Machine Learning Classifiers with Acoustic Features for Prosodic Segmentation in Brazilian Portuguese: A Comprehensive Evaluation

Giovana M. Craveiro[1], Caroline A. Alves[2], Flaviane Svartman[2], Sandra M. Aluísio[1]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo, São Carlos, SP, Brazil

giovana.meloni.craveiro@alumni.usp.br, sandra@icmc.usp.br

[2]Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH)
Universidade de São Paulo, São Paulo, SP, Brazil

{carolalves, flavianesvartman}@usp.br

Abstract. *Spontaneous speech has not yet been widely explored in Brazilian Portuguese (BP) for the task of automatic prosodic segmentation. In this article, we compared seven types of classifiers, considering their performance for various types of speaker profiles (varied genders, ages, education levels, and regions of birth) and environmental impact, and trained the most appropriate one. Thus, we propose a Random Forest classifier, based on acoustic features, with low environmental impact and an F1 score of 0.55% and 0.77%, with binary and macro averages, respectively. Additionally, we are making it publicly available and present a discussion of its efficiency for different speaker profiles, as well as its environmental impact.*

1. Introduction

Prosodic segmentation is the process of dividing spoken language into smaller units (prosodic units) based on prosodic clues such as intonation, intensity, and duration. These units, which do not always correspond to grammatical units, help structure speech and facilitate the comprehension of the spoken message. Between one unit and the following, prosodic boundaries are inserted. Previous studies [Raso et al. 2020] have distinguished between terminal prosodic breaks (TB), which mark completed sequences, that is, they communicate the conclusion of the utterance, constituting the smallest pragmatically autonomous unit of speech, from non-terminal prosodic breaks (NTB), which signal a non-autonomous prosodic unit whose information is not completed within the same utterance.

This task is applied in a variety of areas, including speech technology (both text-to-speech (TTS) and automatic speech recognition (ASR) systems) and linguistic analysis [Liu et al. 2022, Lin et al. 2019, Viola and Madureira 2008, Chen and Hasegawa-Johnson 2004]. Several studies have already addressed automatic prosodic segmentation (see Tables 1 and 2). Many of them have used prepared speech corpora, such as radio news. In these cases, prosodic and syntactic boundaries coincide, since the speaker follows punctuation, which marks the syntactic boundaries, consequently placing a prosodic boundary in the same positions where syntactic boundaries exist. In addition, disfluencies are rare in this scenario [Biron et al. 2021]. However,

studies that focus on spontaneous speech may have more difficulty in achieving high performance due to the presence of disfluencies and less clear prosodic boundaries, since the speaker formulates the text as they speak, unlike a reading task of a previously punctuated text.

The task of automatic prosodic segmentation for spontaneous speech, specifically, is a long-standing problem [Biron et al. 2021], but it remains relevant, given the aforementioned obstacles. Previous approaches include rule-based methods (heuristics), traditional machine learning, and, more recently, deep learning. While there are approaches based exclusively on acoustic signals, some methods have also relied on lexical and syntactic cues and include extensive preparation steps, such as manual tagging. The work proposed here is based solely on acoustic features, uses traditional machine learning, focuses on spontaneous speech, and innovates by presenting a comprehensive evaluation of classification algorithms applied to BP and segmentation bias of speaker profiles across gender, region of birth, age, and education level.

The contributions of this work are as follows.

1. an automatic prosodic segmentation method inspired by the work of [Ananthakrishnan and Narayanan 2008], with spontaneous speech data in Brazilian Portuguese, bridging the gap of automatic prosodic segmentation in BP;
2. evaluation of the method bias in terms of speaker profiles, which vary in gender, age, education level, and dialectal varieties, using corpus MuPe-Diversidades [Craveiro and Galdino 2025];
3. measurement of carbon emissions, energy costs, and duration of performing the task with seven ML classifiers using CodeCarbon[1], in addition to the classic performance measures (binary and macro F1 score, and accuracy)[2], to choose the most appropriate ML method for the task; and
4. provision of the code and model[3] to facilitate the evaluation and replicability of the method, as well as enabling further training and usage with other datasets.

2. Literature Review on Prosodic Segmentation Methods

We present seven automatic prosodic segmentation studies published between 2008 and 2024, chosen because they include a variety of approaches. Tables 1 and 2 summarize them and the method we propose, evaluated in the spontaneous speech dataset MuPe-Diversidades, which covers different BP dialectal varieties and is freely available [4].

[Biron et al. 2021] detects prosodic boundaries in spontaneous English speech (Santa Barbara Corpus) through heuristics based on pause durations and speech rate discontinuities (SRDs)[5], measured through phone durations within 300ms windows. Phonetic alignment[6] is obtained with Montreal Forced Aligner[7], and results are evaluated in Praat. They report an F1 score of 66%.

