Unraveling Emotional Dimensions in Brazilian Portuguese Speech through Deep Learning

Henrique Tibério B. V. Augusto[1], Vinícius P. Gonçalves[1], Edna Dias Canedo[1], Rodolfo Meneguette[2], Gustavo Pessin[3], Geraldo Pereira R. Filho[4]

[1] Universidade de Brasília, Brazil
{henrique.brandao,ednacanedo,vpgvinicius}@unb.br
[2] Universidade de São Paulo, Brazil
meneguette@icmc.usp.br
[3] Instituto Tecnológico Vale, Brazil
gustavo.pessin@itv.org
[4] Universidade Estadual do Sudoeste da Bahia
geraldo.rocha@uesb.edu.br

Abstract. Speech is often our first form of communication and expression of emotions. Speech Emotion Recognition is a complex problem, as emotional expression depends on spoken language, dialect, accent, and the cultural background of individuals. The intensity of this emotion can affect our perception and lead us to interpret information inappropriately, with potential applications in various fields such as: patient monitoring, security, commercial systems, and entertainment. This work performed a Machine Learning task using both Machine Learning and Deep Learning to infer the intensity of emotions in Portuguese speech, employing Domain Fusion with two distinct databases. To do so, an Autoencoder was created to extract features, and then we trained a supervised model to classify the intensities into four classes: (i) weak; (ii) moderate; (iii) high; and (iv) peak intensity. The results indicate the possibility of inferring intensity, although the dataset is limited, even when combining two datasets. Two experimental scenarios were carried out, with analogous architectures, varying the dimensionality of representative features used as input for the models. Additionally, observing the performance metrics, it was possible to note the recurrence of the same class (high) with the lowest variation of F1-Score between both experiments, which raises questions for further studies, while the most distant classes (weak and peak) had the best performance for both experiments.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: brazilian portuguese, deep learning, emotion intensity, machine learning, speech emotion recognition

1. INTRODUCTION

Speech is usually our first form of communication and expression of emotions [Filho et al. 2024]. Since childhood, even before we use our language, we express emotions through non-verbal sounds that have meaning for the speaker. There are even studies investigating the emotions of babies through crying. We continue to express emotions in this way throughout life [Gonçalves et al. 2017]. In a modern world, we are increasingly interacting with, and through, technological tools (*e.g.*: virtual assistants).

The analysis of speech emotion has become a prominent area of research, largely thanks to the increase in computational capacity and the efficiency of algorithms. In this way, the classification of emotions and their intensity plays an important role in the development of science and technology, since the message transmitted can have its semantics altered by emotions [Koolagudi and Rao 2012].

Speech Emotion Recognition (SER) is a complex problem, as emotional expression depends on

the spoken language, dialect, accent and cultural background of individuals. The recognition and evaluation of emotions presents difficulties due to its interdisciplinary nature: The evaluation of emotional intensities are the subject of the sciences of Psychology, the measurement and evaluation of physiological data are related to medical sciences, and the analysis and solution of data and sensors are the subject of mechatronics.

Inferring the intensity of emotion finds potential applications in several areas, such as health [Elsayed et al. 2022], security [Nassif et al. 2022], entertainment (through smart environments [Cook and Das 2004] and smart assistants [Purington et al. 2017]). Works such as [Zhu et al. 2019] seek to understand the intensity of emotion to improve the performance of the synthesis of an emotional vocalization originating from a Speech To Text mechanism.

This work aimed to propose, develop and evaluate a solution for classifying the intensity of emotion in Portuguese speech. It used Deep Learning (DL) techniques for the classification solution and an unsupervised Machine Learning (ML) technique to validate the results. Namely, two DL models, the first for dimensionality reduction and extraction of representative features, the second model to perform intensity prediction and, finally, a ML model to evaluate and interpret our results.

