Robustness and Diversity Evaluation on ProsSegue-ML: a Free Prosodic Segmentation Tool for Brazilian Portuguese

Giovana Meloni Craveiro
ICMC-USP, BRAZIL
giovana.meloni.craveiro@alumni.usp.br

Sandra Maria Aluísio
ICMC-USP, BRAZIL
sandra@icmc.usp.br

Abstract

Prosodic segmentation is the task of dividing a sound unit into smaller units, which can be distinguished between units with a completed idea, marked by TBs, and non-autonomous units, marked by NTBs. Enhancing the performance of ASR and TTs systems is a useful task, and it remains relevant for Brazilian Portuguese due to the diversity of conditions and speaker-related factors that influence its performance. Here, we explore a low-impact, open-source approach based on a Random Forest classifier and a set of features that include fundamental frequency, speech rate, pauses, and energy (Craveiro et al., 2025). We perform a robustness evaluation of the referred ML model, modifying a few conditions on its training, comparing its performance when tested in other datasets, and comparing its results with those of other studies using the same data samples. We experiment with augmenting the training dataset and evaluating how the bias of speaker profile aspects is affected when the size and diversity of the training set are changed. Although we don't achieve statistically significant values in the bias evaluation, we observe that inequalities grow as the training dataset is expanded with a much larger, but less diverse sample of data.

1 Introduction

Information in spoken language is conveyed not only through lexical items, but also through a range of non-segmental features, commonly referred to as prosodic cues, including pitch, intensity, speech rate, rhythm, and timbre. Speech segments delimited by such prosodic cues are capable of expressing coherent messages and fulfilling a variety of linguistic functions, which are realized through different types of utterance (e.g., imperative, interrogative, assertive, or exclamatory). These prosodically delimited segments are typically referred to as intonational phrases or intonation units (IUs). Although IUs are difficult to define precisely, they are generally characterized by the presence of a well-defined (i.e., "single") pitch contour (Biron et al., 2021).

There are studies, such as (Mello et al., 2012; Santos et al., 2022) that distinguish between terminal break (TB) units, which mark complete sequences, that is, they communicate the conclusion of an idea, constituting the smallest pragmatically autonomous unit of speech, and non-terminal break (NTB) units, which signal a non-autonomous unit, whose information is not completed within the same unit. The identification of these boundaries is based on prosodic cues, such as variations in fundamental frequency (F0), segment duration, and the presence of pauses, in addition to inspection of the acoustic signal. In this work, we will follow the same distinction, but we will focus only on terminal boundaries, as we chose a method that does not segment NTBs.

Automatic detection of prosodic boundaries in natural language speech has been extensively investigated in the speech processing literature (Wightman and Ostendorf, 1991; Ananthakrishnan and Narayanan, 2008; Huang et al., 2008; Jeon and Liu, 2009; Kocharov et al., 2017; Biron et al., 2021). Despite substantial progress, this task remains challenging due to the numerous sources of variability inherent in speech signals. These sources include speaker-related factors (e.g., age, gender, and dialectal variation), recording conditions (such as microphone type, room acoustics, and background noise), and production style, ranging from spontaneous to read speech, which can be understood as points along a continuum between unplanned and planned speech production. Furthermore, machine learning methods are prone to biases (Broussard, 2018; Buolamwini and Gebru, 2018; Ruback et al., 2022). In the prosodic segmentation scenario, (Craveiro and Galdino, 2025) argues that it is imperative to select a corpus that is diverse in terms

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 170–180
April 13-16, 2026 ©2026 Association for Computational Linguistics

of accent, gender, age, and educational level.

Accurate prosodic boundary detection has direct implications for both automatic speech recognition (ASR) and text-to-speech (TTS) systems. In ASR, training models for speech excerpts segmented according to IUs have been shown to reduce syllable, character, and word-level error rates (Chen and Hasegawa-Johnson, 2004; Lin et al., 2019). In TTS systems, appropriate modeling of prosodic phenomena, such as pause duration, naturally used by human speakers, contributes to improved speech intelligibility and more effective transmission of meaning (Liu et al., 2022). Consequently, effective automatic identification of prosodic boundaries is expected to (i) facilitate linguistic analysis of spontaneous speech, (ii) support the creation of more informative datasets for ASR and TTS training, and (iii) improve the performance of speech-related applications operating on spontaneous speech (Galdino et al., 2026b,a).

Approaches to automatic prosodic boundary detection range from rule-based or heuristic systems, e.g. (Biron et al., 2021), to supervised machine learning models that integrate lexical and syntactic information with acoustic features, such as (Kocharov et al., 2017). The set of acoustic features differs in each study, but many of them include features related to pauses, speech rate, amplitude, and fundamental frequency (Kocharov et al., 2017; Raso et al., 2020; Biron et al., 2021). Such methods have been predominantly applied to scripted speech, where syntactic and prosodic structures tend to align, and disfluencies are relatively rare. More recently, Roll et al. (2023) proposed fine-tuning Whisper (Radford et al., 2023), a pretrained end-to-end ASR model, to segment spontaneous speech into intonation units, achieving strong performance.

Research on automatic prosodic boundary detection for Brazilian Portuguese has been conducted mainly by the speech processing group of the Federal University of Minas Gerais (Teixeira et al., 2018; Raso et al., 2020; Teixeira, 2022). However, no segmentation tool was made publicly available, hindering the easy application of the method to various studies in this language. In an effort to promote the public availability of tools, (Craveiro et al., 2024, 2025; Galdino et al., 2026b) have made open-source resources and models available for the task of prosodic segmentation of Brazilian Portuguese.

