Accent Classification is Challenging but Pre-training Helps: a case study with novel Brazilian Portuguese datasets

Ariadne Nascimento Matos
ICMC – Universidade de São Paulo, Brazil
ariadnenmtos@usp.br

Gustavo Evangelista Araújo
ICMC – Universidade de São Paulo, Brazil

Arnaldo Candido Junior
IBILCE - Universidade Estadual Paulista

Moacir Antonelli Ponti
ICMC – Universidade de São Paulo, Brazil

Abstract

Accents arise due to variations in pronunciation, intonation, and other speech characteristics caused by geographical, cultural, or linguistic differences. Investigating accent classification methods is a way towards accent-aware speech-processing. This paper evaluates accent classification for spontaneous speech using CNN-LSTM networks and the Wav2vec2 model. We study the importance of dataset size, pre-trained models, and external validation. For that we used 90 hours of data, encompassing 9 accents and involving 204 speakers of Brazilian Portuguese, obtained from manually annotated subsets from Spotify Podcasts [1] and CORAA ASR. Our best results range from 82% (closed-dataset) to 75% (cross-dataset) f1-scores for binary classification. Unless there is speaker leakage from training to testing, accent classification models trained from scratch fail for spontaneous speech data. Therefore, methods should be evaluated using both out-of-speaker and cross-dataset scenarios. We contributed with an experimental protocol for this task with a novel dataset. Finally, our results highlight the value of larger accent-annotated datasets, and the use of larger pretrained-models.

1 Introduction

Speech is a fundamental form of human communication, allowing expressing ideas and information. Automatic methods for processing and understanding speech are a relevant subject of study. Machine learning techniques are shown to be particularly useful in this scenario, becoming the state of the art in many speech processing tasks (Casanova et al., 2023, 2022). The two most remarkable tasks in this context are Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) systems. One of the challenges in ASR and TTS is how to deal with different accents of a given language, which can significantly impact the system's performance.

[1] https://github.com/aryamtos/spotify-subset

Accents arise due to variations in pronunciation, intonation, and other speech characteristics caused by geographical, cultural, or linguistic differences (Lippi-Green, 2012). There are two different types of accents: the first refers to foreignness, which occurs when a person speaks a language using rules and sounds of another language, and the second occurs within the native language itself (Teixeira et al., 1996). This paper aims at the automatic classification of the second type of accent.

Based on the dialectical division proposed by Nascentes (1953), Brazil is divided into two linguistic groups, related to Northern and Southern regions. The Northern region has specific phonological and morphological features, such as pretonic vowels, with a greater oral aperture facilitating airflow, in contrast to the closed vowel pronunciation typical of the Southern and Southeastern regions.

According to Ilari and Basso (2009), the regional characteristics of Brazilian Portuguese are distinguished by various pronunciation features. One notable feature is the absence of palatalization in the pronunciation of /t/ and /d/, a phenomenon widespread throughout Brazil except in São Paulo and the southern region. Additionally, the retroflex pronunciation of /r/ is a distinctive trait observed in the "caipira dialect" (Ilari and Basso, 2009). Previous accent classification methods also follow this definition (Batista et al., 2018; Batista, 2019). We focus on matching the accents within different Brazilian states, prioritizing the ones with the most data available. By that, we expect to offer a model that could fit in different dialectical divisions. When considering states within the North and South, we are offering a more fine-grained classification of Brazilian Portuguese accents.

The variations caused by accents can result in differences in acoustic features, such as the spectral content and timing of speech signals (Hansen et al., 2020). Amplitude modulations of the envelope with different timescales are also associated with

accent variations (Frota et al., 2022).

Accent classification is a relevant problem since it allows to better understand language variations, in particular for low resources languages, such as Portuguese. Also, ASR and TTS methods typically rely on models trained on large annotated speech data to transcribe spoken words and synthesize them, respectively. Improving the quality of accent classification is important towards accent-aware ASR and TTS systems (Deng et al., 2021).

1.1 Goals

We aim to study the difficulty of the accent classification task in Brazilian Portuguese considering realistic scenarios. In particular, we propose the use of novel datasets involving spontaneous speech under different recording setups (based on Spotify Podcasts (Tanaka et al., 2022) and CORAA (Candido Junior et al., 2021) datasets). With those datasets, we evaluate models and strategies often employed in the recent literature under such tasks.

Two scenarios are investigated: closed-dataset validation (training and testing carried out in the same dataset) and cross-dataset validation (training carried out in one dataset, and testing in a different dataset). Those are also referred to as closed-set and cross-dataset scenarios, respectively, by Batista et al. (2018); Batista (2019). For that, we apply different data validation scenarios, aiming to evaluate their generalization capacity.

