Transfer Learning and Data Augmentation Techniques applied to Speech Emotion Recognition in SE&R 2022

Caroline Alves[1], Bruno Carlotto[2], Bruno Dias[1], Anátale Garcia[1], Bruno Gianesi[3], Renan Izaias[1], Maria Luiza de Morais[1], Paula de Oliveira[1], Vinícius G. Santos[1], Rafael Sicoli[1], Flaviane R. Fernandes Svartman[1], Sandra Aluisio[2] and Sidney Leal[2]

[1]*Departamento de Letras Clássicas e Vernáculas, FFLCH-USP*

[2]*Instituto de Ciências Matemáticas e de Computação, ICMC-USP*

[3]*Engenharia Mecatrônica, EESC-USP*

Abstract

In this work, our team ICMC-EESC-FFLCH explores several techniques to address data scarcity and imbalance in SE&R 2022 task dedicated to speech emotion recognition (SER). We evaluate two types of transfer learning models: (i) Multi-task learning, in which two tasks are learned simultaneously, and (ii) Sequential transfer learning where the tasks are learned sequentially. In both models, the auxiliary task is genre classification from speech, using a large dataset with almost 145 hours of speech signals. As for the techniques to balance the training data, we have used the SMOTE (Synthetic Minority Over-sampling Technique) and Praat's Change gender command to over-sampling minority classes. Our Sequential transfer learning architecture, using the two baselines feature sets provided by the shared-task (prosodic audio features and embeddings generated by the Wav2Vec 2.0 model) and the two approaches to balance the training dataset reaches satisfactory performance with a 0.5353 F1-macro, surpassing the prosodic features baseline. On the other hand, our multi-task learning approach using the two baseline features sets and the SMOTE approach to balance the training dataset reaches only a 0.5301 F1-macro. Finally, our worst result is 0.469 F1-macro, obtained with the feature selection experiment (29 prosodic features manually chosen from the literature), using our multi-task learning architecture with the two approaches to balance the training dataset.

Keywords

Deep Learning, Transfer Learning, Data Augmentation, Speech Emotion Recognition

1. Introduction

According to [1], speech emotion recognition (SER) systems are composed of methods, namely feature extraction and emotion classification, that process and classify speech signals to detect the embedded emotions of speech. They can also include a preprocessing step before the extraction of the features used to normalize the signals, for example, the use of noise reduction

Proceedings of the First Workshop on Automatic Speech Recognition for Spontaneous and Prepared Speech & Speech Emotion Recognition in Portuguese, co-located with PROPOR 2022. March 21st, 2022 (Online).

carolalves@usp.br (C. Alves); bruno.baldissera@usp.br (B. Carlotto); brunoadiaspapa1@usp.br (B. Dias); anatale.garcia@usp.br (A. Garcia); brunogianesi@usp.br (B. Gianesi); renan.izaias@usp.br (R. Izaias); marialuizamorais@usp.br (M. L. d. Morais); paulamarindeoliveira@usp.br (P. d. Oliveira); vinicius.santos@alumni.usp.br (V. G. Santos); rafael.pac90@gmail.com (R. Sicoli); flavianesvartman@usp.br (F. R. Fernandes Svartman); sandra@icmc.usp.br (S. Aluisio); sidleal@gmail.com (S. Leal)

techniques. Emotion classes depend on labeled data of the dataset used to create the model; these datasets can be of three types: acted, elicited or natural. While most of the natural datasets are from spontaneous speech recorded in noisy environments, acted speech databases are recorded by professional actors in sound-proof studios. Elicited speech datasets are created by placing speakers in a simulated emotional situation that can stimulate various emotions and can be close to real ones. It is important to notice that, the definition of emotion is an open problem in psychology and there are two models being used in SER systems: discrete and dimensional emotional models. The first one is based on the six primary and culturally independent categories of basic emotions [2]: sadness, happiness, fear, anger, disgust, and surprise, where other emotions are obtained by the combination of the basic ones. Most of the existing SER systems focus on all these basic emotional categories, sometimes including the neutral category (see, for example, [3], a study focusing on Portuguese language), or in a small group of those emotions[1]. The second one, the dimensional emotional model, uses a small number of latent dimensions to define emotions such as: valence, arousal/excitation, control/power. In this model, emotions are not independent of each other, instead, they are analogous to each other in a systematic way. [5] support of the thesis that the three dimensions of pleasure-displeasure (valence), arousal-nonarousal (excitation), and dominance-submissiveness (power/control) are both necessary and sufficient to describe a large variety of emotional states. Specifically, valence describes whether an emotion is positive or negative, and it ranges between unpleasant and pleasant; excitation defines the strength of the felt emotion, ranging from boredom to frantic excitement; and the dimension of control/power refers to the seeming strength of the person (between weak and strong). For example, the third dimension differentiates anger from fear by considering the strength or weakness of the person, respectively; however, as the surprise emotion may have positive or negative valence depending on the context, it is difficult to categorize.