However, the low-impact open-source model from (Craveiro et al., 2025), which is the most recent approach and was trained and tested on a diverse dataset (balanced in gender and relatively diverse in terms of accents, ages, and educational levels), MuPe-Diversidades (Craveiro and Galdino, 2025), is no longer replicable due to an update in the version of the forced phonetic aligner, UFPAlign (Batista et al., 2022), used in the study. In addition to this problem, the machine learning model they evaluated includes a feature based on the difference between the F0 average of a syllable and the F0 average of the TB it belongs to, implying that it requires prior annotation of TBs. Furthermore, the authors removed all the questions uttered by the interviewers, focusing only on the respondents' answers. The authors also performed a bias evaluation on their model, and their results suggest a biased performance depending on the profile of the speaker, but the values obtained did not achieve statistical significance.

This work starts from the scenario above, which impacts the public availability of functional models for the task of prosodic segmentation, and aims to answer 4 research questions, listed below. All the models trained here and the resources used to answer the questions are available on the website: `https://github.com/nilc-nlp/ProsSegue`

1. What is the impact of using the current version of UFPAlign instead of UFPAlign's prior version? What is the impact of removing the feature that requires a previous annotation of TBs, using only 8 features? Also, what is the impact of modifying such a feature by using the F0 average of units separated by silent pauses, instead of the F0 average of units separated by TBs? And, what is the impact of training the model with and without the interviewers' speech? (see Section 4.1)

2. Considering that the test set from MuPe-Diversidades is comprised of speech from the same speakers that are present in the training set, it is especially relevant to assess the robustness of the model. Thus, what is the performance of the best model resulting from Question 1 when tested in different corpora, such as NURC-CM and samples from C-oral I and II? (see Section 4.2)

3. Is the model equally effective for diverse speaker profiles (in terms of gender, age, region of birth, and educational levels) if trained

in a less diverse but larger dataset (NURC-SP MC)? (see Section 4.3)

4. How is bias affected when augmenting a diverse dataset with a significantly larger but less diverse sample of data (MuPe-Diversidades + NURC-SP Minimum Corpus)? (see Section 4.3)

2 Related Work

Various approaches (rule-based, traditional machine learning, and deep learning) have been proposed to address the challenge of automatic prosodic segmentation (see Table 1 for studies that focus on Portuguese and English).

In (Kocharov et al., 2017), intonational units were predicted by combining syntactic and acoustic features using a Random Forest classifier. Applied to American English (Boston University Radio Speech Corpus), the study reported an F1 measure of 76% using prepared speech; for Russian, the language for which the method was originally proposed, it obtained an F1 equal to 91% in the Corpus of Professionally Read Speech. (Biron et al., 2021) used heuristics based on pause duration and speech rate discontinuities to detect prosodic boundaries in spontaneous speech from American English (Santa Barbara Corpus of Spoken American English - SBCSAE). With Montreal Forced Aligner and evaluation in Praat, the study indicated a performance of 66% on the F1 measure. By fine-tuning the Whisper model (Radford et al., 2023), (Roll et al., 2023) proposed a method (named PSST) that integrates prosodic and lexical-syntactic information for the segmentation of spontaneous speech, and functions also as a transcription tool. It achieved 87% F1 measure for American English (SBCSAE) and 73% F1 measure for British English (Intonational Variation in English (IViE) corpus - urban dialects of English spoken in the British Isles). The authors suggest that at least some of the success of PSST is due to the interaction of acoustic and lexico-syntactic information, which arises due to its integration of IU boundary detection with STT transcription.

For European Portuguese, (Hoi et al., 2022) detected boundaries through spectrograms and a convolutional neural network, using prepared speech. The technique achieved 95.6% accuracy and works for any language, but it is based solely on pauses, excluding the possibility of identifying units that do not end with silences.

(Raso et al., 2020) developed a linear discriminant analysis (LDA) classifier applied to spontaneous speech in Brazilian Portuguese, based on acoustic parameters. They use samples from C-ORAL BRASIL I and II, with prosodic boundaries annotated by experts. 111 phonetic-acoustic features were extracted, via Praat script, from the speech signal corresponding to all V-V units in windows centered on the boundaries between phonological words. The extracted features comprised 5 groups of measures: 1) Speech rate and rhythm; 2) Normalized duration; 3) Fundamental frequency; 4) Intensity; 5) Silent pause (presence and duration). Positions at which at least 50% of the annotators indicated a boundary of the same type were considered a boundary. Several models were trained to identify terminal boundaries (TBs) and non-terminal boundaries (NTBs): (i) the TB-b1 model, with pause and F0 as main parameters, was trained on Sample I (balanced), and the test on Sample II had an accuracy of 76.3% for TBs; (ii) the TB-b2 model was trained on Sample II (balanced), and the test on Sample I had an accuracy of **80.8%** for TBs; Features related to pauses and F0 were the main features associated with the identification of terminal boundaries. The best values of accuracy (in bold above) are in Table 1, for TB and NTB models. For spontaneous speech in BP, there is also the method by (Craveiro et al., 2024), which detected prosodic boundaries using the forced phonetic aligner UFPAlign (Batista et al., 2022) and the same heuristics as (Biron et al., 2021). The results indicated an F1 measure of 31%, using a 5-hour excerpt from the NURC-SP Minimal Corpus (MC), which reflects the linguistic variety of São Paulo. (Craveiro et al., 2025), inspired by the work of (Ananthakrishnan and Narayanan, 2008), used nine acoustic-prosodic features to train a Random Forest classifier, and reported binary and macro F1 measures of 55% and 77%, respectively, in the Mupe-Diversidades corpus (speech from 17 Brazilian states).