In terms of the models, we use as a baseline a CNN-1D+LSTM (One-dimensional Convolutional Neural Network with Long-Short Term Memory) which was the winning model as reported by Tostes et al. (2021) and also finetune a pretrained Wav2Vec 2.0 Large XLSR as it was shown potential in other languages (Zuluaga et al., 2023).

1.2 Contributions

The main contributions of this work are: (1) Organization of two subsets of dataset Spotify Podcasts, consisting of approximately 90 hours of audio recordings (spontaneous speech) from 204 speakers representing 9 Brazilian states; (2) the study of different closed-dataset and cross-dataset settings, which allows drawing important conclusions on the difficulty of the task, and provides insights towards better ways to solve the problem under a more realistic scenario.

2 Related Work

The study of Batista (2019); Batista et al. (2018) presented the first neural accent classification model for Brazillian Portuguese. It employed statistical modeling approaches, including Gaussian mixtures and machine learning techniques. The authors developed the Braccent dataset to represent the 7 accents found in Brazil, namely: baiano, carioca, fluminense, mineiro, nordestino, nortista, and sulista. The dataset consisted of 1,757 online-collected read speech audio samples, each ranging from 8 to 14 seconds in duration. Additionally, the same study utilized the Ynoguti dataset (Ynoguti, 1999) to represent the 5 accents (baiano, nordestino, mineiro, fluminense e sulista) and the Forensic Corpus of Brazilian Portuguese (CFPB - Corpus Forense do Português Brasileiro), covering respectively accents of Braccent. The study employed two validation scenarios: closed set and cross-dataset. In the closed-dataset scenario, Batista achieved an f1-score of 91%. In the cross-dataset, most showed results below 50%. The author emphasizes the importance of validating the models using additional datasets to evaluate their performance but did not offer alternatives on how to improve cross-dataset performance.

The work of Tostes et al. (2021); Tostes (2022) applied different architectures for accent classification based on the Braccent and Ynoguti datasets. Their best results were achieved with a hybrid neural network, combining one-dimensional (1D) Convolutional Neural Networks (CNN) and a Long-Short Term Memory Neural Network (LSTM). They obtained an f1-score of approximately 88% in a closed-dataset validation (Tostes, 2022).

Later, de Almeida (2022) compared the results of Tostes et al. (2021) and Batista (2019); Batista et al. (2018) accent classification models for Brazilian Portuguese. They utilized both Multiclass Logistic Regression and fine-tuning of a pret-rained Wav2vec 2.0 base model using the Braccent dataset. The results showed that Wav2vec 2.0 achieved an overall accuracy of 69% and an f1-score of 38%, while Multiclass Logistic Regression only achieved an accuracy of 39%. The authors also carried out an analysis of gender, but could not find performance differences between gender-specific models and gender-agnostic ones. The author emphasized the importance of evaluating these models with other datasets and extending experiments with pretrained models for Portuguese. Interestingly, in

other languages such as English, Italian, German, and Spanish, a recent study found large pre-trained models to be good candidates for transfer learning to the accent classification (Zuluaga et al., 2023).

The limitations of the aforementioned studies include the use of read speech audio samples (not spontaneous) and similar recording setups. Also, only one of them explicitly evaluated a cross-dataset scenario without succeeding in it. Additionally, previous studies evaluated different models and strategies but their conclusions are difficult to generalize into guidelines for future work.

In light of such gaps, our paper proposes a larger dataset, with audio data closer to real-world speaking style, encompassing a more extensive collection of audio data of various accents. This allows for a more comprehensive exploration of accent variations and enhances the robustness of the models. We manually collected audio samples from a diverse dataset (Spotify Podcasts (Tanaka et al., 2022) and CORAA (Candido Junior et al., 2021)). Different than Braccents, Ynoguti's, and CFPB (Corpus Forense do Português Brasileiro) datasets, the accents in Spotify Podcasts and CORAA ASR are not self-declared. Consequently, we do not follow the accent annotation presented in the related works but consider geographic information of the speaker's present state. Also, besides using the best model reported in the literature (CNN1D-LSTM), we apply a pre-trained model Wav2vec 2.0 large XLSR for accents classification. Unlike the Wav2vec 2.0 base used by de Almeida (2022), the XLSR is multilingual and it is larger, which we show to better suit the task at hand.

3 Materials and Methods

Figure 1 illustrates the overall methodology, including preprocessing, model training, and conducting the evaluation, detailed in the following sections.

