



OPEN ACCESS

The whole content of *Cadernos de Linguística* is distributed under Creative Commons Licence CC - BY 4.0.

EDITORS

- Miguel Oliveira, Jr. (UFAL)
- René Almeida (UFRB)

REVIEWERS

- Erik Thomas (NCSU)
- Beatriz Medeiros (USP)
- Paul Reed (UA)
- Marisa Cruz (UL)

ABOUT THE AUTHORS

- **Leônidas Silva Jr.**
Conceptualization; Methodology; Software; Formal Analysis; Data Curation; Visualization; Project Administration; Writing – Original Draft; Writing – Review & Editing.
- **Luciani Ester Tenani**
Investigation; Resources; Validation; Data Curation; Writing – Original Draft; Writing – Review & Editing; Funding Acquisition.
- **João Marcelo Monte**
Software; Validation; Supervision; Writing – Original Draft; Writing – Review & Editing.

Received: 03/21/2026

Accepted: 07/01/2026

Published: 09/21/2026

HOW TO CITE

SILVA JR., L.; TENANI, L. E.; MONTE, J. M. (2026). The Influence of Prosodic-Acoustic Features in Automatic Speech Recognition of Two Inland Brazilian Dialects: A Pilot Study. *Cadernos de Linguística*, v. 7, n. 7, e1016.



CHECK FOR
UPDATES

PILOT STUDY

THE INFLUENCE OF PROSODIC-ACOUSTIC FEATURES IN AUTOMATIC SPEECH RECOGNITION OF TWO INLAND BRAZILIAN DIALECTS: A PILOT STUDY

Leônidas SILVA JR.

Department of Modern Languages – State University of Paraíba (UEPB)
Guarabira, PB, Brazil

Luciani Ester TENANI

Department of Language and Literature Studies – State University of São Paulo (UNESP)
São José do Rio Preto, SP, Brazil

João Marcelo MONTE

Department of Operational Technology – Pernambuco Water and Sewage Company (COMPESA)
Recife, PE, Brazil

ABSTRACT

This study investigates the role of prosodic-acoustic features in distinguishing two inland varieties of Brazilian Portuguese, Paraíba (PB) and São Paulo (SP), and examines how these features influence automatic speech recognition (ASR) performance. Grounded in sociophonetic theory and dialect-aware speech modeling, this pilot study evaluates whether prosodic-acoustic descriptors of duration, fundamental frequency (f0), and intensity provide robust cues for dialect differentiation and to what extent automatic dialect classification can be explained by the relative contribution of these acoustic dimensions. The study integrates two complementary methodological stages. Methodology 1 encompasses corpus design, acoustic extraction procedures, and inferential statistical modeling. Twenty

university students (ten per dialect) produced controlled reading recordings, yielding 460 speech samples. Prosodic-acoustic parameters were extracted and analyzed using Linear Mixed-Effects Models to identify statistically significant dialect differences while controlling for inter-speaker variability. Methodology 2 applies automatic dialect classification using six machine-learning classifiers to evaluate the discriminative capacity of those features. The results are presented in two stages. Results 1 demonstrate that duration-related measures—particularly pause variability—provide the strongest explanatory power for distinguishing PB and SP, followed by intensity-related parameters such as spectral tilt and intensity variation coefficient, whereas f_0 measures show comparatively smaller effects. Results 2 show that classification accuracy ranges from 94.6% to 97.8%, with tree-based models achieving the highest performance. Feature-importance analyses reveal a hierarchical contribution of acoustic parameters dominated by rhythmic-temporal organization. The convergence between explanatory statistical modeling and predictive classification demonstrates that inland dialect differentiation in Brazilian Portuguese is acoustically structured, prosodically organized, and computationally detectable. These findings contribute to the sociophonetic description of inland Brazilian Portuguese varieties and provide empirical grounding for incorporating dialect-sensitive prosodic modeling into ASR systems.

KEYWORDS

Prosodic-Acoustic Features; Dialect Classification; Automatic Speech Recognition; Brazilian Portuguese; Sociophonetics.

RESUMO

Este estudo investiga o papel de parâmetros prosódico-acústicos na distinção de duas variedades dialetais do interior brasileiro, na Paraíba (PB) e em São Paulo (SP), e examina como tais parâmetros influenciam o desempenho de sistemas de reconhecimento automático de fala. Fundamentado em estudos da Sociofonética e na modelagem de fala automática sensível à variação dialetal, este estudo piloto avalia se descritores prosódico-acústicos de duração, frequência fundamental (f_0) e intensidade constituem pistas robustas para a diferenciação dialetal e em que medida a classificação automática de dialetos pode ser explicada pela contribuição relativa dessas dimensões acústicas. A pesquisa integra duas etapas metodológicas complementares. A Metodologia 1 envolve o desenho do corpus, os procedimentos de extração acústica e a modelagem estatística inferencial. Vinte estudantes universitários (dez por dialeto) produziram gravações de leitura

controlada, totalizando 460 amostras de fala. Os parâmetros prosódico-acústicos foram extraídos e analisados por meio de Modelos Lineares de Efeitos Mistos, a fim de identificar diferenças dialetais estatisticamente significativas, controlando a variabilidade interindividual. A Metodologia 2 aplica a classificação automática de dialetos por meio de seis classificadores baseados em aprendizado de máquina, com o objetivo de avaliar a capacidade discriminativa dessas características acústicas. Os resultados são apresentados em duas etapas. Resultados 1 indicam que medidas relacionadas à duração, especialmente a variabilidade de pausas, apresentaram maior poder explicativo na distinção entre PB e SP, seguidas por parâmetros de intensidade, como o declive espectral e o coeficiente de variação da intensidade, enquanto medidas baseadas em f_0 exibiram efeitos comparativamente menores. Resultados 2 mostram que a acurácia da classificação automática variou entre 94,6% e 97,8%, com modelos baseados em árvores apresentando o melhor desempenho. A hierarquização da importância das variáveis confirmou a predominância de parâmetros rítmico-temporais no processo de classificação. A convergência entre modelagem estatística explicativa e classificação preditiva demonstra que a diferenciação dialetal entre variedades do interior do Brasil é acusticamente estruturada, hierarquicamente organizada e computacionalmente detectável. Esses resultados contribuem para a descrição sociofonética das variedades dialetais do interior do português brasileiro e oferecem fundamentação empírica para a incorporação de modelagem prosódica sensível à variação dialetal em sistemas de reconhecimento automático de fala.

PALAVRAS-CHAVE

Parâmetros Prosódico-Acústicos; Classificação Dialetal; Reconhecimento Automático de Fala; Português Brasileiro; Sociofonética.

INTRODUCTION

Regional speech variation constitutes a central dimension of linguistic identity, reflecting regional, social, and interactional histories embedded in speech production. Language varieties (and their regional dialects) are not only differentiated by segmental patterns but also by prosodic organization, including rhythmic structuring, pitch variation, and intensity distribution. These prosodic-acoustic dimensions, in interaction with segmental features and intonational contour patterns, operate as socially meaningful cues, indexing regional belonging and contributing to listeners' perception of dialectal identity. While human listeners are highly sensitive to such variation, Automatic Speech Recognition (ASR) systems frequently struggle to accommodate dialectal diversity, particularly when acoustic models underrepresent structured sociophonetic variability (Braid, 2003; Behravan, 2016; Flege, 1995; Kumar; Singh, 2019; Moyer, 2013).

Recent research in speech technology has shown that performance disparities in ASR systems often reflect insufficient modeling of acoustic-prosodic variation rather than random noise in the speech signal (Wassink; Gansen; Bartholomew, 2022). Temporal organization, intonation dynamics, and voice quality settings may systematically differ across regional varieties, influencing classification accuracy and recognition robustness. Despite these developments, inland Brazilian dialects remain comparatively underrepresented in dialect-sensitive ASR research. This gap motivates the present study, which investigates two inland Brazilian Portuguese varieties: Paraíba (PB) and São Paulo (SP).

In this pilot study, we hypothesize that prosodic-acoustic descriptors based on duration, fundamental frequency (f_0), intensity, and voice quality measures derived from long-term spectral analysis constitute structured acoustic markers distinguishing inland PB and SP dialects of Brazilian Portuguese. A secondary hypothesis holds that these same features provide sufficient discriminative information for automatic dialect classification, with feature-importance analyses expected to reveal a hierarchical contribution of acoustic parameters.

This study addresses the following research questions (RQ):

- **RQ1** – Do prosodic-acoustic descriptors of duration, f_0 , intensity, and long-term spectral voice quality provide robust and reliable cues for distinguishing the inland Brazilian Portuguese dialects PB and SP?
- **RQ2** – To what extent can automatic dialect classification be explained by the relative contribution of prosodic-acoustic features, and how do these features affect ASR performance?

This paper is structured as follows. Section 1 presents the theoretical framework, situating the study within sociophonetic research and dialect-aware ASR modeling. Section 2 details the

methodology, including participants, corpus design, acoustic extraction procedures, statistical modeling, and machine learning (ML) implementation. Section 3 reports the inferential results. Section 4 presents the machine-learning methodology, and Section 5 reports the classification results. Section 6 discusses the findings in relation to the research questions and existing literature. Sections 7 and 8 present the conclusion and limitations, respectively.

1. THEORETICAL FRAMEWORK: PROSODY, SOCIOPHONETIC VARIATION, AND ASR SYSTEMS

Recent research has consistently shown that, despite significant technological advances, ASR systems remain sensitive to dialectal and sociophonetic variation. Rather than reflecting random noise in speech signals, recognition errors often stem from the underrepresentation of socially structured acoustic patterns in acoustic model training (Wassink; Gansen; Bartholomew, 2022). Drawing on the variationist tradition, Wassink, Gansen and Bartholomew (2022) argue that linguistic variables, realized through fine-grained segmental and prosodic differences, are conditioned by regional and social factors. However, such sociophonetic knowledge has rarely been fully integrated into ASR architecture.

A central claim of Wassink, Gansen and Bartholomew (2022) is that dialect-related recognition failures arise primarily at the level of acoustic modeling rather than language modeling. While language models may accommodate non-standard lexical or syntactic patterns, acoustic models are directly dependent on phonetic realization, including vowel quality, rhythmic timing, pitch dynamics, and spectral structure. Importantly, dialects are not treated as discrete, homogeneous entities; instead, recognition errors emerge from clusters of co-occurring acoustic features that may overlap across social and regional boundaries. From this perspective, dialect differentiation is gradient and configuration-based, and ASR failures reflect insufficient modeling of these structured acoustic repertoires.

This sociophonetic account is further reinforced by Torgbi *et al.* (2025), who examined the performance of OpenAI's Whisper large-v3 model on Scottish regional dialects within public service contexts. Their results show that, despite being trained on extensive multilingual and multi-domain datasets, the out-of-the-box model exhibits significantly higher word error rates for regionally accented speech. These findings challenge the assumption that training scale alone ensures dialectal robustness. The authors further demonstrate that fine-tuning the model on region-specific datasets leads to consistent reductions in recognition errors, indicating that dialect-sensitive phonetic and prosodic configurations can be effectively internalized when adequately represented in acoustic modeling. Moreover, their analysis highlights that performance disparities are not merely technical

artifacts but reflect asymmetries in the sociolinguistic distribution of training data. Both works – Wassink, Gansen and Bartholomew (2022) and Torgbi *et al.* (2025) – converge on the view that dialectal variability must be explicitly modeled at the acoustic level and that persistent ASR performance gaps are rooted in structured sociophonetic underrepresentation rather than in random variation.

In addition to the studies discussed above, Koenecke *et al.* (2020) examined the ability of five state-of-the-art commercial ASR systems – developed by Amazon, Apple, Google, IBM, and Microsoft – to transcribe structured sociolinguistic interviews with 42 speakers of General American (GA) English and 73 African American Vernacular English (AAVE) speakers, finding that all five systems exhibited substantial racial disparities, with an average word error rate of 0.35 for AAVE speakers compared with 0.19 for GA speakers. The authors trace these disparities directly to the underlying acoustic models, demonstrating that the performance gap was equally large on a matched subset of identical phrases produced by speakers of both groups, which points to the phonological, phonetic, and prosodic characteristics of AAVE as the primary drivers of recognition failure rather than lexical or grammatical features. Earlier work by Tatman (2017) had already identified similar patterns: an evaluation of YouTube's automatically-generated captions across two genders and five dialect groups demonstrated robust differences in accuracy across both gender and dialect, with systematically lower performance for speakers from Scotland – one of the earliest empirical demonstrations of the relationship between regional speech variety and ASR performance disparities. These findings were further extended by Tatman and Kasten (2017), who compared the accuracy of Bing Speech and YouTube captions across gender, race, and four dialects of American English, finding that the lowest average error rates consistently corresponded to GA speakers.

Markl (2022) investigated the performance of Google and Amazon transcription services across multiple accents of British English and found that both systems perform substantially worse for L2 speakers and for speakers of stigmatized regional varieties of British English, exploring the structural and sociopolitical mechanisms through which speech and language technologies reproduce disadvantage for marginalized language communities. Expanding the scope further, Chan *et al.* (2022) evaluated the effectiveness of automatic captioning for world English varieties, comparing word error rates across English varieties supported and unsupported by a major commercial ASR provider, and finding that English varieties classified as supported consistently yield lower error rates, while unsupported varieties – including those typologically distant from the training data – exhibit systematically higher error rates.

