# Brazilian Portuguese in speech emotion recognition

**Bolsista Neelakshi Joshi (CTI) njoshi@cti.gov.br**

**Coauthors: Pedro Paiva, Murillo Batista, Marcos Cruz, & Josué Ramos**

## Abstract

*For better practical automated experience, attention is towards building emotionally intelligent machines that can recognize real time emotion of the user and can simulate emotions while responding. Speech emotional data is crucial for emotion recognition studies and respective model development goes parallelly with resource availability. Different types of emotional databases are available and known by their acquisition methods such as natural, acted and elicited, among which recognizing emotions from natural speech is the most challenging. This work addresses the scarceness of resources in Brazilian Portuguese by introducing only available two databases. Reviewing related work carried on to address resource scarcity, possible paths are discussed along with their needs and challenges, to progress emotional studies in Brazilian Portuguese. We emphasize the need of the larger in-the-wild multimodal emotional database in Brazilian Portuguese to advance human-robot interaction. Research with adequate resources will be beneficial to build emotion recognition models for practical robotic use in Brazilian society.*

*Palavras-chave: Speech emotion recognition, Speech emotion database, Brazilian Portuguese, Survey.*

## 1. Introduction

Artificial emotional intelligence (AEI) enables machines to process, interpret and simulate emotions to advance human machine interaction by responding more human-like (SCHULLER, 2018). AEI is advancing in diverse domains like marketing, advertising, education, healthcare, etc., knowing that users favor emotionally responding bots and applications over without it (VOGT et al., 2008). Humans express emotions through many modalities such as speech, text, facial expression, body poses, and can be obtained through physiological signals. The field speech emotion recognition (SER) utilizes vocal, speech and non-verbal communication signals to recognize the underlying emotions. Humans do perceive emotion from different languages based on the speaker's vocal modulation, yet the ability is limited by cultural influence on languages and intercultural differences. Machines learn emotions through extracted suprasegmental features that vary within a language, hence is challenging (WANG et al., 2022). In the current data-driven world, as obvious, AEI is advancing in the languages that possess the most varied resources, for e.g. English (SCHULLER, 2018, ZADEH et al., 2020). Brazilian Portuguese (BP), accounts for 81% of the Portuguese-speaking population which is 6$^{th}$ most spoken language in the world (https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native _speakers), yet only two small SER databases are available.

In this work, we want to draw the reader's attention to the scarcity of SER resources in BP. The current work is organized as follows. Section 2 introduces emotional database types. Section 3 reports the earlier work done on transferring emotional learning from one type of database to another. In section 4, we explore the available BP SER databases and summarize

respective previous work. Finally, in section 5, we discuss possible pathways to progress SER in BP.

## 2. Types of speech emotional databases

For SER studies, speech signals are obtained from speech emotional databases (or corpora) and are named based on their acquisition methods as described below.

*Acted* database is recorded in a noise free environment with an expert actor who can simulate predefined text with desired emotions. Predefined texts are semantically neutral to avoid any interference. Each acted corpus is distinct, in terms of number of utterances and emotions. Acted speech is comparatively easier to learn as speech is semantically neutral and noise free but may overlook subtleties due to intense or overemphasized expressions. To overcome limitations in acted corpus and to get closer to natural corpus, desired emotions can be induced in the recording environment artificially to obtain *Elicited* or *Induced* databases. The closeness to natural speech depends on how the induced environment is controlled (VOGT et al., 2008). Acted and elicited databases are also known as non-spontaneous databases as these are produced in controlled environments and by regulating an actor's emotional performance (WANG et al., 2022). Contrary, when a controlled recording environment gives emotion expression liberty to participants to capture emotional data in a more naturalistic way, is known as *Spontaneous* database (e.g. Belfast database, WANG et al., 2022).