3 Methodology

3.1 Datasets

This work uses three datasets of spontaneous speech in Brazilian Portuguese that already contained annotation of prosodic segmentation: MuPe-Diversidades, NURC-SP Minimum Corpus, and samples extracted from C-ORAL BRASIL I and II.

MuPe-Diversidades is described in (Craveiro and

Table 1: Summary of prosodic segmentation research on prepared and spontaneous speech

Source	Language	Corpus	F1 Score/Accuracy	Open code?	Speech
Kocharov et al. (2017)	EN-US*	BURSC (∼10hs)	76%/86.5%	No	prepared
Biron et al. (2021)	EN-US*	SBCSAE (∼20hs)	66%/—	No	spontaneous
Roll et al. (2023)	EN-US EN-GB	SBCSAE (∼20hs) IViE (∼36hs)	87%/96% (SBC) 73%/93% (IViE)	open code	spontaneous
Hoi et al. (2022)	PT-PT*	RTP** (∼33hs)	—/95.6%	No	prepared
Raso et al. (2020)	PT-BR*	C-ORAL BRASIL I and II (∼17min) TB boundary	—/80.8%	No	spontaneous
Raso et al. (2020)	PT-BR	C-ORAL BRASIL I and II (∼17min) NTB boundary	—/75.6%	No	spontaneous
Craveiro et al. (2024)	PT-BR	Part of the NURC-SP MC (∼5hrs)	31%/—	open code	spontaneous
Craveiro et al. (2025)	PT-BR	MuPe-Diversidades (2h30min)	55%/97%	open code	spontaneous

*"EN-US" stands for American English, "EN-GB" for British English, PT-PT for European Portuguese, PT-BR for Brazilian Portuguese. **https://www.rtp.pt/

Galdino, 2025); it contains around 2.5 hours of speech extracted from life interviews of 30 people with diverse speaker profiles. The speakers were born in different cities from one of the 17 states comprised in the dataset: Alagoas, Bahia, Ceará, Paraíba, Pernambuco, Piauí, Sergipe, Pará, Rondônia, Goiás, Mato Grosso do Sul, Espírito Santo, Minas Gerais, Rio de Janeiro, São Paulo, Paraná, and Rio Grande do Sul. Each state present in the dataset is represented by 1 or 2 speakers, with excerpts of 10 or 5 minutes of speech, respectively. The corpus is also balanced in gender and diverse in age (20 to 91 years old) and educational level (no education, incomplete elementary school, complete elementary school, technical education, incomplete bachelor's degree, complete bachelor's degree, and master's degree).

Minimum Corpus is a subset of NURC-SP, composed of 21 audio files, with six formal lectures (EF), six dialogues between two informants (D2), and nine dialogues between one informant and one interviewer (DID). All of its speakers have superior education and are from the capital of São Paulo. Women are represented in 11 of the audios, and men are represented in 10. The speakers were categorized in age groups: group I: 25–35 years old, group II:36–55 years old, and group III:56 to 85 years old. The Minimum Corpus contains speech of 7 people from group I, 9 people from group II, and 5 people from group III. Since the recordings were made in the 1970s, the quality of the audios was also categorized, either as positive (good, very good, audible, clear), negative (low, very low, bass, noisy), or mixed evaluation (Santos et al., 2022). Since a few excerpts were removed due to failure of forced alignment, and one file was separated for the test set, the training set totals 17h35min19s,

C-ORAL BRASIL I and II are corpora of spontaneous speech in Brazilian Portuguese. C-ORAL I is entirely dedicated to informal speech and comprises 139 informal speech texts, and 21:08:52 hours of recording, distributed into family/private (80%) and public (20%) context. It is quite balanced in terms of speakers' gender, age, and school level (Raso and Mello, 2012). C-ORAL BRASIL II is dedicated to formal speech, comprising also a media and a telephonic corpus (Bossaglia and de Almeida Ferrari, 2019; Mello et al.). The prosodically segmented samples consist of fourteen approximately 1.5-minute excerpts of monologic male speech. Seven excerpts are drawn from C-ORAL BRASIL I (hereafter, Sample I), and seven from C-ORAL BRASIL II (Sample II), which include formal and media speech in natural contexts. The speakers represent the cities of Minas Gerais, Rio de Janeiro, Pará, São Paulo, and Santa Catarina. Age and education of these specific speakers were not disclosed. The total duration of the annotated corpus is approximately 17 minutes (Teixeira, 2022).

3.2 Models

The method that was chosen for the segmentation evaluation in this paper was reported in (Craveiro et al., 2025). It is a low-cost, low-impact approach based on a Random Forest classifier, trained with a diverse corpus, which is automatically phonetically aligned to identify the initial and final timestamps of each phone, syllable, and word. It covers solely TBs and considers 9 features at the syllable level, with the following order of importance: pause duration, energy range, difference between maximum and average energy, F0 range, nucleus vowel duration, difference between maximum and average F0, difference between minimum and average F0, difference between minimum and average energy, and difference between average F0 of the syllable and average F0 of the TB unit (f0_avgutt_diff). We assess the impact of performing a few modifications (detailed in Section 4.1) to the original model, generating a model we call MuDi.

We then perform further evaluations with MuDi

to analyze its performance on different corpora and compare its results with two other studies, using the same data samples (see Section 4.2). Finally, we experiment with expanding the training set with Corpus Minimum data. We train a model exclusively on 19 files from Minimum Corpus, excluding SP DID 234, which is separated as the test set, and name it MC. We evaluate MC's effectiveness across age, gender, educational level, and region of birth. We compare the performance of both models, analyzing whether there was any increase or decrease in biases caused by the profile of the speakers present in the training set. We also perform those evaluations on a model trained on a data sample composed of MuPe-Diversidades combined with NURC-MC, which we named MC-MuDi.

