Overview of the Automatic Speech Recognition for Spontaneous and Prepared Speech & Speech Emotion Recognition in Portuguese (SE&R) Shared-tasks at PROPOR 2022

Ricardo Marcacini[1], Arnaldo Candido Junior[2] and Edresson Casanova[1]

[1]*Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos - SP, Brazil*
[2]*Federal Univesity of Technology – Paraná, Avenida Brasil, 4232, Medianeira, Paraná, Brazil*

Abstract

The Automatic Speech Recognition for Spontaneous and Prepared Speech & Speech Emotion Recognition in Portuguese (SE&R 2022) challenge is a workshop consisting on two main tracks: Automatic Speech Recognition (ASR) for spontaneous and prepared speech for Portuguese; and Speech Emotion Recognition (SER) in Portuguese. This language still demands more resources for robust speech processing. To improve the research options, two corpora are proposed: CORAA ASR and CORAA SER. CORAA ASR contains 389 hours of spontaneous and prepared speech while CORAA SER is a 50 minute corpus for sentiment recognition. In this work, we present an overview of the challenge, discuss the submissions and present the obtained results. The best ASR model performance for CORAA ASR achieved an Character Error Rate of 10.98%, while the best model for CORAA SER achived 72.8% Macro-F1.

Keywords

Automatic Speech Recognition, Speech Emotion Recognition, Portuguese Processing.

1. Introduction

In this work, we present the challenge Automatic Speech Recognition for Spontaneous and Prepared Speech & Speech Emotion Recognition in Portuguese (SE&R 2022), a workshop presented in the 15th International Conference on Computational Processing of Portuguese (PROPOR 2022). The workshop consisted on two main tracks: Automatic Speech Recognition (ASR) for spontaneous and prepared speech for Portuguese; and Speech Emotion Recognition (SER) in Portuguese.

Our main objective in proposing this challenge was to promote research in Portuguese audio processing. While some languages as English have many available resources for audio processing, such as corpora, datasets, models and processing tools, other languages still lack in

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).

*Corresponding author.

†These authors contributed equally.

ricardo.marcacini@icmc.usp.br (R. Marcacini); arnaldocatutfprdotedudotbr (A. C. Junior); edresson@usp.br (E. Casanova)

0000-0002-2309-3487 (R. Marcacini); 0000-0002-5647-0891 (A. C. Junior); 0000-0003-0160-7173 (E. Casanova)

CEUR Workshop Proceedings (CEUR-WS.org)

this area. In particular, audio resource availability imposes limitations for conducting research in the Portuguese language. Although this scenario is gradually changing, as new corpora are released [1, 2, 3, 4], research obstacles due to resource scarcity are still a problem.

Two corpora were proposed for use in the challenge: CORAA ASR and CORAA SER. CORAA (Corpus of Annotated Audios) are a group of resources to foster research in spoken Portuguese processing. A focus is given on the Brazilian Portuguese variant. CORAA ASR corpus contains 389 hours of spontaneous and prepared speech, segmented at utterance level, together with the respective transcriptions for each utterance. CORAA SER is a 50 minute corpus for speech emotion recognition containing utterances and their respective polarities or valencies. Three categories were proposed: neural; non-neutral female; non-neutral male.

For the ASR track, participants should submit models to be evaluated against the CORAA ASR test set (11.3 hours). Researches could use CORAA ASR training and development sets, as well as external corpora. A baseline model based on Wav2vec 2.0 [5, 6] was also made available, although participants could also use other models.

For the SER track, participants should submit models to be analyzed against 308 audios belonging in the CORAA SER test set. The remaining of CORAA SER could be used to train and validated the model. As in the ASR track, external resources could also be used. Two baselines were provided as a starting point, allowing authors to improve the models or the use other models.

This work is organized as follows. Section 2 present details about the the SE&R ASR track and received submissions. Section 3 contain information and results for the SE&R SER track. Section 4 presents the concluding remarks.

2. Automatic Speech Recognition

Automatic Speech Recognition is a complex task, presenting many challenges for speech based applications due to a different number of reasons. A first challenge is that modern speech modeling usually requires large portions of data in order to a model perform in a satisfactory way. A second challenge are additional complexities in spoken language when compared to the written variant. For example, utterance borders in spoken text are normally less clear than in written text. A third issue are external factors such as environmental noises and recording quality are also a concern for ASR systems. In the case of spontaneous utterances, a fourth problem are difficulties arising due to phenomena including voice overlapping, laughs, sentence reformulations and disfluencies (such as stuttering, filled pauses and hesitations). Finally, a fifth challenge ASR systems must face is accounting for the mapping of speech phonetics to orthographic written systems. This mapping is demanding because of phenomena as loanwords, acronyms, neologisms and rare proper nouns and orthographic irregularities regarding pronounce.