Whereas most studies on SER deal with simulated, noise-free datasets recorded in sound-proof studios [4], SE&R 2022 brings a small dataset of approximately 50 minutes, with 625 audio segments (training dataset) from the C-ORAL-BRASIL I corpus [6], consisting of audio segments representing Brazilian Portuguese informal spontaneous speech, recorded in natural contexts and noisy environments.

The two baseline feature sets (prosodic audio features for emotion classification [7, 8] and embeddings generated by the Wav2Vec 2.0 model [9]) made available for SE&R 2022 were used in this work. Feature selection was also evaluated, focusing on four small prosodic features sets, manually chosen, with 29, 19, 10, and 8 features, taken from pitch, intensity, and spectrum groups of features. While the first SER systems used machine learning methods with a careful feature engineering (see several examples in [10]), recent approaches use ensembles to learn hybrid acoustic features [11], and deep learning architectures, such as multi-task learning [12, 13], attention mechanisms [14], and transfer learning approaches [15].

Our contribution to SE&R 2022 explores two architectures based on deep neural networks (DNN) aiming at detecting Speech Emotion Recognition in Portuguese audio files. Our proposal evaluates two types of inductive transfer learning: multi-task [16] and sequential transfer learning [17]. In both models, the auxiliary task is genre classification from speech[2]. Since

[1] There are large lists of datasets used for emotion recognition in [1] and [4].

[2] Project's github: https://github.com/BrunoGianesi/Speaker-Gender-Recognition.

DNN-based classifiers have a generalization error problem when trained with limited datasets, we explore two different data augmentation techniques aimed to balance the training data. We have used the SMOTE [18] to create synthetic data for the minority classes and Praat's [19] Change gender command to manipulate the acoustic features in order to create new synthetic data based on the pre-existing ones. The Jupyter notebooks and characterization of the training dataset are publicly available at https://github.com/BrunoBaldissera/ser-transfer.

2. Experimental Framework

First, we present the original dataset for the main task and the dataset used for the auxiliary task of genre classification from speech in both inductive transfer learning architectures (Section 2.1), noting that the original dataset is unbalanced. Therefore, we applied two techniques for data augmentation (Section 2.2). Section 2.3 presents the feature sets we explored in our linguistically motivated selection of prosodic features, based on the literature. Finally, Section 2.4 presents our multi-task and sequential transfer learning architectures.

2.1. Datasets

2.1.1 Primary Task Dataset: official dataset of SE&R shared-task on SER. In the SE&R 2022 shared-task on SER, the audio segments are labeled in three classes: neutral, non-neutral female, and non-neutral male. The neutral class is the majority class (491 samples) and is used to label audio segments with no well-defined emotional state while the non-neutral classes label segments (89 non-neutral-female and 45 non-neutral-male) associated with one of the primary emotional states in the speaker's speech. In order to better understand the training dataset used in this study, seven annotators from our group pursued a qualitative analysis of the dataset. They labeled every audio in the training set with "yes" (meaning presence) or "no" (meaning absence) according to the following categories:

- **Noise**: any sort of noise not related with the primary voice(s)[3], e.g., background chatting, microphone hissing noise, music, children voices, etc.;
- **Voice overlapping**: periods in which there were two primary voices speaking at the exact same moment;
- **Different gender**: the presence of more than one perceived gender in the primary voices of the same audio; and
- **Voices in sequence**: the presence of more than one primary voice in the same audio, but without direct overlapping between them.

Our evaluation is summarized in Figures 1a and 1b. As we can see, there is a lot of noisy audio. Although noise is not a problem for the auxiliary task (Audio Genre Classification) [20] of the neural architectures, only an error analysis can identify possible problems for the SER task as a whole. Also, two complex problems were found: high overlapping rate of voices and audios with different genres, which we believe may have an impact on the classification of the 2 non-neutral classes (male and female). Of the 26 non-neutral audios that have different gender,

[3] We consider primary voices to be the loudest, and secondary voices to be the least prominent in the audio.