3.3 Evaluation

The results are calculated considering false positives, false negatives, true negatives, and true positives. False positives (FP) occur when the method falsely indicates a boundary. False negatives (FN) occur when the model does not identify a boundary that existed in the reference annotation. True positives (TP) occur when the model correctly identifies boundaries, and true negatives (TN) occur when the model correctly indicates a no-boundary position in places where there are no boundaries in the reference annotation. Each study uses a slightly different set of metrics, including a few of the following: accuracy, specificity, sensitivity/recall, precision, SER, macro F1 score, and binary F1 score. The two types of F1 score are differentiated according to the values considered. While the binary F1 score considers the existence of only one class: terminal breaks (TBs), the macro F1 score considers an average of the results of the class TB and the results of a secondary class that considers every position where there are no boundaries (NB). Such a category may not exist, according to the approach. (Craveiro et al., 2024), for instance, only has a class of type TB, but (Craveiro et al., 2025) considers the end of each syllable as a possible position for a boundary, so there are two classes: TBs and NBs.

4 Results and Discussion

4.1 Impact evaluation on updating the ProsSegue-ML model

This section details the results obtained when assessing the impact of making a few changes to the circumstances in which model ProsSegue-ML was trained. Table 2 compares different versions of models similar to the ProsSegue-ML model, but comprising one or more of the changes described below. The first line of the table indicates the model used in (Craveiro et al., 2025) and its last line indicates MuDi, the model we use in our further experiments.

ProsSegue-ML model relies on UFPAlign, a forced phonetic aligner developed for Brazilian Portuguese, which was updated in June of 2025. The update included changing its phonetic transcription and syllabification routines, which were independent from each other and implied a problematic procedure to align them, to a single routine that relies on a many-to-many (m2m) aligner [1] (git, 2025). Thus, the former UFPAlign version, which was used in (Craveiro et al., 2025), became obsolete. Here, we measure the impact of using this new version of UFPAlign to align the dataset (see line 2 of Table 2). There is a decrease of 1% in macro F1, which is acceptable since UFPAlign's former version is now obsolete.

Additionally, one of the 9 features used by (Craveiro et al., 2025), the difference between the average F0 of a syllable and the average F0 of the TB (f0_avg_utt_diff), requires a previous annotation of TBs, limiting the usability of the method to annotated datasets. Thus, we experiment with modifying how f0_avg_utt_diff is calculated by relying on the average F0 of units separated by pauses, instead of the F0 average of TBs. We name this altered feature f0_avg_utt_diff_2 and measure the impact of using it instead of f0_avg_utt_diff (line 3 of Table 2). We also measured the impact of simply removing such a feature (line 4 of Table 2). The table indicates that both of those experiments obtained a macro F1 of 77%, which is 1% higher than the macro F1 of the original set of features when the current version of UFPAlign is also used. We use 8 features with MuDi since it is simpler and actually yielded slightly better results (0.7711) than using f0_avg_utt_diff_2 (0.7698). It is hard to understand why this difference in performance occurred without a qualitative analysis of the segmentation, so, in future work, we intend to perform an analysis of the errors, searching for differences in alignment and segmentation among the versions of the classifier, as well as errors that occurred in each one.

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

Table 2: Table comparing the performance of different versions of the model trained in MuPe-Diversidades. It shows the impact of changing UFPAlign's version, the set of features, and training audios (including and excluding interviewers' speech, which contains several questions). v1 = f0_avg_utt_diff and v2 = f0_avg_utt_diff_2.

Model	Features	UFPAlign	Questions included ?	macro F1
Craveiro et al. 2025 model	9 (v1)	obsolete version	no	0.77
UFPAlign evaluation	9 (v1)	current version	no	0.76
Features evaluation 1	9 (v2)	current version	no	0.77
Features evaluation 2	8	current version	no	0.77
Evaluation 4	8	obsolete version	no	0.77
Evaluation 5	9 (v2)	current version	yes	0.75
MuDi	8	current version	yes	0.75

Furthermore, (Craveiro et al., 2025) uses the corpus MuPe-Diversidades, which is composed of a series of excerpts of interviews. To train ProsSegue-ML, they removed the speech of interviewers in order to preserve the balance across speaker profiles [2]. However, all the questions contained in the speech of the interviewers were, therefore, removed. Thus, we also measure the impact of training a model without removing the speech of interviewers (lines 6 and 7 of Table 2). The cost associated with this change was a decrease of 2% in performance [3] , which is considered acceptable, since it also implies including a more significant amount of TBs that represent questions. We suspect that this decrease may be due to the different characteristics of TBs composed of questions, as they end with a higher intonation, instead of a lower intonation, typical of TBs composed of affirmations. Considering that the impact of these three changes was not considered very significant and considering the circumstances that led us to those changes, the model we use for the following experiments includes all of the changes. We refer to it as ProsSegue-ML-MuDi, or simply MuDi.

4.2 Robustness Evaluation

In (Craveiro et al., 2025), despite the usage of a relatively diverse corpus (detailed in (Craveiro and Galdino, 2025)) to train and test their model, since the speech excerpts of the training and test sets are from the same speakers, the model is relatively biased, as speakers' unique voices are present in both sets [4]. Here, we present a robustness evaluation by testing MuDi, the updated model, in different corpora. Table 3 compares the performance of the model in samples from corpora C-ORAL BRASIL I and II (Mello et al., 2012; Mello et al.), NURC-SP Minimum Corpus (Santos et al., 2022), and MuPe-Diversidades' test set (Craveiro and Galdino, 2025). The macro F1 obtained with the MuPe-Diversidades test set is 14% higher than that obtained with NURC-CM, and 4% higher than the one obtained with samples from C-ORAL I and II. Those results show that the model functions well in C-ORAL BRASIL, which is a small and less complex dataset, but functions consistently worse for NURC-CM. The average macro F1 considering all corpora is 69%. Regarding NURC-SP Minimum Corpus, its results could be lower (binary f1-score of 26% and macro f1-score of 61%) due to the poorer quality of the audios and to the presence of occasional overlaps of voices.

