Speech Emotion Recognition Model for Human-Robot Interaction using Machine Learning

Bolsista Neelakshi Joshi (CTI) njoshi@cti.gov.br
Orientador: Dr. Josué J. G. Ramos

Abstract

Speech is one of the modalities to identify underlying emotions by analyzing vocal and non-verbal modulations of the speech. Humans do perceive emotions 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. Speech emotion recognition (SER) modeled for Brazilian Portuguese can present robots as emotion-aware and perceiver and be helpful in human-robot interaction studies. As with any supervised learning task, it is constrained by data availability. Though Portuguese is ranked as the 8th most spoken language in the world, very few SER corpora are available in Brazilian Portuguese. Moreover, these are smaller and with class imbalance. This work presents the SER analysis of the available corpora and discusses the findings.

Key-words: Speech Emotion Recognition, Machine Learning, HRI, SER.

1. Introduction

Humans express emotions through many modalities such as speech, text, facial expression, body poses. Speech emotion recognition (SER) analysis utilizes vocal, speech, and non-verbal communication signals to recognize the underlying emotions. Identifying emotions experienced by the speaker is challenging. 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 it is more challenging (WANG et al., 2022). In the current data-driven world, to increase the acceptability and usefulness of speech-related technology, attention is already given to pragmatic and paralinguistic aspects of speech to increase the effectiveness of technology. While reviewing SER real-time applications, the authors (VOGT et al., 2008, TÓTH et al., 2008) noted that people prefer responses from emotionally sensitive virtual agents. However, advancement of SER is constrained by the resourcefulness of languages, so very few SER real-world applications are in the market and are limited to well-resourced languages (SCHULLER, 2018).

Different types of emotional databases are available and known by their acquisition methods such as real (or natural), acted, and elicited. Acted database is created with the help of actors who can simulate desired emotions in predefined text and is recorded under a controlled environment. Artificially inducing desired emotions in the recording environment make elicited database different than acted. A real database consists of natural conversations like real life shows, interviews, or social media contents. As conversations happen spontaneously, it is also known as *Spontaneous* or *in-the-wild* corpus. It is the most challenging corpus to obtain and 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.

To identify underlying emotions from speech signals, obtaining optimal features is a crucial step. Among them spectral, prosodic, and temporal categories are widely used. Prosodic features (e.g. stress, intonation, rhythm) reveal how speech signals are perceptible to humans. Temporal features (e.g. intensity, amplitude envelope) inform changes in the signal over time. Energy and intensity can be derived from variation in amplitude of speech signal which varies per emotion. Spectral features (e.g. spectrogram, various cepstral features) are obtained through transforming signal from time domain to frequency domain. The time-frequency transform of a log-spectrum is known as cepstrum. Modeling spectral features on logarithmic scale helps in speech perception as it approximate human hearing sensitivity. Only spectral features can provide crucial information about the vocal tract and hence found to be an integral part of all SER studies. Different statistical measures computed over local features are known as global features. Our earlier study (JOSHI et al., 2022a) has shown that for Brazilian Portuguese (BP) acted SER studies, Mel frequency cepstral coefficients (MFCC) with mean statistics is the robust and optimal feature. Irrespective of database type it is the most widely analyzed feature (WANG et al., 2022, ZHANG et al., 2021).

With the aim of developing the SER model for Brazilian Portuguese, we explored the available SER databases. We found both acted and real types and are small but are very imbalanced among emotional classes. Thus, we experimented with two augmentation strategies to make the train set balance so the model will learn all the emotion classes adequately. This work uses a simple but proven methodology to obtain the BP-SER model. This work is organized as follows. Section 2 introduces databases, augmentation methods, and analysis methodology. Section 3 details the analysis by presenting results. Section 4 concludes the paper by discussing findings.

2. Methodology

This section introduces databases used in this work, the strategy followed to combine them together, implemented augmentation methods and how analysis is performed.

We found four datasets with emotional audios in BP.

- SER dataset by Rosa (2017): ROSA (2017) built the BP SER dataset[1] by extracting audios from films and YouTube videos of a few seconds duration as a part of graduation work. It contains a total of 143 audios with anger, fear, happiness, neutral, sadness, surprise, and tense emotions. The author constructed this dataset as no other SER corpus was available. Also, alerted of possible errors in the annotations, as the author himself annotated the extracted audios.