[1] `https://codecarbon.io`

[2] F1 score is the harmonic mean of precision and recall. While binary average only considers the positive label (TBs), macro F1 considers metrics for each label (TBs and NBs) and measures their unweighted mean.

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

[4] `https://github.com/nilc-nlp/MuPe-Diversidades`

[5] SRDs refer to slowing down of speech rate at the end of a unit along with acceleration at its beginning.

[6] A forced phonetic aligner is a tool that automatically aligns a speech recording with its corresponding phonetic transcription, providing time-aligned boundaries for each phonetic unit.

[7] `https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner`

[Kocharov et al. 2017] proposes predicting intonational units by combining syntactic and acoustic information and a Random Forest classifier. The approach assumes that certain word boundaries, such as between prepositions and nouns, are unlikely to contain prosodic breaks, allowing syntax to narrow down potential boundary locations. With Boston University Radio Speech Corpus (BURNC) for English, it reached an F1 score of 76% and 86.5% accuracy. Key acoustic features include changes in fundamental frequency (F0), speech rate, and intensity. The study found that about 97% of the prosodic boundaries were in syntactically plausible positions and that the remaining 3% could be related to language-specific rules and parsing errors.

[Roll et al. 2023] uses the Santa Barbara Corpus (SBC) to introduce the PSST method, which finetunes OpenAI's Whisper ASR model (764M parameters) to segment speech by integrating prosodic and syntactic cues, functioning also as a transcription tool. They manually revise transcriptions, preserving disfluencies and removing unwanted tokens. Versions with masked syntax or text-only input tested the influence of syntax, but the best model combined acoustic and syntactic information, obtaining 96% accuracy and 87% F1 score. The method is efficient, semi-supervised, and does not require extensive annotations or computational resources, making it practical for NLP applications.

[Raso et al. 2020] presents a Linear Discriminant Analysis (LDA) classifier that automatically identifies prosodic boundaries in spontaneous BP speech using phonetic-acoustic parameters. Data came from the C-ORAL-BRASIL I and II corpora with expert-annotated prosodic boundaries. 42 phonetic-acoustic features were extracted, covering speech rate, segment duration, fundamental frequency (F0)[8], and pauses. The model achieved an F1 score of 81.5% for TBs and 54.5% for NTBs, with an average of 68%. Pauses and F0 were key predictors, while duration was less influential. The model was prone to falsely detecting boundaries due to pause sensitivity. Overall, parameters near boundaries, especially pauses and normalized duration, were the most effective.

[Hoi et al. 2022] proposes a method based on pause detection using spectrograms and a convolutional neural network (CNN). Using 33 hours of transcribed news audio from corpus RTP[9] (15,000 sentences), in European Portuguese, the method detects whether pauses greater than or equal to 250 ms mark terminal or non-terminal breaks. Audio windows (100 ms before + 300 ms after the pause) are classified using a 3-layer CNN. Without relying on phonetic alignment or linguistic features, the model achieved 95.6% accuracy. While efficient and language-agnostic, the method only handles pause-based boundaries and may be influenced by unintended acoustic biases.

[Craveiro et al. 2024] adapted the approach described in [Biron et al. 2021] to Brazilian Portuguese, using the forced phonetic aligner UFPAlign[10] [Batista et al. 2022], designed for BP. Working with lengthy NURC-SP audio recordings (30–90 minutes), they segmented the audios into 10-minute chunks for alignment. Their approach uses a 300 ms time window to detect pause duration and SRDs. It obtained an F1 macro of 31% with a hit threshold of 0.25 seconds. The code is available[11].

[8]Fundamental frequency (F0) refers to the approximate frequency of the (quasi-)periodic structure of voiced speech signals [Bäckström et al. 2020].

[9]https://www.rtp.pt/

[10]https://github.com/falabrasil/ufpalign/

[11]https://github.com/nilc-nlp/ProsSegue

[Ananthakrishnan and Narayanan 2008] uses the Boston University Radio Speech Corpus to explore an LDA, a Gaussian Mixture Classifier (GMM), and a Neural Network (NN), basing their approach on acoustic features, but also on their combination with syntactic and lexical evidence. They extract the following features from each syllable: duration of pauses immediately after syllables (p_dur), nucleus vowel duration (n_dur), F0 range (f0_range), energy range (e_range), difference between minimum and average within-syllable F0 (f0_avgmin_diff), difference between maximum and average within-syllable F0 (f0_maxavg_diff), difference between minimum and average within-syllable energy (e_avgmin_diff), difference between maximum and average within-syllable energy (e_maxavg_diff), and difference between syllable average F0 and average F0 of the utterance it belongs to (f0_avgutt_diff). With their NN classifier, they achieved 91.6% accuracy with the acoustic + syntactic approach, and 89.9% without syntactic features.