These studies establish a robust and cross-linguistically consistent pattern: ASR performance disparities correlate with the acoustic distance between a speaker's variety and the dominant variety in the training corpus, and this distance is structured sociophonetically, reflecting systematic sources of phonetic variation that have yet to be adequately represented in acoustic model architectures, rather than being a product of random variation or recording conditions.

A systematic review of 100 studies on speech synthesis (Galdino *et al.*, 2025) reveals that f0 is by far the most extensively modeled prosodic dimension across text-to-speech systems, present in 95 of the 100 studies surveyed, with duration and intensity following as secondary but consistent correlates. However, the review also demonstrates that the relative weighting and operationalization of these parameters vary substantially across systems, languages, and applications: f0 dominates in synthesis architectures optimized for perceived naturalness and intonational expressiveness, while duration and intensity receive proportionally greater emphasis in rhythm-sensitive and prominence-driven contexts. These results indicate that while f0, duration, and intensity collectively define the parameter space of prosodic modeling, the internal hierarchy among them, that is, which dimension carries the greatest explanatory weight, is neither fixed nor universal across linguistic and dialectal contexts. This observation directly motivates the present investigation: rather than assuming a predetermined hierarchy derived from synthesis research, this study empirically examines how duration, f0, and intensity are hierarchically organized in the specific context of two inland Brazilian Portuguese dialect varieties. Importantly, the review also shows that research in this area remains strongly concentrated in a small set of languages – particularly English, Mandarin, and Japanese – while many languages and regional varieties remain comparatively underexplored. In this sense, extending prosodic-acoustic investigation to underrepresented varieties of Brazilian Portuguese contributes not only to sociophonetic description but also to the broader development of speech technologies capable of modeling prosodic variability across diverse linguistic contexts.

These perspectives directly motivate investigations of regional variation within Brazilian Portuguese. Studies such as Constantini and Barbosa (2015) posit that Brazilian Portuguese exhibits strong prosodic and rhythmic diversity across regions, yet inland varieties remain comparatively underrepresented in speech technology research. The authors investigated the role of prosodic-acoustic parameters in distinguishing regional varieties in seven Brazilian Portuguese dialects and analyzed a set of prosodic descriptors including speech rate, duration variability of vowel-to-vowel units, spectral emphasis, and f0 median. Their statistical analyses revealed significant differences across varieties for parameters related to intensity, pitch, and rhythmic structure, demonstrating that prosody can function as a reliable indicator of dialectal variation. Their research shows that prosodic cues such as duration variability and spectral emphasis can separate regional groupings within Brazilian Portuguese, confirming the view that dialectal identity is reflected not only in segmental features but also in systematic suprasegmental (i.e., prosodic) organization.

More recent research on Brazilian prosody has demonstrated that prosodic parameters, such as duration and f0 contours, function as stable markers of Brazilian Portuguese (Barbosa; Alvarenga, 2024). In addition, Silveira, Albert and Grice (2025) compare the prosodic configuration of Alagoas and São Paulo dialectal varieties, focusing on f0 dynamics in boundary contexts. Using measures of f0 slope and intrasyllabic synchrony, the study identifies systematic regional differences in pitch excursion and contour direction, particularly at terminal and non-terminal boundaries. The results

demonstrate that dialect differentiation is encoded in both continuous and categorical intonational patterns, reinforcing the role of prosodic-acoustic features as stable markers of regional identity.

With respect to automatic dialect recognition, regional speech variation can be systematically modeled through acoustic feature extraction and statistical classification techniques. In the context of Brazilian Portuguese, studies combining speech signal processing and ML have shown that dialect identification can be approached as a pattern recognition problem in which acoustic representations capture structured phonetic and prosodic variability across regions. For instance, Batista (2019) investigates Brazilian regional accent identification in a sentence-reading task using statistical modeling techniques such as Gaussian Mixture Models, iVectors, and Support Vector Machines, demonstrating that dialectal variation can be computationally modeled through acoustic descriptors derived from the speech signal. Importantly, this line of research emphasizes that regional dialects do not represent random deviations but structured linguistic patterns that can be detected through systematic acoustic analysis.

This perspective aligns with recent computational studies showing that prosodic-acoustic information can be reliably extracted and operationalized in speech technology. Craveiro *et al.* (2024), for example, demonstrate that ML classifiers using prosodic-acoustic features can automatically segment prosodic units, highlighting the feasibility of extracting rhythmic and intonational patterns through automated methods. Similarly, Gauy and Finger (2023) show that transformer-based acoustic models pre-trained on Brazilian Portuguese corpora capture subtle acoustic variability, improving downstream classification tasks. Together, these studies reinforce the view that prosodic-acoustic features constitute a computationally tractable dimension of dialect variation and provide a foundation for dialect-sensitive ASR modeling.

The selection of prosodic-acoustic parameters in this study was guided by three complementary criteria. First, the literature reviewed above establishes that duration, f0, and intensity – together with voice quality measures derived from long-term spectral analysis – constitute the principal acoustic correlates of prosodic organization, and their role in dialect differentiation has been documented in Brazilian Portuguese (Constantini; Barbosa, 2015; Barbosa; Alvarenga, 2024). Second, these dimensions correspond precisely to the parameters covered by the third step of *Protosody* protocol, which is the *SpeechRhythmExtractor* pipeline (Silva Jr.; Barbosa; Monte, 2024), adopted here for methodological consistency and reproducibility. Third, within each dimension, the specific measures extracted – pause-based duration measures, spectral tilt, f0 variability and contour slope – were selected because they have been shown to capture supra-segmental and phonatory properties sensitive to regional variation. The set was constrained, consistent with the pilot scope, to ensure statistical stability given the 20-speaker sample. Additional features are identified as a priority for future research in Section 8.

Considering the use of prosodic-acoustic differentiation within a machine-learning framework, this study integrates sociophonetic analysis with ML-based ASR research to investigate inland Brazilian dialects that remain underrepresented in speech corpora. In doing so, it follows the research agenda proposed by Wassink, Gansen and Bartholomew (2022) by examining, in a Brazilian context, how systematic prosodic variability affects dialect-sensitive ASR modeling.

2. METHODOLOGY 1: CORPUS DESIGN, ACOUSTIC PROCEDURES, AND STATISTICAL ANALYSIS

2.1. PARTICIPANTS (PER DIALECT)

The study included 20 undergraduate students enrolled in Letras (Modern Languages) programs, with ten participants from the State University of Paraíba (UEPB) and ten from the State University of São Paulo (UNESP). The sample was balanced for gender (50% female, 50% male). Participants were between 18 and 23 years old ($M = 20.2$; $SD = 2.70$). All participants were native speakers of Brazilian Portuguese and were born and raised in inland municipalities of their respective states. The Paraíba group consisted of students from Guarabira and surrounding cities, representing the regional speech variety commonly referred to as Brejo Paraibano (*the Paraiban Brejo*) speech. The São Paulo group comprised students from São José do Rio Preto and neighboring municipalities, representing the inland São Paulo variety. Initial contact with participants occurred at the beginning of the academic semester, when they were enrolled in the course “Fonética e Fonologia” at both institutions. Recruitment procedures were conducted simultaneously at UEPB and UNESP to ensure comparable academic and instructional contexts across groups.

2.2. DATA COLLECTION

The corpus analyzed in this study was derived from controlled reading tasks based on a scripted text prepared for a podcast in Brazilian Portuguese. The podcast addressed topics in phonetics, phonology, and the potential influence of sociophonetic differences between the PB and SP dialects on foreign language speech. The recording process took approximately two weeks, generating about 1 hour and 30 minutes of audio material. For analytical purposes, each recording was segmented into 23 speech stretches, resulting in an initial dataset of 460 speech samples. Sample balance was controlled across dialect groups, yielding exactly 230 samples for PB and 230 samples for SP.

Participants received the text scripts and were instructed to read them at a natural speaking rate, using their everyday speaking voice – that is, the habitual vocal quality they employ in ordinary conversational contexts – without artificially raising or lowering their pitch, breathiness, or loudness

– provided they felt familiar with the content. This procedure follows recommendations for elicited speech protocols designed to minimize hyperarticulation and prosodic distortion associated with reading tasks (cf. Silva Jr.; Barbosa, 2023). Participants were allowed to rehearse the text, silently or aloud, for as long as needed before the recording session.

Data collection was conducted remotely using voice recording applications installed on participants’ personal mobile phones. Recordings were saved in uncompressed “.wav” format to preserve acoustic quality. Given the remote recording context, participants received detailed guidelines to minimize environmental noise and ensure acoustic consistency. These guidelines included selecting quiet recording environments, positioning themselves near sound-attenuating surfaces (e.g., curtains or upholstered furniture), maintaining a consistent microphone-to-mouth distance of approximately 10–15 cm, and avoiding excessive background reverberation. These procedures were adopted in accordance with best practices for remote acoustic data collection (e.g., C. Zhang *et al.*, 2021). Table 1 provides a synthesis of data collection procedures.

Dialect	Participants	Samples	Recording task	Total recording time
PB	10	230	Controlled reading from text scripts of a Podcast	Approximately 1 hour and 30 minutes
SP	10	230		
Total	20	460		

Table 1. Corpus design for the two inland Brazilian Portuguese dialect groups (PB: inland Paraíba; SP: inland São Paulo). The table reports the number of participants and speech samples per dialect group, along with grand totals. All recordings consisted of controlled reading tasks from podcast text scripts, yielding an approximate total recording time of 1 hour and 30 minutes. ALT – Table 1. Data collection table with 5 columns and 3 rows. Columns: Dialect, Participants, Samples, Recording task, Total recording time. Row 1 – PB (inland Paraíba): 10 participants, 230 samples, controlled reading from podcast text scripts, approximately 1 hour and 30 minutes. Row 2 – SP (inland São Paulo): 10 participants, 230 samples. Row 3 – Total: 20 participants, 460 samples.

2.3. ACOUSTIC ANALYSIS

Acoustic processing was conducted using Praat (Boersma; Weenink, 2024), following the segmentation and feature-extraction guidelines proposed by Silva Jr., Barbosa and Monte (2024), and Silva Jr. and Barbosa (2024). The procedure consists of three sequential stages: (i) phonetic forced alignment, (ii) realignment and relabeling of phonetic units, and (iii) automatic extraction of prosodic-acoustic parameters. This pipeline ensures segmentation consistency, phonetic accuracy, and replicable feature computation across recordings.

2.3.1. STEP 1: PHONETIC FORCED ALIGNMENT

Automatic phonetic alignment was performed using the MAUS (Munich Automatic Segmentation) aligner (Kisler *et al.*, 2017b) via its web-based interface. Each pair of sound and orthographic transcription files (with identical filenames) was submitted for processing. For each audio file, the aligner generated a corresponding TextGrid file compatible with Praat. These files contained

time-aligned phonetic segment boundaries, serving as the basis for subsequent refinement and feature extraction.

2.3.2. STEP 2: REALIGNMENT, RELABELING, AND MANUAL CORRECTION

Following forced alignment in step 1, a script for Praat was employed to implement new segment boundaries and reorganize annotation tiers according to vowel-to-vowel-based rhythmic units. This step ensured consistency in the definition of phonetic units and facilitated the extraction of prosodic measures. The resulting annotation structure comprised the following tiers:

- Tier 1 (V_to_V units): Phonetic syllable units spanning from the onset of one vowel to the onset of the subsequent vowel (Barbosa, 2006).
- Tier 2: Segmental classification, including vowels (V), consonants (C), and silent or filled pauses (#).
- Tier 3: Speech chunks (CH), corresponding to automatically segmented stretches based on pause duration and intensity, a heuristic-based method to detect prosodic boundaries proposed by Craveiro *et al.* (2024). This chunk segmentation was later used for statistical and automatic classification procedures.

Following automatic forced alignment procedures, all TextGrid annotations were manually inspected and corrected by an experienced phonetician. This step was necessary to ensure the consistency and reliability of phonetic segmentation, as automatic alignment tools may introduce inaccuracies, particularly in cases involving coarticulation, segment reduction, or low signal quality. As emphasized by Ladefoged (2003), acoustic measurements obtained through computational methods are not free from error, and careful manual verification is often required to ensure analytical precision.

Manual correction was carried out through the joint analysis of waveform and wide-band spectrogram displays, following the criteria proposed by Barbosa and Madureira (2015). Vowel boundaries were established based on the presence of periodic waveform patterns and stable formant structures, particularly the first (f1) and second (f2) formants, which are central to vowel quality and identity. Consonantal segments were evaluated according to their characteristic acoustic properties. Stops were identified by a closure interval (silence), the possible presence of a voicing bar in voiced contexts, a burst release, and the transition into the following vowel. Fricatives were defined by sustained aperiodic noise, with voicing distinctions determined by the presence or absence of a voicing bar. Laterals and nasals were identified by periodic waveforms of lower amplitude than vowels, a concentration of energy in lower frequencies, and the presence of antiresonances. Trills and taps were characterized by brief interruptions in the waveform caused by rapid articulatory contact.