These challenges have slowed the adoption of voice based interfaces and applications, although advances in the area have been made, for example personal digital assistants and closed caption systems for television and streaming services. However, several languages are lacking in resources to build such systems, or the systems do exist, but they are propriety alternatives. For the Portuguese Language, open resources are becoming available. In 2020, three new datases were released: BRSD v2 [1] ; Multilingual LibriSpeech (MLS), which includes Portuguese [2]; and Common Voice version 6.17 [3]. In 2021, Multilingual TEDx Corpus [4] was released. These resources encompass more than 574 hours of audio in Portuguese. However, there is still the necessity of more data for the ASR task, particularly, regarding spontaneous speech, since the existing resources consists mostly of prepared speech.

CORAA ASR corpus [7] and SE&R 2022 ASR track are initiatives aimed at fomenting speech related research in Portuguese processing. It contains both prepared and spontaneous speech. The second is more challenging for systems due to the characteristics proper of this speaking style.

2.1. Dataset and baseline

CORAA ASR is composed of five corpus: (a) ALIP [8]; (b) C-ORAL-BRASIL I [9]; (c) Nurc-Recife [10]; (d) SP2010 [11]; and (e) TeDx Talks[1] in Portuguese. Regarding composition, TeDx Talks are composed of prepared talks, while Nurc-Recife contains both prepared and spontaneous speech. The remaining corpora contain spontaneous speech.

During CORAA ASR creation, all corpora but TeDx Talks had existing transcriptions. Previous transcriptions were adapted to the ASR task by annotators which manually validated and categorized them, indicating audio quality, presence of noise, more than one speaker in the audio, among other data [7]. When needed, automatic alignment between transcriptions and segmented utterances were performed. Annotators also marked sentences for revision where problems were found. TeDx Talks were transcribed for the first time. In this case, transcriptions of numerals, acronyms, dates, loanwords among other related phenomena were guided to a transcription manual specifically designed for the ASR task.

The resulting dataset contains 289 hours of audio and transcriptions, with more than 2.7 million tokens and 58 thousand types. In total, the corpus have more than 400 thousand segmented sentences with duration on average of 3.4 seconds. For the challenge, we categorized the corpus into prepared and spontaneous speech. Additionally, we also categorized prepared speech into European Portuguese (approximately 4.6 hours) and Brazilian Portuguese (the remaining audios).

For the baseline, we used Wav2Vec 2.0 XLSR-53. The model was fine-tuned for the version 1.1 of CORAA ASR. The corpus was divided into three sets: train (283.6 hours), development (5.7 hours) and test (11.6 hours). We trained the model for 40 epochs freezing its feature extractor. More details of the training phase can be obtained at [7].

2.2. Results

Table 1 presents ASR track submissions, baselines and results. The submitted models were evaluated into four categories. In the mixed category, all CORAA ASR test set were used. Spontaneous speech category used only the audios in the test set labeled as spontaneous speech (only Brazilian Portuguese audios). Finally, the prepared speech category was subdivided into Brazilian Portuguese and European Portuguese and evaluated accordingly. The performance

[1] https://www.ted.com/

was evaluated mainly using CER (Character Error Rate), although WER (Word Error Rate) is also presented.

Table 1
ASR Track Results SE&R 2022.

	Team	CER	WER	Category
1º	GPED	10.9884	24.8916	
2º	dovahkiin	11.1568	21.9077	Mixed
3º	Baseline	11.3593	25.8593	
1º	GPED	3.5503	11.2508	
2º	Baseline	3.5635	11.1955	PT-br Prepared Speech
3º	dovahkiin	4.2229	10.4440	
1º	GPED	14.9288	31.5125	
2º	dovahkiin	16.0329	32.3271	PT-pt Prepared Speech
3º	Baseline	17.0861	39.7575	
1º	dovahkiin	12.1857	22.4298	
2º	Baseline	12.3939	26.2421	PT-br Spontaneous Speech
3º	GPED	12.5115	26.5006	

Two models were submitted to the ASR track: team GPED and team Dovahkiin. Only GPED submitted the paper detailing the model. Overall, GPED performed better, winning in three categories, while Dovahkkiin achieved best results against European Portuguese. We used a strong baseline, which obtained rank two in two evaluated categories. The winning system used open set, being trained on other corpora besides CORAA ASR, and applied the strategy of generating domain specific models for the four proposed categories.