This work is organized as follows. In Section 2, we present a discussion and a comparison of related literature; in Section 3 we present the strategy used in this research; Section 4 presents and discuss the results; and, finally, in Section 5, we present the conclusions of this work.

2. RELATED WORK

Researching the state of the art, we find work recalling the need to improve the naturalness of human-computer interactions and the importance of accurately interpreting emotional information given the ubiquity of automated systems, which increased through time and proposing incorporating more features into the input data to improve performance in SER tasks [Bhargava and Polzehl 2013]; and pointing the use of unsupervised models to try to remedy the data shortage for SER tasks, questioning the feasibility of learning features from datasets of other speech domains and using them to train emotion classification models [Eskimez et al. 2018].

When dealing with SER, spectral characteristics appear to be an important feature for the models, carrying a lot of information about the sample. We were able to observe in [Bhargava and Polzehl 2013] that the absence of Mel-frequency Cepstral Coefficients (MFCCs) resulted in a drop of more than 50% in classification performance.

Works like [Zhou et al. 2022] point out the scarce presence of studies relating to the intensity of emotion. Another point to be noted is that although there are works using more than one dataset, this practice still does not seem to be fully widespread among SER publications, since the scopes and natures (simulated, semi-natural or natural) databases are usually quite different. Also, [Goncalves et al. 2024] underscores the complexity of speech emotion recognition and the challenges in recognizing emotions on lower intensities.

Analyzing the information presented in Table I, we may note that this work is close to the others because it uses spectral features and involves a supervised approach. It also has in common the fact that it uses consolidated ML techniques, while also investigating DL architectures, such as Deep Neural Network (DNN) and Autoencoder (AE). However, this work begins to distance itself from the others when it decides to work with audios in Portuguese. Furthermore, the unsupervised approach will also be used, which is not so common in a same SER task.

Table I. Comparison between this and related work.

Work	Supervised Learning	Unsupervised Learning	*Machine Learning*	*Deep Learning*	Portuguese	Emotion	Intensity
[Zhang et al. 2018]	x		x			x	
[Eskimez et al. 2018]	x	x		x		x	
[Li et al. 2019]	x			x		x	
[Campos and Moutinho 2021]	x				x	x	
[Josh 2021]	x		x	x	x	x	
[Olatinwo et al. 2023]	x	x	x	x		x	
[Goncalves et al. 2024]	x			x		x	x
This work	x	x	x	x	x		x

3. UNRAVELING EMOTIONAL DIMENSIONS IN BRAZILIAN PORTUGUESE SPEECH THROUGH DEEP LEARNING

This section presents an architecture for classifying the intensity of emotion in Portuguese speech. In Figure 1 an overview of the architecture is presented. To this end, two Deep Learning models were created, namely: (i) Autoencoder, responsible for dimensionality reduction and extraction of representative characteristics of the data; and a (ii) Deep Neural Network, to predict the intensity class for a given vocalization. Three main steps will be necessary to recognize the intensity of emotions: (A) Acquisition of information; (B) Feature extraction; and (C) Intensity classification.

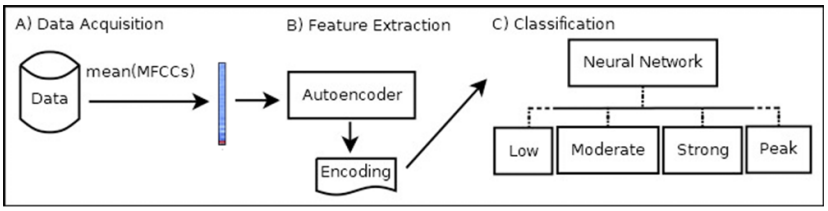

Fig. 1. Proposed architecture for emotional intensity classification.