4.2.1 ProSegue-ML-MuDi x other methods

Tables 4 and 5 compare the results obtained with MuDi when tested in the same corpora used in other works.

(Craveiro et al., 2024) presents a methodology based on three heuristics that identify pauses and differences of speech rate, which is justified since the lengthening of speech rate at the end of a unit, together with the acceleration at its beginning, is a characteristic of prosodic boundaries among units (Biron et al., 2021). Table 4 shows the comparison between the results they reported and the results we obtained by testing ProsSegue-ML-MuDi in the data samples that they used. The machine-learning-based method shows values that range from a decrease of 6% to an increase of 4% in binary F1. The performance, which was lower than expected, could be explained by the different characteristics

[2]While the interviewees were carefully selected according to gender, age, and region of birth, the interviewers were not controlled in such a manner.

[3]This decrease of 2% may have also been affected by the implementation of a correction on the attribution of labels, which avoided wrongly labeling a sequence of syllables as TBs. This type of error occurred only when the last TB of the reference transcription ended before the ending time of a few syllables, as aligned with UFPAlign. In those cases, before the correction, all of those adjacent syllables were being labeled as "TB" instead of "NB".

[4]Note that this bias will occur for all cases that use samples from MuPe-Diversidades as the test set and samples from MuPe-Diversidades in the training set

Table 3: Table comparing the performance of ProSSegue-ML across different corpora. F1, recall, and precision solely considering TBs are at the left, and the average of TBs and NBs (no boundary) are at the right. Acc. = accuracy, bF1 = binary F1 score, mF1 = macro F1 score.

Test Corpus	Size	Gender	Region	Age range	Education	bF1 / mF1	Acc.	Precision	Recall
Mupe-Diversidades	∼30min	balanced	17 states	20-91	varied	53% / 75%	95%	55% / 76%	51% / 74%
C-oral I and II	∼17min	male	5 states	-	-	44% / 71%	95%	32% / 66%	60% / 83%
NURC-CM	∼17.5 hrs	balanced	SP	25-85	higher	26% / 61%	93%	24% / 60%	30% / 63%
Total Avg	∼18.3hrs	-	-	-	-	41% / 69%	94%	37% / 67%	47% / 73%

Table 4: Overall results of the baseline method applied only to TBs from four inquiries from NURC-MC, compared to the machine learning approach performance applied to the same inquiries.

	SP_EF_156		**SP_DID_242**	
	ProsSegue-Baseline	**ProsSegue-ML**	**ProsSegue-Baseline**	**ProsSegue-ML**
F1	0.18	**0.22**	**0.29**	0.23
p	0.12	**0.17**	**0.22**	0.2
r	**0.38**	0.33	**0.41**	0.29
ser	3.55	**2.28**	02.03	**1.89**
	SP_D2_255		**SP_D2_360**	
	ProsSegue-Baseline	**ProsSegue-ML**	**ProsSegue-Baseline**	**ProsSegue-ML**
F1	0.16	**0.19**	**0.2**	0.17
p	0.11	**0.14**	**0.14**	**0.14**
r	**0.32**	0.27	**0.37**	0.21
ser	3.31	**2.34**	2.92	**2.04**

Table 5: Table comparing the performance of Prossegue-ML vs. models TB-b1 and TB-b2, trained in balanced datasets, published at (Teixeira, 2022) (see Appendix A for composition of the Sample I and Sample II).

Article	Model	Train set	Test set	Accuracy	Specificity	Sensitivity
Teixeira 2022	TB-b1	Sample I	Sample II	87.0%	88.0%	74.0%
Teixeira 2022	TB-b2	Sample II	Sample I	91.0%	92.0%	81.0%
Craveiro et al.2025	ProsSegue-ML	MuPe-Diversidades	Sample II	94.6%	95.2%	65.4%
Craveiro et al.2025	ProsSegue-ML	MuPe-Diversidades	Sample I	95.4%	96.2%	71.8%

of the audios from the Minimum Corpus, suggesting that the model may not be very robust and may benefit from training with more representative samples. It also suggests that the model could be relying on similar features, as we know that the most important feature of ProsSegue-ML-MuDi is the duration of pauses, and that ProsSegue-Baseline uses a rule based on silences to identify boundaries.

(Teixeira, 2022) presents an approach based on an LDA classifier trained with a wider set of features that consider pauses, F0, intensity, speech rate, and rhythm. They present several models, and we selected two models that facilitated the comparison between the works. TB-b1 is a model trained in a balanced sample of Sample I with 8 features related to pauses, F0, and articulation rate. TB-b2 is a model trained in a balanced sample of Sample II with 5 acoustic features, including parameters related to pauses and F0 of units adjacent to terminal breaks. The most relevant feature, with a significantly higher importance than the others, in both models, is the presence of pauses.