Table 1. Summary of prosodic segmentation research on prepared and spontaneous speech (1).

Source	Language	Corpus	Gender balanced ?	Segment Types	Domain
Raso et al. (2020)	PT-BR	C-Oral Brasil I C-Oral Brasil II (∼17min)	No only male voices	TB NTB	SPONT
Hoi et al. (2022)	PT-PT	RTP (∼33hs)	No	TB NTB	PREP
Craveiro et al. (2024)	PT-BR	Part of the NURC-SP MC (∼5hrs)	Aprox. 2 male, 4 female	TB NTB	SPONT
Ananthakrishnan et al. (2008)	EN	BURNC (∼3hs)	Yes 3 male, 3 female	IUs	PREP
Kocharov et al. (2017)	EN	BURSC (∼10hs)	Yes 3 male, 3 female	IUs	PREP
Biron et al. (2021)	EN-US	SBC (∼20hs)	Yes	IUs	SPONT
Roll et al. (2023)	EN-US EN-GB	SBC (∼20hs) IViE (∼36hs)	Yes 55% female, 44% male	IUs	SPONT
This Work (2025)	PT-BR	MuPe-Diversidades (2h32m15s)	Yes 53% female, 47% male	TB	SPONT

(1)"EN-US" stands for American English, "EN-GB" for British English, "TB" for terminal prosodic boundaries, "NTB" for non-terminal prosodic boundaries. "SPONT" refers to spontaneous speech, "PREP" to prepared speech.

Table 2. Continuation of Table 1 (1)

Source	F1 score/ Accuracy	Training ?	Features	Approach	Code Availability
Raso et al. (2020)	68%/—	Yes LDA	Speech Rate, Rhythm, Duration, F0, Intensity, Pauses	TML	Not Available
Hoi et al. (2022)	—/95.6%	Yes CNN API	Spectrogram	DL	Not Available
Craveiro et al. (2024)	31%/—%	No	Pauses, SRDs	Heuristics	Open Code
Ananthakrishnan et al. (2008)	—/91.6%	Yes LD, GMM, NN	9 acoustic features; see Section 2	TML	Not Available
Kocharov et al. (2017)	76%/86.5%	Yes Random Forest	Pauses, SRS, Df0C, Intensity	TML	Not Available
Biron et al. (2021)	66%/—	No	Pauses, SRDs	Heuristics	Not Available
Roll et al. (2023)	87%/96% (SBC) 73%/93% (IViE)	Yes IUs Finetuning with Whisper	—	DL	Not Available
This Work (2025)	55%/97% (2) 77% (3)	Yes Random Forest	9 acoustic features; see Section 3	TML	Open Code (4)

(1) The acronym "SRD" stands for speech rate discontinuities, "SRS" for speech rate slowdown at the end of a sentence, "Df0C" for F0 contour decline, "DL" for deep learning, and "TML" for traditional machine learning. (2) F1 score binary average and accuracy. (3) F1 score with macro average. "GMM" stands for Gaussian Mixture Classifiers, "LD" for Linear Discriminant, and "NN" for Neural Network. (4) The code for the proposed method in this paper is available at github.com/nilc-nlp/ProsSegue.

3. Our ML-based Method using Acoustic Features

3.1. Dataset

The publicly available[12] MuPe-Diversidades [Craveiro and Galdino 2025] contains short samples of speech (4-10 minutes), totaling 2hrs32min15s. Its speakers are balanced in gender and state of origin (Alagoas, Bahia, Ceará, Espírito Santo, Goiânia, Minas Gerais, Mato Grosso do Sul, Pará, Paraíba, Paraná, Pernanmbuco, Piauí, Rio de Janeiro, Rio Grande do Sul, Rondônia, Sergipe, and São Paulo), and have varied education levels and ages, which range from 20 to 91 years old.

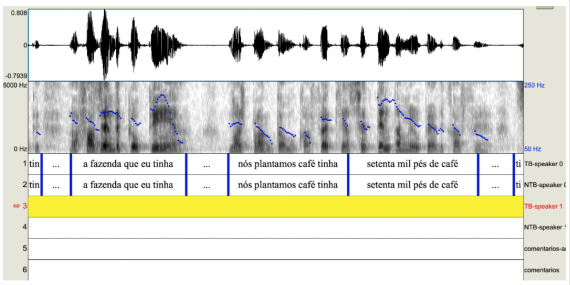

Figure 1. Life story SP1 MuPe - unedited.

