Brazilian Portuguese emotional speech corpus analysis

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

Abstract

Speech emotion recognition analysis identifies correct emotions from speech and plays an important role for better human-machine interaction. Compared to other languages, Brazilian Portuguese database is relatively new. This work aims to explore different features of the database to obtain improved recognition rate. A set of handful spectral, temporal, and prosodic features are found to increase recognition rate, and results are compared with previous works. At least 6% higher accuracy is obtained with support vector machines with radial basis function kernel. Different measures are reported for better understanding of the classifier's performance.

Keywords: Speech Emotion Recognition, Brazilian Portuguese Corpus, Multiclass Classification.

1. Introduction

Speech emotion recognition (SER) analysis identifies correct emotions from speech using machine learning algorithms. SER analysis is helpful in improving human-machine interaction and its applications are found in different fields, like, smart health AI systems can monitor mental and physical condition using SER to provide better medical treatment. SER analysis is challenging as conveying emotion changes with languages, cultures, gender, and speaker (WIERZBICKA, 1999; AKÇAY & OGUZ, 2020). Further challenges in SER analysis include availability of speech emotion databases with desired number of emotions with proper emotional annotations to utterances. A further endeavor is to model the speech signals to extract robust and effective features as analysis is highly feature dependent (ANAGNOSTOPOULOS et al., 2015, AKÇAY & OGUZ, 2020).

SER studies are advanced in various languages but for Brazilian Portuguese (BP), the research is handful. The very first BP SER database, the voice emotion recognition database (VERBO, NETO et al., 2018) is explored in few studies. This work aims to explore the VERBO database, and come with appropriate features to increase the recognition rate of all emotions.

2. Overview of SER system

SER databases are of three types, Acted, elicited or real. Acted database is created with the help of skilled actors who can simulate desired emotions in predefined text. Artificially inducing desired emotions in the recording environment make elicited database different than acted. Real database consists of natural conversations like real life shows or interviews.

Emotions can be studied as a discrete model consisting of six basic emotions -- anger, disgust, fear, happiness, sadness and surprise, and all other emotions can be derived from these basic emotions (EKMAN, 1971). Another emotional model is dimensional which describes inter-

relation of emotions in two or three dimensional planes (RUSSELL, 1980). Following a discrete emotional model, SER can be framed as a supervised learning multiclass-classification problem which can be divided into two main tasks. Extraction of relevant and robust features is the first one and modeling these features to identify the correct emotion, i.e., the classification is the latter one. Each emotion can be identified by its own characteristics like energy, loudness, etc. For example, speech signals with anger emotion may show high energy in the starting whereas with surprise may show higher energy around the end of the speech. These characteristics voiced or unvoiced features play an important role in SER analysis influencing the performance of the classifier.

Spectral, prosodic, temporal are widely reported feature types known as acoustic or local features. Prosodic features (eg. stress, intonation, rhythm) reveal how speech signals are perceptible to humans. Correlation between prosodic features and different emotions is found to be helpful in SER analysis. Temporal features (eg. 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 (eg. spectrogram, various cepstral features) are obtained through transforming signal from time domain to frequency domain. 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 (KOOLAGUDI, 2012). Different statistical measures computed over local features are known as global features. Extracted features should be normalized. For SER, standardization method is found to be more effective (BOCK, 2017).

Extracted features are then modeled with various machine learning algorithms. Some of the widely reported classifiers are support vector machines (SVM), hidden Markov model, Gaussian mixture model, ensemble methods, and different types of neural networks including deep learning.

2.1 Related work

Various studies contributed to identify efficient audio features and various surveys have listed those together. Not limited to feature selection, studies have also reviewed the SER databases, preprocessing methods, data augmentation, classifiers, and the challenges to overcome (VERVERIDIS & KOTROPOULOS , 2003; AKÇAY & OGUZ, 2020; LALITHA et al., 2020; SHAH et al., 2021). One of the widely used feature set is the extended Geneva minimalistic acoustic parameter set (eGeMAPS, EYBEN et al., 2016). This set includes various spectral, prosodic, temporal, and amplitude related features with different statistical measures.

