BAH: Beyond Acoustic Handcrafted features for speech emotion recognition in Portuguese

Larissa Guder
larissa.guder@edu.pucrs.br
School of Technology
Pontifical Catholic University
of Rio Grande do Sul (PUCRS)
Porto Alegre, Brasil
Laboratory of Advanced Research on
Cloud Computing
Faculdade Três de Maio
Três de Maio, Brasil

Luan Dopke
Marcos Kaiser
Dalvan Griebler
dalvan.griebler@pucrs.br
School of Technology
Pontifical Catholic University
of Rio Grande do Sul (PUCRS)
Porto Alegre, Brasil

Felipe Meneguzzi
felipe.meneguzzi@abdn.ac.uk
School of Technology
Pontifical Catholic University
of Rio Grande do Sul (PUCRS)
Porto Alegre, Brasil
University of Aberdeen
Aberdeen, Scotland

ABSTRACT

It is through affective computing that we have the integration of human feelings and computing applications. One affective computing task is Speech Emotion Recognition (SER), which identifies emotions from spoken audio. Even though emotion is a universal aspect of human experience, each culture and language has different ways to express and understand emotions. So, when designing models for SER, it is common to focus on a single language. In this work, we explore VERBO, a Brazilian Portuguese dataset for categorical emotion recognition. Our main objective is to define the best way to extract acoustic features to train a classifier for SER. We compare 18 different methods to generate audio representations, grouped by handcrafted features and audio embeddings. The best representation for VERBO is TRILL embeddings, and with an SVM classifier, we achieved 92% accuracy in VERBO. As far as we know, this was the state of the art for this dataset.

KEYWORDS

Speech Emotion Recognition, Audio Processing, Natural Language Processing

1 INTRODUCTION

Understanding what others are feeling is one of the foundations of relationships, whether in society or the family environment. From an evolutionary perspective, emotions have a direct impact on our sense of survival and well-being. For example, fear is an essential regulator that can help in decision-making. Generally, the demonstration of emotion occurs naturally and subconsciously. They can be perceived by facial and corporal expressions, vocal intonation, pupil dilatation, heart rate, and breathing [25].

For a machine learning model to perform emotion recognition, emotions must be categorized into a finite set of discrete classes. The vast spectrum and subtlety of human emotions make a comprehensive classification system challenging to implement. Due to the large number of existing ones, the most common way to label datasets is using the Ekman [8] definition. Ekman [8] identified six emotions considered essential: anger, disgust, fear, happiness, sadness, and neutral, where the model must classify the input according to the most correlated class.

SER is of paramount importance for several applications. Geetha et al. [11] identifies sectors that are using SER, varying from education and healthcare to entertainment and automotive, as well as specialized fields like human-robot interaction and security. The lack of data for training and testing deep learning models makes it difficult for the field of SER to grow [7], especially in non-English languages. Existing datasets have limited data, are less diverse than necessary, or are too different from real-world data.

In our previous work [14], we explored the use of a dimensional approach using the IEMOCAP dataset, which is in English. In this work, our primary objective is to explore the VERBO dataset [35], laying the groundwork for a more robust multimodal architecture in the future, focusing on the Brazilian Portuguese language. VERBO is one of the few publicly available datasets in Brazilian Portuguese, and the most complete one. VERBO consists of the same person speaking the same utterance for each of seven classes, which follow the Ekman [8] structure, and adds neutral as the seventh emotion. The other two available datasets with audio for SER in Brazilian Portuguese are the CORAA [4], with only three classes: neutral, non-neutral female, and non-neutral male, and emoUERJ [2], with four classes: happiness, anger, sadness, and neutral.

Therefore, our research question is how to improve the SER in Brazilian Portuguese using acoustic representations. To answer this question, we compare 18 different methods for extracting features from audio, grouped by handcrafted features and audio embeddings. To validate our findings, the efficacy of these representations is evaluated across three different classifiers. Consequently, our novel contributions are a deep exploration of how audio embeddings can outperform traditional features in SER, a new state-of-the-art in the VERBO dataset.

This paper is organized as follows. We first provide the necessary background on SER in Section 2 and detail our experimental methodology in Section 3. Next, we present the comprehensive results from our experiments in Section 4. Section 5 presents the related work, while in Section 6 we discuss our results concerning related work. In Section 7 we show the threats to validity present in our work, and finally we present the conclusion in Section 8.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

2 SPEECH EMOTION RECOGNITION

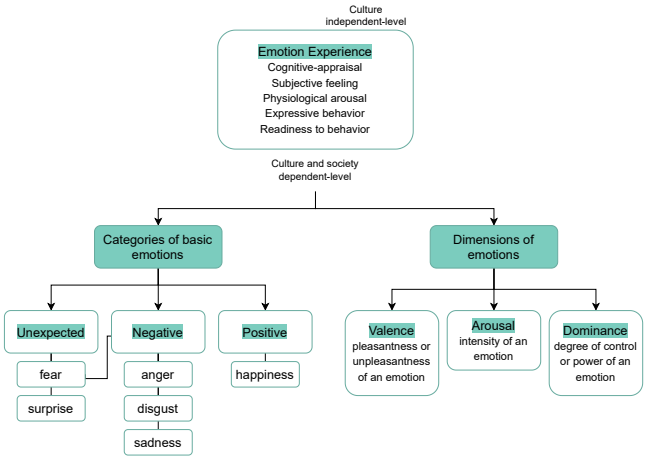


Figure 1: Structure of emotion experience and classification, adapted from Munezero *et al.* [23] and Roberts *et al.* [27]

