Improvements in Brazilian Portuguese Speech Emotion Recognition and its extension to Latin Corpora

Neelakshi Joshi
Cyber physical systems division
CTI Renato Archer
Campinas, Brazil
njoshi@cti.gov.br

Pedro V. V. Paiva
Cyber physical systems division
CTI Renato Archer
Campinas, Brazil
pvpaiva@cti.gov.br

Murillo Batista
Cyber physical systems division
CTI Renato Archer
Campinas, Brazil
mbatista@cti.gov.br

Marcos V. Cruz
Cyber physical systems division
CTI Renato Archer
Campinas, Brazil
mvcruz@cti.gov.br

Josué J. G. Ramos
Cyber physical systems division
CTI Renato Archer
Campinas, Brazil
josue.ramos@cti.gov.br

***Abstract*— Speech emotion recognition (SER) is challenging, language dependent, and, as with any supervised learning task constrained by data availability. An important aspect in SER modeling is the selection of optimal features to have an unbiased approach towards recognizing emotions while retaining accuracy. Brazilian Portuguese (BP) is a dialect of the 6th most spoken language in the world yet BP SER studies are scarce. This work aims to explore the solely available BP SER database, providing better features to increase the recognition rate for all emotions, and then proposing a simple but robust multi-corporal SER model. For all corpora analyzed in this work, an improvement of up to 9% in mean accuracy is achieved and the obtained recognition rate indicates that the proposed feature sets show comparatively less biased behavior towards all emotions. Also, we prescribe a combination of different metrics to be used with two cepstral features to obtain a higher recognition rate. Our proposed state of the art Latin multi-corporal model contains few features, and achieves outperforming results with a classical machine learning classifier, compared to previously exercised complex features, algorithms and architectures, by yielding the best recognition rate.**

Keywords— speech emotion recognition, Brazilian Portuguese, feature extraction, multi-corpora

I. INTRODUCTION

The imprint of emotions on human beings' cognitive actions, decisions, reasoning, and learning processes is indisputable. Hence it is a multidisciplinary study spanning cognitive, physiological, and computer sciences [1,2]. Speech emotion recognition (SER) is a subset of affective computing seeking to develop intelligent systems with human-like capabilities in recognizing emotions by processing speech, followed by interpreting and simulating emotions to respond accordingly. Identifying a user's emotions can be useful in various real-time commercial applications, like in e-tutoring where an e-tutor can alter or modify its teaching methods. By learning from the user's emotions, digital pets can become sociable companions by responding to the user's emotions. AI-based health systems can monitor the mental and physical condition of a user to provide better medical treatment. Other applications can be in audio surveillance, call centers, banking, entertainment, computer gaming, and many more [1,2,3].

SER is challenging since emotions can be conveyed differently across speakers, genders, languages, and cultures. The same language spoken across countries sounds geographically unique due to variations in speech rate, dialect, linguistic and cultural slants [4,5,6,7,8]. Another challenge is collecting speech databases and annotating the utterances with the correct emotions. To learn emotions from speech signals, various features need to be extracted. The SER recognition rate depends on how features unravel the characteristics of the emotions, noting that these are highly language dependent [6,9,17] and hence, identifying the best acoustic features is still an open question.

Among different audio features, spectral, prosodic, and temporal features are widely used. Spectral features are obtained by transforming signals from the time domain to the frequency domain. Modeling spectral features on a logarithmic scale assists in speech perception as it approximates human hearing sensitivity. The time-frequency transform of a log-spectrum is known as a cepstrum that can provide a robust but easy way of obtaining fundamental frequency from the speech signal through the periodicity of harmonic structures, and can provide information on formants by capturing slow variations of the log-spectrum [39]. The peaks in the harmonic structure are known as formants, which help to identify vowels, and hence can segregate different emotions. The macro-level structure of harmonics provides information about the vocal tract, which carries crucial information and can be obtained only through spectral features [3,6,7]. Prosodic features are perceptible to humans like stress, intonation, and rhythm. Correlation between prosodic features and different emotions is found to be useful in SER analysis. Temporal features inform

changes in the signal over time. Timbre can differentiate sound quality, thus informing the physical characteristics.

Multifractal measures are also used as one of the SER features, as the sound production process is a complex, nonlinear, and multifractal phenomenon [34,35]. They provide information on the dynamics of glottal activity responsible for speech production [37,38], and thus for language independent analysis multifractal measures have been used to identify emotions [36].

Following feature extraction, a common approach is to compute various statistical measures for dimension reduction that improve classification accuracy with less computational time [7,8]. For SER modeling, feature normalization using the standardization method ($\mu = 0$, $\sigma = 1$) is found to be more effective [33]. Finally, machine learning models are trained on these features to recognize emotions and this presents the next challenge in choosing appropriate classifier methods with proper train-test splits for SER analysis [6,7,8,9].

Portuguese is ranked as the 6th most spoken language in the world and its dialect, Brazilian Portuguese (BP), is spoken by more than 211 million people, which accounts for around 81% of the population [31], yet BP SER studies are scarce compared to the progress achieved with other languages [1,8]. This can be attributed to a limited available database and/or to a selective study. Only one acted BP SER database [10] is available, of which, among three reported works, two constitute a small part of a larger AI framework [11,12], thus lacking comprehensive study. The third work presented deep learning architecture for BP SER [13]. None of the aforementioned studies discuss the caveats in preparing the data for analysis, which are important for uniformity in the analysis, though they appear trivial.