*Natural* databases is a collection of natural day-to-day conversation, real-world shows, or social media contents. As conversations happen spontaneously, it is also known as *Spontaneous* or *in-the-wild* corpora. It is the most challenging corpora to obtain and to analyze. The situation and temperament of a speaker define the emotion and its intensity. The availability of natural speech with the desired emotions and in the desired language is itself challenging and one needs to be careful while acquiring it so as not to violate legal obligations. As a non-actor speaker, expressions of emotion are unique and can be a blend of many emotions. Natural speech may be lengthy, full of semantic context, unworded vocal expression such as laughter or cry, and can contain background noise. Hence, many such corpora are multimodal containing audio, video and respective text transcripts (e.g. CMU-MOSEI). Discerning the challenges in acquiring spontaneous emotional speech, non-spontaneous sources are immense to learn about emotions, as they are easily available in the desired number of emotions in desired languages (HUANG et al. 2013).

Corpora classifies the emotion embedded in speech mainly using two models, discrete and dimensional. Ekman's discrete emotional model (1971) believes that basic emotions have identical forms across different languages and cultures. The six basic emotions are anger, disgust, fear, happiness, sadness and surprise. Complex emotions can be represented using combinations of basic emotions. Russell's circumplex model (1980) comprises valence and arousal axes. The valence axis scales from negative (displeasure) to positive (pleasure) emotional state along with the arousal axis, which calibrates the intensity of the emotion. Recently, researchers are preferring a latter model as it accommodates a large range of emotions, including complex emotions.

**2.1. Recent *in-the-wild* databases:** Recently in English, the largest multimodal database, CMU-multimodal opinion sentiment and emotion intensity (CMU-MOSEI), was developed by Linag et al. (2018). Though the majority of corpus are in English, the aim was to address pressing issues of size, extemporaneous speech, and speech environment. The Corpus provides data in text, audio, and video modalities for sentiment and emotion analysis and spans ~66 hours. With the same motive, another large-scale, multilingual, and multimodal CMU multimodal opinion sentiment, emotions, and attributes (CMU-MOSEAS) database was

created by Zadeh et al. (2020) in European Spanish, Portuguese, German, and French languages spanning ~69 hours.

## 3. Inter-corporal emotional learning transfer studies

Various detailed reviews exist giving up-to-date information on SER databases, their modalities, and experiments done using different techniques (WANG et al., 2022, ZHANG S. et al., 2021). These reviews emphasized the need for a multimodal natural emotion database in many languages to enhance learning for harmonious HRI interactions. This section notes some important findings from inter-corporal emotional learning transfer studies reported to overcome scarcity of resources.

Küstner et al. (2004) analyzed acted and elicited databases, and studied their respective acquisition methods to find differences. They found that acted speech is easier to classify compared to elicited speech. Environmental, additive, or convoluted noise do affect classifier performance. Thus, the author advised adding such noise-contaminated data to the training set during the analysis. Feature selection does influence the analysis, but variance in utterance length and the type of database are more accountable than the extracted features from them. Huang et al. (2013) introduced an online learning method where acted data is used for offline training and elicited data is used for online training and testing. To transfer such learning, the authors mentioned the need for a larger training set. Schuller et al. (2010) performed cross-corpus SER analysis with three acted and three spontaneous corpora. The authors noted that corpora of Germanic languages (English, German, and Danish) allowed inter-corpus testing, pointing out the similarities in the cultural background of Germanic languages and emphasized the influence of cultural differences on recognizing emotions.

Deschamps-Berger et al. (2021) analyzed elicited and natural corpora, extracting spectrograms and their first and second order derivatives, combining them into a 3-d feature that was fed into the CNN+BiLSTM network. They stated that natural speech is more complex as emotions vary per speaker, and speech contains blends of emotions and shaded feelings. Milner et al. (2019) evaluated transfer learning within inter-corpus analysis with acted, elicited and natural corpora using cross-corpus, multi-domain, and out-of-domain (OOD) analysis. The authors observed the performance reduction with a natural corpus. And suggested that adopting domain adversarial training with OOD may generalize emotion learning across databases.