In addition, speech chunks were identified by linguistic and acoustic-phonetic criteria to ensure the integrity of the input features. Linguistically, chunks were delimited based on prosodic envelopes (i.e., variations in amplitude, duration, and f0 that shape the speech signal acting as cues for perception and comprehension), ensuring that boundaries coincide with natural prosodic breaks to preserve the integrity of the syllable structures or rhythmic units (Barbosa, 2006; Cantoni *et al.*, 2022). This prevented the fragmentation of phonotactic units, such as complex onsets or diphthongs which could introduce spectral artifacts (i.e., distortions or noise that appear in the spectrogram but were not actually produced by the speaker's vocal tract) into the machine learning model. All corrections were applied conservatively, prioritizing clear acoustic evidence and avoiding unnecessary over-segmentation.

2.3.3. STEP 3: AUTOMATIC FEATURE EXTRACTION

Prosodic-acoustic parameters based on f0, duration, and intensity were automatically extracted using the *SpeechRhythmExtractor* script for Praat (Silva Jr., 2026; Silva Jr.; Barbosa; Monte, 2024). This script computes measures from the segmented tiers defined in Step 2 (Section 2.3.2) at the level of individual V-to-V units and speech chunks. F0 measures were computed locally, within each speech chunk. Rising f0 slope (*df0mean_pos*) was operationalized as the mean rate of f0 change (in semitones per second) over intervals in which the contour displays a positive first derivative – that is, rising movements detected locally within each unit. The standard deviation of the rising slope (*df0sd_pos*) captures variability in those local rising movements within a chunk. F0 standard deviation (*f0sd*) and semi-interquartile range (*f0SAQ*) were computed across all voiced frames within each chunk, reflecting overall pitch dispersion. These local values were aggregated at the speaker level using the mean, ensuring comparability across speakers. Intensity measures included the spectral tilt parameter (*sl_LTAS_alpha*), derived from the slope of the long-term average spectrum (LTAS) in the low-frequency domain – an index of phonatory mode and voice quality rather than loudness per se – and the variation coefficient of intensity (*cvint*), computed as the standard deviation of intensity normalized by its mean, reflecting short-term amplitude fluctuations. Duration measures comprised pause standard deviation (*pause_sd*), mean pause duration (*pause_meandur*), and pause rate (*pause_rate*, i.e., number of pauses per second of speech). All measures were computed within the chunk tier (Tier 3), ensuring that each observation corresponds to a delimited prosodic unit.

2.4. DATA SCREENING PROCEDURES

Prior to statistical and computational modeling, the dataset was screened for extreme values, zero values, and potential artifacts arising from acoustic extraction and segmentation procedures. Such

values are commonly observed in automatically extracted prosodic-acoustic datasets and may reflect both linguistically meaningful variation and methodological constraints inherent to speech signal processing (e.g., Boersma; Weenink, 2024; Hirst, 2011; Kisler *et al.*, 2017a; Rybka; Janicki, 2013; Galdino *et al.*, 2025).

Extreme values were manually inspected to determine whether they corresponded to plausible speech behavior (e.g., extended pauses, high pitch excursions, or increased intensity variability) or to extraction anomalies. No observations were removed solely based on magnitude. Instead, values were retained when consistent with known patterns of natural prosodic variability, in line with recommendations emphasizing ecological validity in speech analysis (Barbosa, 2006). Zero values were treated with particular care. In prosodic-acoustic datasets, zero values may arise from at least three sources:

- i. The absence of meaningful variation in the measured phenomenon – for instance, *pause_sd* approaches zero when all pauses within a chunk are of near-identical duration, a condition that may arise in short or highly controlled speech stretches; it is important to note, however, that zero values for *pause_rate* are not expected in this dataset, since the chunk segmentation algorithm (Step 2, section 2.3.2) requires the simultaneous presence of a silent interval of at least 35 ms and an intensity threshold of at least 40 dB for a chunk boundary to be established (cf. Akis; Bell; Simpson, 2025; Silva Jr.; Barbosa; Monte, 2024; among other Voice Activity Detection studies), meaning that every segmented chunk contains at least one pause event by construction.
- ii. Minimal acoustic variation (e.g., flat or near-monotonic *f0* contours yielding near-zero slope values).
- iii. Methodological and signal-processing constraints inherent to automatic feature extraction, particularly in unvoiced segments, low-intensity intervals, or short analysis windows. Notably, within standard extraction frameworks such as in The Geneva Minimalistic Acoustic Parameter Set (GeMAPS – Eyben *et al.*, 2016), *f0* is explicitly defined as zero in unvoiced regions, reflecting the absence of periodic vocal fold vibration rather than an error in processing (Boersma; Weenink, 2024; Hirst, 2011; Kisler *et al.*, 2017a).

In addition, the inclusion of zero and near-zero values aligns with a usage-based and gradient perspective on prosodic variation, in which the absence of certain acoustic events (e.g., pauses or pitch movement) is itself informative for modeling prosodic organization. Therefore, the adopted data screening procedures prioritized interpretability and ecological validity over aggressive filtering, ensuring that the resulting dataset preserves the full range of prosodic-acoustic variation relevant for both inferential and predictive analyses (James *et al.*, 2013; Molnar, 2022).

2.5. STATISTICAL ANALYSIS

For the statistical analysis, Linear Mixed-Effects Models (LMEM) were employed. This modeling approach has demonstrated robustness and flexibility in recent research involving multidimensional comparisons of prosodic-acoustic parameters across different speaker groups, including both L1 and L2 populations (cf. Ortega-Llebaria; Silva Jr.; Nagao, 2023; Silva Jr.; Silva; Meer, 2024; Teixeira, 2023). By incorporating both fixed effects (dialect) and random effects (participant), LMEM enables reliable estimation of dialect-related differences while controlling for individual variability, thereby reducing the risk of Type I Error (Barr *et al.*, 2013; Baayen; Davidson; Bates, 2008).

Furthermore, this modeling framework is especially well-suited for dialect classification and speech research, as it allows for the identification and quantification of specific prosodic-acoustic features that significantly contribute to cross-dialect differentiation, while simultaneously accounting for natural inter-speaker variability (Baayen; Davidson; Bates, 2008; Wieling; Tiede; Tomaschek, 2017). Mixed-effects models explicitly model hierarchical and grouped speech data, enabling more accurate estimation of dialect effects while preserving the structure of speaker-dependent variability.

This approach provides robustness for evaluating which acoustic parameters function as reliable predictors in automatic dialect classification and for interpreting their contribution to ASR-related performance, as mixed-effects models are specifically designed to handle correlated and hierarchical data structures commonly found in speech corpora (Nakagawa; Schielzeth, 2013; Barr *et al.*, 2013).

In addition to statistical robustness, the explanatory power of the models was assessed using the coefficient of determination (R^2), calculated as Marginal R^2 and Conditional R^2 , following the framework proposed by Nakagawa and Schielzeth (2013). Marginal R^2 represents the proportion of variance explained exclusively by fixed effects (i.e., dialect "PB-SP", where "PB" is the reference level), whereas Conditional R^2 reflects the total variance explained by both fixed and random effects (i.e., dialect and inter-speaker variability). This distinction is particularly relevant in speech research, where inter-speaker variability constitutes an inherent and meaningful source of acoustic variation.

The interpretation of R^2 values remains a subject of discussion in the statistical literature. To ensure consistent and theoretically grounded interpretation, this study adopted the normalization procedures proposed by Silva Jr. and Barbosa (2023), that integrate effect-size benchmarks from Cohen (1992), Chin (1998), Henseler; Ringle; Sinkovics (2009), and Moksony (1999). Based on these procedures, effect sizes were interpreted according to the following criteria:

- Strong effect size: $R^2 \geq 0.50$
- Weak-to-satisfactory effect size: $0.21 \leq R^2 \leq 0.49$
- Weak effect size: $R^2 \leq 0.20$

A significance alpha level of 5% was established, meaning that observed differences were considered statistically significant if the probability (p-value) was less than or equal 0.05 ($p \leq 0.05$).

The statistical analysis was performed within the R environment (R Core Team, 2026). Linear Mixed-Effects Models were fitted using the *lme4* package v. 1.1-38 (Bates et al., 2015), which provides functions for specifying fixed and random effects structures via the *lmer()* function. P-values for fixed effects were obtained through the *lmerTest* package v. 3.1-3 (Kuznetsova et al., 2017), which extends *lme4* by appending Satterthwaite-approximated degrees of freedom and significance tests to model summaries. Estimated marginal means and pairwise contrasts were computed using the *emmeans* package v. 2.0.3 (Lenth, 2024), enabling post-hoc comparisons between dialect groups. Marginal R^2 and Conditional R^2 coefficients of determination were calculated using the *MuMIn* package v. 1.48.19 (Bartoń, 2024) via the *r.squaredGLMM()* function, following the procedure proposed by Nakagawa and Schielzeth (2013). Data organization and pre-processing were handled through the *tidyverse* package collection v. 2.0.0 (Wickham et al., 2019).

3. RESULTS 1: STATISTICAL MODELING

This section reports the prosodic-acoustic parameters that showed statistically significant dialect differences, along with their contribution to automatic dialect classification.

Linear mixed-effects models were fitted for each prosodic-acoustic parameter, with ‘Dialect’ as a fixed effect and ‘Speaker’ as the random intercept. Effect sizes are reported as Marginal R^2 (variance explained by fixed effects) and Conditional R^2 (variance explained by fixed and random effects combined) in accordance with Nakagawa and Schielzeth (2013). Table 2 and Figure 1 synthesize the results for all acoustic features.

Acoustic correlate	Prosodic feature	β	SE	$t(18)$	p	Marginal R^2	Conditional R^2
f0	<i>f0sd</i>	0.535	0.249	2.15	.04	0.03	0.12
	<i>f0SAQ</i>	0.465	0.203	2.29	.03	0.02	0.07
	<i>df0mean_pos</i>	1.975	0.694	2.85	.01	0.05	0.12
	<i>df0sd_pos</i>	1.076	0.485	2.22	.04	0.03	0.11
Intensity	<i>sl_LTAS_alpha</i>	5.230	0.582	8.98	<.001	0.27	0.31
	<i>cvint</i>	-10.701	2.986	-3.58	<.001	0.03	0.03
Duration	<i>Pause SD</i>	581.81	97.29	5.98	<.001	0.44	0.68
	<i>Pause mean</i>	176.82	82.97	2.13	.04	0.10	0.54
	<i>Pause rate</i>	0.14	0.055	2.54	.02	0.14	0.56

Table 2. Prosodic-acoustic correlate, prosodic-acoustic features for each correlate, estimate coefficients of the models (β), standard error (SE) values, t -value (degrees of freedom from the fixed effects), p -value, and the Marginal coefficient of determination (R^2 , variance explained by fixed effects) and the Conditional coefficient of determination (R^2 , variance explained by both fixed and random effects), following Nakagawa and Schielzeth (2013).

ALT- Table 2. Statistical results of LMEMs. Reports β , SE, $t(18)$, p -values, Marginal R^2 and Conditional R^2 for all nine acoustic measures showing significant dialect differences.

Across all models ($N = 460$ speech samples; $N = 20$ speakers), dialect significantly affected every prosodic-acoustic parameter analyzed in this section.

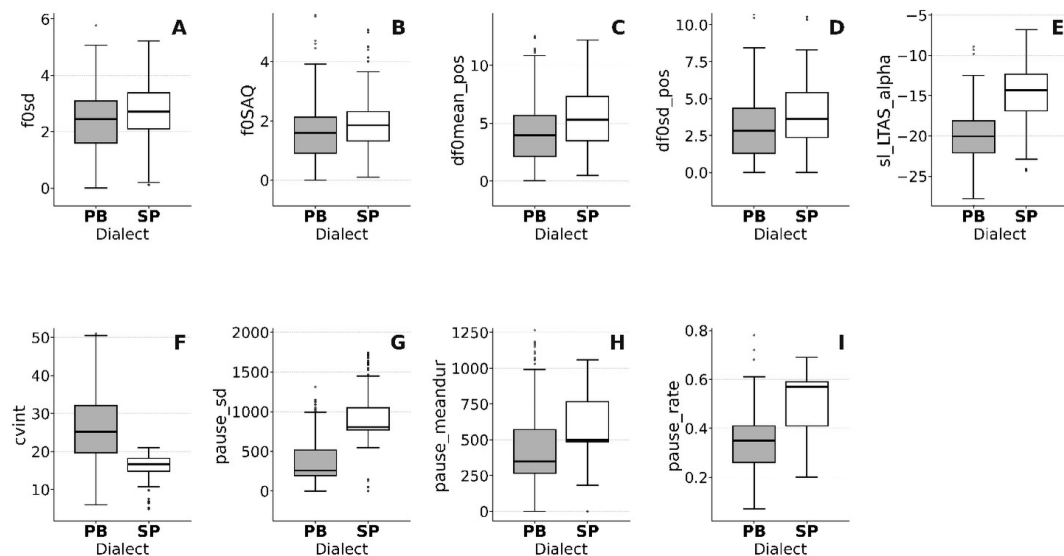


Figure 1. Boxplots of the statistically significant acoustic parameters based on the Mixed-Effects statistics. Nine prosodic-acoustic parameters across three dimensions: **f0 standard deviation (SD)** – *f0sd*; **f0 semi-interquartile range (semi-IQR)** – *f0SAQ*; **rising f0 slope** – *df0mean_pos*; **rising f0 SD slope** – *df0sd_pos* (panels A-D). Intensity: **spectral tilt (voice quality)** – *sl_LTAS_alpha*; **intensity variation coefficient** – *cvint* (panels E, F). Duration: **pause SD** – *pause_sd*; **pause duration mean** – *pause_meandur*; **pause rate** (pauses per second) – *pause_rate* (panels G-I) controlled for two dialects (PB, SP).