In the current study, we perform comprehensive BP SER analysis addressing the cautions while preprocessing, and come up with feature sets to improve previously reported results. SER features are language specific, and BP being one of the Latin languages, we validate our SER approach by analyzing the other three Latin SER databases. The main contributions of this work are as follows:

- Using a simple approach and a few but effective features BP SER results outperform the previous results;
- Proposed SER approach is found to be robust with multi-corporal SER analysis showing improvement over the previous results;
- Based on the multi-corporal analysis, we demonstrate the combination of different metrics to be used with cepstral features - Mel frequency cepstral coefficients (MFCC) and Mel frequency magnitude coefficients (MFMC) to obtain higher recognition rate.

The paper is organized as follows. Section 2 briefs the previous SER work related to Latin languages analyzed here. Section 3 describes our analysis, introducing the sets of features we explore, the SER databases under study, and the implemented methods. Section 4 presents results and work is concluded in section 5.

II. Related Work

In this section, we present the SER former results for four Latin languages - BP, Italian, Spanish, and Canadian French. As a part of an AI health framework, Neto [12] analyzed the BP corpus, VERBO, by extracting spectral and prosodic features such as MFCC, spectral frequency, energy, loudness, jitter, shimmer, and pitch, and reported 10-fold cross-validation (CV) mean accuracy with k-nearest neighbors (KNN) as 76.49% and with support vector machines (SVM) as 75.38%. Da Silva et al. [11] analyzed VERBO to validate their UXmood sentiment analysis tool and reported classifying accuracy as 78.64%. Campos and Moutinho [13] analyzed VERBO by extracting MFCC (13 coefficients), chroma, fundamental frequency, loudness, jitter and shimmer features using openSmile [30] software to train their deep learning architecture. With 70% training and 30% validation set, a higher recognition rate of 85.56% is achieved for surprise emotion and a lower recognition rate of 63.74% for anger emotion, with 76.69% overall average recognition rate.

Ancilin and Milton [14] suggested a new feature, Mel frequency magnitude coefficients (MFMC), and compared it with the other three cepstral features for SER analysis. Along with 12 statistical measures and SVM (linear kernel) classifier, 10-fold CV mean accuracy is reported as 73.30% for an Italian corpus, EMOVO. Sönmez, and Varol [15] proposed a new light framework using one-dimensional local binary and ternary patterns. Using neighborhood component analysis, distinct features were selected and modeled with a cubic polynomial kernel SVM classifier, and reported 10-fold CV mean accuracy of 74.31% for EMOVO. Latif et al. [16] used eGeMAPS and deep belief networks (DBN) to analyze EMOVO and reported 76.22% accuracy with a 25% validation set.

Kerkeni et al. [17] analyzed a Spanish database, INTER1SP considering seven emotions. First, audio signals were decomposed with empirical mode decomposition and used the Teager-Kaiser energy operator, and then different cepstral features were extracted for the analysis. The recursive feature elimination (RFE) method was used to select an effective subset of features. Using recurrent neural networks (RNN), with a 30% validation set, 91.16% accuracy is reported, and using SVM (polynomial kernel), 10-fold CV mean accuracy of 90.36% is reported. Kerkeni et al. [32] analyzed INTER1SP extracting MFCC and modulation spectral features and then choosing a subset of features using RFE. With SVM (polynomial kernel), 10-fold CV recognition rate of 90.05% and with RNN 94.01% is reported.

Ilive et al. [18] performed multi-corporal SER analysis using Canadian French (CaFE), Italian (EMOVO), and North American English (RAVDESS) databases considering only four emotions viz. happy, anger, sad and neutral. Two types of data augmentation methods, pitch tuning and the addition of white noise, have been implemented. Using the MFCC feature and a 1-d convolutional neural network (CNN) for a 25% test set, 71.10% accuracy is reported for CaFE and 79.07% for EMOVO.

Ilive et al. [18] reported that MFCC is an effective feature for SER analysis irrespective of different languages,

cultures, and genders. A similar observation but with the MFMC feature is reported by Ancilin and Milton [14], mentioning it as a robust and sufficient feature. By comparing SVM results with other deep learning methods, Sönmez and Varol [15] demonstrated that by extracting robust and distinct features, a classical machine learning algorithm like SVM can outperform deep learning algorithms.

Two pragmatic aspects learnt from this overview are: (i) cepstral features are prominent features; and (ii) a crucial part in achieving a higher recognition rate is the selection of robust but distinct features. As features are language dependent, it is a substantial task to find the key features for a specific language. In the next section, we describe our approach based on this principle, using cepstral features plus very few additional distinct features.

III. Methodology

This section describes the feature selection process. Introducing the flow of the analysis, we detail the features selected. The SER implementation from preprocessing to modeling the speech signals, and databases explored are described in the following subsections.

A. SER Approach

With the classical SER approach, a feature selection step is included in the audio modeling part. Feature selection methods can be used to reduce feature dimension, and hence computational costs. It is also found to be useful in improving the performance of the model [3,17,32]. In audio modeling, the speech signals are preprocessed to extract various acoustic features, then significant features are chosen using a feature selection method. Selected features are standardized by scaling to unit variance and removing the mean. Finally, for each emotion, the recognition rate is obtained using machine learning algorithms as a multiclass classification task.