3.1 Datasets

Since Braccents, CFPB, and Ynoguti's datasets used by Batista (2019); Batista et al. (2018) and Ynoguti (1999) were not publicly accessible, we look into alternative datasets for pt-BR accent classification. As Batista et al. (2018) emphasizes the importance of validating models in more than one source of data, our study includes two datasets: Spotify Podcasts[2] (Tanaka et al., 2022)

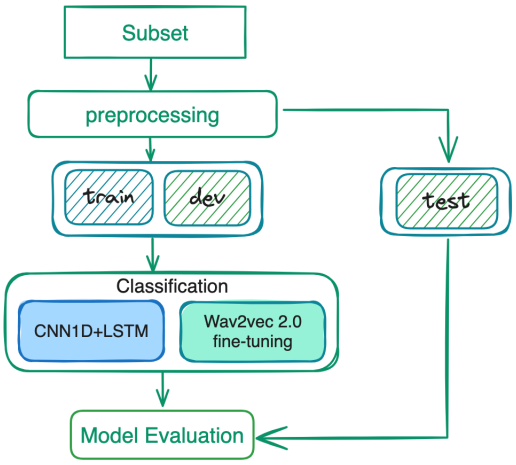

Figure 1: Overall methodology

and CORAA ASR[3] (Candido Junior et al., 2021). Those datasets offer an opportunity to explore new challenges and push the boundaries of accent classification algorithms since those have more data (in hours and speakers), and also have audio recorded under different conditions. Therefore, extensive preprocessing, i.e. selecting and cleaning audio, was necessary in order to make those datasets aligned to our aims.

Spotify Podcasts: proprietary dataset (available for research by request) which consists of around 123,000 episodes in both pt-BR and pt-PT, encompassing more than 76,000 hours of speech audio (Tanaka et al., 2022). This is an interesting dataset since podcasts are a growing form of mass communication and exhibit diverse formats and levels of formality, which can adopt various tones, from formal to informal, and encompass conversational exchanges or monologues. The most popular topics include business, education, and sports. Each audio file has an equivalent XML file that provides more specific metadata information, such as author, episode content, RSS links, and, in some cases, the recording location. Some podcasts are conducted by more than one person, including guests, while others have only one host.

The Corpus of Annotated Audios for ASR (CORAA ASR): public dataset for automatic speech recognition in Brazilian Portuguese (Candido Junior et al., 2021). It contains around 290.77 hours of audio and transcriptions. This dataset is a compilation from five other projects: ALIP (Gonçalves, 2019), C-ORAL Brasil I (Raso and Mello, 2012), NURC Recife (Oliviera Jr et al.,

[2]https://podcastsdataset.byspotify.com/

[3]https://github.com/nilc-nlp/CORAA

2016), SP2010 (Mello et al., 2012) and TEDx talks in Portuguese. CORAA audios were validated by annotators and transcriptions were adapted for ASR. Differently than Spotify Podcasts, it provides annotations for the regions: Minas Gerais (MG), Recife (RE), São Paulo cities (SP), São Paulo capital (sp-SP), or miscellaneous for unidentified accents. The speaking style varies from spontaneous, prepared, and read speech, from the genres: interviews, dialogues, monologues, conversations, conferences, class talks, reading, and stage talks.

3.2 Data subsets

We curated subsets from the Spotify and CORAA datasets to investigate accent classification. For the Spotify Podcasts subset, we manually selected audio episodes likely to feature Brazilian accents based on speaker geographic data in the metadata such as "Rádio Manaus" and "Puc Minas" or idiomatic expressions indicative of a specific accent (e.g., "Bah", "Oxe") in the episode description. The description field was used to confirm speaker location, since it sometimes included guest names. Prior research was conducted to ensure these speakers were indeed from the identified state.

We evaluated two scenarios: one involving a limited number of speakers (Spotify-A) and another with a larger number of speakers (Spotify-B). In both scenarios, we conducted both closed-dataset and cross-dataset evaluations.

Subset Spotify-A: The initial subset, described in Table 1, emphasized episodes with solo speakers to reduce the potential impact of other speakers' accents in the audio recordings. Diarization was not applied to this subset, and all audio clips were trimmed to 10 seconds.

Subset Spotify-B: In this subset, detailed in Table 2, we selected only two classes for model evaluation: São Paulo (SP) and Pernambuco (PE). Since many podcasts featured more than two speakers, diarization was performed, resulting in a significant number of speakers per podcast.