Latif et al. (2018a, 2018b) exercised eGeMAPS to analyze real database in Urdu with SVM and deep belief networks (DBN) as the classifiers. Authors reported five-fold cross-validation (cv) unweighted average recall rate with SVM and 25% validation set accuracy with DBN. Nandan and Vepa (2020) analyzed acted database in Canadian French (CaFe) extracting eGeMAPS features. Zvarevashe and Olugbara (2020) extracted widely used spectral and prosodic features and opted brute force method to select best features. Authors augmented the training dataset by adding Gaussian noise. North American acted databases were analyzed with an ensemble algorithm, Random forest, and reported 10-fold cv accuracy as a metric. Neto (2020) used wavelets filter bank, MFCC, frequency spectrum, energy, loudness, jitter, shimmer, and pitch to analyze Brazilian Portuguese database VERBO as a part of AI health framework. Authors reported 10-fold cv mean accuracy with k-nearest neighbors (KNN) as 76.49% and with support vector machine (SVM) as 75.38% (NETO, 2020). Da Silva et al.

(2019) reported 78.64% accuracy with VERBO corpus which is used for their UXmood sentiment analysis tool. Campos and Moutinho (2020) analyzed VERBO corpus to train their deep learning architecture. With 70% training and 30% validation set, the maximum accuracy is reported for surprise emotion as 85.56% and the least for anger as 63.74%.

Luengo et al. (2010) reported spectral features alone to be sufficient to increase SER accuracy. Tahon and Laurence (2016) found a minimal set of robust acoustic features consisting of cepstral features. Ilive et al. (2020) used cepstral feature, MFCC, to study CaFe. Ancilin and Milton (2021) suggested a new cepstral feature, Mel frequency magnitude coefficients (MFMC), by modifying Mel frequency cepstral coefficients (MFCC) extraction procedure to overcome its drawbacks, wiz. sensitivity to noise factors and vanishing of non dominant phonemes, and shown that only the MFMC feature is sufficient to extract important information for SER analysis.

3. Proposed approach

3.1 Feature selection

Starting with widely reported various spectral, prosodic, temporal, and fractal features and computing ten different statistical measures, a feature vector of length 2141 is obtained. Features considered are MFCC with their derivatives, chroma, root mean square energy (RMSE), spectral contrast, spectral rolloff, spectral flatness, spectral bandwidth, spectral centroid, tonal centroid, amplitude envelope, onset strength, first five harmonics, pitch, formants, harmonic to noise ratio, intensity, power spectral density, spectral entropy and multifractal measures. Statistical measures taken over these features are five quartile, range, mean, standard deviation, skewness and kurtosis. To find effective features, the recursive feature elimination (RFE) method is used. RFE recursively trains the data by discarding less contributing features to find more weighing features. Selected feature set is reduced to 38 long vector consisting of following features with mean statistics.

- Amplitude envelope: captures the changes in amplitude of a signal over time giving idea of onsets, where non-silent part appears in speech signal;
- CQT chromagram (12 chroma): captures harmonic characteristics of speech by classifying pitches on equal tempered scale;
- Intensity: provides information about vocal tract air pressure, and thus of acoustic energy which is useful to differentiate vowels and can perceive intonation.
- MFCC (20 coefficients) : is obtained by computing a discrete cosine transform applied on the logarithm of power spectrum, which is mapped to the Mel scale. On the Mel scale, frequencies are arranged in a way that humans can perceive distance between pitches. MFCC describes the intensity of each Mel band. It is widely used cepstral feature;
- Pitch: provides perceptual information of fundamental frequency of vocal tract vibration which varies as per the speech signal. Another information extracted is the strength of unvoiced sounds of a speaker;
- RMSE: is obtained by computing root mean square of the speech waveform which provides information about loudness;
- Spectral rolloff: returns the frequency below which certain amount of the total spectral energy is contained. With proper cutoff frequency, harmonics can be separated from noise. For current analysis, cutoff frequency is considered as 85%.