Boehner et al. [3] defines that emotions are culturally grounded and dynamically experienced, to some extent, constructed in interaction. Complementary to that, Loderer et al. [20] states that emotions can be perceived differently in different cultures. Loderer et al. [20] findings show that the more similar components across cultures are affective, cognitive, and motivational. On the other hand, the less similar are the physiological and expressive components.

Figure 1 details the structure of an emotion. At the top, we have the culture-independent level, representing how the individual perceives emotions. Then, we have the culture and society-dependent level, which describes how individuals name their feelings.

Applied to machine learning, two classifications are often used: discrete classification and dimensional classification [19]. Ekman [8] proposes what we will call the discrete classification of emotions. His study is an update of a previous work published in 1957. He proposes six basic emotions: anger, disgust, fear, happiness, sadness, and surprise. Anger, disgust, fear, and sadness represent all negative emotions, happiness represents positive emotions, and fear and surprise represent unexpected emotions.

On the other hand, we have the circumplex model of affect proposed by Russell [28], which we will refer to as the dimensional classification model. Two axes represent values for arousal (y-axis) and valence (or pleasure) (x-axis) in a dimensional space. These two values range from -1 to 1, allowing for the determination of emotion. Arousal is related to acoustic features, and valence is related to linguistic features.

In the circumplex model, we can see that we obtain happy and excited emotions with high valence and arousal values. Feelings like gladness and calm can be found when the valence is low and arousal is high. Sad, tired, and bored emotions are related to low valence and arousal values, while we have frustration, anger, and fear emotions for high valence and low arousal.

In addition to valence (pleasure) and arousal dimensions, we have dominance as a third dimension, proposed by Mehrabian [21]. Dominance refers to how emotion influences a person's behavior. Lower levels represent passive or submissive feelings, while high levels are assertive or powerful.

Singh and Goel [34] defines the Speech Emotion Recognition (SER) task as recognizing emotions from speech utterances without using linguistic features. Since speech is one of the most used ways to communicate in society, the SER can be applied in different sectors, like education, healthcare, marketing and advertising, human-robot interaction, security and surveillance, and customer service, sports, entertainment, gaming, and the automotive industry [11]. Also, the use of speech is less intrusive than physiological signals.

Lieskovská et al. [19] summarizes SER into two main parts: feature extraction and classification. The speech signal serves as input to the front-end. In the front-end, we have the preprocessing and feature extraction steps that define a representation for the input signal. This representation can be a handcrafted set of features or audio embeddings. After generating the representation, they are fed to the back-end, which has the ML classifier and the scoring function. As output to the back-end, we obtain an emotional state for the speech signal.

3 METHODOLOGY

In this section, we present our methodology for feature extraction and model training using the VERBO dataset. We present our complete setup of experiments in Figure 2.

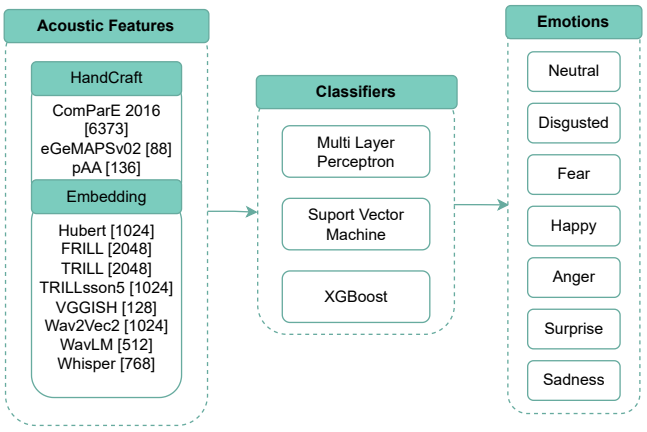
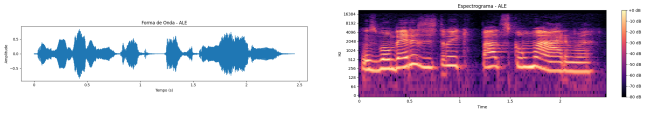


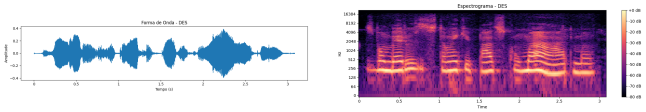
Figure 2: The proposed experiment set. The acoustic features block presents each feature and its respective dimension size. The classifiers block presents the three classifiers evaluated. And the emotions block presents all predicted classes.