For the CORAA ASR subset, detailed in Table 3, we took into account the availability of accent annotations, excluding audio files as miscellaneous accents (unknown classes). This was needed to ensure compatibility between the classes observed in training with the Spotify Podcasts dataset and the subsequent testing phase with CORAA.

As a result, it was possible to obtain audio samples from various locations across Brazil, including Bahia (BA), Amazonas (AM), Maranhão (MA), Mato Grosso do Sul (MS), Minas Gerais (MG), Pernambuco (PE), Rio de Janeiro (RJ), Rio Grande do Sul (RS), and São Paulo capital (sp-SP). Table 1 and Table 2 present specific information such as the number of episodes per accent and duration in hours. To facilitate the reproducibility of the results and provide access to the specific shows and episodes used in the study, a table containing the identifiers of the selected shows and episodes from both the Spotify Podcasts and CORAA ASR datasets is available (omitted due to blind revision). This table serves as a reference for other researchers who seek to replicate the findings or conduct further investigations using the same datasets.

Accent	segments	Hours	Speakers
AM	487	~ 1	$2-3$
BA	625	~ 1	1
MA	1,326	~ 3	$2-3$
MS	109	~ 0.3	$1-2$
MG	2,461	~ 5	$2-3$
PE	1,624	~ 4	$1-3$
RJ	284	~ 0.8	$1-3$
RS	402	~ 1	$1-2$
sp-SP	464	~ 1	$1-3$
Total	7,782	~ 17	~ 23

Table 1: Total subset Spotify-A Information

Accent	segments	Hours	Speakers
PE	14,008	~ 48.23	102
SP	11,906	~ 30.88	85

Table 2: Total subset Spotify-B information

3.3 Preprocessing

Our preprocessing steps were defined to be consistent with related works as best as possible:

(1) Audio Conversion and Resampling: converted `.ogg` audio files into `.wav`, and resampled the audio to a 16kHz sample rate, ensuring uniformity across the dataset;

(2) Data Cleaning: used a threshold-based si-

Accent	Segments	Hours
subset CORAA ASR		
PE	353	~ 0.9
SP	371	~ 1
MG	351	~ 0.6

Table 3: Total Subset CORAA-ASR Information

lence removal step, and employed Spleeter[4] (Hennequin et al., 2020) to separate the speakers' voices from the music, which conveniently offers pretrained models for this purpose;

(3) Audio Trimming: due to computational cost of training models in higher time instances, we trimmed the audio to approximately 10 seconds sentences, following (Tostes et al., 2021);

(4) Diarization and Transcription: since episodes in Spotify-B may have multiple speakers, we used Pyannote[5] with Whisper[6] (Radford et al., 2022) to obtain specific timestamps for each speaker and transcriptions for future ASR work;

(5) Spectrogram generation: via Short-Time Fourier Transform (STFT)[7] using the Librosa library. Specifically, we applied a window size of 3000 frames with a step size of 2000, following the guidelines of Tostes et al. (2021).

3.4 Train/Dev/Test splits

In the process of splitting the data into training, development, and test sets, we took careful consideration of the speakers' identities to prevent any contamination or bias in the evaluation. It is crucial to maintain speaker independence during this partitioning to ensure that the models are tested on unseen speakers, thereby providing a fair assessment of their generalization capabilities. For that, no Spotify-A the train includes 5,665 audio files, while the test has 2,117 audio files featuring different speakers and podcasts.

For the Spotify-B subset, out of the total shown in Table 2, approximately 50 distinct speakers were selected for each class during training. The data was split as follows. For PE class: 8,161 segments for training (train), 2,304 for development (dev) and 534 for testing (test); for SP class: 7,998 segments for training (train), 2,353 for development (dev) and 500 for testing (test).

3.5 Model

For accent classification, we selected two architectures: CNN1D LSTM (One-dimensional Convolutional Neural Network with Long-Short Term Memory) and Wav2vec 2.0 XLSR.

– **CNN1D LSTM**: This model was selected taking into consideration the work by Tostes et al. (2021) and also to assess the model's performance with other datasets. In this architecture, each frequency interval (97 timesteps per 2049 frequencies of the spectrogram) is used as input to a convolution layer, generating feature vectors that serve as the input for the LSTM units. A rate of 0.4 is used in the Dropout layer. Finally, a series of fully connected layers are responsible for the classification. In terms of training strategies, we employed the Adam optimizer with an initial learning rate of 0.0001 and decay of 0.001 using the Cross-Entropy loss. The model was trained with a minibatch size of 64 for a maximum of 50 epochs, employing early stopping with patience 25 for the development loss.