[1] https://github.com/jdarosaj/emotion_portuguese_database

- VERBO: The Voice emotion recognition database (VERBO, TORRES et al., 2018) is an acted SER corpus containing a total of 1167 audios for seven basic emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. It is recorded with 12 professional Brazilian actors (6 females and 6 males) with theater experience of 2–23 years and were chosen from different regions in Brazil. It contains predefined 14 phrases, including all BP linguistic phonemes, with four categories: short sentences, phrases, questions, and nonsense phrases. The clean, semanticless audios are labeled according to the emotions presented in the utterances.

- Emossônico: SER corpus Emossônico[2] (KINGESKI, 2019) contains three datasets, each obtained from different sources. The first dataset contains 142 audios extracted from 14 national films, 1 national series, and 5 dubbed films composed of 85 female and 56 male voices for anger, fear, happiness, neutral, and sadness emotions. The second dataset contains 245 audios for basic seven emotions, which were selected from online videos of ordinary people (5 males). The third dataset, elicited one, was created by the author with 6 female and 9 male participants while they were watching videos (a way to induce emotions). A total of 200 audios in seven emotions were recorded.

- emoEURJ: emoEURJ[3] (BASTOS et al., 2021) was developed at the State University of Rio de Janeiro for the development of SER models acknowledging database scarcity in BP. Ten sentences were recorded by eight actors (4 females and 4 males) in anger, happiness, sadness, and neutral emotions. This acted database comprises 377 audios in total.

The audio distribution per emotion within these databases are listed in Table 1. This shows the distribution of audios within the Emossônico dataset which are recorded (acted) and gathered from the internet (real) separately in light colors. Please note, a row for the Emossônico database shows the total number audios available in the corpus per emotion, considering both, acted and real. SER dataset created by ROSA (2017) comes under real type. Thus, we have two types of corpora, acted and real data types.

As these corpora are imbalanced, we experimented with two augmentation strategies. The first one is a classical way of augmentation where the following three ways were implemented randomly to augment audios till desired numbers.

- ➢ time stretching: this method changes the speed of audio without altering its pitch
- ➢ pitch shift: this method modifies the pitch up or down without altering its tempo
- ➢ shift: this method shifts the audio forwards or backwards, chosen with rollover. Thus original audio length and contents are preserved.

AVCI (2025) reviewed SER augmentation methods and above three methods have been found increasing the accuracy from 1 to 5 % over original (without augmented) dataset.

The next method is well known for its simplicity, easy implementation and robustness (JIN et al., 2025). ZHANG et al. (2018) introduced the mixup augmentation method that constructs

[2] https://www.udesc.br/cct/geb/projetos/emossonico
[3] https://zenodo.org/records/5427549

virtual training samples. Considering (x_i, y_i) and (x_j , y_j) are two samples drawn at random from the training data set, and $\lambda \in [0,1]$:

$x = \lambda\ x_i + (1-\lambda)\ x_j$, where x_i and x_j are raw input vectors

$y = \lambda\ y_i + (1-\lambda)\ y_j$, where y_i and y_j are one-hot label encodings

Thus the mixup method augments the dataset by taking the weighted average of two audio waveforms. The λ coefficient is drawn from Beta distribution, which is a continuous probability distribution defined over interval [0, 1], with hyper-parameter α that controls the value of coefficient.

Along with accuracy and F1 score, a statistical measure, the Matthews correlation coefficient (MCC) is presented as a third measure which does consider all positive and negative instances equally. The MCC measure ranges in between -1 and +1. The best MCC measure +1 indicates that the classifier is performing well over positive as well as negative instances. It is considered as an unbiased measure of the model performance (CHICCO, 2017).

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	Total
VERBO	167	167	166	166	167	167	167	1167
emoUERJ	94	--	--	91	92	100	--	377
Emosônico	111	60	78	95	82	96	65	587
(acted)	34	34	37	34	34	35	37	
(real)	77	26	41	61	48	61	28	
Rosa	37	--	14	27	24	22	10	134
combined	409	227	258	379	365	385	242	2265

Source: Author's production

Table 1 – Total audios per emotion in the databases & number of augmented audios per emotion

3. Analysis & Results

As these databases consist of real and acted audio types, the first step we took is to segregate as per data types. Later both real and acted datasets were split into 2 sets: train set (80%) & test set (20%). As both these datasets are imbalanced in emotion class, both augmentation strategies were implemented to balance the train set, creating two different versions of augmented dataset. Table 2 informs how many audios were augmented for each emotion in the train set of real and acted datasets in order to achieve 88 and 238 audios per emotion respectively. From 20% of the test set, a balanced val set is created with 40 audios per every emotion.