3.1 Dataset

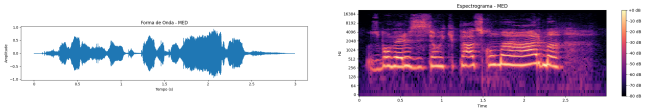
The VERBO (Voice Emotion Recognition dataBase in Portuguese Language) corpus is an emotional speech dataset developed to support the training and evaluation of speech emotion recognition systems in Brazilian Portuguese. It consists of 1167 audio samples recorded by 12 professional Brazilian actors (six male and six female), each interpreting 14 semantically neutral sentences under



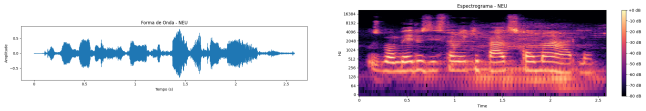
(a) Happiness – Speaker F1, Sentence 1



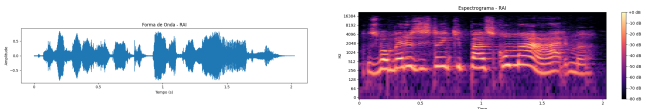
(b) Disgust – Speaker F1, Sentence 1



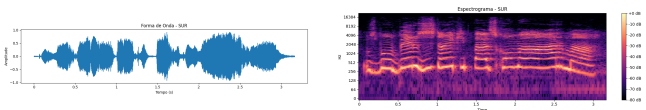
(c) Fear – Speaker F1, Sentence 1



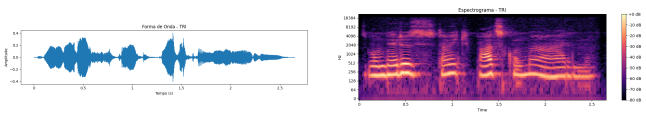
(d) Neutral – Speaker F1, Sentence 1



(e) Anger – Speaker F1, Sentence 1



(f) Surprise – Speaker F1, Sentence 1



(g) Sadness – Speaker F1, Sentence 1

Figure 3: Waveform (left) and spectrograms (right) for each emotion class. We consider the same female speaker for the sentence: "*Os bombeiros estão equipados com uma arma*" (Firefighters are equipped with a gun).

seven emotional conditions: happiness, sadness, anger, fear, surprise, disgust, and neutral. The recordings follow a consistent naming structure in the format `<emotion>-<speaker>-<sentence>.wav`, allowing for controlled selection based on emotion, speaker identity, and the content uttered.

The audio recordings were captured in a studio environment using standardized equipment to ensure signal clarity and acoustic consistency. Each sample has an average duration of approximately 1.5 seconds and was stored in a 44.1 kHz, 16-bit, mono format. The dataset includes both gender balance and phrase diversity. Three expert judges performed a perceptual validation, revealing a 37.7% rate of unanimous agreement in all emotions, with happiness and neutral being the most accurately perceived.

To illustrate the acoustic and spectral variation across different emotions in Figure 3, we selected a controlled subset of the data comprising one female speaker (F1) and one specific sentence (sentence 1). For this configuration, we extracted one audio sample per emotion, resulting in seven distinct recordings from the same speaker and phrase. Using the same speaker for each case allows for a fair comparison. For example, the waveform for sadness (g) had a lower amplitude when compared to the other emotions. When analyzing the spectrogram, happiness (a) had more variation in the frequency than the others.

3.2 Acoustic Feature Extraction

The first stage in SER classification is the pre-processing. The audio cannot be directly fed into the classifier. To extract the features, there are two common ways: handcrafted features and audio embeddings. The handcrafted process is used to get the common information present in the speech, like the fundamental frequency, energy, and quality. The audio embeddings are generated using a machine learning model that extracts the patterns from the audio and converts them into a vector representation.

For both cases, we load the audio as a waveform, converting the signal into 16Hz, which is the expected frequency for the models to generate embeddings. For each embedding case, the models commonly generate a vector at each time x. So to standardize the size for the classifier input, we calculate the average of each multi-dimensional vector to create a one-dimensional representation.

3.2.1 Handcrafted Features. Focusing on determining which combinations of features are best suited for use in tasks involving speech processing, such as Speech Emotion Recognition, we have identified several feature sets, including ComParE [29] and pyAudioAnalysis (pAA) [13]. The Computational Paralinguistics Challenge 2016 (ComParE) feature set includes 6,373 static features derived from the computation of various functionals over low-level descriptor (LLD) contours [29]. These low-level descriptors enclose a wide range of speech-processing parameters. To extract these feature sets from audio, we used OpenSmile [9]. In addition, at feature-level, it is possible to extract the feature sets using three different approaches: (1) only the LLDs, which are calculated over a sliding window; (2) Delta regression of LLDs, and (3) the statistical functionals, which map LLDs values to static values [9].

The pAA permits the use of two different functions: one for short-term features and another for mid-term features. The short-term features use windowing to split the signal into frames and process the features for each frame. This method extracts a total of 34 features. With the mid-term features, it is possible to extract the mean and standard deviation for each short-term feature. Using the mid-term feature extraction, the total number of features is 136.