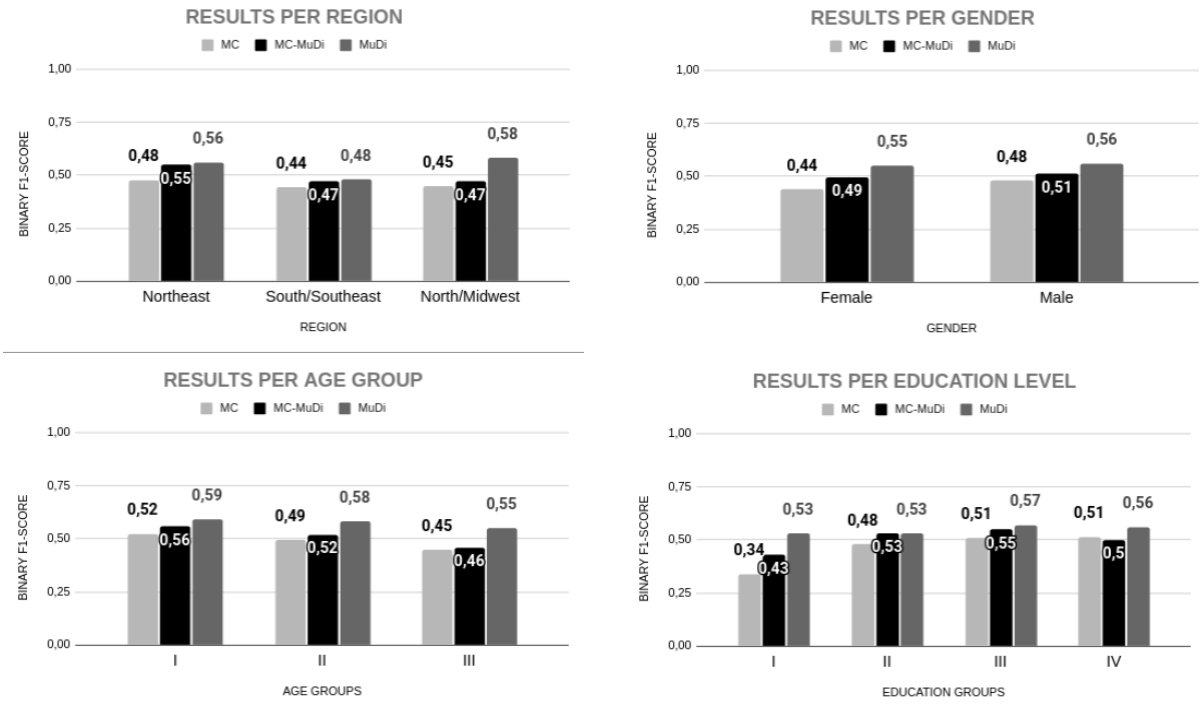
Table 5 shows a comparison of the results that their models obtained and the results obtained with ProSegue-ML-MuDi tested on the same data samples, extracted from corpora C-ORAL BRASIL I and II. ProsSegue-ML-MuDi performs from 4% to 7% higher in accuracy and specificity, while TB-b1 and TB-b2 win by approximately 9% in terms of sensitivity, implying that TB-b1 and TB-b2 are better at not missing TBs, while ProsSegue-ML-MuDi is better at indicating boundaries solely where they actually exist.

4.3 Results per Speaker Profile

Considering how relevant it is that a model is equally effective to all individuals, regardless of their characteristics, we present results per region, gender, age group, and education group, relevant aspects for speech analysis (Craveiro and Galdino, 2025). Figure 1 exhibits results obtained with three different models analyzed under the perspective of each of those aspects. ProsSegue-ML-MC, or simply MC, was trained exclusively with data from the Minimum Corpus, to evaluate whether a much larger dataset (approximately 17 hours) but less diverse (all speakers are from São Paulo and have higher education) suffices for an even performance across diverse speaker profiles. We trained ProsSegue-ML-MuDi with NURC-MC and MuPe-Diversidades to evaluate whether it is worth it to expand a diverse corpus (MuPe-Diversidades) with a larger less diverse corpus (NURC-MC), that is, if the model functions even better for each speaker profile group or if the inequalities of performance also grow.

Regarding gender, ProsSegue-ML-MC performs

Figure 1: Comparison of the performance of the three models (one trained with Minimum Corpus (MC), one trained with MuPe-Diversidades (MuDi), and one trained with a dataset composed of both (MC-MuDi) according to region, gender, age group, and education group, respectively.



4% better for males than for females, and such inequality grew by 3% when we consider results obtained for the model trained exclusively with MuPe-Diversidades train set (MuDi), as reported in (Craveiro et al., 2025). With MC-MuDi, despite the growth of 1% in inequality in respect to MuDi, it is better than MC by 2%. As for region, ProsSegue-ML-MC performed from 8% to 10% worse for speakers from the North and Southeast, when compared to the other regions, which is very curious since its training set contained solely speakers from São Paulo. The difference per region decreased in both new models, reaching 0% to 8% with MC-MuDi, and 1% to 4% with MC, depending on the groups compared. Regarding age group, both new models also favor younger speakers, and the bias grew from 1%-4% with MuDi, to 3%-7% with MC, to 4%-10% with MC-MuDi. It is surprising that the greater bias came from the model trained in both NURC-MC and MuPe-Diversidades. Finally, the difference also grew according to the educational level of the speakers. MC-MuDi and MC seem to strongly favor more educated speakers (groups III and IV), reaching maximum differences of 12% and 17%, respectively, among groups, when speakers with no education are concerned, while MuDi indicated a maximum difference of 4% among different education groups. The performance difference among groups II, III, and IV reaches differences ranging from 0%-5%. Thus, the greater inequality observed, considering all aspects of speaker profiles that we considered, is the bias disfavoring non-educated speakers.