24 have voices overlapping and only 2 have voices in sequence. Of the 56 neutral audios that have different gender, 53 have voices overlapping and only 3 have voices in sequence.

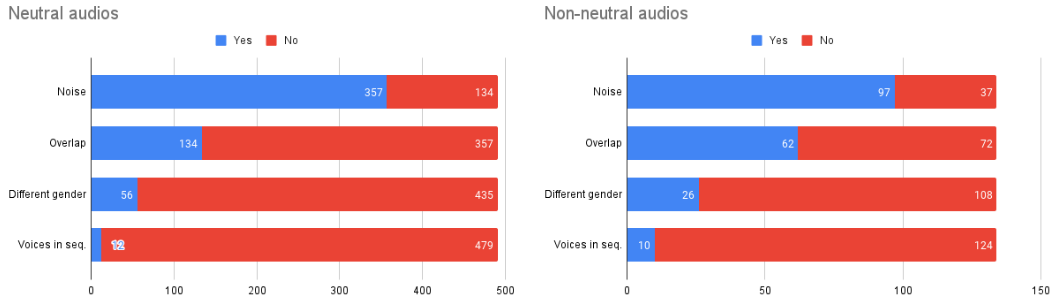


Figure 1: A qualitative analysis of the SER dataset performed by our team. Figures 1a and 1b show a characterization of the training dataset, presenting the number of audios with noise, primary overlapping voices, primary voices with different genders, primary voices in sequence, for both types of classes (neutral and non-neutral) audios.

2.1.2 Auxiliary Task Dataset: CETUC. The task of classifying gender based on voice identifies automatically a voice as male or female, based on the audio features. The gender identification of a given speaker was implemented in an undergrad project of one of the authors [20], to evaluate machine learning methods, such as decision trees, random forest, gradient boosting, support vector machine, multi-layer perceptron and logistic regression, and to compare the use of distinct features and models applied on different datasets. In addition, the study also assessed whether the models generalize to other contexts, such as other languages (English) or noisy environments, when trained on CETUC dataset [21] that was recorded in a controlled environment.

The best performance method (gradient boosting) was trained using the large dataset CETUC, with almost 145 hours of speech signals spoken by 50 male and 50 female speakers[4], each one pronouncing 1,000 phonetically balanced sentences selected from the CETEN-Folha corpus[5]. The best performance model used three sets of features from audio signals, totalling 44 features: (i) 12 statistics extracted from the highest frequency value, after applying the Fourier transform on the audios, divided into time windows of 0.2 seconds, (ii) the fundamental frequency (F0) statistics (12) and (iii) 20 MFCCs (Mel-Frequency Cepstral Coefficients), and reached an accuracy of 94,1%. This model was able to generalize well to audios with noise; it reached an accuracy of 90,8% on the testset MLS [22] with noise.

2.2. Data Augmentation Approaches: SMOTE and Praat’s Change Gender

We used two approaches to balance the training dataset applied specifically on audios of non-neutral male and non-neutral female classes: SMOTE [18] and Praat’s Change gender

[4] https://igormq.github.io/datasets/
[5] https://www.linguateca.pt/cetenfolha/index_info.html

command [19].

It is suggested by the authors of the original SMOTE paper that previously performing a random under-sampling of the majority class followed by over-sampling the minority class tends to yield good results. However, in this work, we have only over-sampled the minority classes, following the work by [23], and using the technique in its simplest implementation. Nonetheless, as the synthesis of new data with SMOTE uses a linear combination of randomly chosen neighbors of the underrepresented instances in the feature space rather than just replicating the given instances, we gave more focus to this augmentation approach in place of the simple oversampling (even though a number of such tests was performed). We have used the Python imbalanced-learn package [24]; all the parameters were set as default.

Praat's Change gender command allow us to manipulate the acoustic features to create new synthetic data based on the preexisting ones. Through this method, we can change the perceived gender of a given voice into the opposite gender. The second method for data augmentation consists in the use of the algorithm for gender conversion available in the software for acoustic analysis Praat. A total of 133 files were used, 45 of them containing male voices, then converted to female ones, and 88 containing female voices, then converted to male ones[6]. The task was undertaken by five annotators and had two phases: attribution of parameters for conversion and quality evaluation of the generated voice. In the quality assessment phase, the annotators changed the previously established default values in order to obtain voices that they judged the most natural as possible. For the conversion process, we first defined the frequency range in which the algorithm parameters were applied, using the values already predefined by the program, with the minimum pitch value being 75 Hz, and the maximum 600 Hz. The algorithm contains four parameters, described below, that can be used for gender conversion, from which we have only used the first two:

- **Formant shift ratio** (default value is 1.0) determines the ratio for proportionally modifying the value of formants, i.e., the sound frequency values at which the highest peaks of intensity occur, resulting from the resonance of the sound wave in its path through the vocal tract, from its production in the vocal folds until the moment of emission. The factor valued 1.0 means there is no alteration. For the task, we established the factor value 1.1 as the standard for male-to-female conversion, used in 30 of 45 files, and 0.8 for female-to-male conversion, used in 72 of 88 files. As mentioned above, these values were altered in some files in order to maintain a perceived natural quality of the converted voice: for the other 15 male-to-female converted files, factors between 1.15 or 1.2 were used, and for the other 16 female-to-male converted files, values between 0.85 or 0.9.
- **New pitch median** (default value is 0.0): a new median for the pitch values is established for each file, which, in turn, is used to compose a factor expressed by the ratio between this new median and the original median pitch. This factor is then used by the algorithm to multiply the original pitch values to obtain new values. In this metric, the value 0.0 represents the default setting, yielding the factor 1.0, which means no alteration. We established as standard values for this assignment the frequency measurement of 300 Hz for male-to-female conversion, for 35 of 45 files, and 140 Hz for female-to-male conversion,

[6] For one of the audios, the algorithm could not produce a successful conversion.

for 58 of 88 files. These values were also altered in some files to achieve a convincing result: for male-to-female conversion, values between 250 Hz and 380 Hz were used for the other 10 files, and for female-to-male conversion, values between 80 Hz and 260 Hz were used for the other 30 files.

- **Pitch range factor** (default value: 1.0) provides for an additional modification in pitch by an extra scaling of the values around the new pitch median, obtained in the previous step. A factor of 1.0 means that no additional pitch modification will occur, and a factor valued as 0.0 monotonizes the new sound to the new pitch median. Considering the essential goal of the project, the default value was kept and no modifications for the pitch range were provided.
- **Duration factor** (default value: 1.0) establishes a factor used for lengthening the sound file. For a factor valued less than 1.0, the resulting sound will be shorter than the original, and a value higher than 3.0 will not work. The default value provided by the software was also maintained, as a change in the duration of the sound is deemed as unnecessary for the development of the task.

2.3. Selection of Prosodic Features for SER

We grouped the 56 prosodic audio features (one of the baseline feature sets) into six classes[7] in order to select those strongly related to the classes defined for SE&R 2022 and evaluate them separately and conjoined: (1) related to voice quality (13 features), including local_jitter and local_shimmer, those from Harmonics-to-Noise Ratio (HNR) and those from Glottal-to-Noise Ratio (GNE); (2) related to intensity (9 features), for example, min_intensity, max_intensity; (3) related to F0 (pitch) (10 features), for example, mean_pitch, stddev_pitch; (4) related to spectrum (10 features), for example, skewness_spectrum, kurtosis_spectrum; (5) related to formants (10 features), for example, formant_dispersion, average_formant; (6) related to vocal tract length (VTL) (4 features), for example, fitch_vtl, vtl_delta_f.

The groups related to intensity (first 9 features), F0 (from 10 to 19), and spectrum (last 10 features), respectively shown in Table 1, were chosen for our feature selection experiment which included the training of 7 multi-task and 5 sequential classifiers, totalling 12 experiments, shown in Section 3.3. The classifiers used 10 (related to spectrum), 19 (intensity and F0) and 29 (spectrum, intensity, and F0) features and also a subset of 8 features, shown in bold in Table 1.

According to [25], energy, pitch, and time are the three perceptual dimensions on which most vocal indicators of various emotions are based. Therefore, the class of acoustic parameters related to F0, intensity, and spectrum were selected because they are reported in the literature as potential correlates of the vocal expression of emotions [25, 26, 27, 28].

F0 (fundamental frequency) is an acoustic correlate of the rate of vocal cords vibration, that is, the number of times a sound wave produced by the vocal cords is repeated during a given period of time. F0 is perceived as the pitch of the voice, and the range of values for this frequency varies according to sex and age[8]. In turn, sound intensity corresponds to the variations in the air pressure of a sound wave and is perceived as the loudness of a sound. Loudness and pitch are,

[7] The feature *voiced_fraction* was allocated in the group of spectrum features, instead of with the pitch group.