B. Audio Features

Starting with widely reported features, various spectral, prosodic, temporal, and multifractal features are extracted, and different statistical measures are computed. The first feature in the list is the cepstral feature, MFCC, which is widely exercised to identify emotions as well as the speaker, and is known as the sole sufficiently robust feature. It provides insight into audio characteristics in the frequency domain over the human auditory range. It is obtained by computing a discrete cosine transform applied to the logarithm of the 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. The first (delta) and second (delta-delta) order derivatives of MFCC are also computed. Derivatives (denoting as Δs) provide information on how the coefficients change over time and reveal dynamic characteristics of static features. As windowing is essential in cepstral feature calculation, delta is defined as

$$\Delta_t = \frac{\sum_{n=1}^{N} n(f_{t+n} - f_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (1)$$

where Δ_t is computed for frame t containing static coefficients f_{t+n} to f_{t-n}. Second derivatives can be obtained in a similar way by considering the coefficients of the first derivative. The next feature is the Constant Q transform (CQT) chroma which uses logarithmic scale to render harmonic frequency components from speech signals. Chroma apprehends harmonic and melodic characteristics by classifying pitches on an equal-tempered scale. The CQT chroma captures the harmonic characteristics of speech.

Other extracted features are as follows: intensity and rms energy (RMSE) provide information about loudness. These are derived from variations in the amplitude of speech signals, which vary per emotion. An amplitude envelope captures the changes in amplitude of a signal over time, giving instants of onset, where a non-silent part appears in a speech signal. Also, onset strength is another feature. Spectral contrast computes the level difference between energy at the peak and valley in a spectrum. Broadband noise like white noise can be identified with low contrast values whereas high contrast indicates a narrow-band signal. Spectral rolloff returns the frequency below which a certain amount of the total spectral energy (here, 85%) is contained. With the proper cutoff frequency, harmonics can be separated from noise. The spectral centroid informs the spectral center of the magnitude spectrum. Spectral flatness can determine the noise in the audio signal. Spectral bandwidth informs the energy spread across the frequency bands, given the variance from the spectral centroid. Tonnetz is obtained as a tonal centroid of chroma.

Harmonic to noise ratio quantifies the relative amount of noise in a speech signal. Pitch provides perceptual information about the fundamental frequency of vocal tract vibration which varies as per the speech signal. Also, the strength of the unvoiced part of the signal is extracted. The number of times the signal crosses the zero value is given by zero crossing rate, which is helpful in differentiating percussive and pitched sounds. Power spectral density informs the distribution of frequencies in the signal along with their corresponding strength, emphasizing the differences between frequency bins. Spectral entropy estimates the spectral density of the signal. Multifractal measures provide information about the dynamics of the glottal activity that is responsible for speech production.

Statistical measures taken over these features are five quartiles, range, mean, standard deviation, skewness, and kurtosis. Altogether, a 2141-length vector is obtained as a feature set *STPF* (spectral, temporal, prosodic, and fractal). To find higher weighing features among *STPF*, RFE is implemented. RFE recursively trains data by discarding less contributing features to find more weighted features. As a result, MFCC, chroma, RMSE, amplitude envelope, spectral rolloff, spectral contrast, spectral centroid, power spectral density, onset strength, pitch, and harmonic to noise ratio with different statistical measures are found to be significant. Selecting efficient features is the key to achieving a higher recognition rate. Hence, in order to find an effective subset of features, using brute-force approach, a feature set *ST* (spectral

PCI/CTI/MCI/CNPq Program and Fapesp Project 2020/07074-3.

and temporal features set) is chosen with MFCC (40 coefficients), CQT chroma (12 chroma), RMSE, amplitude envelope, spectral rolloff and spectral contrast features with their mean statistics.

MFMC is claimed to be a sufficient feature for SER analysis [14]. MFMC differs from MFCC in two aspects. Computing the Fourier spectrum is the first step in obtaining MFCC and MFMC features. Fourier spectrum with a vector of coefficients $X(k)$ of signal x of segment length N can be computed as

$$X(k) = \sum_{n=0}^{N-1} x_n e^{-i2\pi\frac{kn}{N}} \tag{2}$$

MFCC consider energy of the spectrum $X(k)^2$ for subsequent computation processes, whereas MFMC consider the magnitude of spectrum $|X(k)|$ to reduce the effect of outliers which may get introduced in MFCC while computing energy. Next step is to map the energy or magnitude of the spectrum to the Mel scale using a triangular filter bank to render linear frequency to nonlinear Mel frequency. The transform and inverse transform between the frequency, ν (in hertz) and Mel frequency, m can be obtained respectively as follows,

$$m = 2595\log_{10}\left(1+\frac{\nu}{700}\right); \nu = 700\left(10^{\frac{m}{2595}}-1\right) \tag{3}$$

The next step is to take the logarithm of the filterbank energy or magnitude. The obtained log-spectrum is known as the cepstrum, which has a quefrency-axis (time-domain). Low quefrency provides information on formants by capturing slow variations of log-spectrum. The last step in MFCC is to apply discrete cosine transform to decorrelate log energies. This step is excluded in MFMC, and its computation ends with obtaining a nonlinear log magnitude spectrum. We include MFMC as one of the feature sets to corroborate its efficiency, and for comparison, MFCC is also considered as another feature set. Computing derivatives and different statistical measures, MFCC and MFMC are compared. Their comparison is presented in the next section.

C. Data

This subsection introduces four acted Latin language databases which are analyzed in this work. The selected databases follow a discrete emotional model comprising the six basic emotions - anger, disgust, fear, happy, sad, and surprise, along with neutral.

EMOVO [19] is Italian database recorded with 6 professional actors, containing a total of 588 audios. All audio is reported to be 16 bit stereo and sampled at 48000 Hz. This corpus contains semantically neutral phrases covering all phonemes in Italian language, including short, long and nonsense sentences, plus questions.