The first step (A) deals with obtaining the data that will be used in this work and converting it into an representation that can be used by machine learning models. We can describe the data as a set of labeled records that will be used for training and testing the implemented models. The second step (B) deals with extracting features from the converted data. These characteristics will be obtained through an unsupervised model for dimensionality reduction and subsequently used as input for a supervised model for classifying emotion intensity. The third and final step (C) is responsible for inferring intensity. The model receives the features obtained in the previous step and trains a classification model according to four possible classes: (i) Low; (ii) Moderate; (iii) Strong; and (iv) Peak intensity.

3.1 Data Acquisition

To model the intensity of an emotion, one of the difficulties is the lack of labeled data [Zhou et al. 2022]. Traditionally, in areas such as computer vision or speech recognition, datasets can have millions of records, such as: ImageNet (image) with more than 14 million and Google AudioSet (audio) with more than 2 million of samples and average duration $\approx 10s$.

Emotion	VERBO	VIVAE	Total
Achievement/Joy	166	161	327
Fear	166	176	342
Anger	167	174	341
Surprise	167	187	354

Table II. Samples per class per dataset.

Given this scenario, we can utilize a technique that allows us to merge knowledge from multiple datasets organically for a machine learning task. Domain Fusion [Zheng 2015] is a technique for leveraging more databases and producing more robust and useful information than that provided by a single data source individually. An example of using domain fusion can be seen in [Liu et al. 2022] to improve the generalization of the model, which saw an improvement in performance on unseen data (distinct from training and test samples).

One of the Domain Fusion methodologies is called Transfer Learning-Based Data Fusion [Zheng 2015]. And one of the possibilities that this methodology addresses comprises the fusion of databases of similar nature when the task is the same but the domain (starting point) and the counter-domain (arrival point) are different. In this work, we will use two data sets, VERBO and VIVAE, which meet the requirements of being in Portuguese and presenting cataloged emotions (VERBO); and has labels for emotions and intensities (VIVAE).

VERBO [Neto et al. 2018] is a database with 1176 files in *.wav* format, published in 2018, made up of audio files in Brazilian Portuguese, labeled with emotions. It is the first [Josh 2021] dataset for SER in Brazilian Portuguese. The audios last between 2 and 5 seconds, recorded by twelve Brazilian actors - six men and six women - of different ages and regions of the country. It comprises fourteen utterances validated by a linguistic professional, accommodating all phonemes of the Portuguese language. It has examples of the 6 basic emotions proposed for Russel: (1) Joy; (2) Disgust; (3) Fear; (4) Anger; (5) Surprise; (6) Sadness. Finally, a seventh emotional state was added, called (7) Neutral.

VIVAE [N Holz 2021] is a database with 1085 files in *.wav* format, published in 2020, created by German and American researchers, formed by non-verbal vocalizations. The audios were recorded by eleven people, comprising three positive and three negative feelings and an average duration of approximately one second. The positive ones being: (1) Achievement; (2) Sexual Pleasure; and (3) Surprise. And the negative ones: (1) Anger; (2) Fear; and (3) Physical Pain. All were recorded with intensity varying between low, moderate, strong and peak emotion.

Making an intersection between the classes of datasets, we notice only four classes in common: Joy, fear, anger and surprise. Performing a count of the classes in common, in both databases, we will have 1364 samples, detailed on Table II.

3.2 Feature extraction

To perform ML tasks from audio files, it is necessary to convert them to a form that can be ingested by the model. The database files are in *.wav* format (shortened to WAVEform), which does not compress the digital sound, keeping it closer to the expression of natural sound. We need a way to transform data into a representation that preserves its characteristics.

We can understand a signal as the variation of a quantity over time. In our case, the variation in air pressure. Air pressure samples are taken over time, at a certain frequency, then we have a one-dimensional signal that maps the amplitude of the sound distributed over time. To transport a signal from the time domain to the frequency domain, the Fourier Transform is used, which will decompose the signal into its frequency components, carried out computationally through Fast Fourier Transform (FFT).