In the first augmentation strategy, shift with rollover, time stretch, and pitch shift were implemented with 0.5 probability. For augmentation, audio from the train dataset and one of the methods to implement were randomly selected till desired number of augmented emotions reached.

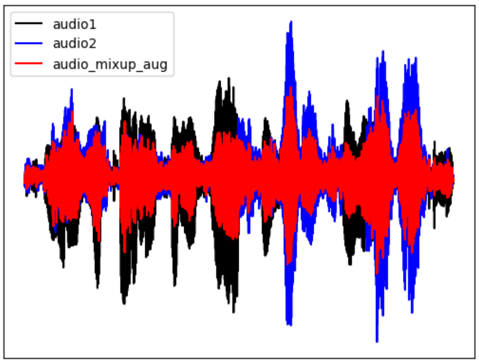
Though mixup strategy works with combining data with different labels, we applied within the same labels only. Per each emotion, two speech audios were randomly selected for augmentation and newly created audio was labeled with that same emotion. To generate coefficient λ, hyper-parameter α was chosen as 0.4. Figure 1 shows an example of mixed-up augmented audio for fear emotion. The train set for both acted and real datasets is balanced but the test sets are highly imbalanced. The distribution among train and test sets is shown in Figure 2.

	anger	**disgust**	**fear**	**happiness**	**neutral**	**sadness**	**surprise**
real	0	60	32	18	18	18	50
acted	0	84	87	4	15	0	82

Source: Author’ creation

Table 2 – Number of augmented audios per emotion in real and acted datasets

For analysis, various spectral, prosodic, temporal features (MFCC, chroma, tonnetz, amplitude envelope, onset strength, and tempogram and many) were obtained using the *Librosa* library.



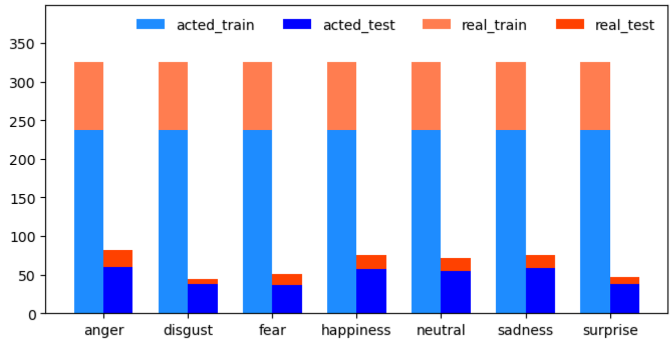


Figure 1 – Augmented audio waveform using mixup method for fear emotion

Figure 2 – Distribution of acted and real audios per emotions in train-test set

Different statistical measures including mean, standard deviation, variance, skewness, kurtosis, inter quartile range were computed over each feature. Next, these features were modeled with different machine learning algorithms using *scikit-learn* library. For SER modeling, features were normalized using the standardization method ($\mu = 0$, $\sigma = 1$). The optimal result was obtained with mean statistics taken over the MFCC feature with 40 coefficients and support vector machines using the radial basis function kernel. Analysis was performed with a combined dataset of real and acted audios, without augmentation, then with augmentation implementing both strategies. Table 3 presents the results of the balanced val set evaluation for all three datasets. 20% test set evaluation for all three datasets showed the similar trend. The dataset augmented with the mixup method has slightly higher measures than the other two datasets. Certainly augmentation has improved the classification.

dataset	accuracy	F1 score	MCC
original dataset	65.71%	65.89%	0.603
dataset with classicalAug	66.79%	66.85%	0.613
dataset with MixUpAug	67.50%	67.51%	0.623

Source: Author' production

Table 3 – Metrics obtained with original and augmented combined dataset

Figure 3 presents the confusion matrix of the evaluated val set of the original unaugmented combined dataset. The highest recognition rate is observed for sadness emotion which has second largest audio samples in the dataset. The second highest recognition rate is for the anger emotion, having the largest samples in the dataset. The lowest recognition rate is obtained for surprise and happiness emotion whereas disgust emotion has the lowest samples.