3.2.2 Audio Embeddings. In complement to the handcrafted features, we also explored using audio embeddings, such as VGGish, TRILL, and their derivations FRILL and TRILLSSON5, as well as Wav2Vec2, Hubert, WavLM, and Whisper. VGGish [16] is a modification of the VGG16 architecture [33], a popular convolutional neural network. The authors trained the VGGish model on a large YouTube dataset. The input of the VGGish is a numerical representation of the audio waveform. This audio can have a maximum

duration of 10 seconds. The output of the VGGish is an audio embedding representation with 128 dimensions.

Wav2Vec2, HuBERT, and WavLM are part of a family of self-supervised models developed by Meta AI to extract deep and context-rich audio representations. Wav2Vec2 [1] is trained via contrastive loss, predicting masked portions of raw audio and later fine-tuned for supervised tasks like Automatic Speech Recognition (ASR) or SER. HuBERT (Hidden Unit BERT) [17] improves upon this by using clustering-based target labels derived from the audio data itself, which helps the model capture more robust structural representations. WavLM (Wav2Vec-based Language Model) [5] extends the capabilities of previous models by incorporating multi-task training objectives and larger-scale training, leading to improved performance in both ASR and emotion recognition tasks, including enhanced robustness to noise and lower latency variants.

The TRIpLet Loss network (TRILL) is a self-supervised model trained on the AudioSet dataset, developed with a focus on non-semantic tasks — that is, tasks that do not consider the meaning or presence of words in speech [31]. TRILL leverages a variant of the ResNet-50 [15] architecture and outputs 2048-dimensional audio embeddings. These embeddings are extracted as a temporal sequence. They are commonly aggregated using statistical operations such as mean or pooling to produce a fixed-size vector suitable for downstream tasks, including SER. TRILL has proven effective in capturing speaker characteristics, intonation, prosody, and emotional cues embedded in speech signals.

Derived from TRILL, TRILLsson5 [32] is a lightweight version of the original model, optimized for deployment in resource-constrained environments, such as mobile devices or embedded systems. This version reduces computational requirements while maintaining competitive performance in capturing paralinguistic features. Similarly, FRILL (Filtered Representation from Intermediate Layers of TRILL) [24] is an extension that extracts and filters intermediate representations from TRILL's architecture. This approach enhances sensitivity to emotionally relevant acoustic variations, making FRILL particularly valuable for fine-grained emotion classification.

The Whisper is a Transformer-based encoder-decoder model developed by OpenAI for automatic speech recognition (ASR), trained on 680,000 hours of supervised multilingual and multitask speech data. Although initially designed for ASR, its encoder representations have been successfully repurposed as audio embeddings for tasks such as Speech Emotion Recognition. Whisper is robust to accents, background noise, and varied speaking styles. In this work, we use the "whisper-small" model and extract the hidden states from the encoder. These embeddings are averaged across time to yield a single vector per utterance, which captures both low-level acoustic features and higher-level prosodic patterns.

3.3 Classifiers

To evaluate each acoustic feature, we used three different classifiers: Multi-layer Perceptron (MLP), Support Vector Machines (SVM), and XGBoost. MLP and SVM are already used for this task [18]. The MLP model is built using the Keras library. Considering that each acoustic feature has a different dimension size, our input layer is defined as the same size as the feature dimension. The first layer is a dense layer with 128 units, utilizing the ReLU activation function. After that, we applied a dropout of 0.3. Next, we have another dense layer that reduces the dimensions to 64 units, followed by ReLU activation, and then a dropout layer with a dropout rate of 0.3. Our output has seven units, and we apply softmax as an activation function.

Adam was applied as the optimizer, with a learning rate of 0.0005. We trained the model for over 100 epochs with a batch size of 8. We chose a lower batch size due to the dataset size. The loss function used is the categorical cross-entropy, defined in Equation 1. Where C is the number of classes, p_i is the probability that the model predicts for each class, and y_i is the correct answer.

$$L = -\sum_{i=1}^{C} y_i \log(p_i) \tag{1}$$

The SVM was implemented through the Scikit-learn library, using the function C-Support Vector Classification (SVC). We applied a variation of the regularization parameter with 0.001, 0.01, 0.1, and 100. The kernel is linear, and the class weight is either none or balanced. The XGBoost was implemented using the default configurations, with the histogram-based tree.

We used 80% of the data to train and 20% to test our model. We applied the stratified strategy, and 42 was the random state.

The standard scaler from sklearn was applied to standardize our features. We applied the label encoder to encode the labels. For MLP, we used the categorical function from Keras to convert into a binary class matrix.

4 RESULTS

In this paper, our primary objective is to evaluate various types of audio representations. TRILL embeddings achieve the best result for all classifier models. Different from VGGish (0.81 vs 0.57 mean accuracy), which was trained on other types of audio, TRILL was trained on speech data, creating non-semantic speech representations. Unlike models that aim to understand the content of speech, such as Whisper, WavLM, Hubert, and Wav2vec2, TRILL is engineered to capture the nuances of how something is spoken.

Even when compared to classic approaches for SER, such as eGeMAPS, pAA, and ComParE, which also focus on non-semantic speech representations, all the TRILL derivations (FRILL and TRILsson) achieve better results. The complete comparison is in Figure 4.