The voice emotion recognition database (VERBO, [10]) is the first speech emotion corpus in the Brazilian Portuguese language. VERBO is recorded with 12 professional Brazilian actors (6 females and 6 males) containing 14 phrases and 1167 total audios. VERBO is based on the EMOVO in a way to include all the Portuguese linguistic phonemes in all the predefined emotions. It has mixed types of audio with different sampling rates. Of the 1167 audios, 1062 are mono and the remaining 105 are stereo. Moreover, 771 audios have a sampling rate of 44100 Hz, 300 audios have a sampling rate of 48000 Hz, and the remaining 96 audios have a sampling rate of 16000 Hz. Also, we found one duplicate audio file. This information is overlooked in [10] and not mentioned in earlier VERBO studies. For the analysis to have uniformity among the data, we fill in this information.

A Canadian French Emotional Speech Dataset (CaFE, [21]) is recorded by 12 actors (6 female and 6 male) and contains six basic emotions recorded in two intensities along with a neutral one. For this work, filtered and downsampled (16 bit and 48000 Hz) audios have been used.

INTER1SP speech emotional database [20] in the Spanish language is a part of the INTERFACE project. This database is recorded with 2 professional actors (1 male and 1 female) and contains 6041 audios with seven emotions at normal pace and only neutral emotion at slow, fast, soft and loud pace. All audios are reported to be 16 bit with sampling rate of 16000 Hz in l16 format. Speech types include sentences, paragraphs, digits and words. In this work INTER1SP is analyzed. For consistency in our analysis with the above databases, we consider only normal paced phrases in seven emotions constituting 3728 utterances and referred to as SP_Phrases.

D. Classifiers and measures

Six different machine learning classifiers and ensemble algorithms, namely, SVM, multilayer perceptrons (MLP), KNN, ensemble SVM (ESVM), random forest (RF), and histogram gradient boost (HGB) classifiers are implemented to model the feature sets. To overcome bias with train-test split and over fitting, a 10-fold CV method is used. Among four databases only the EMOVO is balanced, containing the same numbers of utterances for all seven emotions. Hence, a stratified approach is implemented while splitting the data, to confirm that each emotion has been adequately represented in each fold.

To evaluate the performance of models three different measures are reported. Prior studies of these corpora reported mean accuracy, and hence, for comparison, we calculate mean accuracy as a first measure. To account for the imbalance in data, a weighted F1-score is calculated as a second measure which considers true instances for each label while averaging. Accuracy and F1 score measures are calculated by computing how many instances are correctly and incorrectly classified. These are categorized as true positive (TP), true negative (TN), false positive (FP), and false negative (FN), and are represented as a confusion matrix. These measures do consider positively classified instances but not all negative instances. Hence in addition, a statistical measure, the Matthews correlation coefficient (MCC) is calculated as a third measure which does consider all positive and negative instances equally. Accuracy, F1 score and MCC are computed as follows.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FP} \tag{4}$$

$$F1score = \frac{2TP}{2TP + FP + FN} \quad (5)$$

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP).(TP + FN).(TN + FP).(TN + FN)}} \quad (6)$$

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. Chicco [22] advised to use MCC over other standard measures to evaluate the performance of any machine learning model.

Analyses are done in Python (3.8.5) and with the help of the following libraries: Librosa (0.8.0) [23], NumPy (1.19.2) [24], SciPy(1.5.2) [25], Scikit-learn (0.23.2) [26], Pandas(1.3.0) [27], MatPlotLib(3.3.2)v[28], and Parselmouth (0.4.0) [29].

IV. Results and Evaluation

This section presents first, the set of experiments performed with the VERBO corpus. In the preceding section, the first feature set, *ST*, is defined. Here, we start by comparing MFCC and MFMC, with the objective of defining the remaining two feature sets. Then we determine the best classifier using the feature sets. Finally, we extend this analysis to multi-corpora. All speech signals from the VERBO corpus are first converted to mono, and then are resampled to 22050 Hz to maintain uniformity. The duplicated audio file is removed. For the other three Latin corpora, speech signals are analyzed with their respective sampling rates.

A. Comparing MFCC and MFMC

MFCC and MFMC features are obtained using a Hann window with frames of 20 ms with 50% overlapping. A single statistical measure, mean, is calculated for both features. Then ten different statistics (five quartiles, range, mean, standard deviation, skewness and kurtosis) are computed. First and second order derivatives (Δs) along with the ten statistics are computed next. Table I summarizes stratified 10-fold CV mean accuracies, highlighting higher accuracy in bold font. For VERBO, MFCC without derivatives and with single statistics

TABLE I. Comparing stratified 10-fold CV accuracies for MFCC and MFMC features with mean, 10 stats without and with derivatives (Δs).

Feature	Statistics	10-fold CV mean accuracy (%)			
		VERBO	***EMOVO***	***CaFE***	***SP_Phrases***
MFCC	mean	**87.56**	**86.21**	**65.73**	**96.91**
	10 stats	55.45	73.98	59.33	96.88
	Δs+10 stats	44.80	59.69	51.19	95.81
MFMC	mean	82.26	55.27	47.52	93.72
	10 stats	82.36	73.82	52.46	97.42
	Δs+10 stats	**84.04**	**80.23**	**62.47**	**98.04**

increase accuracy while MFMC with derivatives and more statistical measures are found to be more effective. This observation also holds true for other three corpora. Hence, we define the other two feature sets to be MFCC with mean as set *MFCC*; and MFMC with derivatives and ten statistics as set *MFMC*.

B. Choosing ML Classifier

Fig. 1 shows a stratified 10-fold CV weighted F1 score for VERBO analyzed with different classifiers- SVM, ESVM, MLP, KNN, HGB, and RF with all three feature sets. The best results are obtained with SVM classifier, henceforth, we report analysis performed with SVM classifier only. SKlearn's SVC classifier with one-versus-one approach and radial basis function (RBF) kernel is used for analyses.