– **Wav2vec 2.0 XLSR-53**: This model was chosen to assess its performance in the classification task, especially considering that de Almeida (2022) work utilized the Wav2vec 2.0 base model. We aimed to determine if a model specifically fine-tuned with Portuguese data could yield improved results. Consequently, we conducted fine-tuning based on previous research.

Due to computational constraints, we limited fine-tuning of these models on Spotify-B to 5 epochs. Following the methodologies of Gris et al. (2021) and Conneau et al. (2020), we chose to keep the base model frozen. We introduced a dense layer with 1024 neurons and tanh activation followed by a classification head. The training was carried out on GPU NVIDIA Titan RTX, with batch size 16, gradient accumulation over 4 steps, learning rate of 3e-5, and the Adam optimizer. Checkpoints were saved at regular intervals. The selection of the best checkpoint was based on the model's performance on a validation dataset.

3.6 Evaluation

We used normalized confusion matrices and the f1-score (weighted for the multiclass results). Each model was trained 5 times using different and fixed seeds (42, 101, 123, 1, 5) for binary classification using CNN-LSTM. All reported results are means and standard deviations of those 5 runs. For fine-tuning, we employed fixed seed 42.

4 Results and Discussion

We evaluate the models using closed-dataset validation, where training and testing occur on the same dataset, and cross-dataset validation, where testing is conducted on a dataset that was not part of the training data. We employed both the CNN1D LSTM architecture and Wav2vec 2.0 XLSR-53.

[4] https://github.com/deezer/spleeter
[5] https://github.com/pyannote/pyannote-audio
[6] https://github.com/openai/whisper
[7] https://librosa.org/doc/main/generated/librosa.stft.html

In the first part, we examined two scenarios using the CNN1D LSTM model: scenario A with a more limited number of speakers and multiple classes (9), and scenario B focusing on a binary classification task with a more extensive set of speakers. This way we can assess the impact of the number of available speakers in the results.

In the second part, only for the binary classification task, we employed the Wav2vec 2.0 XLSR-53 model with the Spotify-B dataset.

4.1 Experiments with Fewer Speakers (Spotify-A)

– **Random train/test split**: this experiment uses the whole subset Spotify-A (9 classes) – recall such subset has fewer speakers, ranging from 1 to 3 for each accent –, where each instance has an audio clip of 10 seconds. Then, we randomly defined the training and testing datasets without caring about the speaker, that is, different segments of a given speaker may fall in both training and testing sets. It is evident that, when the model sees all speakers during training instances, even if different segments are used in the testing stage, the performance is high. This result may not reflect the models' ability to learn the accents, but other features related to the speaker and the recording.

– **Out-of-speaker train/test split**: in order to evaluate the model's ability to generalize to unseen speakers within the same accent variation, we used the same Spotify-A subset, but now ensuring different podcasts and speakers are in the training, development and testing sets. This way we ensure that there is no leakage of speaker or recording. We carried out three experiments, varying the number of classes: (i) all available 9 classes, (ii) 3 classes: MG, SP, and PE, and (iii) binary SP, PE.

In Figure 2 we show the confusion matrices for the multiclass test results. Overall, the same model that had a great performance in the previous experiment, now cannot generalize in any scenario, achieving f1-scores 24% for the 9-class, 53% for the 3-class, and $34 \pm 11\%$ for the binary one. The results show a bias towards classes with a greater number of audio samples, like MG and PE, across all three experiments, with very few accurate predictions for the SP class.

Table 4 presents the results of binary classification using 5 different seeds. The accuracy for the "PE" accent is relatively high $(83 \pm 9)\%$, indicating that most positive classifications are correct. However, while positive classification is accurate, many

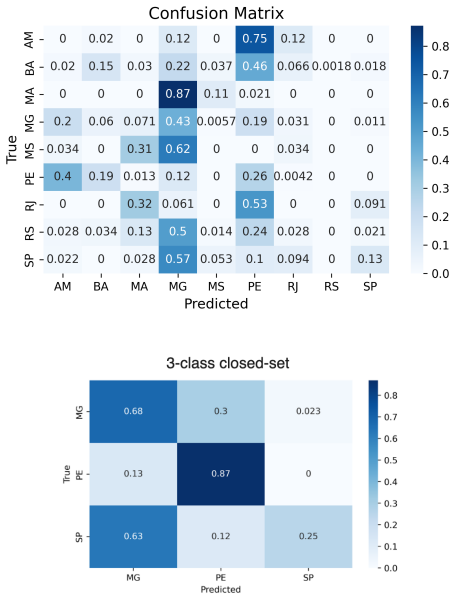

Figure 2: Confusion matrices for closed-dataset validation with unseen speakers and recordings using a test set of Spotify-A. Top: 9 classes, Bottom: 3 classes (MG, SP, PE)

class	Precision	Recall	F1-score
PE	$83 \pm 9\%$	$20 \pm 20\%$	28 ± 24
SP	$26 \pm 2\%$	$87 \pm 17\%$	40 ± 4
Overall			$34 \pm 11\%$