[8] For instance, 80–200 Hz for adult males, 180–400 Hz for adult females [29], and higher ranges for children. The mean values change for older ages.

Table 1
Features used in the classifiers of the feature selection experiment.

1	**Min_intensity**	16	Q1_pitch
2	Relative_min_intensity_time	17	Q3_pitch
3	**Max_intensity**	18	Mean_absolute_pitch_slope
4	Relative_max_intensity_time	19	Pitch_slope_without_octave_jumps
5	**Mean_intensity**	20	Center_of_gravity_spectrum
6	**Stddev_intensity**	21	Stddev_spectrum
7	Q1_intensity	22	Skewness_spectrum
8	Median_intensity	23	Kurtosis_spectrum
9	Q3_intensity	24	Central_moment_spectrum
10	**Min_pitch**	25	Voiced_fraction
11	Relative_min_pitch_time	26	Band_energy
12	**Max_pitch**	27	Band_density
13	Relative_max_pitch_time	28	Band_energy_difference
14	**Mean_pitch**	29	Band_density_difference
15	**Stddev_pitch**		

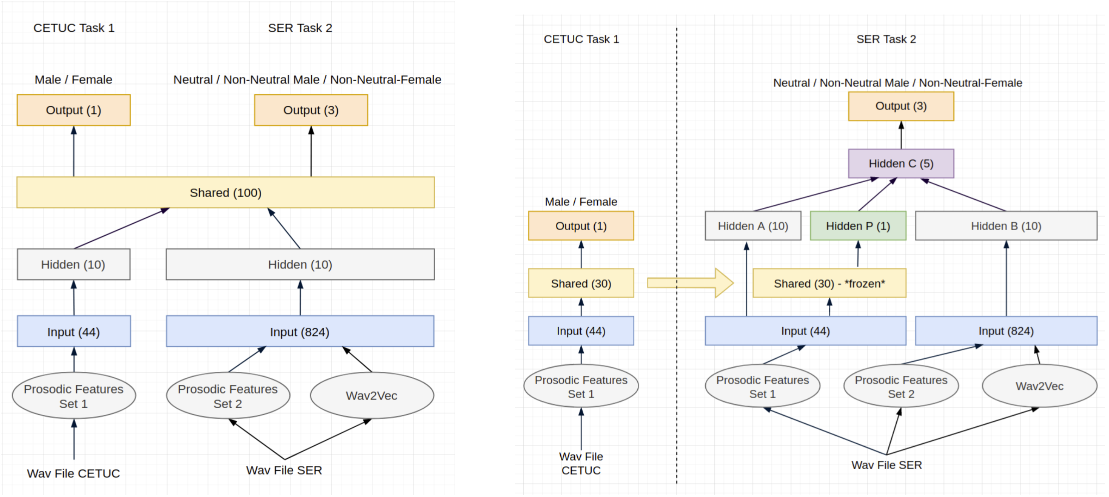
in fact, elementary domains of the auditory signal and changes in sound intensity and F0 seem to be relevant to emotion analysis: higher and wider pitch ranges and higher sound intensity are typically associated with high arousal emotions (e.g., fear, anger, joy) compared to neutral speech, while lower and narrower pitch ranges and lower sound intensity are more associated with low arousal emotions (e.g., sadness, boredom, calmness) [25, 30, 31, 32, 33]. Studies have also shown that emotion affects the distribution of spectral energy across the range of sound frequencies: for example, stronger energy in higher frequency bands is usually associated with high arousal emotions, while weaker energy in the same band is more associated with low arousal emotions [31][9].

2.4. Neural Architectures: multi-task and sequential transfer learning

Transfer Learning is a machine learning approach that transfers weights trained in one task, domain, or language to a different one, with the aim of improving the learning generalization [17]. In this work, two Transfer Learning techniques were used: Multi-task and Sequential Transfer Learning. In the first one, the training of the two tasks is performed simultaneously, sharing a layer of weights between the two tasks [16]. In the second, the weights trained in the first task are transferred to the second, sequentially [34]. Figure 2 presents the two architectures.

For the Multi-task architecture, two MultiLayer Perceptron (MLP) neural networks were used, with 4 layers each, sharing a common layer with 100 neurons. The first one focused on the binary gender prediction task, using the CETUC dataset, with 44 neurons in the input layer and one neuron in the output layer. The second (main task), focused on the prediction of the three

[9] Many of these studies used speech audios recorded in sound-proof booths with controlled scenarios. Spontaneous speech recorded in natural contexts and noisy environments like SER shared-task dataset interferes with extracted features results, as the acoustic signal is affected by sound sources competing with the target signal, the performance of pitch detection algorithms degrades as the noise level increases, and even the speech signal energy depends on the distance and position between the speaker's mouth and microphone. Therefore, in future work, at least methods for noise incorporation/reduction will be explored to assess the impact of noise on data.