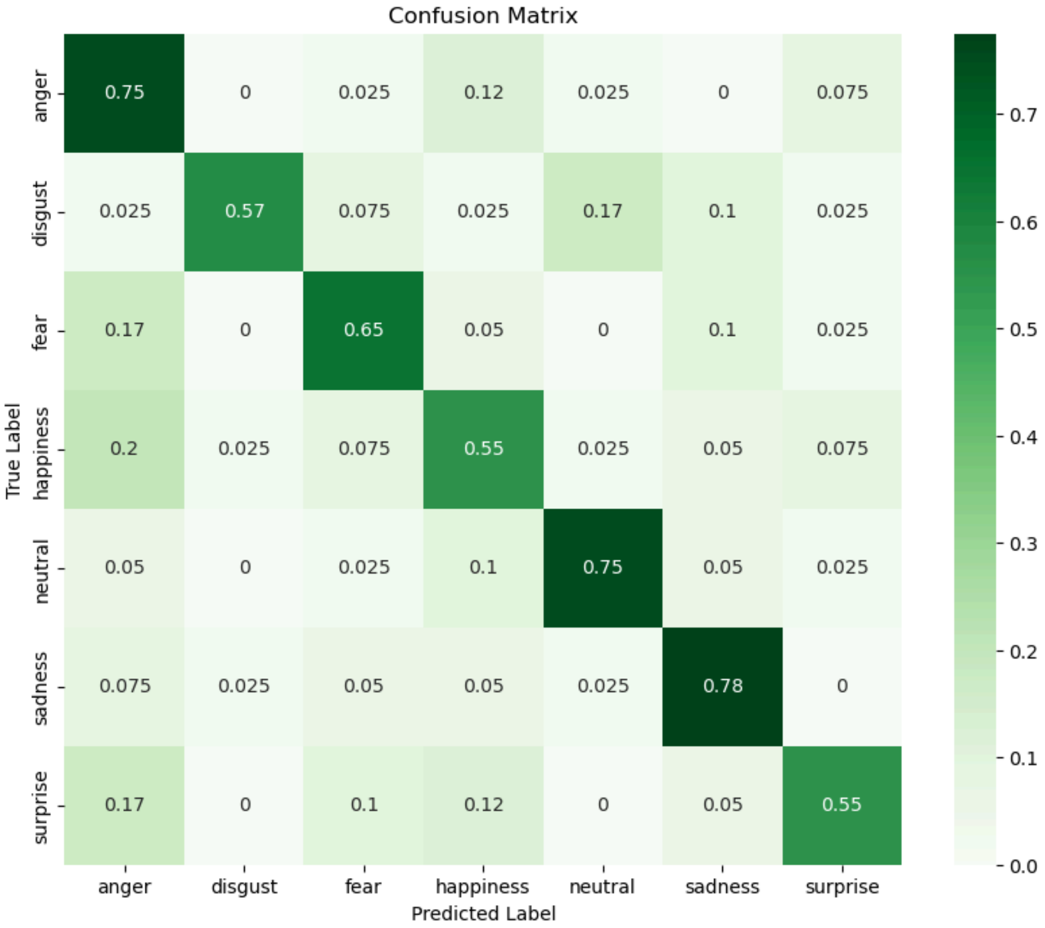


Figure 3 – Confusion matrix for the val set of combined original dataset without augmentation