Typically, speech emotional corpora are smaller, and being language-specific, it is more difficult to acquire more utterances. To overcome it, some researchers used pre-trained deep learning models which are trained on very large datasets. Alexnet is a deep Convolutional neural network (DCNN) model which is trained on imagenet dataset. Badshah et al. (2017) used pre-trained Alexnet and their own CNN to train spectrograms extracted from an acted speech corpus. They found that training a fresh network was easier than fine-tuning a pre-trained model as data types are different. Zhang H. et al. (2021) and Zhang et al. (2022) used Alexnet to extract features to learn emotions from acted, elicited speech and from spontaneous speech respectively. Both authors built and trained their own deep learning model using learnt features. Recently, Sharma (2022) reported a multilingual and multi-task learning SER system, using two pre-trained models: PANN and the multi-lingual wav2vec 2.0. PANN is pre-trained on the audioset containing speech, non-speech, and various environmental noises. The wav2vec 2.0 model is pre-trained on the multilingual corpora XLSR-53, containing 53 languages. The author observed that the multi-lingual model outperformed the single task model, and the wav2vec 2.0 model provided better results than the PANN model. The author noted that due to resourcefulness English was the dominant language in his multilingual study.

## 4. BP SER databases and related work

I. Acted database : *Voice Emotion Recognition (VERBO)*

The voice emotion recognition (VERBO) database is the first acted speech emotion corpus available in BP (NETO et al., 2018). VERBO was recorded with 12 professional Brazilian actors. It includes 14 phrases covering all of the BP linguistic phonemes. It follows a discrete emotional model and thus comprises the six primary emotions: anger, disgust, fear, happiness, sadness, and surprise, along with neutral emotion. Clean, semanticless audios are labeled as per the emotion that is presented in utterances.

II. Spontaneous database : *CORAA SER 1.0*

CORAA SER version 1.0 (https://github.com/rmarcacini/ser-coraa-pt-br) is a collection of spontaneous BP utterances collected from corpus C-ORAL-BRASIL I (RASO & MELLO, 2012) designed to study spoken BP language. This corpus was made available as a part of the SE&R contest. It contains a total of 933 audio files (625 as training set + 308 as test set). These audios are labeled into three categories: neutral, non-neutral-female, and non-neutral-male, depending upon the presence of paralinguistic elements in them. Audio segments without well-defined emotions are categorized as neutral, whereas audio segments with the presence of one of the primary emotions are categorized as non-neutral. The corpus contains noise and paralinguistic elements like laughter and cry, and is highly imbalanced among labels. In addition, two baseline macro F1 scores of 55% using wav2vec features and 50% using prosodic features were provided.

Table 1 details the characteristics of both databases.

| characteristics | VERBO | CORAA |
|---|---|---|
| database type | acted | spontaneous |
| total recording time | 47 min | 60.21 min |
| total audio signals | 1167 | 930 |
| minimum audio duration | 1.19 sec | 2 sec |
| maximum audio duration | 5.47 sec | 15 sec |
| number of emotion classes | 7 | 3 |

Source: Author's production

Table 1 – characteristics of VERBO and CORAA databases

### 4.1. Related Work and Experiments

We analyzed the VERBO corpus and proposed the SER model (JOSHI et al., 2022) which provided the best recognition rate for all emotions compared to the previous studies, presented in Table 2. The model comprised six spectral and temporal features: - Mel frequency cepstral coefficients (MFCC), constant Q transform chroma, RMSE, amplitude envelope, spectral rolloff and spectral contrast with their mean statistics, and support vector machines (SVM) as a classifier. Also, a single spectral feature, MFCC is found to be sufficient and obtained mean accuracy of 87.56%. Comparing two spectral features, MFCC and its variant, MFMC, we demonstrated how specific statistical measures can improve results.

| Features | Classification Method | Accuracy | Reference |
|---|---|---|---|
| Spectral, Prosodic | K-nearest neighbors | 76.49 % | Neto (2020) |
| Spectral, Prosodic | not mentioned | 78.64 % | Silva et al. (2019) |
| Spectral, Prosodic | Convolutional Neural Network | 76.69 % | Campos & Moutinho (2020) |
| Spectral, Temporal | SVM | 87.32 % | this work (Joshi et al., 2022) |

Source: Author's production

Table 2 – comparing our SER model results with earlier work done with VERBO database

Further, we extended this analysis as multi-corporal including three more Romance family languages, Italian (EMOVO), Spanish (INTER1SP), Canadian French (CaFE) along with Brazilian Portuguese. The proposed model improved results compared to previous analysis for all four corpora. The work was presented in the International Joint Conference on Neural Networks (IJCNN) conference and the article is in the press. Recently, Sharma (2022) analyzed VERBO using pre-trained, fine-tuned PANN and wav2vec models, and reported mean weighted F1 score of 45.1% and 44.7% respectively.