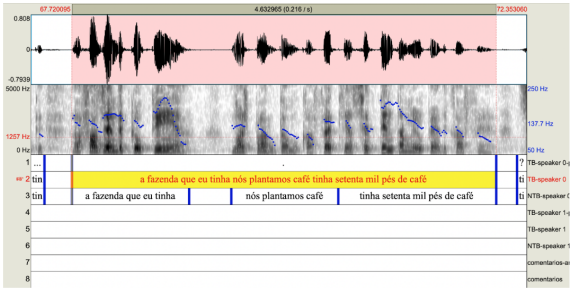

Figure 2. Life story SP1 MuPe - revised.

MuPe-Diversidades includes not only audio files but also multilevel transcriptions aligned to the audio. The revised segmentation data for this corpus, made by an experienced linguist (Version 1), was obtained by first applying a baseline automatic segmentation method (Version 0). The multilevel transcriptions consist of the following main interval tiers annotated in the speech analysis software Praat [Boersma and Weenink 2025] (see Figures 1 and 2 for an illustration).

- 2 tiers (TB-, NTB-) in which the speakers' speech excerpts (speaker 0, speaker 1) are segmented into prosodic units and transcribed orthographically;
- 1 tier for comments (com) about the audio and the annotation;
- 1 tier containing the punctuation (-period) that ends each TB.

Regarding prosodic units, the concept used here is based on the principles of the C-ORAL-BRASIL prosodic segmentation study [Raso and Mello 2012]. In the flow of speech, unit boundaries with terminal or non-terminal values are recognized. The identification of prosodic breaks is based mainly on the perceptual (auditory) relevance of prosodic clues but also on visual inspection of the acoustic signal synthesis provided by Praat. The main clues to a prosodic break in BP are the insertion of pauses and changes related to fundamental frequency and duration [Serra 2009, Raso et al. 2020].

3.2. Acoustic features and pipeline

The method we propose here is inspired by the work of [Ananthakrishnan and Narayanan 2008], as it measures the same nine acoustic features (see Section 2 for details) at a syllable level. We opted for a model based solely on acoustic features as they obtained only a slight improvement (1.6%) when they added syntactic evidence to their acoustic model. To measure f0_avg_utt, we considered

[12] https://github.com/nilc-nlp/MuPe-Diversidades

utterances as all the text between annotated TBs, which we also used to attribute labels to the syllables (TB, indicating a terminal boundary right after the syllable, or NB, indicating no immediate boundary after it). We normalized the syllable nucleus duration per speaker and per vowel-type [13] to normalize the data against variation among speakers and due to vowel-intrinsic properties, while preserving variations produced by boundary cues. We also had to adjust pitch parameters[14] during feature extraction to obtain valid values in all voiced frames.

Our pipeline, illustrated in Figure 3, consists of three phases: forced phonetic alignment with UFPAlign [Batista et al. 2022] to obtain initial and final timestamps of each phone, syllable, and word; extraction of acoustic features with library parselmouth[15]; and segmentation using the Random Forest model (see Section 4 for details).

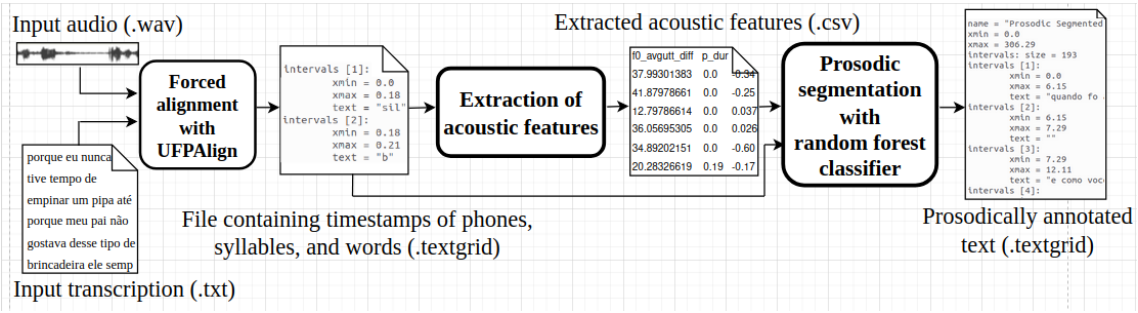

Figure 3. Pipeline of the proposed method

As input for the first phase, UFPAlign requires a WAV audio with 16kHz, and a monophonic signal, along with its transcription, which should contain all words separated by a single space, and ideally no overlaps. Our final output is a textgrid with different intervals containing the resulting utterances, separated prosodically by the classifier. Additionally, UFPAlign was designed to process short audios (e.g., 30 seconds). We dealt with examples of 4-10 minutes, which were mostly processed smoothly, but five of them[16] required a version of UFPAlign that uses M2M aligner[17].