With the result of the FFT of a sample, we can calculate its spectrogram: A representation of

the density of frequencies over time. However, as humans do not perceive the entire sound spectrum [Purves et al. 2001], a normalization will be applied to the frequencies, and we will calculate the Mel Spectrogram, adjusted by the Mel-Scale, which is a perceptual scale of pitches judged by listeners to be equal in distance from one another, and widely used in SER problems [Zhang et al. 2018][Latif et al. 2017].

Another attribute observed in the literature are the Mel Frequency Cepstral Coefficients (MFCCs). A MFCC is a representation of short-term the power spectrum of a sound. Calculating the MFCCs consists of applying the Discrete Cosine Transform (DCT) to the Mel-Spectrogram. We can understand MFCCs as a compression [Bui et al. 2020] of a Mel-Spectrogram. The relevance of this attribute is also found in works such as [Bhargava and Polzehl 2013], which explored the combination of spectral, tonal and rhythmic characteristics to observe their contribution to the performance of their classification models, and concluded that using only MFCCs generated results almost twice as good as adding more features.

Thus, let the two datasets be VERBO and VIVAE, so that VERBO is made up of {sample, class} pairs and VIVAE is made up of {sample, class, intensity} pairs, where the classes are the emotion label attributed to that sample, and the intensity is the intensity label of that sample. We will represent the VERBO samples by X_{VERBO} and VIVAE by X_{VIVAE}.

Let Y_{VERBO} be the set of classes (emotions) $y_i, \forall x_i \in X_{VERBO}$ and let us define the set of intensities by Z. Since intensities are only present in VIVAE, we will use its four classes (low, moderate, strong, peak). We know that $Y = \{joy, fear, anger, surprise\}$ and $Z = \{low, moderate, strong, peak\}$.

Defining y_i as the class of x_i, redefine $X_{VERBO} = \{x_i \in X \mid \exists y_i \in Y\}$, analogously to X_{VIVAE}. Our domain will be $X = X_{VERBO} \bigcap X_{VIVAE}$, as $\forall x_i \in X, \exists y_i \in Y$ and $\forall x_j \in X_{VIVAE}, \exists z_j \in Z$ such that z_j is the intensity of x_j.

Then build an Autoencoder (AE), a model which tries to reproduce an identity function. An Autoencoder is a neural network that tries to reproduce an input as output. We can describe an AE as a set of functions f, f' such that, given an input x, we want $f'(f(x)) = x' \approx x$, where f performs the encoding of x and f' performs the decoding of the result $f(x)$. By definition, an AE is composed of an encoder function (f_e) and a decoder function (f_d), so that $AE : M \to M'$ does $AE(x) = f_d(f_e(x)) = x' \approx x$. Works such as [Eskimez et al. 2018] which used an Autoencoder together with spectral features to learn features from datasets in order to remedy data scarcity.

The Autoencoder will be trained and validated based on a 75%/25% stratified split of X, respectively. In its latent space ($f_e(x)$) we will have a representation of the input data with reduced dimensionality, preserving its characteristics sufficiently so that it can be reconstructed (x') when applying the decoding function.

3.3 Intensity Classification

Deep Neural networks are capable of learning complex relationships between features in data, which can be especially beneficial for the task of classification. Unlike more traditional methods, neural networks are not limited by linear assumptions or specific pre-defined characteristics, allowing for more flexible and adaptive modeling of data.

With the trained Autoencoder, we'll split X_{VIVAE} into two sets for training ($X_{VIVAE_{training}}$) and testing ($X_{VIVAE_{test}}$), containing 75% and 25% of this data, respectively. This will be the data used by our supervised model. For the sake of this work, we will use the F1-Score as our evaluation metric, since it balances Precision and Recall.

To this end, due to the challenge of demand, ratified by the scarcity of data, we took advantage of Data Fusion to create an unsupervised solution that can extract characteristics that are representative

enough so that, even with reduced dimensionality in comparison to the original input, we can reconstruct it; and subsequently use this compressed data, consisting of sufficiently relevant attributes, to develop a supervised intensity inference model.