3.2 Speech Emotion Recognition

3.2.1 VERBO Database

The voice emotion recognition database (VERBO, NETO 2018) is the first speech emotion corpus in Brazilian Portuguese language. This acted database is recorded with 12 professional Brazilian actors (6 females and 6 males) and contains 5 long sentences, 2 short sentences, 2 questions, and 5 nonsense phrases, summing to 14 phrases, in a way to include all the Portuguese linguistic phonemes. It follows a discrete emotional model containing six basic emotions: anger (167), disgust (167), fear (166), happy (166), sad (167), surprise (167) and seventh one is neutral (167) comprising total 1167 utterances.

3.2.2 Methodology

VERBO has mix types of audios with different sampling rates. Of the 1167 audios, 1062 are mono and the remaining 105 are stereo. Moreover, 771 audios have sampling rate of 44100 Hz, 300 audios with 48000 Hz sampling rate, and remaining 96 audios have sampling rate of 16000 Hz. This information was not mentioned anywhere but influences the result, hence is reported here for the first time. As a part of preprocessing, all speech signals are converted to mono and resampled to 22050 Hz. The listed features are extracted and normalized with standardization method. Classifiers support vector machines (SVM), multilayer perceptrons (MLP), k-nearest neighbors (KNN), and random forest (RF) are used to model the features.

To overcome bias with train-test split and over fitting, 10–fold CV is used. Stratified approach is implemented while splitting the data to confirm each emotion has been represented adequately in each fold. Mean accuracy, weighted precision, weighted recall rate, weighted F1 score, and Matthew's correlation coefficient (MCC) are reported to evaluate the model's performance. Recall and precision focus on classified positive instances and F1 score too as it is harmonic ratio of precision and recall. Statistical measure MCC computation takes into consideration all positive and negative instances. Chicco (2017) suggests using MCC to evaluate performance of any machine learning model over accuracy, recall, precision, and F1 score measures.

For analyses following python libraries are used: Librosa (0.8.0), Praat-Parselmouth (0.4.0), NumPy (1.19.2), SciPy (1.5.2), Scikit-learn (0.23.2), Pandas (1.3.0), and MatPlotLib (3.3.2). Scikit-learn metrics have been used to evaluate model's performance.

4 . Results

Extracted 38 features are analyzed with five classifiers among which the best recognition rate is obtained with SVM with radial basis function (RBF) kernel, considering one-versus-one approach. Fig 1 presents the comparison of weighted F1 score for all five classifiers and Table 1 summarizes all mentioned measures. To know the recognition rate for each emotion, the confusion matrix is plotted in Figure 2 and weighted recall rate for each emotion is mentioned. Sad emotion has the highest recognition rate of 93% and disgust emotion has the least recognition rate with 79%. Compared to the previous results, 10-fold CV mean accuracy reported by Neto (2020) is 76.49% with KNN and Da Silva et al. (2019) reported accuracy as 78.64%. This work achieves at least 6% higher 10-fold CV mean accuracy as 84.8%. Campos and Moutinho (2020) reported maximum and minimum accuracy of 85% and 63% with 30% validation set using deep learning algorithm. The current analysis achieves the highest and least recognition rate of 93% to 79%.