(a) Multi-task

(b) Sequential

Figure 2: Transfer Learning architectures: a) Multi-task: 2 MLP's with 4 layers (1 shared); and b) Sequential: the second MLP with 5 layers uses a frozen layer from the first. Prosodic Features Set 1 is composed of 44 features described in the work developed by [20] while Prosodic Features Set 2 is composed of 56 features provided by the SE&R shared-task on SER and described in Section 2.3.

SER classes, with the number of neurons in the input layer varying from 8 to 824 (according to the features used) and three neurons in the output layer. Both use a previous layer of 10 neurons before the common layer. For the Sequential architecture, two MLP's were also used, but they were trained sequentially. The first for the binary gender prediction task with 44 neurons in the input, a hidden layer of 30 neurons and one neuron in the output. The hidden layer was then frozen and transferred to the second MLP, whose input layer ranged from 73 to 868 (according to the features used) and with three neurons in the output layer (one for each class) of the second task. The frozen layer acted by predicting the gender of the samples (auxiliary task) and passing this prediction as a new internal feature to a layer of 5 neurons before the output (for models with more features this layer was changed to 10 neurons).

3. Experiments

All the 26 models described in Sections 3.1, 3.2 and 3.3 were trained using a batch size of 100 and 300 epochs.

3.1 Sequential Learning Results. Table 2 presents the results, in crescent order of F1-macro values, for the experiments with the sequential learning architecture.

3.2 Multi-task Learning Results. Table 3 presents the results, in crescent order of F1-macro values, for the experiments with the multi-task learning architecture.

Table 2
Sequential Learning results using 5-fold cross-validation. We indicate in the model's name which feature set was used and whether a data augmentation technique was used (+) or was not used (-). The last line indicates the value of F1-macro for the submitted model, using the full dataset.

	Average scores for all 5 folds		
Model name/feature sets/data aug techniques	**F1-macro**	**Accuracy**	**Loss**
(1) Seq. wav2vec - SMOTE + CG	0.4139	79.0240	0.1122
(2) Seq. all prosodic + SMOTE - CG	0.4653	53.4022	0.2001
(3) Seq. all prosodic + SMOTE + CG	0.5344	61.8574	0.1815
(4) Seq. all prosodic and wav2vec + SMOTE - CG	0.5621	64.0564	0.1566
(5) Seq. wav2vec + SMOTE - CG	0.7043	74.8603	0.1379
(6) Seq. wav2vec + SMOTE + CG	0.7067	75.4375	0.1170
(7) Seq. all prosodic and wav2vec + SMOTE + CG	**0.8035**	82.4077	0.0846
1st Submission - Model (7) (full dataset)	**0.5353**		

Table 3
Multi-task Learning results using 5-fold cross-validation. We indicate in the model's name which feature set was used and whether a data augmentation technique was used (+) or was not used (-). The last line indicates the value of F1-macro for the submitted model, using the full dataset.

	Average scores for all 5 folds		
Model name/feature sets/data augmentation techniques	**F1-macro**	**Accuracy**	**Loss**
(1) Multi-task wav2vec - SMOTE + CG	0.7492	9.3794	0.3163
(2) Multi-task all prosodic + SMOTE + CG	0.8145	8.3135	0.3054
(3) Multi-task all prosodic + SMOTE - CG	0.8234	8.9146	0.3115
(4) Multi-task wav2vec + SMOTE - CG	0.8498	7.3750	0.3413
(5) Multi-task wav2vec + SMOTE + CG	0.8882	5.6187	0.2786
(6) Multi-task all prosodic and wav2vec + SMOTE + CG	0.8941	5.3127	0.2770
(7) Multi-task all prosodic and wav2vec + SMOTE - CG	**0.9052**	4.9903	0.2733
2nd submission - Model (7) (full dataset)	**0.5301**		

3.3 Feature Selection Results. We focused on twelve experiments to evaluate small and focused feature sets, shown on Table 4, in crescent order of F1-macro values.