Two experimental scenarios were carried out, using 64 and 128 MFCCs, respectively. In both, the AE was trained with X data and then we trained a dense neural network (DNN) to classify the intensity, trained and tested only on the portion of the data resulting from applying f to X_{VIVAE}, since X_{VERBO} does not have labels for intensities (Z).

Having none, we need to investigate whether there is any meaning in the results when we apply the classification to this data, which has not yet been seen by the classifier. Therefore, a way is needed to analyze the records of X_{VIVAE} and X_{VERBO} regarding intensities and the prediction of these intensities, respectively. We can use Principal Component Analysis (PCA) to reduce the dimensionality of records and observe the classification behavior. While the use of PCA finds records in the literature applied to voice and emotions, such as [Ververidis et al. 2004] that despite making a previous feature selection, chose to use PCA to reduce the dimensionality of the data to a two-dimensional visualization of the records; and [You et al. 2006] who developed an emotion recognition system for noisy audio, using 64-dimensional data that would be reduced to a 6-dimensional space.

The Autoencoder training was carried out using the Mean Square Error (MSE) as a Loss function and the Adaptive Moment Estimation (ADAM) as an optimization function, while the classifier training was carried out using Categorical Cross Entropy as a Loss function and ADAM as an optimization function.

4. RESULTS

On both experiments, we applied the classifier to the data from X_{VERBO} and then we performed a PCA with 2 components on the encoding result of X (Figure 2 and 3), where the colors represent the class: Original labels for the $VIVAE$ data and classifier prediction labels for the $VERBO$ data.

Doubling the number of features from the first experiment to the second made the Test Loss of our Autoencoder improve almost eight times, going from $6,40$ (for 64 MFCCs) to $0,84$ (for 128 MFCCs). And based on the results on Table III, we verified that the first experiment obtained superior performance in terms of the selected metric for the intensity classification model, having a *F1-Score* superior to that of the first experiment in three of the four classes. And, for both experiments, the class with the worst classifier result was similar, the Strong class. And the best performing classes are Weak and Peak intensity, respectively.

Intensity	F1-Score for 64 MFCCs	F1-Score for 128 MFCCs
Low	0,68	0,58
Moderate	0,53	0,51
Strong	0,48	0,49
Peak	0,66	0,64

Table III. *F1-Score* for both experiments.

On both figures, we can observe the formation of two large *clusters*, one on the left and one on the right, representing each dataset, respectively, which indicates that the encoding vectors had enough representative characteristics as not to scramble the data. As well as that the data from VIVAE seems to have more variance as the data from VERBO, as its cluster is more spread on both experiments.

We may also notice that there are more predictions for the Moderate class for 64 MFCCs while the predictions for the Low class are predominant for the 124 MFCCs experiment. And for each scenario, both of this classes had the second highest F1-Score.

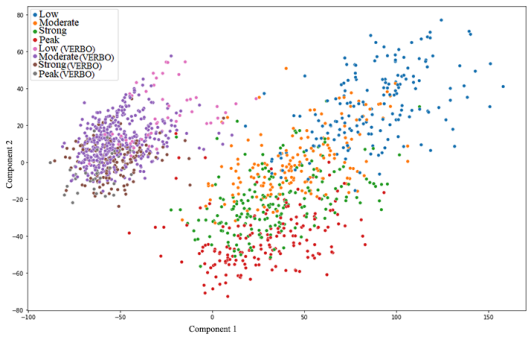

Fig. 2. PCA with 2 components applied to the encoding result of the first experiment.

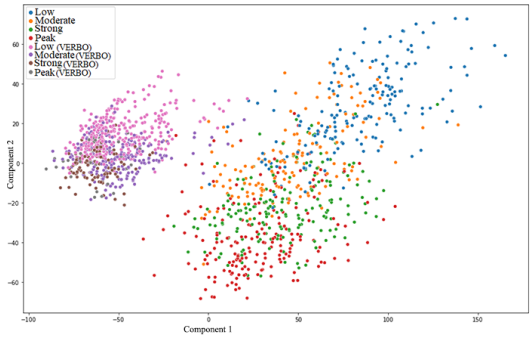

Fig. 3. PCA with 2 components applied to the encoding result of the second experiment.