ALT – Figure 1: Boxplots comparing PB and SP dialects across nine prosodic-acoustic parameters related to pitch, intensity, and pause timing. The plots show higher values for SP in the f0-based measures (panels A-D), spectral tilt (panel E), and pause-related measures (panels G-I). SP shows lower values for the intensity variation coefficient (panel F). PB shows lower values overall for pause-related measures. The figure indicates systematic dialectal differences in pitch dynamics, voice quality, and rhythmic-temporal organization.

3.1. FUNDAMENTAL FREQUENCY (F0)

With respect to f0 measures, dialect exerted a consistent, though generally weak-to-satisfactory, effect across variability and contour-related parameters. f0 standard deviation (*f0sd*, Figure 1A) differed significantly between dialects ($\beta = 0.535$, $SE = 0.249$, $t(18) = 2.15$, $p = .04$), with SP speakers displaying greater pitch variability. However, the Marginal R^2 (henceforth, R^2m) was 0.03 and the Conditional R^2 (henceforth, R^2c) reached 0.12, indicating that dialect accounted for a relatively small portion of the total variance in f0 dispersion. A similar pattern emerged for the semi-amplitude interquartile range (i.e., semi-IQR) (*f0SAQ*, Figure 1B), which was also significantly higher in the SP group ($\beta = 0.465$, $SE = 0.203$, $t(18) = 2.29$, $p = .03$). Nonetheless, effect sizes remained small ($R^2m = 0.02$; $R^2c = 0.07$), suggesting limited explanatory strength.

Contour dynamics revealed comparable tendencies. The mean f0 rising slope (*df0mean_pos*, Figure 1C) was significantly greater in the SP dialect ($\beta = 1.975$, $SE = 0.694$, $t(18) = 2.85$, $p = .01$),

although the R^2m and R^2c values (0.05 and 0.12, respectively) also indicate a weak explanatory power. Likewise, the standard deviation of the f_0 rising slope (df_0sd_pos , Figure 1D), which captures the dispersion of rising contour movements, showed a significant dialect effect ($\beta = 1.076$, $SE = 0.485$, $t(18) = 2.22$, $p = .04$), but with weak R^2m (.03) and R^2c (0.11).

Taken together, these findings suggest that while the two dialects differ systematically in pitch variability and contour dynamics, particularly with SP speakers exhibiting greater dispersion and steeper rising movements, f_0 -based measures contribute relatively modest proportions of explained variance. Although pitch dynamics participate in dialect differentiation, they appear to function as secondary cues rather than primary discriminators in the present data.

3.2. INTENSITY

It is important to note at the outset that the two intensity-related parameters in this study tap into distinct phonatory phenomena and should not be conflated under a single interpretation of "intensity." The spectral tilt parameter (sl_LTAS_alpha) reflects differences in phonation mode – specifically, the distribution of vocal energy across the frequency spectrum as captured by the LTAS slope in the low-frequency domain – and is best understood as a voice quality index rather than a measure of loudness per se (Esling; Edmondson, 2011; Garellek, 2022). The variation coefficient of intensity ($cvint$), by contrast, captures short-term fluctuations in the amplitude envelope and is more directly related to differences in loudness variability across the speech signal, reflecting how consistently or variably speakers modulate their overall vocal effort. This distinction is essential for the correct interpretation of the results that follow.

Intensity-based measures revealed a markedly stronger dialect effect, particularly in long-term spectral configuration. Spectral tilt in the low-frequency domain (sl_LTAS_alpha , Figure 1E), which is related to vocal effort, was robustly predicted by dialect ($\beta = 5.230$, $SE = 0.582$, $t(18) = 8.98$, $p < .001$), with SP speakers exhibiting significantly higher values. The R^2m of 0.27 indicates that dialect alone accounted for 27% of the variance, while the R^2c (0.31) indicates a weak-to-satisfactory effect for the combined fixed and random effects, suggesting that a substantial portion of the total variance is also attributable to speaker-specific variability. This result positions spectral tilt as one of the reliable acoustic markers of dialectal differentiation in the corpus.

In contrast, the variation coefficient (VC) of intensity ($cvint$, Figure 1F) also reached statistical significance ($\beta = -10.701$, $SE = 2.986$, $t(18) = -3.584$, $p < .001$), with SP speakers showing lower variability. However, both $R^2m|c$ values were 0.03, and the model exhibited a singular fit, indicating that the practical contribution of this parameter is limited and weak despite statistical significance. Nevertheless, $cvint$ also contributes as a complementary indicator of dialect-specific loudness variability, capturing differences in how consistently PB and SP speakers modulate amplitude across

the speech signal. Its inclusion in the feature set therefore reflects the multidimensional nature of the intensity domain rather than a redundancy with *sl_LTAS_alpha*.

Overall, the intensity domain demonstrates that dialectal differences are strongly encoded in long-term spectral balance rather than in short-term amplitude variability. It is important to clarify that intensity, as an acoustic dimension, does not carry direct linguistic meaning by itself; rather, its relevance emerges through specific derived measures that tap into distinct phonatory and voice-quality phenomena. Spectral tilt (*sl_LTAS_alpha*) is linked to global phonatory effort and the distribution of vocal energy across frequency bands, reflecting habitual voice quality settings that index regional identity (Esling; Edmondson, 2011; Thomas, 2011). The variation coefficient of intensity (*cvint*), by contrast, captures short-term fluctuations in amplitude, which may reflect differences in prominence signaling and speech rhythm across dialect groups.

3.3. DURATION

Temporal parameters displayed the most substantial dialectal contrasts, particularly in pausing behavior. Pause duration variability (*pause_sd*, Figure 1G) was strongly predicted by dialect ($\beta = 581.81$, $SE = 97.29$, $t(18) = 5.98$, $p < .001$), with SP speakers exhibiting markedly greater dispersion. The R^2m of 0.443 indicates a strong fixed-effect contribution, and the R^2c of 0.676 reflects considerable overall explanatory power. This suggests that differences in pause distribution may play a central role in the Brazilian Portuguese inland dialectal distinction presented here.

Although the current analysis does not permit a precise statement about the syntactic or discourse positions at which pauses occur – given that chunks were defined by prosodic rather than syntactic criteria – the results indicate that SP speakers exhibit not only longer but also more variable pauses, suggesting greater within-speaker inconsistency in pause duration across the utterance. This pattern is consistent with a speech planning profile in which SP speakers distribute silences less uniformly, possibly reflecting differences in discourse pacing or prosodic chunking strategies relative to PB speakers. Whether these pauses cluster at syntactic boundaries, between prosodic phrases, or across other discourse units remains a question for future investigation employing syntactically annotated corpora.

Mean pause duration (*pause_meandur*, Figure 1H) was also significantly longer in the SP dialect ($\beta = 176.82$, $SE = 82.97$, $t(18) = 2.13$, $p = .04$). Although the R^2m (0.10) indicates a weak direct contribution of dialect, the R^2c (0.54) reveals substantial speaker-related variability. Similarly, pause rate (*pause_rate*, Figure 1I – i.e., pauses per second) differed significantly between dialects ($\beta = 0.141$, $SE = 0.055$, $t(18) = 2.54$, $p = .02$), with SP speakers producing pauses more frequently. The R^2m of 0.14 also suggests a weak fixed-effect contribution, while the R^2c of 0.56 again highlights meaningful inter-speaker differences.

It should be noted that pause duration variability (*pause_sd*) reflects dispersion in the length of pauses rather than their positional distribution within the utterance. A high *pause_sd* value indicates that a speaker's pauses vary considerably in duration – some very brief, others markedly extended – rather than that pauses occur at syntactic positions. This measure therefore captures a temporal rhythm signature of the dialect, characterizing the degree of rhythmic variability in the speaker's pausing behavior, and is analytically distinct from measures of pause frequency (*pause_rate*) or average duration (*pause_meandur*).

Collectively, these temporal findings indicate that dialectal differentiation is particularly salient in the organization of speech rhythm and pausing strategies. The magnitude of the effects, especially for pause variability, demonstrates that temporal structuring functions as a primary acoustic dimension of distinction between the dialects. Compared to f0 measures, duration-related parameters provide substantially greater explanatory power, underscoring the importance of rhythmic organization in capturing dialect-specific prosodic patterns.

4. METHODOLOGY 2: AUTOMATIC DIALECT CLASSIFICATION PROCEDURES

An automatic dialect classification task was implemented using a binary classification framework comprising six machine-learning models. The framework was developed in Python (Van Rossum; Drake, 2009). Data handling and pre-processing were performed using *pandas* v. 3.0.2 (McKinney, 2010), which provides tabular data structures for loading and organizing the feature matrix, and *NumPy* v. 2.4.4 (Harris *et al.*, 2020), which supports numerical array operations underlying all model computations. The dataset was partitioned into training (80%) and test (20%) subsets prior to any scaling using the stratified `train_test_split()` function from *scikit-learn* v. 1.8.0 (Pedregosa *et al.*, 2011), ensuring proportional dialect representation in both subsets. Feature normalization was applied using `MinMaxScaler()`, fitted exclusively on the training set and subsequently applied to the test set, to prevent data leakage.

All six classifiers – Random Forest (`RandomForestClassifier`), Decision Tree (`DecisionTreeClassifier`), k-Nearest Neighbors (`KNeighborsClassifier`), Linear Discriminant Analysis (`LinearDiscriminantAnalysis`), Support Vector Machine (`SVC`), and Gradient Boosting Machine (`GradientBoostingClassifier`) – were implemented through *scikit-learn* v. 1.8.0 (Pedregosa *et al.*, 2011). Hyperparameter optimization for each model was conducted via five-fold cross-validated grid search using `GridSearchCV()`, also from *scikit-learn*. Model evaluation relied on `accuracy_score()`, `confusion_matrix()`, and `classification_report()` from *scikit-learn*. Feature-importance scores were extracted via `.feature_importances_` attribute, which returns mean decrease in impurity values normalized to sum to 1 (Hastie *et al.*, 2009). Visualizations were

produced using *Matplotlib* v. 3.10.1 (Hunter, 2007). The framework was designed to classify speech samples according to dialect origin, identifying whether a given recording belonged to PB or SP.

The classification was based exclusively on prosodic-acoustic features extracted from each participant's speech. For the purposes of this pilot study, the ML models were trained only on the subset of acoustic parameters that showed statistical significance in the LMEM (Table 2, section 3), ensuring methodological consistency between the inferential and predictive stages of the analysis.

Six ML techniques were implemented and compared to one another using Accuracy and Error Rate (ER) as performance metrics. The six classifiers are not treated as competing alternatives in a model-selection sense; rather, they are implemented as complementary analytical lenses that differ in their decision mechanisms. By comparing their outputs, this stage evaluates whether the prosodic-acoustic hierarchy identified in the inferential phase is robust across distinct algorithmic approaches, from linear (LDA) to non-linear ensemble methods (RF, GBM).

These techniques (cf. section 4.1-4.6) differ in how they model relationships between acoustic variables and dialect categories, allowing for a comprehensive assessment of how prosodic-acoustic features contribute to automatic dialect classification (Jassim; Abdulmohsin, 2025; James *et al.*, 2013; Nielsen, 2021; Silva *et al.*, 2023; Silva Jr.; Barbosa, 2022).

4.1. RANDOM FOREST (RF)

Random Forest is an ensemble learning technique that combines multiple Decision Trees to produce a single classification outcome. Each tree is trained using a random subset of acoustic observations and acoustic variables, and the final classification is determined through a majority voting procedure. This structure reduces overfitting and increases model robustness, particularly in speech datasets where acoustic variables may be correlated. In the context of dialect classification proposed in this study (and based on Table 2), Random Forest evaluates how combinations of prosodic-acoustic features – such as pause descriptors, spectral tilt, variation coefficient (VC) of intensity, and pitch variability – jointly contribute to distinguishing PB and SP speakers. Because each tree may prioritize different acoustic features or speakers, the model captures both consistent dialect patterns and natural inter-speaker variability. This makes Random Forest particularly suitable for identifying complex, multidimensional acoustic patterns associated with dialect-specific prosodic organization.