The CERs and WERs observed tend to be higher than systems in other languages or in other corpora for Portuguese. It is important to note that some of our subcorpora consists of noisy audios, imposing some limitations in system performance. Pt-br prepared speech lead to the smaller errors, as these speech style is easier to be processed than spontaneous. Pt-pt prepared speech lead to higher errors. This is probably due the few audio examples for this language variant. Spontaneous speech lead to mixed results despite the fact of being more challenging and presenting more noisy. This implies the models adapted well for this speech style.

3. Speech Emotion Recognition for Brazilian Portuguese

Speech Emotion Recognition (SER) is an increasingly relevant task for Human-Computer Interaction [12] and an active research area in information retrieval and natural language processing. The general idea is to promote the interaction between machines through voice conversation [13], which is potentially useful for medical applications, call centers, autonomous vehicles, and personal digital assistants, among others. Recognizing the speaker's emotional state is a critical bridge that challenges the interaction between man-machine because the speech information can be interpreted in different ways according to the state of the speaker's voice, such as surprise, anger, joy, and sadness [14].

Although Speech Emotion Recognition has made promising advances in the English language, which has many resources and labeled corpus for training models, we observe that this task is still underexplored in the Portuguese language due to the lack of labeled corpus [15]. Another important aspect is the type of Speech Emotion Recognition, which can be prepared speech or spontaneous speech. In a prepared speech, actors and actresses record their voices from a studio, without noise and using pre-defined scripts, usually the exact phrase being spoken with different categories of emotion. Voices are generally recorded considering six types of primary emotions: happiness, sadness, disgust, anger, fear, and surprise. However, models trained on these corpora usually fail in real-world applications with ambient noise, pronunciation, and accents. Spontaneous speech corpora are relevant in these scenarios as they have these patterns intrinsic to real-world applications. On the other hand, it is more challenging to annotate audio segments due to the difficulty in finding various emotional states of the speaker.

We aim to mitigate the lack of spontaneous speech corpus for Brazilian Portuguese. To the best of our knowledge, we present the first initiative of an annotated corpus of spontaneous speech for Brazilian Portuguese. We used the C-ORAL Brasil I corpus [16] and its paralinguistic annotations, such as laughter, crying, screaming, etc., to identify potential audio snippets with an emotional state other than neutral. In addition, we also use gender metadata, such as male and female, to identify the speaker's gender when switching from a non-neutral to a neutral emotional state.

We also proposed the Brazilian Portuguese Speech Emotion Recognition (SER) Task to encourage the first models to be trained with the CORAA SER v1 corpus. This task aims to motivate research for SER in our community, mainly to discuss theoretical and practical aspects of Speech Emotion Recognition, audio pre-processing, feature extraction, and machine learning models for Brazilian Portuguese. We provide a dataset called CORAA SER version 1.0[2] composed of approximately 50 minutes of audio segments labeled in three classes: neutral, non-neutral female, and non-neutral male. While the neutral class represents audio segments with no well-defined emotional state, the non-neutral classes represent segments associated with one of the primary emotional states in the speaker's speech.

3.1. Dataset and Baseline Models

We provide a training corpus with audio segments labeled in three categories: neutral (491 audios), non-neutral-female (89 audios), and non-neutral-male (45 audios).

The test file has 308 audios, organized in the same categories, with ground truth labels available only after the shared task: Neutral (248 audio files), non-neutral-female (37 audio files), and Non-neutral-male (23 audio files).

Contestants received the original dataset (raw wave files) as well as two pre-processed versions:

- **Prosodic features:** we use features related to physical characteristics of speech, such as intonation, rhythms, pitch, time, loudness, etc. This type of pre-processing is traditionally used in early versions of emotion recognition systems [17]. In total, 56 prosodic features were made available.

[2] https://github.com/rmarcacini/ser-coraa-pt-br

- **Wav2Vec features:** we explore unsupervised pre-training for speech recognition to extract features (i.e., embeddings) from the audio segments [18]. These features can be used for training a speech emotion recognition classifier. Wav2Vec has been a state-of-the-art deep learning model for the last five years. This model is trained on a large corpus of unlabeled audios through a noise contrastive binary classification task strategy.

To provide the baselines to the competitors, we trained two MLP classifiers (Multilayer Perceptron), one for each type of audio representation (prosodic features and wav2vec embeddings).

More details about the corpus are publicly available at https://github.com/rmarcacini/ser-coraa-pt-br.

3.2. Results

Table 2 presents emotion recognition performance results (Macro-F1 measure) for four competing teams, as well as two baseline models and one null/dummy model. It is worth mentioning that each team reported details of its implementation in its own paper.