Table 4: Closed-dataset f1-scores for the Spotify-A dataset and the CNN-LSTM model (binary classification - PE, SP)

real examples of the "PE" accent are not correctly identified. On the other hand, for the "SP" class, we observe a high recall rate $(87 \pm 17\%)$, meaning that most real examples of the "SP" accent are correctly identified. However, the precision for this class is low. The overall F1-score for this classification was 34%, indicating an imbalance between the recall rate and precision.

– **Cross-dataset**: in this experiment, we train with all available Spotify-A data, and evaluate it on CORAA ASR as the test set. In Figure 3, the results showed a contrasting pattern compared to the cross-dataset validation for 3-class.

The model confuses Pernambuco (PE) with Minas Gerais(MG) and vice versa with an f1-score of 27%. Among the possible hypotheses to consider, besides class imbalance, are the characteristics of the speakers in each dataset. In the binary classification scenario (Table 5), the model misclassified PE as SP, with most results concentrated in that

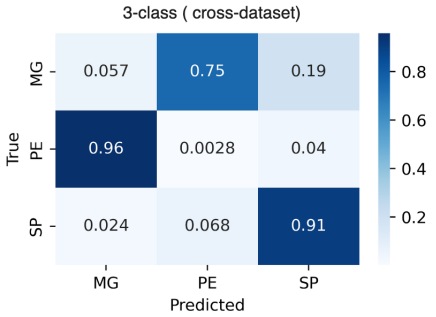

Figure 3: Confusion matrices for cross-dataset validation with unseen speakers and recordings using CORAA ASR as test set: (i) 3-class (MG, SP, PE).

class	Precision	Recall	F1-score
PE	$35 \pm 10\%$	$10 \pm 3\%$	15 ± 5
SP	$48 \pm 2\%$	$81 \pm 6\%$	60 ± 3
Overall			$38 \pm 3\%$

Table 5: Cross-dataset f1-scores for the Spotify-A dataset and the CNN-LSTM model (binary classification - PE/SP)

class, achieving f1-scores $38 \pm 3\%$.

In the Spotify Podcast dataset, many speakers had some knowledge about the topics they discussed, whereas in CORAA, there are interviews with everyday people on diverse topics, and the presence of audio noise is notable. Another hypothesis is that in both states (MG, PE), despite their distinctiveness, there is a tendency to frequently use diminutives in language and exhibit a slightly more melodic and musical intonation.

4.2 Experiments with More Speakers (Spotify-B)

The Spotify-B subset presents a significantly superior amount of speakers concerning the Spotify-A. Spotify-B has 102 distinct speakers from Pernambuco and 85 from São Paulo.

– **Closed-dataset Validation Out-of-speaker**: we trained the models using audio data from the training and development sets described in Table 2. Samples were balanced so that the training set has 51 and 52 speakers from Recife and São Paulo, respectively. For the development set, we employed 16 speakers from São Paulo and 25 from Recife. For testing, 11 distinct speakers were chosen for each condition from various podcasts, also balancing instances for each condition.

Table 6 presents the results, where the precision rate is slightly higher than the recall rate for the PE class, while the opposite scenario occurs for the SP class. This indicates that, even after balancing the number of audio samples for each class during training, the model performs slightly better for class PE. Furthermore, the inclusion of a greater variety of speakers led to a better balance between precision and recall for both classes. For the PE accent, although precision was slightly lower $(61 \pm 3)\%$, we observed a significant increase in recall $(58 \pm 7)\%$ compared to previous results $(20 \pm 20)\%$. Similarly, for the SP accent, there was an improvement in precision $(57 \pm 3)\%$ compared to previous results with a smaller number of speakers. The overall F1-score was $59 \pm 2\%$, indicating an enhanced balance between precision and recall compared to the previous results in Table 4. Moreover, the results with a larger number of speakers showed less variability compared to the Subset-A results in Spotify.

class	Precision	Recall	F1-score
PE	$61 \pm 3\%$	$58 \pm 7\%$	59 ± 4
SP	$57 \pm 3\%$	$60 \pm 7\%$	58 ± 4
Overall			$59 \pm 2\%$