4.2. DECISION TREE (DT)

Decision Trees classify data by recursively splitting observations according to decision rules based on acoustic feature values. Each split selects the acoustic parameter that best separates the dialect groups, based on statistical criteria such as entropy or the Gini index. This model provides a transparent representation of dialect differentiation, as it explicitly reveals which acoustic variables,

and which thresholds, most effectively distinguish PB and SP speakers. For example, the model may identify specific pause duration thresholds or spectral tilt values that consistently separate dialect groups. This interpretability makes DT particularly useful for identifying which individual prosodic-acoustic features function as primary dialect markers.

4.3. K-NEAREST NEIGHBORS (K-NN)

The k-Nearest Neighbors algorithm classifies speech samples based on their similarity to other samples in the acoustic feature space. Classification is determined by comparing each speech sample to its closest neighbors, using distance metrics such as Euclidean distance. In dialect classification, kNN evaluates how acoustically similar a given participant's speech is to other speakers within each dialect group. If a speaker's acoustic profile—based on features such as f0 variability, pause rate, or intensity patterns – closely resembles those of PB speakers, the sample will be classified accordingly. This approach reflects the principle that dialect membership corresponds to clustering in acoustic space, with speakers from the same dialect exhibiting similar prosodic-acoustic profiles.

4.4. LINEAR DISCRIMINANT ANALYSIS (LDA)

Linear Discriminant Analysis is a statistical classification method that identifies linear combinations of acoustic features that maximize separation between dialect groups. It assumes that acoustic variables follow approximately normal distributions and that group variances are comparable. In this study, LDA identifies which combinations of prosodic-acoustic parameters most effectively differentiate PB and SP speakers. For example, it may reveal that a specific combination of pause variability, spectral tilt, and pitch dynamics provides optimal separation between dialects. This method is particularly useful for understanding the relative contribution of multiple acoustic dimensions in defining dialect-specific prosodic features.

4.5. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine is a classification algorithm that identifies the optimal boundary separating dialect groups in the acoustic feature space. It maximizes the distance between the closest observations from each dialect category, thereby improving classification reliability. In dialect classification, SVM evaluates how consistently prosodic-acoustic features distinguish between PB and SP speakers. It is particularly effective when dialect differences are subtle and involve complex interactions between acoustic variables. For instance, if SP speakers systematically exhibit longer pause durations and higher spectral tilt values than PB speakers, SVM identifies the optimal boundary in the acoustic feature space that separates these two groups. When a new speech sample is introduced, the model evaluates its prosodic-acoustic profile relative to this boundary and classifies it as PB or SP depending on which side of the boundary it falls.

4.6. GRADIENT BOOSTING MACHINE (GBM)

Gradient Boosting Machine (GBM) is an ensemble ML technique that builds a sequence of decision trees in a stepwise manner, where each successive tree is trained to correct the classification errors made by the previous ones. Unlike RF, which builds trees independently, GBM constructs trees sequentially, allowing the model to progressively refine its classification performance by focusing on the most difficult cases. In the context of the present research, GBM evaluates how combinations of prosodic-acoustic features contribute incrementally to distinguishing PB and SP speakers. By iteratively prioritizing acoustic observations that are more difficult to classify, the model becomes particularly sensitive to subtle dialect differences reflected in parameters such as pause variability, spectral tilt, and pitch dynamics. This makes GBM especially effective in identifying fine-grained acoustic patterns that may not be captured by simpler models. For example, if most SP speakers are correctly classified based on pause duration alone, but some overlap with PB speakers due to similar pitch variability, GBM will iteratively focus on those ambiguous cases and incorporate additional features – such as spectral tilt – to improve separation. In this way, the model progressively refines classification by correcting previous errors.

4.7. IMPLEMENTATION PROCEDURES

For the development of the AI-based classifier, a total of 460 speech samples (Table 1, section 2.2) was used. The dataset was randomly partitioned by the dialect classifier into two subsets: 80% of the samples were allocated to training and cross-validation, and the remaining 20% were reserved for testing. In addition, the random selection procedure preserved representation from both dialect groups (PB and SP) across the two subsets. This train-test split is widely adopted in ML research to ensure model generalizability and to prevent overfitting, as it allows the system to learn classification patterns from one portion of the data and evaluate performance on previously unseen samples (James *et al.*, 2013; Brownlee, 2020).

It is worth highlighting that, in the context of this pilot study, the train-test split was conducted at the sample level rather than at the speaker level. This choice reflects the primary objective of the predictive stage, namely, to assess whether the prosodic-acoustic variables identified as statistically significant in the inferential analysis provide sufficient information to discriminate between PB and SP dialects at the level of individual speech samples. Within this framework, the unit of prediction is the segmented speech sample rather than the unseen speaker. Given that each recording was segmented into multiple speech stretches, yielding a balanced number of observations per dialect, this procedure enables a controlled evaluation of how prosodic-acoustic variation operates across samples under comparable conditions. As such, the present design is consistent with a proof-of-

concept approach, in which the focus lies on testing the sensitivity and discriminative capacity of selected acoustic variables within a predictive framework.

At the same time, it is important to acknowledge that the dataset includes multiple observations derived from the same speakers, which implies that a random split at the sample level may result in observations from a given speaker appearing in both the training and test sets. From a methodological standpoint, this configuration may relax the assumption of independence between subsets and may lead to optimistic estimates of classification performance. As discussed in the literature, random partitioning procedures are appropriate when observations can be treated as independent; however, when related samples are distributed across training and test sets, evaluation metrics may partially reflect shared structure rather than strict generalization (Bishop, 2006; James *et al.*, 2013). In speech processing, this issue is particularly salient, as speaker-dependent and speaker-independent evaluation protocols constitute distinct methodological configurations, with the former typically yielding higher performance due to the presence of speaker-specific information across subsets (Hastie; Tibshirani; Friedman, 2009; Jassim; Abdulmohsin, 2025). Empirical studies in speech classification further demonstrate that evaluation settings which do not control for speaker overlap tend to produce inflated accuracy when compared to speaker-independent setups (Rybka; Janicki, 2013; Atmaja; Sasou, 2022; Sasse *et al.*, 2025).

Accordingly, the classification results reported in this study should be interpreted as reflecting within-corpus, sample-level performance, rather than a fully speaker-independent estimate of dialect generalization. Importantly, this design remains appropriate for the present stage of research. By aligning predictive modeling with the inferential results obtained through mixed-effects analysis, the current implementation provides an initial validation of the discriminative capacity of prosodic-acoustic features. In this sense, the methodological choice follows the distinction between explanatory and predictive modeling (Shmueli, 2010), in which the present analysis prioritizes feature evaluation and model sensitivity, while recognizing that more restrictive evaluation protocols are required to assess generalization in applied contexts.

Each classifier was evaluated using Accuracy, computed as the proportion of correctly predicted dialect labels relative to the total number of observations. The complementary Error Rate ($1 - \text{Accuracy}$) was also reported to indicate the proportion of misclassified samples. Accuracy is one of the most widely adopted performance metrics in binary classification tasks, particularly when class distributions are balanced (James *et al.*, 2013; Hastie *et al.*, 2009).

For interpretative purposes, feature-importance values were descriptively categorized according to their empirical distribution within each classifier. Features with scores above 0.30 were treated as dominant, as they represented a substantial proportion of the models' decision-making process. Values ranging from 0.10 to 0.30 were interpreted as moderately important, indicating predictors that contributed meaningfully to classification but were not independently decisive. Features with values below 0.10 were considered weak contributors, reflecting limited relative

influence within the model. It is important to emphasize that these thresholds do not constitute theoretical cutoffs; rather, they serve as analytical conventions adopted to facilitate interpretation of the internal ranking of predictors. As discussed in the literature of model interpretability, feature-importance measures are inherently relative and model-dependent, reflecting each predictor's contribution within a specific trained model rather than absolute explanatory power (Molnar, 2022).

5. RESULTS 2: AUTOMATIC CLASSIFICATION PERFORMANCE

Figure 2 presents six confusion matrices illustrating the performance of each machine-learning model in classifying PB and SP speech samples. As mentioned in section 4.7, the testing phase corresponded to 20% of the corpus and was randomly partitioned into **92 samples** speech samples ($460 \text{ samples} \times 20\% = 92 \text{ samples}$), comprising **46 PB** samples and **46 SP** samples.

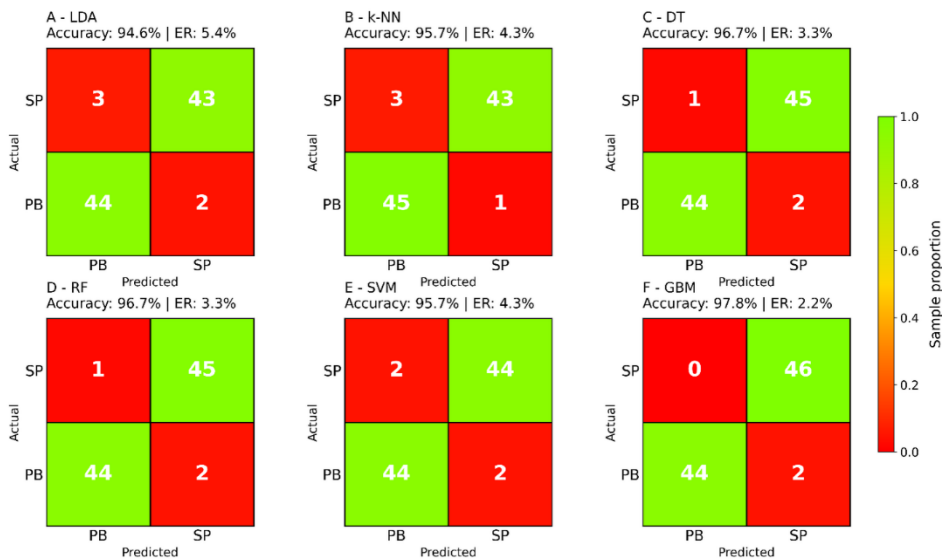


Figure 2. Confusion Matrices for the AI-based binary dialectal classification (PB vs. SP), and the evaluated metrics (Accuracy | Error Rate - ER) based on the *Predictive* values (dialect classification - X-axis), and the *Actual* values (what the samples are in reality - Y-axis). Cells in which the actual and predicted dialect labels match represent correct classifications; cells in which they do not match represent misclassifications.

ALT - Figure 2. Six confusion matrices showing the performance of different machine-learning models in classifying PB and SP speech samples. Cells in which the actual and predicted dialect labels match represent correct classifications; cells in which they differ represent misclassifications. Overall accuracy ranged from 94.6% to 97.8% (2 to 5 errors out of 92 test samples).

Results shown in Figure 2 indicate highly consistent discrimination between dialect groups, suggesting that tree-based modeling (GBM, RF and DT) effectively capture the relevant prosodic-acoustic distinctions. It is also important to highlight that identifying a single best classifier is not an objective of this study. Rather, the six models serve as complementary evidence of feature

robustness. The convergent performance across all six classifiers, despite their different decision mechanisms, reinforces the conclusion that the identified prosodic-acoustic features provide structurally stable dialect cues rather than algorithm-specific artifacts.

Five out of the six ML techniques implemented reached higher than 95% accuracy values in the classification task. The highest classification performance was achieved by the Gradient Boosting Machine (GBM – Figure 2F) with an accuracy of 97.8% (ER = 2.2%) followed by Decision Tree (DT – Figure 2C) and Random Forest (RF – Figure 2D) models, both reaching an accuracy of 96.7% (ER = 3.3%). In the sequence, results show *k*-Nearest Neighbors (*k*-NN – Figure 2B) and Support Vector Machine (SVM – Figure 2E) also presenting high accuracy values (95.7%; ER = 4.3%). The lowest – but still high – accuracy values came from LDA with a near-95% rate (94.6%; ER = 5.4%). In GBM, 44 out of 46 PB samples and all SP samples were correctly classified. In DT and RF, 44 out of 46 PB samples and 45 out of 46 SP samples were also correctly classified. For *k*-NN and SVM – respectively – 45 out of 46 PB; 43 out of 46 SP samples, and 44 out of 46 PB/SP samples were correctly classified.

Overall, classification accuracy ranged from 94.6% to 97.8%, with tree models (GBM, RF and DT) consistently yielding the strongest performances. These results indicate that non-linear interactions among prosodic-acoustic features play a substantial role in distinguishing inland PB and SP dialects, reinforcing the robustness of automatic dialect classification in the present dataset based on multiple prosodic-acoustic descriptors.

Figure 3 presents the relative feature-importance scores derived from the best-performing tree-based models using the Gini impurity reduction criterion. These values reflect the proportional contribution of each prosodic-acoustic parameter to the decision process from the model.

