Professor in the Department of Psychiatry at the Faculty of Medicine, University of São Paulo (USP), and Advisor in the Graduate Program in Neuroscience and Behavior at USP. He is also a Professor and Advisor at the Paradigma Center for Behavior Analysis.

Fátima L. S. Nunes is a Full Professor of the University of São Paulo (USP). Bachelor in Computer Science (Universidade Estadual Paulista Júlio de Mesquita Filho), Master in Electrical Engineering (USP) and PhD in Science (Computational Physics) (USP). Associate Professor (University of São Paulo) in the Graphics area. Her research is predominantly in Computer Science, with emphasis on Virtual Reality, Image Processing, Content-based multimedia data retrieval, and Data Science.

Ariane Machado-Lima is an Associate Professor in the Information Systems Bachelor's and Post-Graduate Program at the University of São Paulo, Brazil, where she has worked since 2009. She holds a Bachelor's and Master's degree in Computer Science, a PhD in Bioinformatics and completed her postdoctoral research at the Institute of Psychiatry, University of São Paulo. Her research is predominantly in Computer Science, with an emphasis on Machine Learning.