When comparing MLP to the XGBoost model, the results in some cases are quite similar, particularly for eGeMAPS (0.63 vs. 0.62), WavLM (0.38 vs. 0.37), and pAA (0.60 and 0.58). The only case where XGBoost achieves the best result was with the ComParE set. ComParE has a larger dimension size, with 6,373 audio features. Considering this size, the best result is due to the XGBoost being a gradient-boosted tree structure, which performs implicit feature selection during the training process. This mechanism is well-suited to handle the high-dimensional and highly correlated nature of the ComParE feature set.

MLP outperforms SVC with TRILLsson5, @hisper, Hubert, and Wav2Vec2. However, the use of SVC as a classifier and TRILL as an audio embedding achieves the best overall result, with an accuracy of 0.88. This is an expected result because the TRILL model was trained to generate acoustic representations of speech. VGGish, for

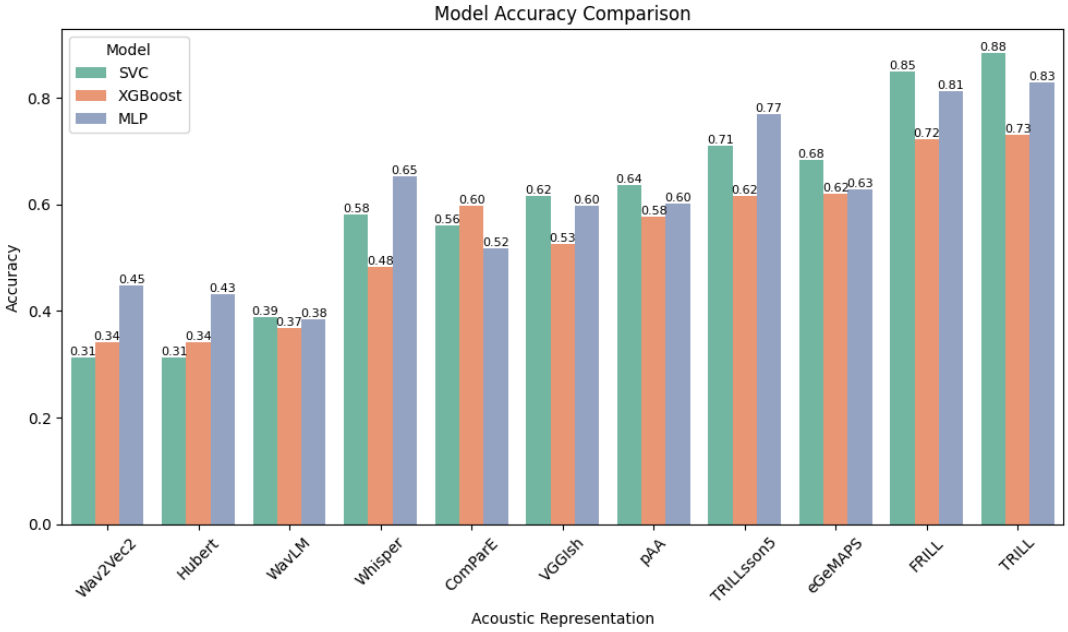


Figure 4: Model accuracy comparison for each acoustic representation. We compare SVC, MLP, and XGBoost classifiers.

example, focuses on audio classification in general. Models like Wav2Vec focused on acoustic and phonetic aspects, while Whisper focused on semantic, contextual, and phonetic aspects.

Parameter	Values
Regularization parameter C	0.001 0.01 0.1 **100**
Kernel type	rbf linear **poly** sigmoid
Gamma	**scale** auto 0.01 0.1 1
Degree	**2** 3 4
Coef0	0.0 **0.1** 0.5
Shrinking	**True** False
Probability	**True** False
Class Weight	**None** balanced
Decision Function Shape	**One-vs-rest** one-vs-one
Break Ties	**True** False
Cache Size	**200**

Table 1: Hyperparameter variations used on grid search for C-Support Vector Classification from Scikit-learn. The elected parameters are in bold.

To explore the full potential of the SVC, we perform a grid search varying the hyperparameters. As we used the scikit-learn implementation, the detailed combinations of hyperparameters are presented in Table 1. We highlight the best parameter choice in bold. We used the same split process, with an 80/20 ratio. With the grid search, we improve our results from 0.88 in accuracy to 0.92.

5 RELATED WORK

Sharma [30] creates a benchmark to evaluate the Wav2vec 2.0 [1] in 25 open source datasets focused on SER task. The proposed approach is based on four techniques: (1) Multi-Lingual (MLi) Learning, (2) Multi-Task Learning (MTL), (3) Transfer Learning, and (4) Fine-tuning. To address the relatively small size of aggregated SER datasets, a common strategy of using transfer learning is applied. This involves pre-training a model on a different, larger dataset and then fine-tuning it for the specific SER task. A pre-trained model (wav2vec 2.0) is fine-tuned on an aggregated multilingual speech emotion corpus for emotion classification.

Datasets are divided into train/test/validation sets (75%/15%/10% respectively) using stratified sampling across gender and emotions to maintain similar emotion ratios. These datasets cover 13 languages. The language distribution is highly skewed towards en-US, as most publicly available datasets are in English. The models are fine-tuned on seven major emotion categories. The corpus contains both natural and acted human speech emotions.