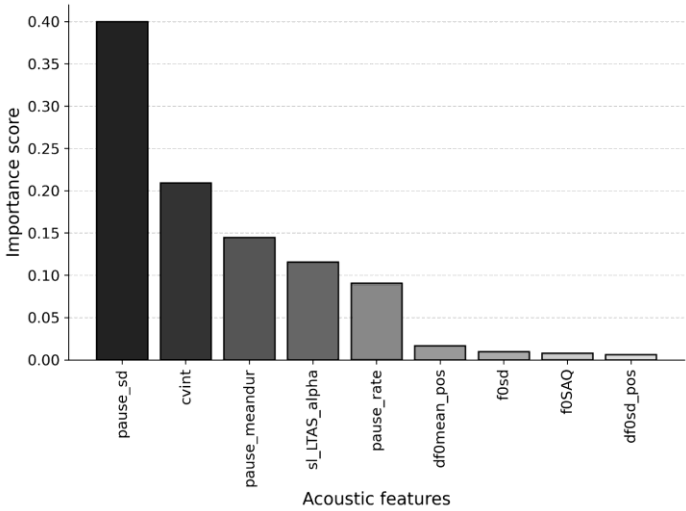


Figure 3. Feature-importance scores based on tree-based models. The prosodic-acoustic features that were statistically significant in the inferential phase (X-axis, section 3), and importance scores (Y-axis) normalized to sum to 1.

ALT – Figure 3. Bar chart displaying the relative importance of prosodic-acoustic features used in dialect classification. Pause variability has the highest importance score, followed by intensity-related measures and other duration features, while pitch-based measures show lower importance.

The ranking reveals a highly asymmetrical distribution of importance across features. Pause SD (*pause_sd*) emerges as the dominant predictor, accounting for approximately 0.40 of the total importance weight. This value alone represents a disproportionately large share of the classifier's decision-making process, indicating that variability in pause duration constitutes the most robust acoustic cue distinguishing between PB and SP speakers in the testing set. This finding directly aligns with the high overall accuracy observed in the tree models (GBM = 97.8%; RF/DT = 96.7%), suggesting that much of the classification success is driven by rhythmic-temporal variability rather than by isolated segmental cues.

The second most important feature is intensity variation coefficient (*cvint* ~0.22), followed by mean pause duration (*pause_meandur* ~0.14), spectral tilt (*sl_LTAS_alpha* ~0.12), and pause rate (*pause_rate* ~0.08). Together, these parameters reinforce the centrality of rhythmic organization in speech planning (Krivokapić, 2007), and long-term spectral balance (i.e., voice quality) in cross-dialect discrimination. Voice quality in this analysis reflects the degree to which speakers modulate vocal energy across the speech signal. Specifically, the intensity coefficient and spectral tilt capture how PB and SP speakers adjust loudness patterns to signal prominence, emphasis, or expressive variation in speech (Garellek, 2022; Silva Jr.; Barbosa, 2023).

Notably, the cumulative contribution of pause-related measures exceeds 0.60 of the total importance weight, indicating that rhythmic structuring of speech provides the primary acoustic foundation for dialect separation in the present dataset as suggested by Esling and Edmondson (2011) and Yaeger-Dror and Fagyal (2011). In practical terms, these measures capture how speakers organize silence within the speech stream. The standard deviation of pause duration reflects how consistently or variably speakers distribute pauses; the mean pause duration indicates the typical length of these silent intervals; and pause rate (pauses per second) represents how frequently speakers interrupt speech with pauses. These parameters describe dialect-specific patterns of speech segmentation and pacing, suggesting that PB and SP speakers differ not only in how long they pause but also in how regularly and how often pauses occur throughout the utterance. In contrast, f0-based parameters (*f0sd*, *df0sd_pos*, *f0SAQ* and *df0mean_pos*) exhibit comparatively low importance values (all below 0.03). This distribution suggests that pitch centrality and dynamics alone do not provide sufficiently strong discriminative power to sustain high classification accuracy in this dataset.

This hierarchy of feature importance helps to explain why tree-based ensemble methods slightly outperformed linear classifiers. Algorithms such as GBM, RF and DT can prioritize dominant predictors (e.g., *pause_sd*) while modeling non-linear interactions among secondary cues, thereby capitalizing on the weighted acoustic structure of the dataset. The classification results therefore do not merely indicate that PB and SP differ acoustically; they reveal how those differences are hierarchically organized within the prosodic system. The heuristic method for using speech feature importance in this pilot study is partially aligned with Z. Zhang (2021), who proposes a hybrid system

(traditional ML techniques like the ones used in this study, and Neural Networks) for speech emotion recognition. Z. Zhang's study aimed to obtain the importance of each analyzed acoustic feature for dimensionality reduction (of feature) in speech analysis.

Therefore, the feature-importance analysis clarifies the internal mechanics of the ASR classification phase: dialect discrimination is primarily structured around rhythmic-temporal pause-based variability, with measures of spectral tilt and vocal effort providing complementary support, and f0-based parameters functioning as less prominent secondary cues. This internal weighting pattern further corroborates the mixed-effects findings discussed earlier, reinforcing the conclusion that dialect differentiation in this research is grounded in rhythmic and intensive organization rather than predominantly in pitch modulation.

6. DISCUSSION

This study addressed two central research questions concerning the role of prosodic-acoustic features in inland Brazilian Portuguese dialect differentiation and their contribution to automatic dialect classification.

- **RQ1** – Do prosodic-acoustic descriptors of duration, f0, intensity, and long-term spectral voice quality provide robust and reliable cues for distinguishing the inland Brazilian Portuguese dialects PB and SP?

Yes. The present results provide evidence of dialect differentiation. These findings are consistent with earlier prosodic investigations of Brazilian Portuguese dialects. Constantini and Barbosa (2015), for example, analyzed multiple regional varieties using prosodic descriptors derived from vowel-to-vowel rhythmic units and found that several parameters related to duration variability, spectral emphasis, and f0 median significantly differentiated dialect groups. Their results demonstrated that rhythmic parameters, especially variability and asymmetry in duration distributions, play an important role in distinguishing varieties, while intensity-related measures such as spectral emphasis and pitch statistics contribute complementary cues to dialect identification. The authors showed that dialectal contrasts emerged not from a single acoustic dimension but from the interaction of rhythmic, melodic, and intensity-related parameters, a pattern that parallels the hierarchical distribution observed in this work. While the present study's results highlight pause variability and other duration-related measures as dominant predictors in both inferential and machine-learning analyses, the contribution of spectral balance and vocal effort, and pitch parameters remain consistent with the multidimensional prosodic differentiation reported by Constantini and Barbosa (2015). Such results strengthen the sociophonetic interpretation that

regional varieties of Brazilian Portuguese are encoded, at least in part, through structured prosodic-acoustic configurations. Since the present study did not include a systematic examination of segmental features, no comparative claim about the relative primacy of prosodic over segmental cues can be drawn from these data alone; the findings demonstrate that prosodic-acoustic dimensions are sufficient to sustain statistically robust dialect differentiation and high-accuracy computational classification within this dataset. These results support the view that rhythmic organization, pitch behavior, and spectral energy distribution are among the dimensions that jointly contribute to dialectal identity, while leaving open the question of how prosodic and segmental dimensions interact in the broader acoustic characterization of these varieties.

In this respect, the mixed-effects analysis demonstrated that prosodic-acoustic descriptors do indeed provide statistically robust cues for dialect differentiation. By modeling dialect as a fixed effect and speaker as a random effect, the LMEM framework isolated regional variation while controlling for inter-speaker variability (Baayen; Davidson; Bates, 2008; Mirman, 2014). This approach ensures that the observed differences reflect structured dialectal patterning rather than individual performance variability. Across the acoustic domains, duration-related parameters, especially pause variability, displayed the strongest marginal explanatory power. From a sociophonetic perspective, this indicates that rhythmic (pause-based) organization functions as a primary indexical resource distinguishing PB and SP speakers.

Pause variability reflects discourse segmentation, speech planning strategies, and interactional pacing, dimensions that are socially learned and community specific. Cross-dialect research in other languages has shown that speech rhythm and temporal structuring frequently serve as salient regional markers, both for native and foreign dialects (Esling; Edmondson, 2011; Krivokapić, 2007; Silva Jr.; Barbosa, 2019; White; Mattys, 2007). In Brazilian Portuguese, rhythmic patterns and variation have been documented across several regional varieties (Cagliari; Abaurre, 1986; Meireles; Gambarini, 2012), but inland dialects have received comparatively limited empirical attention. The present findings therefore extend Brazilian sociophonetics by demonstrating that temporal variability is not merely stylistic but structurally embedded in inland dialect differentiation.

Intensity-related measures, particularly spectral tilt and vocal effort, also significantly contributed to dialect distinction. Both spectral tilt and vocal effort capture long-term spectral balance and habitual phonatory settings, indexing voice quality differences between groups (i.e., vocal energy distribution, prominence, expressive modulation etc.). Voice quality has increasingly been recognized as a sociophonetic marker capable of signaling regional and social identity (Thomas, 2011; Yaeger-Dror; Fagyal, 2011). The present findings suggest that PB and SP differ not only in rhythm but also in stable phonation configurations, reinforcing the multidimensional nature of dialect identity. In contrast, f0-based measures, although statistically significant, displayed comparatively smaller R^2_{mm} values, indicating that pitch variability and dynamics function as secondary but systematic cues. This

hierarchy aligns with cross-dialect findings in English and Spanish, where pitch scaling differences coexist with more dominant rhythmic patterns (Clopper; Smiljanić, 2011).

It is also worth noting that the presence of zero and near-zero values in the dataset should not be interpreted solely as artifacts, but rather as an intrinsic characteristic of acoustic-prosodic representation. Within standard feature extraction frameworks such as GeMAPS (Eyben *et al.*, 2016), zero values are used to encode unvoiced regions (e.g., $f_0 = 0$), thereby capturing the absence of phonatory activity. In this sense, zero values contribute directly to the modeling of rhythmic and, more modestly, melodic structures, particularly in pause-related measures. This is consistent with the present findings, in which duration-based parameters emerged as the strongest predictors of dialectal variation between PB and SP dialects. In relation to the extreme values observed in the dataset, they were retained because acoustic-prosodic measures are inherently sensitive to both linguistic variability and signal-dependent factors. In speech analysis, domains such as duration, intensity, and f_0 may naturally exhibit wide ranges due to differences in speech rate, pausing behavior, and intensity emphasis. Values consistent with natural speech production, such as extended pauses or abrupt pitch movements, were retained as they contribute to the representation of prosodic structure.

- **RQ2** – To what extent can automatic dialect classification be explained by the relative contribution of prosodic-acoustic features, and how do these features affect ASR performance?

The automatic classification results further clarify how these acoustic dimensions operate in applied computational systems. The high classification accuracy rates observed in the predictive phase (ranging from 94.6% to 97.8%) demonstrate that the dialectal patterns identified through statistical modeling are sufficiently structured to allow reliable prediction of dialect membership in previously unseen speech samples. Importantly, the feature-importance ranking derived from the Gini impurity reduction method mirrors the inferential results obtained in the mixed-effects analysis. Pause variability emerged as the dominant predictor (~ 0.43 importance weight), followed by additional duration and intensity-related measures, whereas f_0 -based parameters occupied comparatively lower positions in the hierarchy of predictive relevance. This convergence between inferential effect sizes and predictive feature importance strengthens the construct validity of the prosodic-acoustic parameters identified as dialect markers and confirms their explanatory and predictive relevance for dialect classification tasks (Bishop, 2006; Emara; Shaker, 2024; Molnar, 2022). As discussed previously, this alignment reflects the complementary goals of explanatory and predictive modeling frameworks (Shmueli, 2010; Yarkoni; Westfall, 2017). The inferential phase established which acoustic dimensions significantly account for cross-dialect variance while

controlling for inter-speaker variability, whereas the predictive phase evaluated whether these same dimensions generalize effectively in classification tasks.

The fact that duration-related parameters both explain dialect variance and dominate feature-importance rankings suggests that rhythmic structuring constitutes the primary acoustic dimension influencing ASR performance in the present dataset, and for these two inland Brazilian Portuguese varieties specifically. The primacy of rhythmic parameters observed here reflects the prosodic organization of PB and SP as documented in this corpus; it should not be assumed that the same hierarchy applies to other dialectal contrasts or other languages. Cross-linguistic research has consistently shown that the relative contribution of *f*₀, duration, and intensity to dialect and accent differentiation varies considerably depending on the specific varieties under comparison, the phonological systems involved, and the acoustic dimensions that carry social meaning in each community (Thomas, 2011; White; Mattys, 2007). Whether durational parameters maintain their dominant role in other Brazilian Portuguese regional contrasts, or whether *f*₀ or voice quality dimensions assume greater prominence in other variety pairs, remains an open empirical question that future research must address.

From an applied perspective, this finding indicates that systems trained on Brazilian Portuguese must account for dialect-specific (or dialect-grouped) rhythmic organization to achieve consistent recognition performance. In this sense, the observed temporal parameters may function as potential sociophonetic markers of dialect differentiation, reflecting systematic differences in rhythmic organization between PB and SP speakers. However, given the pilot nature of this research, and based on a limited dataset, these acoustic patterns should be interpreted cautiously and require validation through larger and more heterogeneous corpora. Future research should examine whether such cues remain stable across multiple dimensions of intra-dialect variation: across age groups, sex and gender, socioeconomic strata, rural-urban continua, and different speech registers or task types within the same regional variety. The present pilot does not permit such distinctions, given that all participants were young university students from comparable academic and institutional contexts. Establishing the stability of these prosodic markers across sociolinguistic stratification is a necessary step toward validating their status as robust dialect-level features rather than community- or context-specific patterns.