3.4 Preliminary Evaluation of the Selected Models. Table 5 shows the confusion matrices for the first fold (20% of data), related to the three selected models. In the matrices, rows are termed as actual/true class and columns are termed as a predicted class. For the three selected models, the neutral class had the worst performance. It seems that the auxiliary task (genre classification from speech) has helped in classifying non-neutral male and non-neutral female classes.

4. Conclusions and Future Work

In this work, we evaluate 26 DNN models, using 5-fold cross-validation over the training dataset, and submitted our best models, i.e. those with higher F1-macro, for each group of experiments in Sections 3.1, 3.2, and 3.3. One of the submitted models surpassed the prosodic features baseline, reaching 0.5353 F1-macro. As a future work, we will perform an error analysis to

Table 4

Feature Selection results using 5-fold cross-validation. We indicate in the model's name which feature set was used and whether a data augmentation technique was used (+) or was not used (-). The last line indicates the value of F1-macro for the submitted model, using the full dataset.

	Average scores for all 5 folds		
Model name/feature sets/data aug techniques	**F1-macro**	**Accuracy**	**Loss**
(1) Seq. 8 prosodic - SMOTE + CG	0.2917	77.9714	0.1230
(2) Seq. 8 prosodic + SMOTE + CG	0.3216	43.5904	0.2356
(3) Seq. 19 prosodic + SMOTE - CG	0.4022	52.5366	0.2077
(4) Seq. 8 prosodic + SMOTE - CG	0.4109	48.8061	0.2382
(5) Seq. 19 prosodic + SMOTE + CG	0.5315	0.1684	0.1684
(6) Multi-task. 8 prosodic - SMOTE + CG	0.7261	9.4466	0.3167
(7) Multi-task 8 prosodic + SMOTE - CG	0.7440	11.7683	0.3622
(8) Multi-task 19 prosodic + SMOTE - CG	0.7835	10.5195	0.3275
(9) Multi-task. 8 prosodic + SMOTE + CG	0.7917	9.3354	0.3156
(10) Multi-task 19 prosodic + SMOTE + CG	0.8172	8.6319	0.3086
(11) Multi-task 10 prosodic + SMOTE + CG	0.8214	8.4761	0.3070
(12) Multi-task 29 prosodic + SMOTE + CG	**0.8266**	8.2014	0.3265
3rd submission Model (12) (full dataset)	**0.4696**		

Table 5

Confusion Matrices generated in the first iteration for the first fold (20% of data), during the training of the models. (N = Neutral, M = Non-Neutral Male, F = Non-Neutral Female).

	N	**F**	**M**
Seq. all prosodic and wav2vec + SMOTE + CG:			
N	71	27	8
F	2	87	1
M	1	11	87
Multi-task all prosodic and wav2vec + SMOTE - CG:			
N	82	20	5
F	5	88	3
M	1	0	90
Multi-task 29 prosodic + SMOTE + CG:			
N	68	18	9
F	8	100	0
M	2	5	85

understand why our best submitted model had a good performance on the training dataset, but only a 0.5353 F1-macro value on the test set.

Acknowledgments

This research was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant 2019/07665-4) and by the IBM Corporation.

References

[1] M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication 116 (2020) 56–76. doi:https://doi.org/10.1016/j.specom.2019.12.001.

[2] P. Ekman, H. Oster, Facial expressions of emotion, Annual Review of Psychology 30 (1979) 527–554.

[3] G. A. Campos, L. da S. Moutinho, DEEP: Uma arquitetura para reconhecer emoção com base no espectro sonoro da voz de falantes da língua portuguesa, 2020. URL: https://bdm.unb.br/handle/10483/27583, january 18, 2022.

[4] S. Zhang, R. Liu, X. Tao, X. Zhao, Deep cross-corpus speech emotion recognition: Recent advances and perspectives, Frontiers in Neurorobotics 15 (2021).

[5] J. A. Russell, A. Mehrabian, Evidence for a three-factor theory of emotions, Journal of research in Personality 11 (1977) 273–294.

[6] T. Raso, H. Mello, M. M. Mittmann, The C-ORAL-BRASIL I: Reference corpus for spoken Brazilian Portuguese, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 106–113.

[7] I. Luengo, E. Navas, I. Hernáez, J. Sánchez, Automatic emotion recognition using prosodic parameters, in: INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005, ISCA, 2005, pp. 493–496. URL: http://www.isca-speech.org/archive/interspeech_2005/i05_0493.html.

[8] K. S. Rao, S. G. Koolagudi, R. R. Vempada, Emotion recognition from speech using global and local prosodic features, Int. J. Speech Technol. 16 (2013) 143–160.