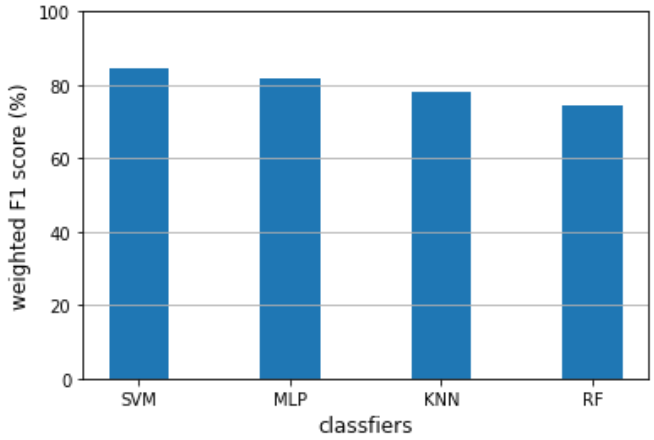

Figure 1 – 10-fold CVWeighted F1 score for different classifiers

Classifiers	Mean Accuracy (%)	Weighted Precision (%)	Weighted Recall (%)	Weighted F1 score (%)	MCC
SVM	84.8	84.9	84.7	84.7	0.82
MLP	81.9	81.9	81.9	81.8	0.79
KNN	78.2	78.6	78.2	78.0	0.75
RF	74.5	75.0	74.6	74.6	0.70

Source: created by author

Table 1 – VERBO SER analysis - classifiers with their respective 10-fold cv measures

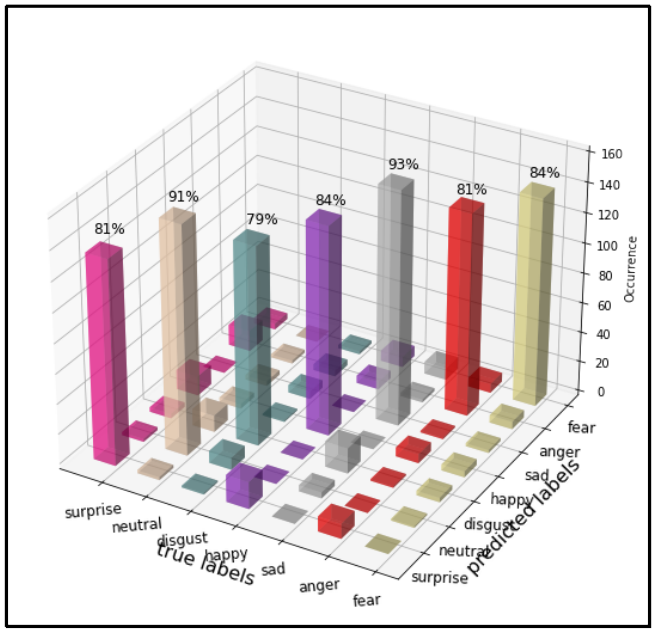

Figure 2 – Confusion matrix for SVM classifier

5. Conclusion

With the aim to explore the BP VERBO database consisting of seven emotions, anger, disgust, happy, neutral, sad, and surprise; and to achieve higher recognition rate, four classifiers, namely, SVM, MLP, KNN, and RF are used for SER analysis. Three spectral (MFCC, CQT chroma and spectral rolloff), two temporal (RMSE and amplitude envelope), and two prosodic (intensity and pitch) are found to be the optimal features. This work has shown promising results with improved recognition rate of minimum 6% using SVM classifier. Highest recognition rate of 93% is obtained for sad emotion. MCC with 0.82 indicates that the SVM model's result is reliable.

References

ANAGNOSTOPOULOS, C. & ILIOU, T. & GIANNOUKOS, I. *Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011*. Artificial Intelligence Review. Vol. 43, p. 155–177, 2015.

AKÇAY, M. & OGUZ, K. *Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers.* Speech Communication. Vol. 116, p. 56–76, 2020.

ANCILIN, J & MILTON, A. *Improved speech emotion recognition with Mel frequency magnitude coefficient*. Applied Acoustics. Vol. 179, p. 108046, 2021.

BOCK, R. & EGOROW, O. & SIEGRET, I. & WENDEMUTH, A. *Comparative study on normalisation in emotion recognition from speech*. Intelligent Human Computer Interaction. pp. 189–201, 2017.