Table 6: Closed-dataset f1-scores for the Spotify-B dataset and the CNN-LSTM model

class	Precision	Recall	F1-score
PE	$50 \pm 1\%$	$96 \pm 4\%$	66 ± 9
SP	$73 \pm 7\%$	$11 \pm 4\%$	19 ± 6
Overall			$43 \pm 3\%$

Table 7: Cross-dataset f1-scores using Spotify-B to train and CORAA to test and the CNN-LSTM model

– **Cross-dataset Validation**: for cross-dataset validation, the results corroborate the conclusions highlighted by Batista, where the models used have difficulty generalizing to other datasets. Table 7 presents the results for cross-dataset. In particular, the model's predictions favored the PE (Recife) class, however presenting many false PE classifications. On the other hand, class SP has better precision but low recall. In comparison with the scenario with fewer speakers, for the PE class, there's a significant improvement in its detection capability, with a considerably higher recall rate $(96 \pm 4)\%$ compared to the Spotify-A$(10 \pm 3\%.)$ This indicates that the model is much better at correctly identifying the PE accent.

For the SP class, although precision has in-

class	Precision	Recall	F1-score
PE	90%	72%	80%
SP	75%	92%	83%

Table 8: Closed-dataset f1-scores for the finetuning of Wav2Vec model using Spotify-B

class	Precision	Recall	F1-score
PE	68%	99%	80%
SP	99%	55%	70%

Table 9: Cross-dataset f1-scores for the finetuning of Wav2Vec model using Spotify-B and testing with CORAA

creased ($73 \pm 7\%$), the ability to accurately identify the SP accent has decreased significantly($11 \pm 4\%$). This means that despite the improved precision, the model struggles to detect the SP accent. The overall F1-score with Spotify-B is slightly better at $43 \pm \%$. This is primarily due to the improvement in both precision and recall for the PE accent.

Therefore, the classification of different variations proves to be challenging across different datasets, as reported by Batista et al. (2018); Batista (2019), even when increasing the amount and variety of training speakers.

– **Wav2Vec Finetuning Closed-dataset and Cross-dataset**: the Wav2vec 2.0 XLSR pre-trained model was fine-tuned with the Spotify-B subset (binary classification PE, SP). Table 8 presents the results for a closed-dataset scenario and Table 9 the cross-dataset scenario. The results are remarkable in comparison with the previous model, reaching an F1-score of 82% for the closed-dataset scenario, and 75% for the cross-dataset, demonstrating the potential of using pre-trained models for this task.

Even with the superior metrics, we noticed a similar effect of favoring precision and recall on different classes and tasks (e.g. closed-dataset task is more precise on PE, while the cross-dataset is more precise on SP). The fact it happened for the same classes, indicates there are probably a set of examples or patterns that the model hardly learns.

In summary, it was observed that utilizing a dataset with a larger and balanced set of speakers for fine-tuning with the Wav2vec 2.0 XLSR-53 model can have a considerable impact on the model's performance for accent classification. In addition to an increased number of speakers providing greater linguistic variability, the recording conditions in Spotify-B, characterized by minimal noise compared to other datasets, play a significant role. It is important to note that the success in the classification task depends on other factors, such as consistent data preprocessing, and the use of additional datasets for model evaluation.

5 Conclusions

Our results show that accent classification is still an open problem, with challenges going beyond the use of different datasets, as reported by Batista et al. (2018); Batista (2019). In fact, our results indicate that both models: CNN1D LSTM and Wav2vec 2.0 XLSR may be learning spurious features, e.g. related to the speakers and/or the recording conditions. This raises questions about the ability of the models to learn accent attributes. We believe the Spotify podcasts dataset is valuable in this context since it has subtle speakers and recording variations within the same dataset.

When comparing results with preatrained models, de Almeida (2022) could not reach good results with Wav2Vec 2.0 base (trained just using English language), when evaluating the closed-dataset scenario on a different dataset. Our choice of the Wav2Vec 2.0 XLSR multilingual model achieved results superior to those using a CNN1D-LSTM trained from scratch. This indicates a larger and multilingual model may be more effective.

In general, we found two main guidelines for improving results in the accent classification tasks. First, improving the resources for a given language is paramount, i.e. increasing the number of speakers to cover the accent characteristics better. Secondly, using larger and pre-trained models appears to excel training from scratch. Nevertheless, a more in-depth analysis is still needed to understand what the models are truly learning, in particular biases related to individual speakers or recordings.