4. Experiments and Results

4.1. Evaluation of Classification Algorithms

Before settling on a Random Forest classifier for our approach, we explored seven types of ML classifiers from scikit-learn, version 1.6.1. To compare them, a K-Fold cross-validation was performed (k=5). Firstly, we removed interviewers' speech to avoid bias and separated the entire corpus (MuPe-Diversidades) into a train set (80%) and a test set (20%). This division was performed for each of the speakers to ensure we would have a

[13] We used the following list of phones to calculate phone average durations of all possible nucleus vowels that UFPAlign indicated: "a", "e", "i", "o", "u", "a~", "e~", "i~", "o~", "u~", "E", "O". We found a few cases of "j" and "w" indicated as nucleus vowels, which we treated as "i" and "u", respectively.

[14] We used a pitch floor of 50Hz, a pitch ceiling of 800Hz, a time step of 0.001, an octave jump cost of 0.4, a voicing threshold of 0.2, and default values for all other parameters.

[15] `https://parselmouth.readthedocs.io/en/stable/`

[16] In two of those, one word could not be properly aligned, representing approximately 0.0015% of its audio. In those cases, two adjacent syllables were concatenated and treated as one by our method.

[17] `https://github.com/letter-to-phoneme/m2m-aligner`

representation of their speaker profile in both the training and test set, making it possible to evaluate the method's efficiency for their specific profile (region of birth, gender, age, and education level) further on. Then, we extracted acoustic features individually for each speaker and transformed NAN values to 0 to feed them to the classifiers.

Table 3. This table presents the average and standard deviation obtained considering three seeds (17, 42, 79) at the K-Fold cross-validation for each classifier.

Model		F1 binary	F1 macro	Accuracy
Linear Discriminant Analysis (LDA)	Avg	0.51	0.745	0.965
	Stddev	0.0013	0.00068	5e-05
Multi-layer Perceptron (MLP)	Avg	0.53	0.76	**0.97**
	Stddev	0.004	0.002	0.0002
Random Forest (RF)	Avg	**0.55**	**0.77**	**0.97**
	Stddev	0.0037	0.0019	0.00013
Logistic Regression (LR)	Avg	0.53	0.75	0.96
	Stddev	0.0005	0.00026	4e-05
Gradient Boosting (GB)	Avg	0.49	0.74	**0.97**
	Stddev	0.004	0.002	9e-05
Decision Tree (DT)	Avg	0.46	0.715	0.95
	Stddev	0.011	0.006	0.0012
Support Vector Classifier (SVC)	Avg	0.54	0.76	0.96
	Stddev	1.7e-04	9e-05	3e-05

We also performed a parameter search, whose grid varied for each classifier (maintaining 500 and 200 as the maximum number of iterations for LR and MLP, respectively, as well as Adam solver for the latter, and kernel rbf for SVC), followed by a class weight search, whose grid of NBxTB (1x1,1x2,1x3,1x5,1x10,1x15,1x30,1x35,1x40,1x50, balanced) was maintained when present (RF, LR, DT, and SVC). With the optimal parameters[18], we performed a cross-validation with three different seeds (17, 42, 79) for each classifier and measured the average and standard deviation of each, which are presented in Table 3. Using CodeCarbon, we measured total carbon emissions as CO2-equivalents (CO2eq) in kg, CO2eq emissions rate (measured as emissions per duration) in kg/s, energy consumed (as the sum of CPU energy, GPU energy, and RAM energy) in kWh, and duration of these phases in seconds, summing the values obtained in parameter search and cross-validation stage. These values are illustrated in the charts presented in Figure 4.

As can be seen in Table 3, the Random Forest classifier, with 0.55, outperforms the others by at least 0.01. SVC, MLP, and LR classifiers follow it closely with 0.54 and 0.53, respectively. However, the graphs presented in Figure 4 show that the SVC classifier has a significantly greater environmental impact than all the others, which implies that it should be avoided if we wish to preserve good scalability. Also from this perspective, the LDA seems to be the least costly classifier to escalate. In fact, LDA, LR, and DT classifiers have a smaller environmental impact overall and seem to be ideal choices if we were dealing with a huge amount of data for training. But in our case, since our best performing classifier (RF) consumed approximately 0.3g/CO2-eq, the equivalent amount of CO2 necessary to send one email from laptop to laptop, or about 2 seconds worth of the 5-tonne lifestyle[19] recommended by [Berners-Lee 2020], we can safely choose it. Our Random Forest classifier was trained with a maximum depth of 20, a minimum sample split of 5, 100 estimators, a seed of 42, and class weights of 1(NB) x 30 (TB).