Sociophonetic research has demonstrated that dialectal identity emerges from gradient distributions of phonetic features that may vary according to social and geographic factors rather than from single categorical markers (Thomas, 2011; Yaeger-Dror; Fagyal, 2011). Consequently, while the present results suggest that rhythmic parameters may serve as reliable indicators of PB-SP differentiation, further investigation is necessary to determine the extent to which these patterns generalize across speakers and sociolinguistic contexts. Such observations are consistent with previous research demonstrating that automatic dialect classification is particularly sensitive to

sociophonetic variability in speech production (Jassim; Abdulmohsin, 2025; Wassink; Gansen; Bartholomew, 2022).

The present results are also consistent with previous work on automatic classification of Brazilian regional accents, which has shown that acoustic modeling techniques can successfully capture structured dialectal variation in speech signals. In a comprehensive investigation of accent classification in Brazilian Portuguese, Batista (2019) demonstrated that statistical and machine learning approaches can achieve reliable classification of regional speech varieties by modeling acoustic representations extracted from speech corpora. Her work also highlights that classification performance may vary depending on the evaluation scenario, with cross-dataset testing presenting greater challenges due to differences in recording conditions and speaker populations. This observation aligns with the current results suggesting that dialect-sensitive acoustic patterns must be sufficiently stable to generalize across speakers and recording contexts. In this sense, the high classification accuracy observed further supports the argument that regional dialect differences are encoded in systematic prosodic-acoustic configurations that can be detected not only through inferential statistical modeling but also through predictive computational systems.

In broader terms, the integration of mixed-effects modeling and machine learning classification strengthens the theoretical interpretation of dialect differentiation (Hay, 2011). As noted in sociophonetic research, combining inferential statistical approaches with computational classification techniques provides a particularly productive methodological toolkit for analyzing structured phonetic variation. Mixed-effects models allow researchers to estimate the influence of linguistic predictors while accounting for speaker-level variability, whereas tree-based and other ML methods can detect complex interactions among acoustic features that may characterize dialectal patterns. In line with broader methodological discussions distinguishing explanatory and predictive modeling in quantitative research (Hay, 2011; Shmueli, 2010; Yarkoni; Westfall, 2017), this work integrates both approaches through two complementary analytical phases: an inferential stage based on corpus design, acoustic extraction procedures, and mixed-effects statistical analysis (Methodology 1, section 2), and a predictive stage based on automatic dialect classification using ML techniques (Methodology 2, section 4). This dual framework identifies statistically significant acoustic variables of dialect variation and evaluates their discriminative power in classification tasks.

Moreover, rather than treating dialect categories as discrete phonological entities, the present findings support a gradient, usage-based perspective in which dialect identity emerges from weighted distributions of prosodic-acoustic features. Within this hierarchy, rhythmic organization appears as the dominant dimension of sociophonetic differentiation, followed by spectral-intensity patterns and more subtle pitch-related variability.

Although the present study focuses on automatic dialect classification rather than speech synthesis, the findings can be interpreted in light of recent work on prosody modeling in speech technologies. As mentioned in section 1, Galdino *et al.* (2025) conducted a systematic review of

prosody evaluation in speech synthesis systems showing that most computational approaches rely primarily on f0-based descriptors, followed by duration and intensity, reflecting a strong emphasis on pitch-related modeling in current speech technologies. In contrast, the results obtained in this research suggest a different hierarchical organization for the dialectal differentiation investigated here, in which durational-rhythmic parameters, particularly pause-based measures, emerge as the most informative predictors, followed by intensity-related measures and more secondary contributions from f0. This contrast highlights the importance of investigating prosodic-acoustic structure in specific linguistic and dialectal contexts. In particular, the literature on speech synthesis also emphasizes that many languages and dialects remain underrepresented in prosody modeling, largely due to data scarcity and limited annotated speech corpora. The present results therefore contribute to addressing this gap by providing empirical evidence on how prosodic-acoustic cues operate in two inland varieties of Brazilian Portuguese, offering insights that may inform both dialect-aware speech technologies and sociophonetic descriptions of regional speech variation (Mehralian; Van Hamme, 2025).

It is essential to note that a limitation of the present classification concerns the use of a sample-level train-test split, which may allow observations derived from the same speaker to be included in both the training and test sets. Because the dataset consists of multiple segmented samples per speaker, this design introduces a degree of dependency between subsets and therefore relaxes the assumption of independence underlying standard evaluation procedures. Consequently, the reported classification accuracy should be interpreted with caution, as it may represent an upper-bound estimate of performance when compared to a strictly speaker-independent evaluation. Methodological research has consistently shown that when related observations are shared across partitions, performance metrics may be inflated due to partial overlap in speaker-specific characteristics (Bishop, 2006; James *et al.*, 2013). In speech-related tasks, this distinction is particularly well established, with speaker-dependent configurations consistently outperforming speaker-independent ones because they do not fully control for inter-speaker variability (Hastie; Tibshirani; Friedman, 2009; Jassim; Abdulmohsin, 2025). Empirical evidence from speech classification studies reinforces this point by demonstrating that evaluation procedures allowing speaker overlap between training and test sets can yield substantially higher accuracy than those enforcing speaker independence (Rybka; Janicki, 2013; Atmaja; Sasou, 2022; Sasse *et al.*, 2025).

Despite this limitation, the present findings remain informative within the scope of a pilot investigation. The use of a sample-level split was motivated by the need to evaluate whether the prosodic-acoustic variables identified as statistically significant in the inferential stage, especially those related to rhythmic organization, are sufficiently robust to support dialect classification in a predictive setting. The observed convergence between mixed-effects modeling and feature-importance rankings provides consistent evidence that these variables encode structured dialectal

differences. However, the interpretation of classification performance must remain methodologically cautious, as the current design prioritizes feature sensitivity over strict generalization.

The results addressing RQ1 and RQ2 support the hypotheses proposed in this study. The analyses demonstrate that prosodic-acoustic descriptors derived from duration, intensity, and f0 provide statistically consistent cues for distinguishing the inland Brazilian Portuguese dialects PB and SP. Feature-importance analyses revealed a hierarchical contribution of acoustic parameters, with pause-based rhythmic variability emerging as the dominant dimension in dialect classification. The convergence between inferential statistical modeling and predictive machine-learning results indicates that dialect differentiation in inland Brazilian Portuguese is not only statistically observable but also computationally detectable.

These findings therefore confirm the initial hypothesis that prosodic-acoustic features provide reliable cues for automatic dialect classification and that their relative contribution follows a structured hierarchy across acoustic domains. In broader terms, by providing converging explanatory and predictive evidence, this study reinforces the technological relevance of dialect-sensitive acoustic modeling for ASR systems while also highlighting the sociophonetic importance of documenting and analyzing minoritized and under-researched inland dialects, which remain underrepresented in the development of intelligent speech technologies.

7. CONCLUSION

This study makes two complementary contributions: a sociophonetic characterization of the prosodic-acoustic structure distinguishing two inland Brazilian Portuguese varieties (PB-SP), and a computational validation of that structure through automatic dialect classification. By integrating inferential statistical mixed-effects modeling with predictive machine-learning classification, the study demonstrated that dialect differentiation between PB and SP is acoustically structured, hierarchically organized, and computationally detectable. It is worth emphasizing that neither stage is subordinate to the other; together, they constitute a dual inferential-predictive framework in which statistical evidence and machine-learning performance point in the same analytical direction, mutually reinforcing the overall interpretation of the findings.

The results demonstrate that rhythmic organization, particularly pause variability, emerges as the primary acoustic dimension distinguishing the two inland dialects, within this dataset and for these two Brazilian Portuguese varieties specifically. This hierarchical ordering should not be generalized to other language varieties or dialectal contrasts without independent empirical validation, as the relative weighting of prosodic dimensions in dialect differentiation is known to vary across linguistic systems and sociolinguistic contexts. Spectral tilt and intensity-related measures contribute meaningfully, suggesting that habitual phonatory settings, voice quality, and vocal effort

also encode dialect identity. Although f_0 -based parameters show statistically significant differences, they function as secondary cues within the overall hierarchy of prosodic weighting. The convergence between mixed-effects results and feature-importance rankings reinforces the construct validity of these acoustic parameters and confirms that the same features explaining dialect variance also sustain high predictive performance in ASR classification.

From a sociophonetic standpoint, the results underscore that inland dialect identity in Brazilian Portuguese is encoded not merely in segmental differences but also in structured prosodic configurations. The dominance of temporal organization suggests that rhythm operates as a socially embedded indexical resource, contributing to regional differentiation. By focusing on inland varieties that remain comparatively underrepresented in speech research and technology development, this study expands the empirical coverage of Brazilian Portuguese dialectology and highlights the importance of incorporating minority and under-researched dialects into computational modeling.

Technologically, the findings demonstrate that dialect-aware systems benefit from integrating rhythm-sensitive acoustic features (Mehralian; Van Hamme, 2025). The high classification accuracy achieved across models confirms that prosodic-acoustic variation is stable enough to support reliable dialect discrimination. This dual inferential-predictive validation strengthens both the sociophonetic interpretation of the results and their practical implications for intelligent speech technology.

8. LIMITATIONS AND FUTURE DIRECTIONS

Despite the robustness of these initial findings, several limitations must be acknowledged. First, the sample size was relatively modest – 20 speakers yielding approximately 1 hour and 30 minutes of recorded audio used to train the classifiers, and although the use of mixed-effects modeling accounts for speaker-level variability, broader sampling across age groups, educational backgrounds, and sociolinguistic profiles would enhance generalizability. Future studies should incorporate larger and more socially diverse speaker groups to further test the stability of the identified acoustic hierarchies.

Second, the corpus consisted of reading speech derived from a podcast script. While this ensured phonetic control and comparability across speakers, it may not fully capture spontaneous prosodic variation. Given that rhythm and pause organization are sensitive to discourse context, future research should examine whether similar hierarchical patterns emerge in semi-spontaneous or conversational speech.

Third, the classification phase relied on a predefined set of prosodic-acoustic features. Although the feature-importance ranking revealed a clear hierarchy, additional acoustic dimensions, such as segmental features, spectral moments, dynamic intonational contours, voice quality features, or duration-based descriptors could further refine dialect discrimination. Furthermore, incorporating

articulatory measures, as suggested in recent dialect research (e.g., Wieling; Tiede; Tomaschek, 2017), may provide complementary insights into regional variation.

Fourth, while this work focused on binary dialect classification (PB vs. SP), Brazilian Portuguese exhibits a rich continuum of regional varieties. Expanding the framework to multi-dialect classification tasks would allow testing whether the hierarchical dominance of rhythmic organization observed here generalizes across broader dialectal comparisons. Such an expansion would also help identify and validate sociophonetic markers that can differentiate regional speech patterns, providing a clearer understanding of how dialect identity is encoded in prosodic-acoustic structures. Beyond theoretical implications, incorporating these markers into computational models could significantly improve dialect-aware systems operating in linguistically diverse environments (Mehralian; Van Hamme, 2025).

Finally, future research should extend the present framework by implementing speaker-level partitioning strategies for the train-test split, in which all samples from a given speaker are assigned exclusively to either the training or test set. Such an approach would provide a more conservative and ecologically valid evaluation of how well the identified prosodic-acoustic features generalize to unseen speakers, which is essential for real-world ASR applications. In addition, expanding the dataset to include a larger and more diverse pool of speakers, including intra-dialectal variation (e.g., rural and urban accents), would further strengthen the robustness of classification models. Within this perspective, the present study should be understood as an initial step that establishes the acoustic and hierarchical structure of dialect differentiation, while also delineating the methodological pathway for subsequent, more stringent evaluations.

If future research addresses the limitations outlined above, particularly through larger and more socially diverse samples, spontaneous speech data, and expanded dialect comparisons, the findings may offer significant potential for both forensic and pedagogical applications. In forensic scenarios, the identification of hierarchically weighted prosodic-acoustic (or segmental) markers, such as pause variability and spectral measures of voice quality, may contribute to more reliable speaker profiling and regional attribution in cases involving dialect-sensitive analysis. From a pedagogical perspective, understanding which prosodic dimensions systematically differentiate inland Brazilian Portuguese dialects can promote greater awareness of linguistic diversity in educational contexts, reducing stigmatization of non-metropolitan varieties and supporting accent-inclusive language teaching practices.

ADDITIONAL INFORMATION

CONFLICT OF INTEREST

The authors declare no competing interests.

STATEMENT OF DATA AVAILABILITY

The data, code, and materials that support the findings of this study are openly available in Zenodo at <https://doi.org/10.5281/zenodo.19421930>. The repository contains the following components: (i) two datasets used for both the inferential and predictive analyses; (ii) Praat scripts for the complete acoustic processing pipeline, covering phonetic forced alignment, acoustic feature extraction, and annotation segmentation; (iii) an R function script for running the Linear Mixed-Effects Models used in the inferential analysis; and (iv) Python scripts for running each machine-learning algorithm presented in this study. For items (ii), (iii), and (iv), usage instructions are provided in two formats: (a) a detailed step-by-step, line-by-line User Manual; and (b) a README.md file accompanying each script.