In Portuguese, the dataset used is VERBO. The weighted average recall for approaches using the emotion classification with Wav2vec 2.0 is 0.498. For the approach that combines emotion classification with Wav2vec 2.0, gender classification, language classification, and mean deviation for F0 regression, the weighted average recall is 0.566.

da Silva Franco et al. [6], on the other hand, focused on the applied use of a model to predict emotions using multimodality. The application uses a five-second audio chunk processing. MFCC features are extracted with the librosa library, and then a CNN model is used to predict the emotion. With this approach, the accuracy is 78.64%. Google API is used to transcribe the five-second chunks, and then feed the model proposed by Tripathi and Beigi [36]. For

Approach	Audio Feature	Model	Accuracy	Recall
Sharma [30] (2022)	Wav2Vec2 + Em + Ge + La + F0 mean	Wav2Vec2		0.56
da Silva Franco et al. [6] (2019)	MFCC	CNN	0.78	
Filho et al. [10] (2024)	MFCCs + chromatic + prosodic features	CNN specialist	~0.67*	
Joshi et al. [18] (2022)	MFCC mean	SVM	0.87	
BAH (our) (2025)	**TRILL**	**SVC**	**0.92**	**0.92**

Table 2: Comparative results from prior research, detailing the features, classification method, emotions, accuracy, and recall for each approach. The top result is shown in bold. *Em denotes Emotion Classification, Ge Gender classification, La Language classification, and F0 mean deviation for F0 regression. ** The authors did not provide the average accuracy result. Only accuracy per class.

the image processing, three frames per second are extracted. A CNN model is used to predict on the RAVDESS dataset. The authors utilize weights of 7, 38, and 55 in a new classification model for text, audio, and image data.

Filho et al. [10] propose the DEEP architecture, using specialized CNNs for each emotion class. VERBO was split into 70% for training and 30% for testing. The input features are combined into different Low-Level Descriptors (LLDs) as input for its expert models, with the MFCCs, chromatic, and prosodic features. The MFCCs are extracted into 13 coefficients per frame, capturing the power spectrum via framing, Hamming windowing, FFT, and a Mel Filter Bank. The chromatic features consider 12 weighted values per frame, providing information about the tonal classification of audio signals over time. Extraction involves FFT, Chromagram Calculation, and Normalization. Finally, the prosodic features use six values per frame (including loudness, jitter, and shimmer), which contain information about intonation, stress, tremor, and rhythm.

After the feature extraction process, the set of specialist CNN models is trained. The main difference in this work is that the authors used one specialist model for each emotion. These models operate in parallel.

The results presented in the paper are divided by emotions. The accuracy and F1-score for each emotion are 0.68 and 0.70 for neutral, 0.74 and 0.76 for disgusted, 0.56 and 0.69 for fear, 0.66 and 0.72 for happy, 0.67 and 0.68 for anger, 0.77 and 0.77 for surprise, and finally 0.66 and 0.69 for sadness.

Joshi et al. [18] proposed a speech emotion recognition model specifically for Brazilian Portuguese, using the VERBO corpus. Their study stands out for performing an extensive extraction of acoustic features, grouped into three main sets: MFCC, MFMC, and ST (Spectral-Temporal), as well as a broader initial vector called STPF, which contains over 2,000 features. These included cepstral, spectral, prosodic, temporal, and multifractal components, with statistical descriptors such as mean, standard deviation, kurtosis, and skewness computed on them. After applying Recursive Feature Elimination (RFE) for feature selection, the authors trained several traditional classifiers (SVM, MLP, KNN, RF, among others), with the SVM using an RBF kernel achieving the best results.

On the VERBO dataset, the model reached 87.56% mean accuracy and 87.55% weighted F1-score using only MFCCs with mean statistics, outperforming more complex deep learning approaches. The study also highlights that the MFMC set provided more balanced recognition across emotion classes, with reduced bias.

Our approach focuses on evaluating different lightweight and efficient non-semantic audio embeddings on the VERBO dataset to identify the best way to create representations for Portuguese. We also consider three different classifiers for this task. The previous works focus on handcrafted approaches, and only Sharma [30] used an audio embedding approach.

6 DISCUSSION

We compare our work with the related work in Table 2. The use of handcrafted features is more predominant on the VERBO dataset. The MFCC is explored in three different approaches. While the CNN model is used as the classifier by da Silva Franco et al. [6] and Filho et al. [10], Joshi et al. [18] used an SVM model. The best result in this case is using the SVM. But it is not the only factor; the authors also explored the use of the mean of MFCC. In this work, we aim to explore the use of audio embeddings for extracting acoustic features. With this approach, we get more accuracy than the previous state-of-the-art approach in the VERBO dataset. Joshi et al. [18] achieves 0.87 of accuracy. Using the same model, but with different hyperparameters, we achieved 0.92 accuracy.

Cross-lingual approaches, like [30], tend to have lower results than specific ones. This is because each culture has different emotional cues, which can affect the performance [12]. When comparing to our results using Wav2Vec2, Sharma's [30] model outperforms the recall, with 0.56, against our 0.48.