Observing how the Low and Peak classes occupy the extremes of the clusters, it can make their separation easier compared to the class samples of Moderate and Strong classes, which can be seen more amalgamated in the central region. This distribution seems to be in line with our results, since the two best performances - in both scenarios - were from the Low and Peak classes, and the worst from the Moderate and Strong classes.

Given the correct labels for the VIVAE data, this visualization provides another insight of how we are able to draw a line on $Component2 = 0$ in a way that data for both experiments could be split into two categories: Lighter (Low and Moderate) and Heavier (Strong and Peak). So that, for both experiments, given a data point $x_i = (i_{component_1}, i_{component_2})$, if $i_{component_2} >= 0$ it would belong to class Lighter, while if $i_{component_2} < 0$ it would belong to class Heavier. This naive interpretation raises the observation that Component 2 is almost exclusively responsible for the intensity of a given utterance.

5. CONCLUSIONS

The results indicate that it appears to be possible to infer intensity. However, the dataset is still quite sparse. We are also not aware of a speech dataset in Portuguese that presents both emotions and their intensities. Although Portuguese is a language spoken by the sixth largest population and by the ninth largest economy in the world, when we compare VERBO with data sets such as the AudioSet, we notice the enormous distance both in number of samples ($\approx 2,000,000$) and in average duration ($\approx 10s$). It is expected that a more robust database will improve research performance. Despite improving tremendously when using more features, the performance obtained by the Autoencoder did not mean improving the Classifier, which had better performance when using less features. The nature of the results suggests the continuous need for improvements in the methodologies adopted, as well as the exploration of new approaches and data to enhance the accuracy of predictions. Furthermore, this study emphasizes the importance of considering linguistic and cultural nuances specific to Portuguese when developing models of emotional inference in voice.

For future work, we intend to implement a Recurrent Neural Network (RNN) to evaluate the data along the time axis and try to improve the performance of the classification model. We also intend to carry out exploratory analyzes to understand whether the model is presenting any type of bias (such as the Strong class having the worst performance in both experiments), and if so, understand how to mitigate it.

REFERENCES

Bhargava, M. and Polzehl, T. Improving automatic emotion recognition from speech using rhythm and temporal feature, 2013.

Bui, K.-H. N., Oh, H., and Yi, H. Traffic density classification using sound datasets: An empirical study on traffic flow at asymmetric roads. *IEEE Access* vol. 8, pp. 125671–125679, 2020. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9136653.

Campos, G. A. and Moutinho, L. d. S. Deep: uma arquitetura para reconhecer emoção com base no espectro sonoro da voz de falantes da língua portuguesa, 2021. https://bdm.unb.br/handle/10483/27583.

Cook, D. and Das, S. K. *Smart environments: technology, protocols, and applications*. Vol. 43. John Wiley & Sons, 2004.

Elsayed, N., ElSayed, Z., Asadizanjani, N., Ozer, M., Abdelgawad, A., and Bayoumi, M. Speech emotion recognition using supervised deep recurrent system for mental health monitoring, 2022.

Eskimez, S. E., Duan, Z., and Heinzelman, W. Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 5099–5103, 2018.

Filho, G. P. R., Meneguette, R. I., Mendonça, F. L. L. d., Enamoto, L., Pessin, G., and Gonçalves, V. P. Toward an emotion efficient architecture based on the sound spectrum from the voice of portuguese speakers. *Neural Computing and Applications*, 2024.

Goncalves, L., Salman, A. N., Naini, A. R., Velazquez, L. M., Thebaud, T., Garcia, L. P., Dehak, N., Sisman, B., and Busso, C. Odyssey 2024-speech emotion recognition challenge: Dataset, baseline framework, and results. *Development* 10 (9,290): 4–54, 2024.