C. Performance evaluation

With the feature sets, classifier, and measures being defined, we extend the analysis to the other three corpora. Table II lists the measures for the multi-corporal analysis. The metrics used are mean accuracy (MA in %), weighted F1 score (WF1 in %), and MCC. Along with our results, for better and quick comparison, the previous best result is mentioned below the database name.

For VERBO, maximum mean accuracy is obtained as 87.56% with the *MFCC* set and 87.32% with the set *ST*. A higher weighted F1 score is obtained with *MFCC*. MCC with 0.85 consolidates the performance of the model. Please note that *ST* includes MFCC plus the other five features and despite

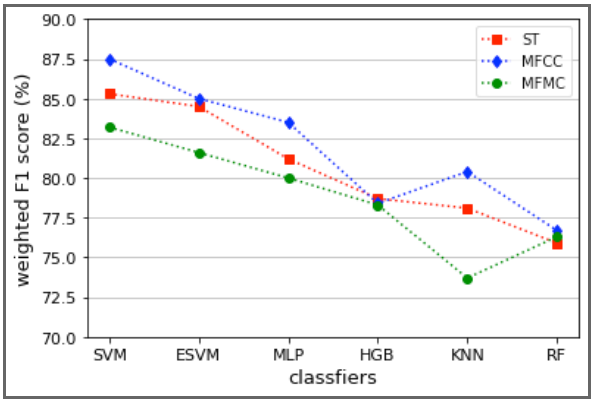

Fig. 1. Stratified 10-fold CV weighted F1 score for three feature sets with six classifiers are plotted for VERBO database.

TABLE II. Stratified 10–fold CV mean accuracy (MA, in %), weighted F1 score (WF1, in %) and Matthew' correlation coefficient (MCC) are listed for all databases (DB) under study. Best results (MA and WF1) are highlighted. Previous results are briefed in Section II and summarized in Table III.

DB/ previous best result	ST			MFCC			MFMC		
	MA	***WF1***	***MCC***	***MA***	***WF1***	***MCC***	***MA***	***WF1***	***MCC***
VERBO 78.64	87.32	87.28	0.85	**87.56**	**87.55**	0.85	84.04	84.06	0.81
EMOVO 79.07[a]	85.88	85.95	0.84	**86.21**	**86.22**	0.84	80.23	80.27	0.77
CaFE 71.10[b]	**68.85**	**68.26**	0.63	65.73	64.94	0.59	62.47	62.74	0.56
INTER1SP 94.01[c]	97.45	96.88	0.97	97.21	96.64	0.96	**97.60**	**97.42**	0.97
SP_Phrases	96.75	96.75	0.96	96.91	96.92	0.96	**98.04**	**98.04**	**0.98**

[a] 25% test accuracy for happy, anger, sad, and neutral emotions
[b] 10-fold CV for happy, anger, sad, and neutral emotions
[c] four different paced neutral emotions are omitted.

that, no significant difference in the recognition rate achieved. This implies that MFCC is predominant. An improvement of 8.9% in accuracy is achieved with the *MFCC* set compared to the previous result of 78.64% [11]. Similarly, the *MFMC* set yields a 5.4% improvement over the previous results [11] thereby demonstrating it to be a sufficient feature.

For EMOVO, a higher accuracy of 86.21% is obtained with set *MFCC* and is 7.1% higher than the formerly reported 25% test set accuracy of 79.07% [18]. Ancilin and Milton [14] reported a 10-fold CV mean accuracy of 53.40% using MFCC (30 coefficients and 12 statistics). Our analysis using MFCC (30 coefficients and 10 statistics) yielded a 10-fold CV mean accuracy of 75.50%, but when computing only 1 statistic (mean) instead of ten, accuracy is improved to 83.32%. This confirms our findings in subsection A that with a lower number of statistics, MFCC feature performance is better.

With *MFMC* (40 coefficients) set, the mean accuracy achieved is 80.27%. To compare results using the *MFMC* set, we want to draw the reader's attention to the following. Previous result [14] reported a 10-fold CV mean accuracy of 73.30% with MFMC (30 coefficients and 12 statistics). We obtained a similar accuracy of 73.82% with MFMC (30 coefficients and 10 statistics). However, when derivatives are included in the set, mean accuracy increases to 76.88%. We want to highlight this improvement achieved owing to the inclusion of derivatives in the MFMC feature. Another observation is that by increasing coefficients from 30 to 40, and along with derivatives and statistics, accuracy is increased further to 80.27%. This last observation is in accordance with the authors [14], who reported that an incrementing number of coefficients results in an improved accuracy.

For CaFE, set *ST* outperforms achieving a higher accuracy of 68.85%. A previous study [18] considered only four emotions, neutral, happy, sad, and anger, and reported 71.10% accuracy with a 25% test set using CNN. To compare, we performed an experiment with similar four emotions only. 10-fold CV mean accuracy of 79.51% is obtained, surpassing [18] by 8.4%. Only for CaFE corpus, the mean accuracy is improved to 73.61% for all seven emotions and 80.04% for above mentioned four emotions, when pitch and intensity are added to *ST* set along with statistical measures, sum, minimum, maximum, standard deviation, and skewness in addition to mean.

For the Spanish corpus, the *MFMC* set outperforms. For INTER1SP the highest accuracy of 97.60% is obtained achieving a 3.6% higher accuracy than the previous result of 94.01% [32] which omitted different paced neutral emotional utterances but included all types of utterances in seven emotions. Feature set ST and MFCC provide similar accuracy and show no significant differences. With SP_Phrases (considering only phrases in normal paced seven emotions) higher results are obtained with a set MFMC of 98.04%.