Future work may devote efforts to investigating the explainability of models, as well as gathering more data from different sources. Exploring other pre-trained models is also a matter of future studies.

Acknowledgements

This work was carried out at the Artificial Intelligence Center (C4AI-USP), with support from the São Paulo Research Foundation (FAPESP grant nº 2019/07665-4) and IBM Corporation. It was also supported by the Ministry of Science, Technology

and Innovation, with resources from Law nº 8,248, Oct 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published Residência no TIC 13, DOU 01245.010222/2022-44.

References

Nathalia Batista, Lee Ling, Tiago Fernandes Tavares, and Plinio Barbosa. 2018. Detecção automática de sotaques regionais brasileiros: A importância da validação cross-datasets.

Nathalia Alves Rocha Batista. 2019. Estudo sobre identificação automática de sotaques regionais brasileiros baseada em modelagens estatísticas e técnicas de aprendizado de máquina. Master's thesis, Universidade Estadual de Campinas, Campinas.

Arnaldo Candido Junior, Edresson Casanova, Anderson da Silva Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, and Sandra Maria Aluísio. 2021. CORAA: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese. *CoRR*, abs/2110.15731.

Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Aluísio, and Moacir Antonelli Ponti. 2023. ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion. In *Proc. INTERSPEECH 2023*, pages 1244–1248.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Diego Ribeiro de Almeida. 2022. Comparação entre modelos com diferentes abordagens para classificação de sotaques brasileiros.

Keqi Deng, Songjun Cao, and Long Ma. 2021. Improving accent identification and accented speech recognition under a framework of self-supervised learning. pages 1504–1508.

Sonia Frota, Marina Vigário, Marisa Cruz, Friederike Hohl, and Bettina Braun. 2022. Amplitude envelope modulations across languages reflect prosody. In *Speech Prosody 2022*, pages 688–692.

Sebastião Carlos Leite Gonçalves. 2019. Projeto alip (amostra linguística do interior paulista) e banco de dados iboruna: 10 anos de contribuição com a descrição do português brasileiro. *Estudos Linguísticos (São Paulo. 1978)*, 48(1):276–297.

Lucas Rafael Stefanel Gris, Edresson Casanova, Frederico Santos de Oliveira, Anderson da Silva Soares, and Arnaldo Candido Junior. 2021. Brazilian portuguese speech recognition using wav2vec 2.0.

John Hansen, Marigona Bokshi, and Soheil Khorram. 2020. Speech variability: A cross-language study on acoustic variations of speaking versus untrained singing. *The Journal of the Acoustical Society of America*, 148:829–844.

Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154. Deezer Research.

Rodolfo Ilari and Renato Basso. 2009. O português da gente: a língua que estudamos, a língua que falamos. *(No Title)*, 2:167–168.

Rosina Lippi-Green. 2012. *English with an Accent: Language, Ideology, and Discrimination in the United States*. Routledge.

Heliana Mello, Massimo Pettorino, and Tommaso Raso. 2012. *Proceedings of the VIIth GSCP International Conference. Speech and Corpora*. Firenze University Press.

Antenor Nascentes. 1953. Études dialectologiques du brésil. *ORBIS-Bulletin International de Documentat ion Linguistique, Louvain*, 2(2):438–444.

Miguel Oliviera Jr et al. 2016. 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(2):149–174.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Tommaso Raso and Heliana Mello. 2012. *C-ORAL-BRASIL: corpus de referência do português brasileiro falado informal. I*. Editora UFMG.

Edgar Tanaka, Ann Clifton, Joana Correia, Sharmistha Jat, Rosie Jones, Jussi Karlgren, and Winstead Zhu. 2022. Cem mil podcasts: A spoken portuguese document corpus.

Carlos Teixeira, Isabel Trancoso, and António Serralheiro. 1996. Accent identification. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1784–1787. IEEE.

Wagner A Tostes, Francisco A Boldt, Karin S Komati, and Filipe Mutz. 2021. Classificaçao de sotaques brasileiros usando redes neurais profundas. In *Simpósio Brasileiro de Automação Inteligente-SBAI*, volume 1.

Wagner Arca Tostes. 2022. Arquiteturas de redes neurais profundas para classificação de dialetos e sotaques.

CA Ynoguti. 1999. Reconhecimento de fala contínua utilizando modelos ocultos de markov. *Faculdade de Engenharia Elétrica–UNICAMP*.

Juan Pablo Zuluaga, Sara Ahmed, Danielius Visockas, and Cem Subakan. 2023. Commonaccent: Exploring large acoustic pretrained models for accent classification based on common voice. pages 5291–5295.