[18]Details on parameters can be found at `https://github.com/nilc-nkp/ProsSegue`.

[19]A lifestyle that causes 5 tonnes of CO2e per year, recommended as a possible and necessary personal goal on the journey to a low-carbon world.

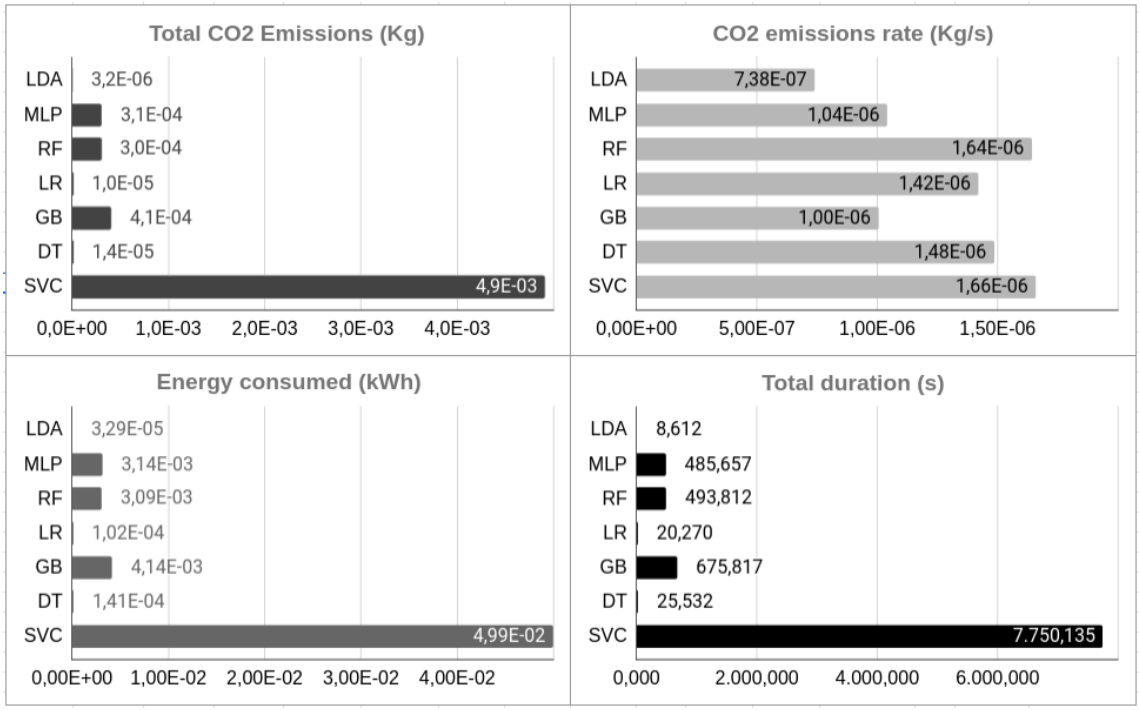

Figure 4. These four charts represent the comparison among the seven classifiers in terms of total CO2eq emissions, CO2eq emissions rate, and energy consumed during parameter search and K-fold cross-validation, and the total duration of the process for each classifier.

In the work by [Ananthakrishnan and Narayanan 2008], feature importance was ranked as follows: p_dur, n_dur, f0_maxavg_diff, f0_range, e_range, f0_avgmin_diff, e_maxavg_diff, e_avgmin_diff, f0_avgutt_diff[20]. With a statistical test, they conclude that all features are helpful, but emphasize the role that F0 range, energy range, and especially pause and nucleus duration played as indicators of boundary events. In our work, pause duration is also the most relevant feature, with an importance of almost 0.5. In contrast, the others differ significantly in order (p_dur, f0_avgutt_diff, e_maxavg_diff, n_dur, e_range, f0_maxavg_diff, e_avgmin_diff, f0_avgmin_diff, f0_range)[21] and have decreasing importances that range approximately from 0.05 to 0.08.

Finally, including the phonetic alignment and extraction of features from the MuPe-Diversidades corpus, parameter search, training, and prediction of results, the total CO2 emission was approximately 8.2 grams, which is equivalent to the CO2eq used to send 28 short emails from laptop to laptop [Berners-Lee 2020]. The total energy consumed was approximately 75 Wh, and the total duration of the process was approximately 5,34 hours (the extraction of features took 5,2 hours).

4.2. Bias Evaluation in Segmentation for BP Spontaneous Speech

We prioritize binary F1, since we understand that placing TBs correctly is the focus of the task, rather than identifying NB locations, especially as NBs are significantly more numerous and could indicate misleading results. However, we also report macro F1 for

[20]These nine features were presented in Section 2.