As part of SE&R competition, the CORAA corpus has been analyzed using different techniques including transfer learning. CORAA is small and has highly imbalanced labels. Alves et al. (2022) inspected the CORAA training set very carefully and categorized each audio as per the presence of noise, voice overlapping, different genders' voice in the same audio, and same voice in sequence. Their observation was that the *CORAA training set audios are noisy, have a higher overlapping rate, some audio contains different gender voices, and very few audios are in the same voice in sequence*.

Alves et al. (2022) used a synthetic minority over-sampling technique and Praat's gender change command to oversample minority classes. A pre-trained model for genre classification trained on CETEN-Folha corpus was used for transfer learning in multi-tasking and sequential ways. Different statistics obtained over various voice quality, prosodic, and spectral features were chosen as input to their model. They achieved a 53.53% test score. Also, the scarcity of data in BP SER is addressed in their work. Perin and Matsubara (2022) introduced a new approach using transductive ensemble learning with a graph convolutional network. With this approach, the CORAA test macro F1 score was 53%. Scaranti et al. (2022) preprocessed CORAA audios to filter noise. Prosodic and spectral features were modeled with a multi-layer perceptron (MLP) classifier and achieved a 55% test score.

Gauy and Finger (2022) used PANN and Transformer models for transfer learning, and SpecAugment method for data augmentation. PANN-Cnn10 the authors achieved a 73% score on the test set, surpassing the baseline score. They observed that due to smaller training size, deep models like PANN-Cnn14 and Transformers were overfitted. They expressed the need for larger training data. The pre-trained Transformers data didn't contain the paralinguistic elements. Hence, to use Transformers, one needs to train it on CORAA-similar data sets.

**Preliminary analysis:** To explore CORAA that is smaller in size, we adopted the classical approach for the analysis. MFCC is a self-sufficient feature and robust against noise (FAHAD et al. 2021), hence first we analyzed CORAA data with MFCC as a feature and SVM as a classifier. CORAA's training set was trained and a model was evaluated on the test set. For given 3 emotion classes, a macro F1 score of 54% is obtained which is higher than a base score provided with prosodic features.

Alves et al. (2022) have shown how biased the CORAA corpus is in labeling audios gender-wise. Hence, for inter-corpus analysis within VERBO and CORAA, we selected two common labels, neutral and non-neutral, considering intersection of two very different databases and labels. Further, all presented experiments have been performed using these two labels. VERBO has 167 neutral and 1000 non-neutral utterances. CORAA contains a total of 739 neutral utterances (491 from train set + 248 from test set) and 194 non-neutral utterances (134 from train set + 60 from test set). Figure 1 shows how imbalanced these labels are in both datasets. Table 3 lists three different inter-corpus preliminary analyses performed with VERBO and CORAA with MFCC as a feature and with SVM and MLP classifiers which performed well among different classifiers. (i) the model is trained only on acted speech (VERBO) and tested on spontaneous speech (CORAA). Poor results confirm the earlier finding that acted speech learning can't be transferred to spontaneous speech. SVM worked better with acted speech. (ii) Then a model is trained on combining VERBO and the training

set of CORAA. It is evaluated on the test set of CORAA. Comparatively, results are improved affirming earlier finding that the training set must include test data type samples for learning. MLP is a better classifier than SVM on spontaneous data.