[9] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, CoRR abs/2006.11477 (2020). arXiv:2006.11477.

[10] B. J. Abbaschian, D. Sierra-Sosa, A. Elmaghraby, Deep learning techniques for speech emotion recognition, from databases to models, Sensors 21 (2021).

[11] K. Zvarevashe, O. Olugbara, Ensemble learning of hybrid acoustic features for speech emotion recognition, Algorithms 13 (2020).

[12] X. Cai, J. Yuan, R. Zheng, L. Huang, K. Church, Speech Emotion Recognition with Multi-Task Learning, in: Proc. Interspeech 2021, 2021, pp. 4508–4512.

[13] N. K. Kim, J. Lee, H. K. Ha, G. W. Lee, J. H. Lee, H. K. Kim, Speech emotion recognition based on multi-task learning using a convolutional neural network, in: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 704–707. doi:10.1109/APSIPA.2017.8282123.

[14] Y. Li, T. Zhao, T. Kawahara, Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning, in: Proc. Interspeech 2019, 2019, pp. 2803–2807. doi:10.21437/Interspeech.2019-2594.

[15] M. Lech, M. Stolar, C. Best, R. Bolia, Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding, Frontiers in Computer Science 2 (2020).

[16] R. Caruana, Multitask learning, Machine Learning - Special issue on inductive transfer - Volume 28 (1997) 41–75.

[17] S. Ruder, M. E. Peters, S. Swayamdipta, T. Wolf, Transfer learning in natural language processing, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 15–18. doi:10.18653/v1/N19-5004.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: Synthetic minority over-sampling technique, J. Artif. Int. Res. 16 (2002) 321–357.

[19] P. Boersma, D. Weenink, Praat: Doing phonetics by computer, 2010. URL: http://www.praat.org/.

[20] B. Gianesi, S. Aluisio, Classificação de gênero via análise de áudio utilizando métodos de aprendizado de máquina tradicionais, 2021. URL: https://github.com/BrunoGianesi/Speaker-Gender-Recognition, To appear in https://eesc.usp.br/biblioteca/.

[21] V. F. S. Alencar, A. Alcaim, LSF and LPC - Derived Features for Large Vocabulary Distributed Continuous Speech Recognition in Brazilian Portuguese, in: 2008 42nd Asilomar Conference on Signals, Systems and Computers, 2008, pp. 1237–1241.

[22] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, R. Collobert, MLS: A Large-Scale Multilingual Dataset for Speech Research, in: Proc. Interspeech 2020, 2020, pp. 2757–2761.

[23] D. Liang, E. Thomaz, Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 3 (2019). URL: https://doi.org/10.1145/3314404. doi:10.1145/3314404.

[24] G. Lemaître, F. Nogueira, C. K. Aridas, Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (2017) 559–563.

[25] J. Pittam, K. R. Scherer, Vocal expression and communication of emotion, in: M. Lewis, J. M. Haviland (Eds.), Handbook of emotions, The Guilford Press, New York, 1993, pp. 185–198.

[26] K. R. Scherer, Vocal affect expression: a review and a model for future research, Psychological Bulletin 99 (1986) 143–165.

[27] P. A. Barbosa, Detecting changes in speech expressiveness in participants of a radio program, in: INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, ISCA, 2009, pp. 2155–2158.

[28] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44 (2011) 572–587.

[29] J. t'Hart, R. Collier, A. Cohen, A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody, Cambridge Studies in Speech Science and Communication, Cambridge University Press, 1990. doi:10.1017/CBO9780511627743.

[30] R. Banse, K. R. Scherer, Acoustic profiles in vocal emotion expression., Journal of personality and social psychology 70 (1996) 614–36.

[31] T. Johnstone, K. R. Scherer, Vocal communication of emotion, in: M. Lewis, J. M. Haviland-Jones (Eds.), Handbook of emotions, 2 ed., The Guilford Press, New York, 2000, pp. 220–235.

[32] P. N. Juslin, P. Laukka, Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion, Emotion 1 (2001) 381–412.

[33] D. Guo, H. Yu, A. Hu, Y. Ding, Statistical analysis of acoustic characteristics of tibetan lhasa dialect speech emotion, SHS Web of Conferences 25 (2016) 1–5.

[34] S. Ruder, Neural Transfer Learning for Natural Language Processing, Ph.D. thesis, National University of Ireland, Galway, 2019.