[21]We measured feature importance with Random Forest Feature Importance (MDI) from sklearn.

an overall perspective, as not inserting breaks at wrong locations is also relevant. Considering the entire test set, we obtained an F1 score with a binary average of 55%, a macro average of 77%, and an accuracy of 97%. However, it is important to further analyze performance considering different speaker profiles. To account for that, we compared the performance of speakers from different regions, ages, genders, and education levels. Thus, we stratified our data, grouping speakers from different states into regions. We grouped speakers from the North and Midwest regions of Brazil, totaling seven speakers, and also speakers from the South and Southeast regions of Brazil, totaling 11 speakers, due to limited representation for each region. We also separated our speakers into three age groups (I:20-35; II:35-55; III:56+), which contain 5, 10, and 15 speakers, respectively, and four education groups (I: no education; II: incomplete elementary school, complete elementary school; III: technical school, incomplete bachelor's degree; IV: complete bachelor's degree, master's degree), which contain 8, 8, 8, and 6 speakers, respectively.

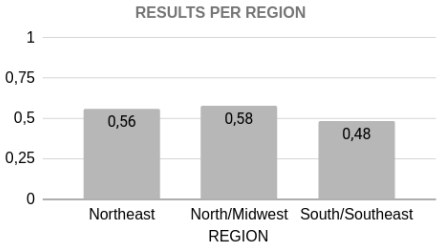

Figure 5. Region results

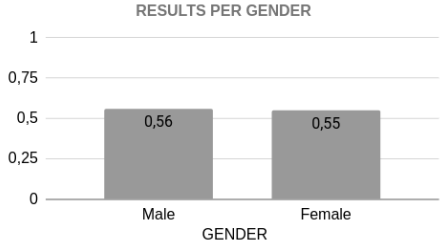

Figure 6. Gender results

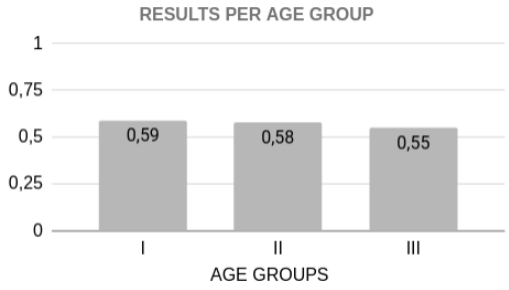

Figure 7. Age group results

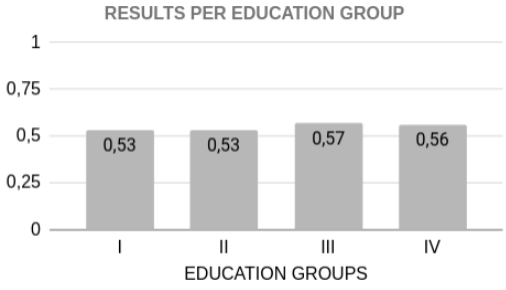

Figure 8. Education results

Once we predicted results for those specific groups, we performed a statistical significance analysis, measuring one-way ANOVA[22], with the SciPy library[23]. Although we did not find statistically significant differences between the groups in comparisons of the F1 score (groups of different genres, ages, regions, and education levels), we report the numbers we obtained.

As can be seen in Figure 6, there is a difference of 1% between male and female speakers (p-value = 0.49). In Figure 5, we see a difference of 8-10% between the F1 scores for speakers of the southern and southeastern regions and speakers from other areas

[22]Analysis of variance (ANOVA) is a commonly used statistical test to determine whether two population means are different. It indicates statistical significance if the p-value obtained is under 0.05.

[23]`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html`

(p-value = 0.68). Our classifier also seems to favor younger speakers (groups I and II), reaching 58%-59%, compared to reaching 55% with older speakers (p-value = 0.5), and disfavor speakers with lower education levels (2% below average for education groups I and II (p-value = 0.86). However, to reliably affirm any of those possible biases, we would need to test the model with more data.

5. Conclusions

In this paper, we propose a low-impact (8.2g CO2-eq for development, and an estimation of 7.9g CO2-eq for usage with a corpus of approximately 2.5 hours) automatic prosodic segmentation method based on acoustic features (pauses, duration, F0, and energy), which are measured for each syllable. It is evaluated with spontaneous speech in BP, aiming to bridge the gap of automatic prosodic segmentation methods explored for this type of speech. We report an F1 score of 55% and 77%, with averages binary and macro, respectively, and an accuracy of 97%. The model is available[24] as well as a model trained without feature f0_avg_utt (as it requires annotated TBs) with the entire dataset for users who wish to segment their datasets. We also present a comprehensive evaluation of classification algorithms explored to choose the best candidate for the task (Random Forest), which includes classic performance metrics, as well as environmental impact. The two major downsides of our approach are the dependency on UFPAlign, a third party software, and the extraction of features, which could be costly according to the size of the user's dataset.