Table 2
Overview of the results obtained in the Speech Emotion Recognition shared task for Spontaneous Speech in Brazilian Portuguese.

Team Name	Open-set	Macro-F1
IME-USP-FINGER	true	0.728
ICMC-EESC-FFLCH	true	0.535
LIA-UFMS	true	0.525
SofiaFala	false	0.509
MLP-Wav2Vec	false	0.536
MLP-Prosodic	false	0.534
Dummy classifier	false	0.300

The winning team (IME-USP-FINGER) proposed a deep neural model based on pre-training and transfer learning. In this case, pre-training and transfer learning are promising ways to mitigate the small number of labeled audios. Moreover, each team was successful in an area of action relevant to the SER community, such as statistical analysis of the corpus (ICMC-EESC-FFLCH), committee evaluation (LIA-UFMS), and potential practical applications in speech disorder (SofiaFala).

4. Concluding Remarks

In this work we presented SE&R 2022, a challenge to stimulate research in Portuguese Speech processing. Two tracks were defines: Automatic Speech Recognition for spontaneous and prepared speech for Portuguese and Speech Emotion Recognition. Two corpora, CORAA ASR and CORAA SER were available for participants. The best ASR model performance for CORAA ASR achieved an Character Error Rate of 10.98%, while the best model for CORAA SER achived 72.8% Macro-F1.

We believe Portuguese speech processing it is an important and active area, and that initiatives like CORAA ASR and CORAA SER will help to develop the area. As future work, we plan to increase the presented corpora by collecting and annotating new audios.

Acknowledgments

This work 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] 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. doi:10.1109/ACSSC.2008.5074614.

[2] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, R. Collobert, Mls: A large-scale multilingual dataset for speech research, Interspeech 2020 (2020). URL: http://dx.doi.org/10.21437/Interspeech.2020-2826. doi:10.21437/interspeech.2020-2826.

[3] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4218–4222. URL: https://www.aclweb.org/anthology/2020.lrec-1.520.

[4] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, M. Post, The multilingual tedx corpus for speech recognition and translation, CoRR abs/2102.01757 (2021). URL: https://arxiv.org/abs/2102.01757. arXiv:2102.01757.

[5] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, É. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.

[6] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in Neural Information Processing Systems 33 (2020).

[7] A. Candido Junior, E. Casanova, A. Soares, F. S. de Oliveira, L. Oliveira, R. C. F. Junior, D. P. P. da Silva, F. G. Fayet, B. B. Carlotto, L. R. S. Gris, et al., Coraa: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese, arXiv preprint arXiv:2110.15731 (2021).

[8] S. C. L. Gonçalves, Projeto ALIP (amostra linguística do interior paulista) e banco de dados iboruna: 10 anos de contribuição com a descrição do português brasileiro, Revista Estudos Linguísticos 48 (2019) 276–297.

[9] T. Raso, H. Mello, C-oral - Brasil I: Corpus de Referência do Português Brasileiro Falado Informal, Editora UFMG, Belo Horizonte, MG, 2012.

[10] M. Oliviera Jr., Nurc digital um protocolo para a digitalização, anotação, arquivamento e disseminação do material do projeto da norma urbana linguística culta (nurc), CHIMERA:

Revista de Corpus de Lenguas Romances y Estudios Lingüísticos 3 (2016) 149–174. URL: https://revistas.uam.es/chimera/article/view/6519.

[11] R. B. Mendes, L. Oushiro, Mapping paulistano portuguese: the sp2010 project, in: Proceedings of the VIIth GSCP International Conference: Speech and Corpora, Fizenze University Press, Firenze, Italy, 2012, pp. 459–463.

[12] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, T. Alhussain, Speech emotion recognition using deep learning techniques: A review, IEEE Access 7 (2019) 117327–117345.

[13] H. M. Fayek, M. Lech, L. Cavedon, Evaluating deep learning architectures for speech emotion recognition, Neural Networks 92 (2017) 60–68.

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

[15] J. R. Torres Neto, L. Y. Mano, J. Ueyama, et al., Verbo: voice emotion recognition database in portuguese language, Journal of Computer Science 14 (2018) 1420–1430.

[16] 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), 2012, pp. 106–113.

[17] K. S. Rao, S. G. Koolagudi, R. R. Vempada, Emotion recognition from speech using global and local prosodic features, International journal of speech technology 16 (2013) 143–160.

[18] S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: Unsupervised pre-training for speech recognition, Proc. Interspeech 2019 (2019) 3465–3469.