We list the previous best results in Table III using the same experimental corpus, number of emotions, evaluation metrics, extracted features, implemented classifier, and split method along with the best result we achieved. We obtained better results for all corpora, compared to the previous results, with all feature sets.

TABLE III. ESULTS: SUMMARIZING EARLIER STUDIES WITH EXTRACTED FEATURES, CLASSIFICATION METHOD, ACCURACY, AND SPLIT SET. BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT.

DB	Features	Classification Method	Accuracy (%)	Split	Ref
VERBO	Spectral and prosodic	KNN	76.49	10-fold CV	[12]
	Spectral and prosodic	Not mentioned (nm)	78.64	nm	[11]
	MFCC, chroma, F0, loudness, jitter, shimmer	CNN	76.69	30% test set	[13]
	MFCC	SVM (RBF kernel)	**87.56**	10-fold CV	this work
EMOVO	MFMC	SVM (linear kernel)	73.30	10-fold CV	[14]
	one dimensional local binary and ternary pattern (1BTPDN)	SVM (cubic kernel)	74.31	10-fold CV	[15]
	eGeMAPS	DBN	76.22	25% test set	[16]
	MFCC (with happy, anger, sad, and neutral emotions)	CNN	79.07	25% test set	[18]
	MFCC (with all emotions)	SVM (RBF kernel)	**86.21**	10-fold CV	this work
CaFE	MFCC	CNN	71.10	25% test set	[18]
	Spectral and Temporal (set *ST*)	SVM (RBF kernel)	**79.51**	10-fold CV	this work
INTER1SP [a] (with main seven emotions [b] with seven emotions plus neutral with varied pace)	empirical mode decomposition, Teager-Kaiser Energy Operator and cepstral features[a]	RNN	91.16	30% test set	[17]
		SVM (polynomial kernel)	90.36	10-fold CV	
	MFCC and Modulation Spectral features[a]	RNN	94.01		[32]
		SVM (polynomial kernel)	90.05	10-fold CV	
	MFMC[b]	SVM (RBF kernel)	**97.60**	10-fold CV	this work

D. Recognition Rate

Though higher accuracy is achieved with all feature sets, it is important to understand how well these sets are able to identify emotions. To examine the performance of the features across databases in recognizing emotions, Table IV presents the recognition rate for all emotions with all feature sets for all four databases. In the table, higher recognition rates are shown in red, whereas lower rates are in blue.

For VERBO corpus all sets could identify sad emotion with a higher recognition rate, with the highest being 96.41% using the *MFCC* set. The lowest rate of 79.04% is obtained for anger emotion when using sets *ST* and *MFCC*, whereas happy emotion is the least recognised with a rate of 80% using *MFMC*. With the *MFMC* set, the difference between the higher (87.43%) and lower (80.12%) recognition rate is 7%, less than the other two sets which have a range of around 17%. The previous study [13] reported a range of higher (85.56%) to lower (63.74%) recognition rate with a difference of 21.82%.

Comparatively, we achieved improvement in recognizing all emotions using all sets and the least difference between the highest and lowest recognition rate indicate that our proposed feature sets are unbiased towards any specific emotions.

To know how feature selection can improve the recognition rate for all emotions, we are also presenting results with the *STPF* feature set for the VERBO database. Analyzing the VERBO database by extracting feature set *STPF* and normalizing with the standardization method (μ = 0, σ = 1), with SVM classifier mean accuracy is found to be 79.43%, F1 score of 79.44%, and MCC of 0.76. A higher recognition rate of 87.43% is achieved for sad emotion and lower (75.45%) for anger emotion with a difference of 11.98%. When comparing the recognition rate of *STPF* set with that of *ST*, improvement is noticeable for all emotions when feature selection is implemented.

TABLE IV. RECOGNITION RATE (IN %) FOR ALL EMOTIONS WITH ALL FEATURE SETS AND FOR ALL DATABASES. HIGHER RECOGNITION RATE IS HIGHLIGHTED IN RED AND LOWER IN BLUE. DB STANDS FOR DATABASES. NEU STANDS FOR NEUTRAL.

DB	Set	Emotions										
		Anger	Disgust	Fear	Happy	Sad	Surprise	Neutral	Loud Neu	Soft Neu	Slow Neu	Fast Neu
VERBO	*ST*	79.04	89.82	87.95	84.94	95.81	82.64	91.02				
	MFCC	79.04	91.02	86.15	83.64	96.41	84.43	92.22				
	MFMC	85.03	81.44	86.75	80.0	87.43	82.04	85.63				
	STPF	75.45	76.05	83.13	79.52	87.43	76.05	82.63				
EMOVO	*ST*	83.33	85.71	84.52	77.38	95.24	82.14	92.86				
	MFCC	84.52	90.36	76.19	85.71	92.86	79.76	94.05				
	MFMC	66.67	91.67	91.67	80.95	84.52	77.11	69.05				
CAFE	*ST*	63.19	72.22	62.5	58.33	75.0	74.31	76.39				
	MFCC	58.04	68.06	61.81	52.78	72.22	72.22	75.0				
	MFMC	73.43	65.97	51.39	59.72	67.36	59.72	59.72				
SP Ph Ra ses	*ST*	95.84	96.81	95.14	94.36	96.05	100	99.07				
	MFCC	95.27	97.0	95.89	94.55	96.62	100	99.07				
	MFMC	97.54	97.0	97.20	97.37	97.56	100	99.63				
INTERI SP	*ST*	95.31	97.40	94.83	93.44	95.88	99.32	99.18	99.57	98.71	100	98.26
	MFCC	95.45	96.72	93.74	94.13	96.02	99.18	98.37	99.57	98.71	99.13	98.26
	MFMC	97.24	96.72	95.65	95.49	97.39	99.73	98.77	99.14	99.14	96.96	97.39