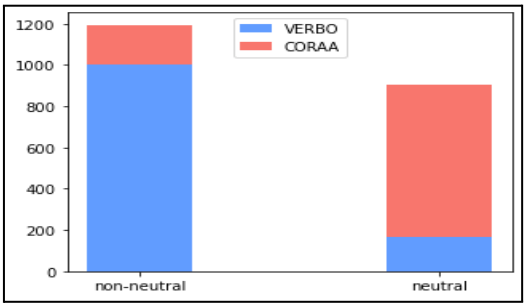


Figure 1 – distribution of class labels in VERBO and CORAA

(iii) To test further, entire CORAA and VERBO data is combined and splitted in training-validation-test sets, such that the test set will contain more spontaneous data. VERBO is splitted in proportion of 80-15-5% and CORAA of 80-10-10% with a stratified approach.

| # | training set | validation / test set | classifier | F1 score |
|---|---|---|---|---|
| i | VERBO | CORAA (train+test) | SVM | 21% |
| | | | MLP | 19% |
| ii | VERBO + CORAA (train) | CORAA (test) | SVM | 48% |
| | | | MLP | 61% |
| iii | 80% training set: combined VERBO & CORAA | 12.5% validation set 7.5% test set | MLP | 87.71% 81.69% |

Source: Author' production

Table 3 – Inter-corpus experiments performed with VERBO and CORAA

Table 4 shows the confusion matrix for the test set. Recognition rate for non-neutral emotions is higher compared to neutral emotions. This is attributed to the larger number of samples available for non-neutral class and the 5 times larger share of clean audios compared to the noisy ones.

| True Labels | Predicted Labels | |
|---|---|---|
| | non-neutral | neutral |
| non-neutral | 61 | 10 |
| neutral | 18 | 64 |

Source: Author's production

Table 4 – Confusion matrix for 7.5% test set (5% VERBO and 10% CORAA)

## 5. Discussion

This work introduces the available SER corpora in BP, related work and experiments performed. Also presentes a survey done on the various inter-corpus learning experiments to overcome scarcity of resources. All reviewed studies agreed on (i) complexity involved in natural SER analysis, (ii) acted speech learning can't be transferred to natural emotion learning, (iii) test data-type samples inclusion in training phase is necessity, (iv) the urgent need for a larger diverse training set. It is indisputable that to achieve AEI, a prime necessity is the availability of a natural or spontaneous emotional speech database, which should be large and diverse enough to learn the subtleties underlying complex real-world emotions. On the contrary, the number and volume of BP SER corpora are limited. Only one acted and one spontaneous SER corpus is available. The analysis presented in Section 4.1 emphasizes the need of addressing data scarcity in BP. The primary objective of this work is to assess the feasibility of natural SER in BP given data scarcity. From the above literature review, a few

pathways can be inferred – data augmentation, multilingual study, transfer learning, and building a new in-the-wild emotion database.

Data scarcity can be addressed by data augmentation. Various augmentation methods are available and have shown mixed results (ALVES et al.,2022, GAUY & FINGER, 2022). Augmented methods are to be chosen wisely such that deep learning algorithms won't learn to overfit (ZHANG et al., 2021). Another pathway is to choose natural SER corpora from a family of languages similar to BP. Such experiments were done only with acted databases which were found to be compatible for inter-learning (SCHULLER et al., 2010, JOSHI et al., 2022). Though it is an interesting option, the question is do large natural corpora available. Recently created CMU-MOSEAS is the largest in-the-wild available dataset but yet labels are not made available for further study.

The next feasible option is to transfer the learning from a pre-trained model. There are caveats though. The availability of the pre-trained model trained on a data type similar to the one considered in the experiment is a must. For CORAA, noise could be filtered using the PANN pre-trained model (GAUY & FINGER, 2022), but with VERBO it yielded poor results (SHARMA, 2022). Next, a deep pre-trained model gives best results with a larger training set; on the contrary, models overfit for a smaller training set. The wav2vec with VERBO as well as Transformer and PANN-Cnn14 with CORAA were unable to produce the desired results. Gauy & Finger (2022) mentioned the need for a large training set to transfer learning from these models. Current status of BP SER urges on developing a multimodal baseline database of natural emotions, which is the next pathway. However, the task of building such a database is substantial.