Furthermore, for our RF classifier, we present an analysis focused on the segmentation bias according to speaker profile. The results indicate a difference of 1% in terms of gender, differences ranging from 3% to 4% in terms of age groups and educational levels, and differences ranging from 8% to 10% in terms of regions of birth. However, since these values are not statistically significant, we intend to expand the test set in future work to further analyze bias. We also plan to expand the training set with data from other corpora and evaluate the difference in overall efficiency and efficiency according to each speaker profile.

Regarding our method's features, we are working on a new version of feature f0_avg_utt, which considers utterances to be the text between silences indicated by UFPAlign when calculating utterance averages, and no longer requires manually annotated TB references. Moreover, we consider including syntactic, semantic, or pragmatic features to better distinguish the conclusive aspect of TBs. And as our current segmentation method only deals with TBs, in future work, we intend to focus on NTBs, to make the technique closer to the manual segmentation task, which distinguishes both breaks.

6. 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. This project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law No. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published Residence in TIC 13, DOU 01245.010222/2022-44.

[24] https://github.com/nilc-nlp/ProsSegue

References

Ananthakrishnan, S. and Narayanan, S. S. (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):216–228.

Bäckström, T., Räsänen, O., Zewoudie, A., Perez Zarazaga, P., Das, S., et al. (2020). *Introduction to speech processing*. Library of Open Educational Resources.

Batista, C., Dias, A. L., and Neto, N. (2022). Free resources for forced phonetic alignment in brazilian portuguese based on kaldi toolkit. *EURASIP Journal on Advances in Signal Processing*, 2022(1):11.

Berners-Lee, M. (2020). *How bad are bananas?: the carbon footprint of everything*. Profile Books.

Biron, T., Baum, D., Freche, D., Matalon, N., Ehrmann, N., Weinreb, E., Biron, D., and Moses, E. (2021). Automatic detection of prosodic boundaries in spontaneous speech. *PLoS ONE*, 16(5):1–21.

Boersma, P. and Weenink, D. (2025). Praat: doing phonetics by computer [Computer program]. Version 2025.

Chen, K. and Hasegawa-Johnson, M. A. (2004). How prosody improves word recognition. In *Speech Prosody 2004*.

Craveiro, G. M. and Galdino, J. C. (2025). Diversity in data for speech processing in brazilian portuguese. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 122–136, Cham. Springer Nature Switzerland.

Craveiro, G. M., Santos, V. G., Dalalana, G. J. P., Svartman, F. R. F., and Aluísio, S. M. (2024). Simple and fast automatic prosodic segmentation of Brazilian Portuguese spontaneous speech. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 32–44, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Hoi, L. M., Sun, Y., and Im, S. K. (2022). An automatic speech segmentation algorithm of portuguese based on spectrogram windowing. In *2022 IEEE World AI IoT Congress (AIIoT)*, pages 290–295.

Kocharov, D., Kachkovskaia, T., and Skrelin, P. (2017). Eliciting Meaningful Units from Speech. In *Proc. Interspeech 2017*, pages 2128–2132.

Lin, C.-H., You, C.-L., Chiang, C.-Y., Wang, Y.-R., and Chen, S.-H. (2019). Hierarchical prosody modeling for Mandarin spontaneous speech. *The Journal of the Acoustical Society of America*, 145(4):2576–2596.

Liu, S., Nakajima, Y., Chen, L., Arndt, S., Kakizoe, M., Elliott, M. A., and Remijn, G. B. (2022). How pause duration influences impressions of english speech: Comparison between native and non-native speakers. *Frontiers in Psychology*, 13.

Raso, T. and Mello, H. (2012). *C-ORAL–BRASIL I: corpus de referência do português brasileiro falado informal*. Editora UFMG, Belo Horizonte. 332 p. : il + 1 DVD-ROM.

Raso, T., Teixeira, B., and Barbosa, P. (2020). Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech. *Journal of Speech Sciences*, 9:105–128.

Roll, N., Graham, C., and Todd, S. (2023). Psst! prosodic speech segmentation with transformers.

Serra, C. R. (2009). *Realização e percepção de fronteiras prosódicas no português do Brasil: fala espontânea e leitura*. PhD thesis, Federal University of Rio de Janeiro.

Viola, I. C. and Madureira, S. (2008). The roles of pause in speech expression. In *Speech Prosody 2008*, pages 721–724.