EMOVO provides mixed results, giving a higher recognition rate for sad emotion using the *ST* feature set. Sad emotion is identified with a higher recognition rate of 95.24% and happy has the least recognition rate of 77.38%, with a difference being 17.86%. Previous study [14] could identify sad emotion the most with 84.52% and fear had the least recognition rate of 65.48% with a difference of 19.04%. CaFE corpus provides mixed results, recognizing neutral emotion the most followed by sad emotion using *ST* features, and the least identified emotion is happy with a difference between them being 18.06%.

SP_Phrases identify surprise emotion perfectly (100%) using all three sets. *MFMC* set yields a better recognition rate compared with the other two sets and differs marginally in identifying the other five basic emotions with a 3% difference between the highest and lowest recognition rate. Disgust is least identified by the *MFMC* set, whereas happy is least identified by the *ST* and *MFCC* sets. INTER1SP correctly identified slow-paced neutral emotion with *ST* set, and the least identified is happy (93.44%). Using *MFMC*, surprise is identified the most with 99.73% and the least identified emotion is happy with 95.49% with a difference of 4.24%. Previous study [32] achieved a higher recognition rate of 97.5% for neutral emotion followed by sad emotion with 95.65% and the least for anger emotion with 91.86% with a difference of 5.64%.

To summarize, the *MFCC* set provides better results for VERBO and EMOVO; the *MFMC* set for INTER1SP; and the *ST* set for CaFE in terms of accuracy, weighted F1 score and also recognition rate.

V. CONCLUSION

This study presents a simple yet robust multi-corporal SER model, with *ST*, *MFCC*, and *MFMC* sets, containing one and upto six features. While the sets *MFCC* and *MFMC* are singular, *ST* consists of MFCC, CQT chroma, RMSE, amplitude envelope, spectral rolloff, and spectral contrast with mean statistics. BP SER corpus VERBO yields a 5-9% improvement in mean accuracy and an improved recognition rate for the three sets using SVM classifier. Next, this analysis is extended to three Latin language databases (Italian, Spanish, and Canadian French) to substantiate the robustness of the model. Higher recognition rates and mean accuracy are obtained for Latin corpora, compared to previous studies.

Key observations for MFCC and MFMC features are: (1) including derivatives and computing various statistical measures with MFMC; while (2) excluding derivatives and computing lesser statistical measures with MFCC, provide robustness and the ability to distinguish emotions effectively. Evidently selected spectral and temporal features are effective for all analyzed databases, but prosodic features are also important for CaFE. In addition to mean accuracy, weighted F1-score and MCC are also reported confirming reliability of the model.

Using a classical machine learning classifier our results outperform previously exercised complex features, algorithms, and architectures, in terms of accuracy. And also, it is computationally less expensive. Agreeing with Sönmez and Varol [15], we believe our proposed feature sets are optimal evidencing that traditional machine learning algorithms are able to surpass advanced deep learning algorithms, achieving higher recognition rate. The differences in recognition rates (20%) for all corpora analyzed in this work indicate that the proposed feature sets exhibit comparatively less biased behavior towards all emotions. We conclude by proposing a state of the art model for SER in Latin languages and to our best knowledge, this is the first article exploring BP SER in detail addressing the preprocessing issues.

ACKNOWLEDGMENT

This work was partially sponsored by PCI/CTI/MCI/CNPq Program and by the Fapesp Project 2020/07074-3

REFERENCES

[1] D. Schuller and B. Schuller, “A Review on Five Recent and Near-Future Developments in Computational Processing of Emotion in the Human Voice,” Emotion Review, vol. 13(1), pp. 44–50, 2021.

[2] J. Tao and T. Tan, “Affective Computing: A Review,” Affective Computing and Intelligent Interaction, LNCS 3784, pp. 981–995, Springer, 2005.

[3] M. Akçay, and K. Oguz, “Speech emotion recognition: Emotional models, databases,features, preprocessing methods, supporting modalities, and classifiers,” Speech Communication, vol. 116, pp. 56-76 , 2020.

[4] R. Banse and K.R. Scherer, “Acoustic profiles in vocal emotion expression,” Journal of personality and social psychology, vol. 70(3), pp. 614, 1996.

[5] A. Wierzbicka, Emotions across languages and cultures: Diversity and universals, Cambridge University Press, New York, 1999.

[6] C-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, “Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011,” Artificial Intelligence Review, vol. 43, pp. 155–177, 2015.

[7] M. Ayadi, M. Kamel, and K. Fakhri., “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition, vol. 44, pp. 572-587, 2011.

[8] M. Shah Fahad, A. Rajan, J. Yadav, and Deepak, A. “A survey of speech emotion recognition in natural environment,” Digital Signal Processing, vol. 110, pp. 102951, 2021.

[9] C-N. Anagnostopoulos, and T. Iliou, “Towards emotion recognition from speech: Definition, problems and the materials of research,” In: Semantics in Adaptive and Personalized Services, vol. 279. Springer, Berlin, 2010.

[10] J.R.T. Neto, G. Filho, L. Mano, and J. Ueyama, “Verbo: voice emotion recognition database in Portuguese language,” J. Comput Sci., vol. 14(11), pp. 1420–1430, 2018.