Finally, among other affects, emotion is distinct, being intense but available for a short duration. Hence, it is crucial to use all possible modalities (audio, text, video, psycho-physio signals) with spontaneous or nonspontaneous environments and labeled them in discrete and dimensional emotional classes to endow emotional intelligence to machines (SCHULLER, 2018, WANG et al., 2022). From the above discussion and experiments, it can be concluded that increasing the size of BP databases or developing a new, possibly multimodal, natural emotion database to address the scarcity is an essential and critical factor to achieve AEI for HRI in BP.

## References

**ALVES, C. & CARLOTTO, B. & DIAS, B. & GARCIA, A. & GIANESI, B. & IZAIAS, R. & et al.** *Transfer Learning and Data Augmentation Techniques applied to Speech Emotion Recognition*. In Computational Processing of the Portuguese Language, PROPOR proceedings, 2022.

**FAHAD, M.S. & RAJAN, A. & YADAV, J. & DEEPAK, A.** *A Survey of Speech Emotion Recognition in Natural Environment*. Digital Signal Processing, volume 110, p. 102951, 2021.

**GAUY, M. M. & FINGER, M.** *Pretrained audio neural networks for Speech emotion recognition in Portuguese* In Computational Processing of the Portuguese Language, PROPOR proceedings, 2022.

**HUANG, C. & LIANG, R. & WANG, Q. & XI, J. & ZHA, C. & ZHAO, L.** *Practical Speech Emotion Recognition based on Online Learning: From Acted data to Elicited data*. Mathematical Problems in Engineering, 2013.

**JOSHI, N. & PAIVA, P.P. & BATISTA, M. & CRUZ, M.V. & RAMOS, J.J.G.** *Improvements in Brazilian Portuguese Speech Emotion Recognition and its extension to Latin Corpora*. Article accepted in International Joint Conference on Neural Networks, IEEE, 2022, in press. https://github.com/neelakshij/Speech-Emotion-Recognition-in-Brazilian-Portuguese

**NETO, J.R.T. & FILHO, G. & MANO, L. & UEYAMA, J.** *Verbo: Voice Emotion Recognition database in Portuguese Language*. Journal of Comput Science, volume 14, no. 11, p. 1420–1430, 2018.

**RASO, T. & MELLO, H.** *The C-ORAL-BRASIL I: Reference Corpus for Informal Spoken Brazilian Portuguese*. In International Conference on Computational Processing of the Portuguese Language, p. 362–367, Springer, 2012.

**SCHULLER, B.W.** *Speech Emotion Recognition: Two decades in a Nutshell, Benchmarks, and Ongoing Trends*. Communications of the ACM, Vol. 61, n. 5, p. 90-99, 2018.

**VOGT, T. & ANDRÉ, E. & WAGNER, J.** *Automatic Recognition of Emotions from Speech: a Review of the Lterature and Recommendations for Practical Realisation*. Affect and emotion in human-computer interaction, pp.75-91, 2008.

**WANG, Y. & SONG, W. & TAO, W. & LIOTTA, A. & YANG, D. & LI, X. & et al.** *A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances*. Information Fusion, 2022.

**ZADEH, A. & CAO, Y.S. & HESSNER, S. & LIANG, P.P. & PORIA, S. & MORENCY, L.P.** *CMU-MOSEAS: A Multimodal Language Dataset for Spanish, Portuguese, German and French*. Proc Conf Empir Methods Nat Lang Process, 2020.

**ZHANG, H. & RUOYUN, G. & SHANG, J. & SHEN, F. & WU, Y. & DAI, G.** *Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition*. Frontiers in Physiology, volume 12, article 643202, 2021.

**ZHANG, S. & LIU, R. & TAO, X. & ZHAO, X.** *Deep Cross-Corpus Speech Emotion Recognition: Recent Advances and Perspectives.* Frontiers in Neurorobotics, volume 15, 2021.

**ZHANG, S. & ZHAO, X. & TIAN, Q.** *Spontaneous Speech Emotion Recognition using Multiscale Deep Convolutional LSTM.* IEEE Transactions on Affective Computing, volume 13, 2022.