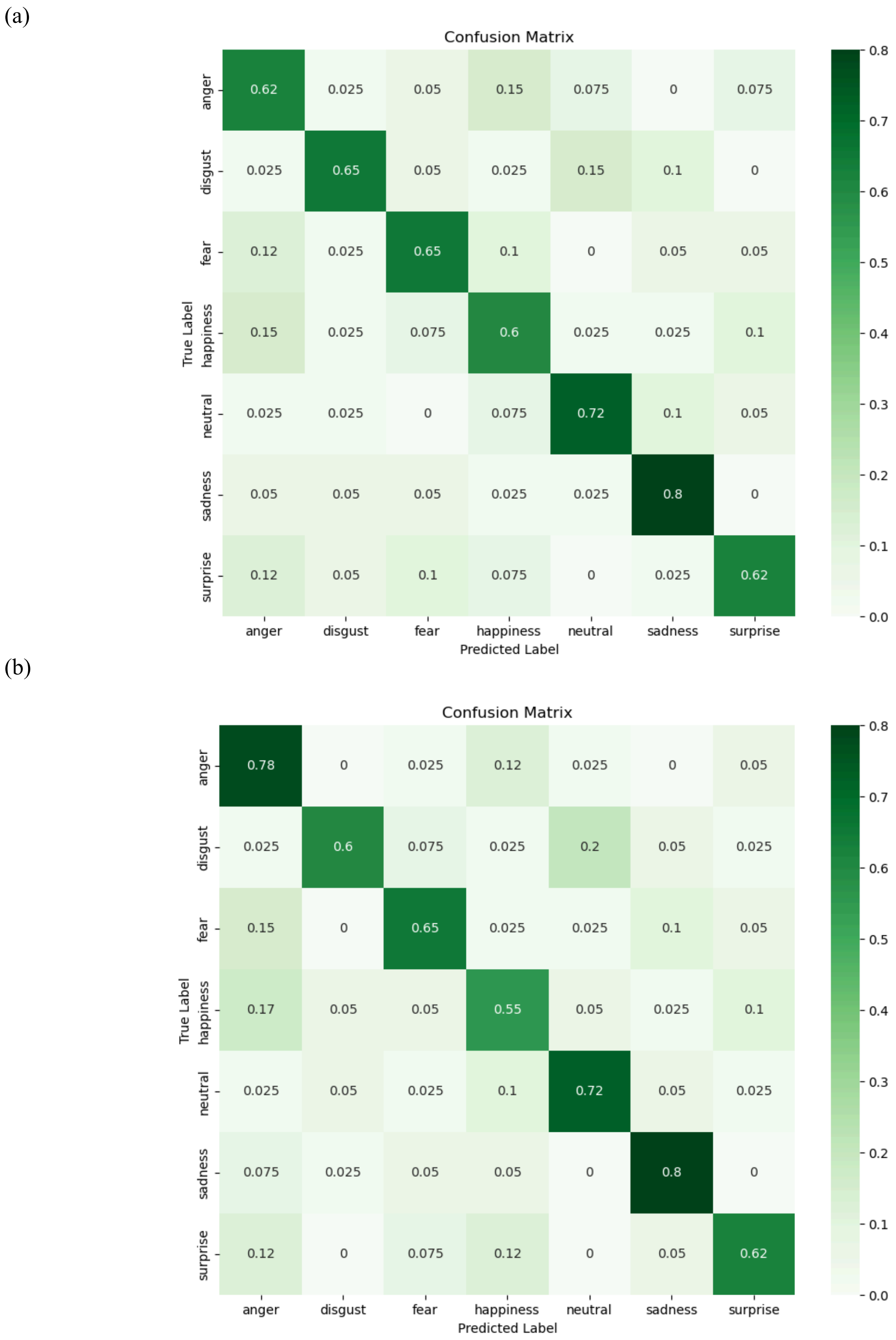


Figure 4 – Confusion matrix for the val set evaluation of combined dataset augmented with (a) Classical methods & (b) with MixUp method

Figure 4 presents the confusion matrix for the val set evaluation of augmented combined dataset. Figure 4 (a) represents a dataset augmented with classical methods and (b) represents a dataset augmented with a mixup method. The highest and the lowest recognition rate is observed for sadness and for happiness emotions, same as observed for the original dataset. Figure 4 (b) matches the next observation with the original dataset, of having the second highest recognition rate for anger emotion which has the largest samples in the datasets. Figure 4 (a) shows less variation among recognition rates for other five emotions except sadness and neutral.

4. Discussion / Conclusion

The objective of this work is to model BP SER dataset. We found four small datasets which are of two types, acted and real. Modeling of real dataset is challenging compared to acted type. Our earlier work (JOSHI et al., 2022b) demonstrated this by working with one real (CORA) and one acted (VERBO) database. In this work we did not consider the CORA dataset as it contains only two emotional categories: neutral and non-neutral, which are not adequate for the analysis. The scarcity of resources still persists though the latest dataset, emoEURJ, is created to mitigate this fact. As emotions per database varies, class imbalance is high.

To reduce the bias of class imbalance so the model can learn to generalize SER better, we implemented well-known simple but robust methods. The metrics presented in Table 3 shows that the augmentation methods do help the model to generalize data a bit more compared to the original dataset. The same is also evident in the Confusion matrices in Figure 3 and Figure 4 (b).