Unimodal SER is resource-efficient but lacks context [22]. When using with dimensional SER, the text directly improves the valence dimension [14]. For categorical SER, an interesting approach is to combine text sentiment with SER. The gains into EmoUERJ (Portuguese) [2] by [26] where the average accuracy is 0.95 using only handcrafted features for SER, 0.85 for sentiment analysis, and 0.96 for the fused approach.

7 THREATS TO VALIDITY

The current study is subject to certain limitations that offer avenues for future research. First, the evaluation was conducted exclusively on the VERBO dataset, which is a full-acted dataset and exhibits low speaker variability. To ascertain the model's generalizability and robustness, future work should involve validation on datasets recorded under naturalistic conditions. This step is crucial for evaluating the model's performance in realistic, real-world scenarios.

Second, our analysis was restricted to a specific subset of speech representation models (Whisper, WavLM, Wav2Vec, Hubert), considering only one variation size for each case. These selections do not necessarily represent the optimal choice for this task, and exploring larger model architectures could achieve different results.

Finally, this work relies on pre-trained models sourced directly from public repositories like TensorFlow Hub [1] and Kaggle Hub [2]. Consequently, the long-term reproducibility of our findings is contingent on the stability and versioning of these external models. Any subsequent updates or modifications to the model architectures within these hubs will affect our results.

8 CONCLUSION

In this paper, we provided a comprehensive evaluation of acoustic features for speech emotion recognition in Brazilian Portuguese. With our experiments, we show that the TRILL model is an excellent alternative to the traditional MFCC features.

We compare 18 different methods for extracting features from audio, grouped by handcrafted and audio embeddings. The handcrafted are: ComParE, eGeMAPS, and pAA. The audio embeddings are: Hubert, FRILL, TRILL, TRILLsson5, VGGIsh, Wav2Vec2, WavLM, and Whisper. We delve deeper into each method used in Section 3.2. Our proposal is an SVM model, which, compared with XGBoost, achieves the best accuracy in 17 cases. The TRILL embedding achieves the best result, with an accuracy of 92.00%.

Our main contribution is a complete experiment set focusing on the Brazilian Portuguese language. The same approach can be applied to different languages, but it does not guarantee that the same acoustic features will yield the best results in feature selection.

As future directions, we aim to explore bi-modality by combining text features with acoustic ones for categorical emotion, thereby enhancing the context of the SER. We already tested this with dimensional SER [14], but when we work with a categorical task, the impact of the text on the classification results works differently. We aim to evaluate our trained model on the emoUERJ dataset to ensure it is not biased on VERBO. Finally, we will try an approach to compare the similarity among the embeddings generated by TRILL to see how different emotions can be clustered. With this, we can map the emotions closely to the circumplex model of affect. With this, it will be possible to generalize the model for new emotions.

ACKNOWLEDGMENTS

This work was partially supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, FAPERGS 09/2023 PqG (Nº 24/2551-0001400-4), CNPq Research Program (Nº311012/2025-6), and PUCRS institutional program to incentivize Stricto Sensu graduate studies - PRO-Stricto.

REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12449–12460. https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[2] Rodrigo Gregory Bastos Germano, Michel Pompeu Tcheou, Felipe da Rocha Henriques, and Sergio Pinto Gomes Junior. 2021. *emoUERJ: an emotional speech database in Portuguese.* https://doi.org/10.5281/zenodo.5427549

[3] Kirsten Boehner, Rogério DePaula, Paul Dourish, and Phoebe Sengers. 2005. Affect: From Information to Interaction. In *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility* (Aarhus, Denmark) *(CC '05)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/1094562.1094570

[4] Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, et al. 2022. CORAA ASR: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. *Language Resources and Evaluation* (2022), 1–33.

[5] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing* 16, 6 (2022), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

[6] Roberto Yuri da Silva Franco, Rodrigo Santos do Amor Divino Lima, Rafael do Monte Paixão, Carlos Gustavo Resque dos Santos, and Bianchi Serique Meiguins. 2019. UXmood — A Sentiment Analysis and Information Visualization Tool to Support the Evaluation of Usability and User Experience. *Information* 10, 12 (2019). https://doi.org/10.3390/info10120366

[7] Javier de Lope and Manuel Graña. 2023. An ongoing review of speech emotion recognition. *Neurocomputing* 528 (April 2023), 1–11. https://doi.org/10.1016/j.neucom.2023.01.002

[8] Paul Ekman. 1999. *Basic Emotions.* John Wiley and Sons, Ltd, Chapter 3, 45–60. https://doi.org/10.1002/0470013494.ch3

[9] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery, New York, NY, USA, 1459–1462. https://doi.org/10.1145/1873951.1874246

[10] Geraldo P. Rocha Filho, Rodolfo I. Meneguette, Fábio Lúcio Lopes de Mendonça, Liriam Enamoto, Gustavo Pessin, and Vinícius P. Gonçalves. 2024. Toward an emotion efficient architecture based on the sound spectrum from the voice of Portuguese speakers. *Neural Computing and Applications* 36, 32 (Aug. 2024), 19939–19950. https://doi.org/10.1007/s00521-024-10249-4