A decrease in performance was already expected in both new models. However, despite the much larger dataset (around 15 hours) and the 19 different speakers comprised in NURC-CM, the difference in performance among different speaker profiles grew, suggesting it may be more valuable to prefer diversity of speakers over quantity of hours of audio available of a more restricted and less diverse set of speakers. Nonetheless, we emphasize that none of those bias evaluations is statistically relevant. The p-value of MC results per region, gender, age group, and education group is, respectively, 0.22, 0.08, 0.19, and 0.34. As for MC-Mudi, the p-values were 0.08, 0.09, 0.09, and 0.61, respectively. We reinforce that a qualitative analysis would be very beneficial to understand what kinds of segmentation errors occurred, as well as to infer limitations of the model. And although we cannot know the reasons for the difference in performance across different speaker profiles or which of their specific aspects were favored or disfavored without

such analysis, we know that certain individual characteristics of the speakers might have influenced the classifier's decisions as the performance of the model varied significantly according to the speaker evaluated. Thus, we definitely need to train using more speakers to improve the model's generalization capability.

5 Final Considerations

In this study, we focused on a low-impact, open-source, automatic prosodic segmentation approach published in (Craveiro et al., 2025). We trained an updated model, MuDi, with the aim of increasing accessibility by updating the version of the forced phonetic aligner, experimenting with the set of features to simplify the requirements of the approach, and assessing the impact of our changes. We explored how MuDi behaves when tested in other corpora, how it compares to other studies, and experimented with the size and diversity of the training set to analyze how bias could be affected. We observed that it maintained a macro F1 above 70% when tested in MuPe-Diversidades teste, and in samples from C-ORAL BRASIL I and II, but had a decrease in performance, reaching 61% of macro F1, when tested in NURC-CM. We also observed that MuDi was probably overspecializing in MuPe-Diversidades, since when we tested MC-MuDi, expanding the dataset with NURC-CM, the performance when testing in MuPe-Diversidades decreased. We could also observe that the results of macro F1 achieved values from 34% to 52% with MC, and values from 43% to 56% with MC-MuDi, with a difference ranging from 1% to 17% in performance, according to the examined speaker profile aspect, values higher than the 1% to maximum 10% obtained training exclusively with MuDi, suggesting that it is worthwhile to prioritize speaker diversity over the quantity of speaking hours. However, it is worth recalling that the bias evaluation did not achieve statistical relevance. For future work, we intend to perform a qualitative evaluation of the segmentation and to test the model on more datasets, including possibly an extended version of MuPe-Diversidades with more speakers from each state and speakers representing all states of Brazil. Also, it would be beneficial to follow the example of the PSST! approach (Roll et al., 2023), a recent study that obtained excellent results, by finetuning ecologically efficient neural models but with large and diverse Brazilian Portuguese datasets that are manually annotated with prosodic segmentation, a work that could begin with NURC-SP Minimum Corpus.

Limitations

We emphasize that our model is still highly misrepresented. It still lacks training with a huge amount of speech excerpts that represent different aspects of Brazilian speech. Our 30 speakers are also a small number to perform statistically relevant diversity tests. This limitation of data is due to the low resource scenario on manual prosodic segmentation annotation, which is a very demanding and long process, limiting us to datasets that were already annotated.

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.

References

2025. Error when generating syllphones tier · Issue #19 · falabrasil/ufpalign — github.com. https://github.com/falabrasil/ufpalign/issues/19. Accessed at 02-02-2026.

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

Cassio Batista, Ana Larissa Dias, and Nelson Neto. 2022. Free resources for forced phonetic alignment in brazilian portuguese based on kaldi toolkit. *EURASIP Journal on Advances in Signal Processing*, 2022(1):11.

Tirza Biron, Daniel Baum, Dominik Freche, Nadav Matalon, Netanel Ehrmann, Eyal Weinreb, David Biron, and Elisha Moses. 2021. Automatic detection of prosodic boundaries in spontaneous speech. *PLoS ONE*, 16(5):1–21.

Giulia Bossaglia and Lucia de Almeida Ferrari. 2019. The c-oral-brasil project: varied resources for the study of spoken brazilian portuguese. *Journal of speech sciences*.

Meredith Broussard. 2018. *Artificial unintelligence: How computers misunderstand the world*. mit Press.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Ken Chen and Mark Hasegawa-Johnson. 2004. How prosody improves word recognition. In *Proc. Speech Prosody 2004*, pages 583–586.

Giovana Craveiro, Caroline Alves, Flaviane Svartman, and Sandra Aluísio. 2025. Machine learning classifiers with acoustic features for prosodic segmentation in brazilian portuguese: A comprehensive evaluation. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 113–124. SBC.

Giovana Meloni Craveiro and Julio Cesar Galdino. 2025. Diversity in data for speech processing in brazilian portuguese. In *Intelligent Systems*, pages 122–136, Cham. Springer Nature Switzerland.

Giovana Meloni Craveiro, Vinicius Gonçalves Santos, Gabriel Jose Pellisser Dalalana, Flaviane R. Fernandes Svartman, and Sandra Maria Aluísio. 2024. Simple and fast automatic prosodic segmentation of Brazilian Portuguese spontaneous speech. In *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. Available at https://aclanthology.org/2024.propor-1.4/.

Julio Galdino, Sidney Leal, Leticia de Souza, Rodrigo Lima, Antonio Moreira, Arnaldo Candido, Miguel Oliveira, Edresson Casanova, and Sandra Aluísio. 2026a. The impact of prosodic segmentation on speech synthesis of spontaneous speech. In *Intelligent Systems*, pages 547–561, Cham. Springer Nature Switzerland.

Julio Cesar Galdino, Rian Pereira Fernandes, Giovana Meloni Craveiro, Caroline Adriane Alves, Sidney Evaldo Leal, Arnaldo Candido Junior, Flaviane Romani Fernandes-Svartman, and Sandra Maria Aluisio. 2026b. Investigating the effect of automatic prosodic segmentation on speech synthesis for brazilian portuguese. Accepted at Speech Prosody 2026.