[11] F.RY. da Silva, D.L.R Santos do Amor, R.D. Monte Paixão, C.G. Resque dos Santos, and B. Serique Meiguins, “UXmood—A Sentiment Analysis and Information Visualization Tool to Support the Evaluation of Usability and User Experience,” Information, vol. 10(12), pp. 366, 2019.

[12] J.R.T. Neto, “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.

[13] G.A. Campos, and L.D.S. Moutinho, “DEEP: Uma arquitetura para reconhecer emoçõa com base no espectro sonoro da voz de falantes”. Ph.D. thesis, Universidade de Brasília, 2020.

[14] J. Ancilin and A. Milton, “Improved speech emotion recognition with mel frequency magnitude coefficient,” Applied Acoustics, 179, 108046, 2021.

[15] Y. Sönmez and A. Varol, “A speech emotion recognition model based on multi-level local binary and local ternary patterns,” IEEE Access, pp. 190784-190796, 2020.

[16] S. Latif, A. Qayyum, M. Usman, and J. Qadir, “Transfer Learning for Improving Speech Emotion Classification Accuracy,” In: 19th Annual Conference of the International Speech Communication Association: Speech Research for Emerging Markets in Multilingual Societies (INTERSPEECH 2018), 2018.

[17] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M.A. Mahjoub, and C. Cleder, “Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO,” Speech Communication, vol. 114, pp. 22-35, 2019.

[18] A. Iliev, A. Mote, and A. Manoharan, “Cross-cultural emotion recognition and comparison using convolutional neural networks,” Digital Presentation and Preservation of Cultural and Scientific Heritage, vol. 10, pp. 87-101, 2020.

[19] G. Costantini, I. Iadarola, A.Paoloni, and M. Todisco, “Emovo corpus: an Italian emotional speech database,” In: Proceedings of the Ninth International Conference on Language Resources and Evaluation, 3501-3504. European Language Resources Association, 2014. [I. Iadarola, “Emovo, database di parlato emotivo per l’italiano,” In: Atti del 4° Convegno Nazionale dell’Associazione Italiana di Scienze della Voce, 2009]

[20] Emotional speech synthesis database, ELRA catalogue (http://catalog.elra.info), ISLRN: 477-238-467-792-9, ELRA ID: ELRA-S0329

[21] P. Gournay, O. Lahaie, and R. Lefebvre, “A Canadian French Emotional Speech Dataset (1.1) [Data set],” ACM Multimedia Systems Conference (MMSys 2018), Amsterdam, The Netherlands. Zenodo. https://doi.org/10.5281/zenodo.1478765

[22] D. Chicco, “Ten quick tips for machine learning in computational biology,” BioData mining, vol. 10, pp. 35, 2017.

[23] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, et al. “librosa: Audio and music signal analysis in python,” In: Proceedings of the 14th python in science conference, 2015.

[24] C.R.Harris, K.J.Millman, S.J.vander Walt, R.Gommers, P.Virtanen, et al. “Array programming with NumPy,” Nature, vol. 585, pp.357-362, 2020.

[25] P. Virtanen, R. Gommers, T.Oliphant, M. Haberland, T. Reddy, et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” Nature Methods, vol. 17, pp. 261-272, 2020.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. “Scikit-learn: Machine learning in Python,” Journal of machine learning research, pp. 2825-2830, 2011.

[27] Wes McKinney, “Data structures for statistical computing in python,” In: Proceedings of the 9th Python in Science Conference, pp. 51-56, 2010.

[28] J.D. Hunter, “Matplotlib: A 2D graphics environment,” Computing in science &amp, engineering, vol. 9, pp. 90-95, 2007.

[29] Y. Jadoul, B. Thompson, and B. de Boer, “Introducing Parselmouth: A Python interface to Praat,” Journal of Phonetics, vol. 71, pp. 1-15, 2018.

[30] F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, A. Elisabeth, et al. “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” IEEE Transactions on Affective Computing, vol. 7, pp. 190-202, 2016.

[31] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[32] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M.A. Mahjoub, and C. Cleder, “Automatic speech emotion recognition using machine learning,” Social media and machine learning, IntechOpen, 2019.

[33] R. Böck, O. Egorow, I. Siegret, and A. Wendemuth, “Comparative study on normalisation in emotion recognition from speech,” Intelligent Human Computer Interaction, pp. 189-201, 2017.

[34] F. Cavallo, F. Semeraro, L. Fiorini, G. Magyar, C. Sincak, P. Dario, “Emotion Modelling for Social Robotics Applications: A Review,” J Bionic Eng, 15, 185–203, 2018.

[35] R.F. Voss, and J. Clarke. "1/f noise in speech and music," Nature vol. 258, pp. 317-318, 1975.

[36] U. Sarkar, S. Nag, C. Bhattacharyaa, S. Sanyal, A. Banerjee, R. Sengupta, and D. Ghosh, “Language Independent Emotion Quantification using Nonlinear Modeling of Speech,” Journal of Image Processing & Pattern Recognition Progress, vol. 6(3), pp. 24–30, 2019.

[37] G. J. Lal, E.A. Gopalakrishnan, and D. Govind, “Glottal Activity Detection from the Speech Signal Using Multifractal Analysis,” Circuits, Systems, and Signal Processing, vol. 39, pp. 2118–2150, 2020.

[38] H. Liu and W. Zhang, “Mandarin emotion recognition based on multifractal theory towards human-robot interaction,” IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 593–598, 2013.

[39] T. Bäckström, O. Räsänen, A. Zewoudie, and P. Zarazaga, Introduction to Speech Processing, aalto wikibook.