Gonçalves, V. P., Giancristofaro, G. T., Filho, G. P., Johnson, T., Carvalho, V., Pessin, G., Neris, V. P. d. A., and Ueyama, J. Assessing users' emotion at interaction time: a multimodal approach with multiple sensors. *Soft Computing* vol. 21, pp. 5309–5323, 2017.

Josh, N. Brazilian portuguese emotional speech corpus analysis. *X Seminário em TI do PCI/CT*, 2021. https://www.gov.br/cti/pt-br/publicacoes/producao-cientifica/seminario-pci/xi_seminario_pci-2021/pdf/seminario-2021_paper_29.pdf.

Koolagudi, S. G. and Rao, K. S. Emotion recognition from speech: a review. *Int J Speech Technol* vol. 15, pp. 99—117, 2012. https://link.springer.com/article/10.1007/s10772-011-9125-1.

Latif, S., Rana, R., Qadir, J., and Epps, J. Variational autoencoders for learning latent representations of speech emotion: A preliminary study, 2017.

Li, Y., Zhao, T., and Kawahara, T. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *Interspeech*. pp. 2803–2807, 2019.

Liu, R., Sisman, B., Schuller, B., Gao, G., and Li, H. Accurate Emotion Strength Assessment for Seen and Unseen Speech Based on Data-Driven Deep Learning. In *Proc. Interspeech 2022*. pp. 5493–5497, 2022.

N Holz, P. L.-M. . D. P. The paradoxical role of emotional intensity in the perception of vocal affect. *Sci Rep* 11 (9663), 2021. https://www.nature.com/articles/s41598-021-88431-0.

Nassif, A. B., Shahin, I., Elnagar, A., Velayudhan, D., Alhudhaif, A., and Polat, K. Emotional speaker identification using a novel capsule nets model. *Expert Systems with Applications* vol. 193, pp. 116469, 2022.

Neto, J. T., Filho, G. P., Mano, L. Y., and Ueyama, J. Verbo: Voice emotion recognition database in portuguese language. *Journal of Computer Science* 14 (11): 1420–1430, Nov, 2018.

Olatinwo, D. D., Abu-Mahfouz, A., Hancke, G., and Myburgh, H. Iot-enabled wban and machine learning for speech emotion recognition in patients. *Sensors* 23 (6), 2023.

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., and Taylor, S. H. " alexa is my new bff" social roles, user satisfaction, and personification of the amazon echo. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. pp. 2853–2859, 2017.

Purves, D., GJ, G. J. A., D, D. F., and et al. *Neuroscience*. Sunderland (MA): Sinauer Associates, 2001. https://www.ncbi.nlm.nih.gov/books/NBK10924.

Ververidis, D., Kotropoulos, C., and Pitas, I. Automatic emotional speech classification. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. pp. I–593, 2004. https://ieeexplore.ieee.org/document/1326055.

You, M., Chen, C., Bu, J., Liu, J., and Tao, J. Emotion recognition from noisy speech. In *2006 IEEE International Conference on Multimedia and Expo*. pp. 1653–1656, 2006. https://ieeexplore.ieee.org/document/4036934.

Zhang, S., Zhang, S., Huang, T., and Gao, W. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia* 20 (6): 1576–1590, 2018.

Zheng, Y. Methodologies for cross-domain data fusion: An overview. *IEEE Transactions on Big Data* 1 (1): 16–34, 2015.

Zhou, K., Sisman, B., Rana, R., Schuller, B. W., and Li, H. Emotion intensity and its control for emotional voice conversion. *IEEE Transactions on Affective Computing*, 2022.

Zhu, X., Yang, S., Yang, G., and Xie, L. Controlling emotion strength with relative attribute for end-to-end speech synthesis. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. pp. 192–199, 2019.