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

CAMPOS, G.A. & MOUTINHO, L.da S. *DEEP: Uma arquitetura para reconhecer emoção com base no espectro sonoro da voz de falantes da língua portuguesa*. Thesis, Universidade de Brasília, 2020.

DA SILVA, F. R.Y. & SANTOS, DO A.D. L. R.& MONTE, P.R.D. & RESQUE, DOS S.C.G. & SERIQUE, M. B. *UXmood—A Sentiment Analysis and Information Visualization Tool to Support the Evaluation of Usability and User Experience*. Information. Vol. 10, n. 12, p. 366, 2019.

EKMAN, P. & FRIESEN, W.V. *Constants across cultures in the face and emotion*. Journal of Personality and Social Psychology. Vol. 17, n. 2, p. 124-129, 1971.

EYBEN, F. & SCHERER, K. & SCHULLER, B. & SUNDBERG, J. & ANDRE, E. & BUSSO, C. & DEVILLERS, L. & EPPS, J. & LAUKKA, P. & NARAYANAN, S. & TRUONG, K. *The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing.* IEEE Transactions on Affective Computing. Vol. 7, n. 2, p. 190–202, 2016.

KOOLAGUDI, S. & RAO, K. *Emotion recognition from speech: a review*. Int. J. Speech Technology, Vol. 15, n. 2, p. 99-117, 2012.

LALITHA, S. & GUPTA, D. & ZAKARIAH, M. & ALOTAIBI, Y.A. *Investigation of multilingual and mixed-lingual emotion recognition using enhanced cues with data augmentation.* Applied Acoustics. Vol. 170, p. 107519, 2020.

LATIF, S. & QAYYUM, A. & USMAN, M. & QADIR, J. *Cross lingual speech emotion recognition: Urdu vs. western languages*. International Conference on Frontiers of Information Technology. p. 88-93, 2018a.

LATIF, S. & RANA, R. & YOUNIS, S. & QADIR, J. & EPPS, J. *Transfer learning for improving speech emotion classification accuracy.* Proc. Inter speech, 2018b.

LUENGO, I. & NAVAS, E. & HERNÁEZ, I. *Feature Analysis and Evaluation for Automatic Emotion Identification in Speech*. IEEE Transactions on Multimedia. Vol. 12, n. 6, p. 490-501, 2010.

NANDAN, A. & VEPA, J. *Language agnostic speech embeddings for emotion classification.* Self-supervision in Audio and Speech, the 37 th International Conference on Machine Learning. Vienna, Austria, 2020.

NETO, J.R.T. *Descarga adaptativa em ambiente com névoa heterogênea: estudo de caso para a área da saúde.* Ph.D. thesis, University of São Paulo, São Carlos, 2020.

NETO, J. & FILHO, G. & MANO, L. & UEYAMA, J. *Verbo: voice emotion recognition database in Portuguese language*. Journal of Computational Science. Vol. 14, n. 11, p. 1420-1430, 2018.

RUSSELL J.A. *A circumplex model of affect.* Journal of Personality and Social Psychology. Vol.39, n.6, p.1161-1178, 1980.

SHAH FAHAD, M. & RAJAN, A. & YADAV, J. & DEEPAK, A. *A survey of speech emotion recognition in natural environment*. Digital Signal Processing. Vol. 110, p. 102951, 2021.

TAHON, M. & DEVILLERS, L. *Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges*. IEEE/ACM Transactions on Audio, Speech, and Language Processing. Vol. 24, n.1, p.16-28, 2016.

VERVERIDIS, D. & KOTROPOULOS, K. *A state of the art review on emotional speech databases.* Proc. of 1st Richmedia Conf. 2003.

WIERZBICKA, A. *Emotions across languages and cultures: Diversity and universals.* Cambridge University Press, New York, US, 1999.

ZVAREVASHE, K. & OLUGBARA, O. *Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition*. Algorithms. Vol. 13, n. 3, 2020.