Lap Man Hoi, Yuqi Sun, and Sio Kei Im. 2022. An automatic speech segmentation algorithm of portuguese based on spectrogram windowing. In *2022 IEEE World AI IoT Congress (AIIoT)*, pages 290–295.

Jui-Ting Huang, Mark Hasegawa-Johnson, and Chilin Shih. 2008. Unsupervised prosodic break detection in Mandarin speech. In *Proc. Speech Prosody 2008*, pages 165–168.

Je Hun Jeon and Yang Liu. 2009. Semi-supervised learning for automatic prosodic event detection using co-training algorithm. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 540–548, Suntec, Singapore. Association for Computational Linguistics. Available at https://aclanthology.org/P09-1061.

Daniil Kocharov, Tatiana Kachkovskaia, and Pavel Skrelin. 2017. Eliciting Meaningful Units from Speech. In *Proc. Interspeech 2017*, pages 2128–2132.

Cheng-Hsien Lin, Chung-Long You, Chen-Yu Chiang, Yih-Ru Wang, and Sin-Horng Chen. 2019. Hierarchical prosody modeling for Mandarin spontaneous speech. *The Journal of the Acoustical Society of America*, 145(4):2576–2596.

Shimeng Liu, Yoshitaka Nakajima, Lihan Chen, Sophia Arndt, Maki Kakizoe, Mark A. Elliott, and Gerard B. Remijn. 2022. How pause duration influences impressions of english speech: Comparison between native and non-native speakers. *Frontiers in Psychology*, 13.

Heliana Mello, Maryualê Malvessi Mittmann, H. P. Vale, and P.O. Cortes. 2012. Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação. In *CORAL-BRASIL I: corpus de referência do português brasileiro falado informal*. Editora UFMG.

Heliana Mello, Tommaso Raso, Lúcia de Almeida Ferrari, and Bruno Neves Rati de Melo Rocha. C-ORAL–Brasil II: Corpus de referência do português brasileiro falado falado formal, mídia e telefone. Available at http://c-oral-brasil.org/c-oral-brasil-ii_N.php. Accessed at 02-02-2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Tommaso Raso and Heliana Mello. 2012. The c-oral-brasil i: reference corpus for informal spoken brazilian portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 362–367. Springer.

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

Nathan Roll, Calbert Graham, and Simon Todd. 2023. PSST! prosodic speech segmentation with transformers. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 476–487, Singapore. Association for Computational Linguistics. Available at https://aclanthology.org/2023.conll-1.31/.

Lívia Ruback, Denise Carvalho, and Sandra Avila. 2022. Mitigating bias in machine learning: A socio-technical analysis. *iSys - Brazilian Journal of Information Systems*, 15(1):23:1–23:31.

Vinícius G. Santos, Caroline Adriane Alves, Bruno Baldissera Carlotto, Bruno Angelo Papa Dias, Lucas Rafael Stefanel Gris, Renan de Lima Izaias, Maria Luiza Azevedo de Morais, Paula Marin de Oliveira, Rafael Sicoli, Flaviane Romani Fernandes-Svartman, Marli Quadros Leite, and Sandra Maria Aluísio. 2022. CORAA NURC-SP Minimal Corpus: a manually annotated corpus of Brazilian Portuguese spontaneous speech. In *Proc. IberSPEECH 2022*, pages 161–165.

Bárbara Teixeira, Plínio Barbosa, and Tommaso Raso. 2018. Automatic detection of prosodic boundaries in Brazilian Portuguese spontaneous speech. In *Computational Processing of the Portuguese Language - 13th International Conference, PROPOR 2018*, pages 429–437, Cham. Springer International Publishing.

Bárbara Helohá Falcão Teixeira. 2022. *Detecção automática de fronteiras prosódicas na fala espontânea*. Ph.D. thesis, Universidade Federal de Minas Gerais, Belo Horizonte.

Colin W. Wightman and Mari Ostendorf. 1991. Automatic recognition of prosodic phrases. *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, 1:321–324.

A Appendix

A.1 Conflicts in the attribution of labels

In order to test the models in the new corpora, it was necessary to attribute a label to every syllable uttered in each audio. In the process, a few alignment conflicts were found, so we explain how they were dealt with. Ideally, the last syllable of TBs would receive the label "TB"and all others would receive "NB", meaning "no boundary".

However, as we used UFPAlign to identify the initial and final time of the syllables, there are moments when the timestamps of syllables and TBs are not perfectly aligned, that is, UFPAlign might have indicated that a certain syllable ended at 8,09 seconds, for instance as happened in SP D2 360, while the reference annotation indicates that the same syllable was the last syllable of a TB ending in 11,11 seconds, implying that such syllable ended at 11,11 seconds instead of 8,09 seconds. The timestamps of the final times of TBs extracted from the reference annotation are always prioritized, meaning that in cases like this one, this syllable would have received a label "NB", despite being the last syllable of the TB, since it ended before the time indicated at the reference file. And that the syllable that ended near 11,11 seconds, according to UFPAlign's forced alignment, was the one that received the label "TB". In cases where these circumstances were found at the end of the file, the last syllable of the last TB was labeled as "NB". There were also cases where the contrary occurred, that is, UFPAlign's alignment indicated that the final syllable of a TB ended after the final time of the TB as indicated by the annotation. And there were also a few cases where a sequence of syllables had its starting and ending time after the final time of the last TB. In those cases, the syllable that received the label "TB" that corresponds to the indication of the final TB was the syllable with the most approximate time to the end of the TB, and all those "extra" syllables that occurred later were labeled "NB". The code also favors including the first syllable of the following TB at the current TB in case the syllable starts before the current TB ends. Thus, in such cases, the initial syllable of the following TB would be labeled as TB instead of the final syllable of the current TB.