Among two augmented methods, the *augmix* gave a little better results than the classical methods. Comparing recognition rate for all the emotions, the classical method seems to generalize better as the difference between them is less, as shown in Figure 4 (a).

The interesting observation is that though disgust emotion was less representative in the samples, its recognition rate is not the lowest and though anger has more samples, its recognition rate is not the highest. One of the possible reasons can be how features are characterized for real and acted audios. Another may be a possibility of misclassification among the datasets. One of the authors (ROSA, 2017) already cautioned about this in his dataset.

To study human emotions during human-robot interaction, the current model evaluation of 67% seems a good starting step noting scarcity of the data. The future work will look into the possible ways to overcome the limitations of the current study. These include obtaining more SER data in Brazilian Portuguese, instead of using hard-core basic emotions, go for soft or complex emotions, and improve SER modeling.

References

AVCI, Umut *A Comprehensive Analysis of Data Augmentation Methods for Speech Emotion Recognition*. IEEE Access, 2025.

BASTOS, G. R. G. TCHEOU, M. P., da ROCHA, H. F. & PINTO, G. J. S. *emouerj: an emotional speech database in portuguese*. 1.0.0) [Data set], Zenodo, 2021. doi:10.5281/zenodo.5427549.

CHICCO, D. *Ten quick tips for machine learning in computational biology*. BioData mining, vol. 10, pp. 35, 2017.

JOSHI, N. PAIVA, PVV. BATISTA, M. CRUZ, M. & RAMOS, JJG. *Improvements in Brazilian Portuguese Speech Emotion Recognition and its extension to Latin Corpora*. In International Joint Conference on Neural Networks (IJCNN), 2022a. doi:10.1109/IJCNN55064.2022.9892110

JOSHI, N. PAIVA, PVV. BATISTA, M. CRUZ, M. & RAMOS, JJG. *Brazilian Portuguese in speech emotion recognition*. In CTI Renato Archer XII PCI seminar, 2022b. https://www.gov.br/cti/pt-br/publicacoes/producao-cientifica/seminario-pci/xii_seminario_pci-2022/pdf/seminario-2022_paper_21.pdf

KINGESKI, R. *Desenvolvimento de um banco de dados de voz com emoções em idioma português brasileiro*. Master thesis, Center of Technological Sciences, Program in Electrical Engineering, State University of Santa Catarina. Joinville, 2019.

ROSA, J. J. *Reconhecimento automático de emoções através da voz.* Graduation Dissertation, Information Systems, Federal University of Santa Catarina, Florianópolis, 2017.

SCHULLER, B. W. *Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends*. Communications of the ACM, 61(5), 90–99, 2018.

TORRES, N. J. R. MANO, L. Y. & UEYAMA, J. Verbo: voice emotion recognition database in portuguese language. Journal of Computer Science, 14(11):1420–1430, 2018. doi:10.3844/jcssp.2018.1420.1430.

TÓTH, SZ. L. SZTAHÓ, D. & VICSI, K. *Speech Emotion Perception by Human and Machine*. In Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction, pp. 213–224, 2008.

VOGT, T. ANDRÉ, E. & WAGNER, J. *Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation*. Affect and emotion in human-computer interaction. 75–91, 2008.

WANG, Y. SONG, W. TAO, W. LIOTTA, A. YANG, D. LI, X. GAO, S. SUN, Y. GE, W. ZHANG, W. & ZHANG, W. *A systematic review on affective computing: Emotion models, databases, and recent advances*. Information Fusion, 2022.

XIN, J. ZHU, H. LI, S. WANG, Z. LIU, Z. TIAN, J. YU, C. QIN, H. & LI, SZ. *A survey on mixup augmentations and beyond*. arXiv preprint arXiv:2409.05202, 2024.

ZHANG, H. CISSE, M. DAUPHIN, YN. LOPEZ-PAZ, D. *mixup: Beyond empirical risk minimization***.** In International Conference on Learning Representations, vol. 3. 2018.

ZHNAG, S. LIU, R. TAO, X. & ZHAO, X. *Deep Cross-Corpus Speech Emotion Recognition: Recent Advances and Perspectives*. Front. Neurorobot. 15:784514, 2021. doi: 10.3389/fnbot.2021.784514