[11] A.V. Geetha, T. Mala, D. Priyanka, and E. Uma. 2024. Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions. *Information Fusion* 105 (March 2024), 102–218. https://doi.org/10.1016/j.inffus.2023.102218

[12] K Ghaayathri Devi, Kolluru Likhitha, J Akshaya, Rfj Gokul, and G Jyothish Lal. 2024. Multi-Lingual Speech Emotion Recognition: Investigating Similarities between English and German Languages. In *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. 1–10. https://doi.org/10.1109/ACCAI61061.2024.10601715

[13] Theodoros Giannakopoulos. 2015. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PLOS ONE* 10 (12 2015), 1–17. https://doi.org/10.1371/journal.pone.0144610

[14] Larissa Guder, João Paulo Aires, Felipe Meneguzzi, and Dalvan Griebler. 2024. Dimensional Speech Emotion Recognition from Bimodal Features. In *Brazilian Symposium on Computing Applied to Health*. Brazilian Computing Society, 12.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/ARXIV.1512.03385

[16] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA, USA). IEEE Press, 131–135. https://doi.org/10.1109/ICASSP.2017.7952132

[17] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 29 (Oct. 2021), 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

[18] Neelakshi Joshi, Pedro V. V. Paiva, Murillo Batista, Marcos V. Cruz, and Josué J. G. Ramos. 2022. Improvements in Brazilian Portuguese Speech Emotion Recognition and its extension to Latin Corpora. In *2022 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892110

[19] Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. 2021. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* 10 (January 2021), 1163. https://doi.org/10.3390/electronics10101163

[1] Tensorflow Hub is available on http://tensorflow.org/hub

[2] Kaggle Hub is available on https://www.kaggle.com/models

[20] Kristina Loderer, Kornelia Gentsch, Melissa C. Duffy, Mingjing Zhu, Xiyao Xie, Jason A. Chavarría, Elisabeth Vogl, Cristina Soriano, Klaus R. Scherer, and Reinhard Pekrun. 2020. Are concepts of achievement-related emotions universal across cultures? A semantic profiling approach. *Cognition and Emotion* 34 (March 2020), 1480–1488. https://doi.org/10.1080/02699931.2020.1748577

[21] Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology* 14 (December 1996), 261–292. https://doi.org/10.1007/BF02686918

[22] Nuzhat Mobassara, Nur Alam, and Nursadul Mamun. 2025. A Comprehensive Review of Speech Emotions Recognition using Machine Learning. In *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. 1–6. https://doi.org/10.1109/ECCE64574.2025.11013787

[23] Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Transactions on Affective Computing* 5 (April 2014), 101–111. https://doi.org/10.1109/TAFFC.2014.2317187

[24] Jacob Peplinski, Joel Shor, Sachin Joglekar, Jake Garrison, and Shwetak Patel. 2021. FRILL: A Non-Semantic Speech Embedding for Mobile Devices. In *Interspeech 2021 (interspeech 2021)*. ISCA, 1204–1208. https://doi.org/10.21437/interspeech.2021-2070

[25] Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA.

[26] Diego Resende Faria, Abraham Itzhak Weinberg, and Pedro Paulo Ayrosa. 2024. Multimodal Affective Communication Analysis: Fusing Speech Emotion and Text Sentiment Using Machine Learning. *Applied Sciences* 14, 15 (2024). https://doi.org/10.3390/app14156631

[27] Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and Detecting Emotions on Twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, 3806–3813. http://www.lrec-conf.org/proceedings/lrec2012/pdf/201_Paper.pdf

[28] J.A. Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39 (December 1980), 1161–1178.

[29] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. 2016. The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity and native language. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 08-12-September-2016. 2001 – 2005. https://doi.org/10.21437/Interspeech.2016-129

[30] Mayank Sharma. 2022. Multi-Lingual Multi-Task Speech Emotion Recognition Using wav2vec 2.0. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6907–6911. https://doi.org/10.1109/ICASSP43922.2022.9747417

[31] Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Félix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. 2020. Towards Learning a Universal Non-Semantic Representation of Speech. In *Interspeech*. ISCA, 140–144. https://doi.org/10.21437/interspeech.2020-1242

[32] Joel Shor and Subhashini Venugopalan. 2022. TRILLsson: Distilled Universal Paralinguistic Speech Representations. In *Interspeech 2022 (interspeech 2022)*. ISCA. https://doi.org/10.21437/interspeech.2022-118

[33] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://arxiv.org/abs/1409.1556

[34] Youddha Beer Singh and Shivani Goel. 2022. A systematic literature review of speech emotion recognition approaches. *Neurocomputing* 492 (July 2022), 245–263. https://doi.org/10.1016/j.neucom.2022.04.028

[35] José; R. Torres Neto, Geraldo P.R. Filho, Leandro Y. Mano, and João; Ueyama. 2018. VERBO: Voice Emotion Recognition dataBase in Portuguese Language. *Journal of Computer Science* 14, 11 (Nov 2018), 1420–1430. https://doi.org/10.3844/jcssp.2018.1420.1430

[36] Samarth Tripathi and Homayoon S. M. Beigi. 2018. Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning. arXiv:1804.05788 http://arxiv.org/abs/1804.05788