The prosodic-acoustic dataset used in this study (BrAWCorp) is openly available to registered researchers. Registration requests must be submitted to: rl1307@unicamp.br.

AI USE STATEMENT

The authors used six machine-learning models—Random Forest, Decision Tree, k-Nearest Neighbors, Linear Discriminant Analysis, Support Vector Machine, and Gradient Boosting Machine—for automatic binary dialect classification and feature-importance analysis in the predictive stage described in Methodology 2. No generative artificial intelligence tools were used in the preparation of this manuscript. All model outputs were reviewed and interpreted by the authors, who take full responsibility for the content of this work.

ETHICS AND CONSENT

The raw audio files from which the BrAWCorp dataset was derived are governed by the access policy described in the Data Availability Statement. Controlled access is available to registered researchers upon submission of a registration request to rl1307@unicamp.br, subject to the ethical terms established under approval number CAAE 67001923.9.1001.8142, approved on 17 May 2023. For further details on access conditions, see the Data Availability Statement.

All participants provided informed consent prior to participation. The consent procedure was administered by the second author, a researcher formally affiliated with the approved project. Participants authorized the recording and use of their speech data for scientific research purposes. For the purposes of this study, all speech samples were anonymized prior to analysis: speaker identities were replaced by codes, and no personally identifiable information is reported or disclosed in this manuscript or in the associated open materials deposited in Zenodo (<https://doi.org/10.5281/zenodo.19421930>).

The raw audio recordings cannot be shared publicly due to participant confidentiality requirements established by the ethics committee approval (CAAE 67001923.9.1001.8142). Access to restricted data may be requested by contacting Luciani Ester Tenani (UNESP) at luciani.tenani@unesp.br, subject to the approval of the requesting researcher's institutional ethics committee and the signing of a data access agreement. These files are part of the BrAMCorp database (UNICAMP – FAPESP), which is currently under development and expected to become publicly available in 2027.

The database as a whole has not yet completed the full anonymization procedures required for unrestricted public release. For this reason, the raw audio files remain restricted at this stage. Controlled external access may be requested by contacting Luciani Ester Tenani (UNESP) at luciani.tenani@unesp.br, subject to the approval of the requesting researcher's institutional ethics committee and the signing of a data-access agreement, in accordance with the ethical terms of the approved project. Public access is expected to become available in 2027 via the *Letramento Contemporâneo* platform, once the required anonymization and platform development processes are complete. The second author, as a member of that research team, authorized the use of the relevant materials in this study under the conditions established in the approved project, ensuring full compliance with the approved ethical guidelines and the non-disclosure of participants' identifiable data. For further details on access conditions, see the Data Availability Statement.

FUNDING SOURCES

This research received financial support from the National Council for Scientific and Technological Development (CNPq), Brazil. Leônidas Silva Jr. was supported by CNPq under grant 307010/2022-8, and Luciani E. Tenani was supported by CNPq under grant 304567/2024-8 and by FAPESP under grant 2022/05908-0.

REVIEW AND AUTHORS' REPLY

Review: <https://doi.org/10.25189/2675-4916.2026.V7.N7.ID1016.R>

Author's Reply: <https://doi.org/10.25189/2675-4916.2026.V7.N7.ID1016.A>

REFERENCES

AKIS, E.; BELL, S.; SIMPSON, D. Objective measures of auditory temporal resolution with ABR. *International Journal of Audiology*, v. 64, n. 11, p. 1095-1105, 2025. DOI: <https://doi.org/10.1080/14992027.2025.2486847>.

ATMAJA, B. T.; SASOU, A. Effect of different splitting criteria on the performance of speech emotion recognition. In: *IEEE REGION 10 CONFERENCE (IEEE TENCON 2022)*. Hong Kong: IEEE, 2022. p. 760-764.

BAAYEN, R. H.; DAVIDSON, D. J.; BATES, D. M. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, v. 59, n. 4, p. 390-412, 2008. DOI: <https://doi.org/10.1016/j.jml.2007.12.005>.

BARBOSA, P. A. *Incursões em torno do ritmo da fala*. Campinas: Pontes, 2006.

BARBOSA, P. A.; ALVARENGA, L. The interplay between syllabic duration and melody to signal prosodic functions in reading and story retelling in Brazilian Portuguese. *Languages*, v. 9, n. 268, p. 1-18, 2024. DOI: <https://doi.org/10.3390/languages9080268>.

BARBOSA, P. A.; MADUREIRA, S. *Manual de fonética acústica experimental: aplicações a dados do português brasileiro*. São Paulo: Cortez, 2015.

BARR, D. J.; LEVY, R.; SCHEEPERS, C.; TILY, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, v. 68, n. 3, p. 255-278, 2013. DOI: <https://doi.org/10.1016/j.jml.2012.11.001>.

BARTOŃ, K. *MuMIn: Multi-Model Inference*. R package version 1.48.19. 2024. Available at: <https://CRAN.R-project.org/package=MumIn>. Accessed on: 14 March 2026.

BATES, D.; MÄCHLER, M.; BOLKER, B.; WALKER, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, v. 67, n. 1, p. 1-48, 2015. DOI: <https://doi.org/10.18637/jss.v067.i01>. Accessed on: 14 March 2026.

BATISTA, N. *Estudo sobre identificação automática de sotaques regionais brasileiros baseada em modelagens estatísticas e técnicas de aprendizado de máquina*. 2019. Dissertação (Mestrado em Engenharia Elétrica – Telecomunicações e Telemática) – Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, 2019.

BEHRAVAN, H. *Advances in automatic foreign accent recognition*. 2016. PhD Dissertation – University of Eastern Finland, Joensuu, 2016.

BISHOP, C. M. *Pattern recognition and machine learning*. New York: Springer, 2006.

BOERSMA, P.; WEENINK, D. *Praat: doing phonetics by computer*. Version 6.1.38. Computer software. Amsterdam: University of Amsterdam, 2024. Available at: <https://www.praat.org/> Accessed on: 24 January 2026.

BRAID, A. *Fonética forense*. 2. ed. Campinas: Millennium, 2003.

BROWNLEE, J. *Train-test split for evaluating machine learning algorithms*. Python Machine Learning, v. 79, p. 1-12, 2020.

CAGLIARI, L.; ABAURRE, M. Elementos para uma investigação instrumental das relações entre padrões rítmicos e processos fonológicos no português brasileiro. *Caderno de Estudos Linguísticos*, v. 10, p. 39-57, 1986.

CANTONI, M. M.; OLIVEIRA, B. G.; NEVADO, H. M. Introdução à análise acústica da fala com o Praat. *Texto Livre: Linguagem e Tecnologia*, v. 15, p. e37947, 2022. DOI: <https://doi.org/10.35699/1983-3652.2022.37947>.

CHAN, M. P. Y.; CHOE, J.; LI, A.; CHEN, Y.; GAO, X.; HOLLIDAY, N. Training and typological bias in ASR performance for world Englishes. In: INTERSPEECH 2022, 2022, Incheon. *Proceedings of Interspeech 2022*. [S. l.]: International Speech Communication Association, 2022. p. 1273-1277. DOI: <https://doi.org/10.21437/Interspeech.2022-10869>.

CHIN, W. W. Issues and opinion on structural equation modeling. *MIS Quarterly*, v. 22, n. 1, p. vii-xvi, 1998. Available at: <http://www.jstor.org/stable/249674> Accessed on: 25 February 2026.

CLOPPER, C. G.; SMILJANIĆ, R. Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics*, v. 39, n. 2, p. 237-245, 2011. DOI: <https://doi.org/10.1016/j.wocn.2011.02.006>.

COHEN, J. A power primer. *Psychological Bulletin*, v. 112, n. 1, p. 155-159, 1992. DOI: <https://doi.org/10.1037/0033-2909.112.1.155>.

CONSTANTINI, A.; BARBOSA, P. Prosodic characteristics of different varieties of Brazilian Portuguese. *Revista Brasileira de Criminalística*, v. 4, n. 3, p. 44-53, 2015. DOI: <https://doi.org/10.15260/rbc.v4i3.103>.

CRAVEIRO, G.; SANTOS, V.; DALALANA, G.; SVARTMAN, F.; ALUÍSIO, S. Simple and fast automatic prosodic segmentation of Brazilian Portuguese spontaneous speech. In: *Proceedings of the 16th International Conference on Computational Processing of*

Portuguese – v. 1. Associação Portuguesa de Linguística / Association for Computational Linguistics. São Paulo, 2024. p. 32–44.

EMARA, I. F.; SHAKER, N. H. The impact of non-native English speakers' phonological and prosodic features on automatic speech recognition accuracy. *Speech Communication*, v. 157, Article 103038, p. 1–12, 2024. DOI: <https://doi.org/10.1016/j.specom.2024.103038>.

ESLING, J.; EDMONDSON, J. Acoustical analysis of voice quality for sociophonetic purposes. In: DI PAOLO, M.; YAEGER-DROR, M. (ed.). *Sociophonetics: a student's guide*. London: Routledge, p. 131–148, 2011.

EYBEN, F.; SCHERER, K.; SCHULLER, B.; SUNDBERG, J.; ANDRÉ, E.; BUSO, C.; DEVILLERS, L.; EPPS, J.; LAUKKA, P.; NARAYANAN, S.; TRUONG, K. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, v. 7, n. 2, p. 190–202, 2016. DOI: <https://doi.org/10.1109/TAFFC.2015.2457417>.

FLEGE, J. Second language speech learning: theory, findings and problems. In: STRANGE, W. (ed.). *Speech perception and linguistic experience: issues in cross-language research*. Timonium, MD: York Press, 1995. p. 233–277.

GALDINO, J.; MATOS, A.; SVARTMAN, F.; ALUÍSIO, S. The evaluation of prosody in speech synthesis: a systematic review. *Journal of the Brazilian Computer Society*, v. 31, n. 1, 2025. DOI: <https://doi.org/10.5753/jbcs.2025.5468>.

GARELLEK, M. Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics*, v. 94, p. 1–22, 2022. DOI: <https://doi.org/10.1016/j.wocn.2022.101155>.

GAUY, M.; FINGER, M. Acoustic models of Brazilian Portuguese speech based on neural transformers. *arXiv preprint arXiv:2312.09265*, 2023. DOI: <https://doi.org/10.48550/arXiv.2312.09265>.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference, and prediction*. 2. ed. New York: Springer, 2009.

HARRIS, C. R.; MILLMAN, K. J.; VAN DER WALT, S. J.; GOMMERS, R.; VIRTANEN, P.; COUNAPEAU, D.; WIESER, E.; TAYLOR, J.; BERG, S.; SMITH, N. J.; KERN, R.; PICUS, M.; HOYER, S.; VAN KERKWIJK, M. H.; BRETT, M.; HALDANE, A.; DEL RÍO, J. F.; WIEBE, M.; PETERSON, P.; GÉRARD-MARCHANT, P.; SHEPPARD, K.; REDDY, T.; WECKESSER, W.; ABBASI, H.; GOHLKE, C.; OLIPHANT, T. E. Array programming with NumPy. *Nature*, v. 585, p. 357–362, 2020. DOI: <https://doi.org/10.1038/s41586-020-2649-2>.

HAY, J. Statistical analysis. In: DI PAOLO, M.; YAEGER-DROR, M. (ed.). *Sociophonetics: a student's guide*. London: Routledge, p. 198–214, 2011.

HENSELER, J.; RINGLE, C.; SINKOVICS, R. The use of partial least squares path modeling in international marketing. In: SINKOVICS, R.; GHOURI, P. (ed.). *New challenges to international marketing*. Bingley: Emerald Group Publishing, 2009. p. 277–320. DOI: [https://doi.org/10.1108/S1474-7979\(2009\)0000020014](https://doi.org/10.1108/S1474-7979(2009)0000020014).

HIRST, D. The analysis by synthesis of speech melody: from data to models. *Journal of Speech Sciences*, v. 1, n. 1, p. 55–83, 2011. DOI: <https://doi.org/10.20396/joss.v1i1.15011>.

HUNTER, J. D. Matplotlib: a 2D graphics environment. *Computing in Science & Engineering*, v. 9, n. 3, p. 90–95, 2007. DOI: <https://doi.org/10.1109/MCSE.2007.55>.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An introduction to statistical learning: with applications in R*. New York: Springer, 2013.

JASSIM, S.; ABDULMOHSIN, H. Accent Classification Using Machine Learning Techniques: A Review. *International Journal of Computer Information Systems and Industrial Management Applications*, v. 17, p. 421–451, 2025. DOI: <https://doi.org/10.70917/ijcisim-2025-0028>.

KISLER, T.; REICHEL, U. D.; SCHIEL, F. Multilingual processing of speech via web services. *Computer Speech & Language*, v. 45, p. 326–347, 2017a. DOI: <https://doi.org/10.1016/j.csl.2016.11.005>.

- KISLER, T.; REICHEL, U. D.; SCHIEL, F. *WebMAUS Basic*: web service for automatic phoneme segmentation. Munich: BAS Web Services, Ludwig Maximilian University of Munich, 2017b. Available at: <https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic>. Accessed on: 14 January 2026.
- KOENECKE, A.; NAM, A.; LAKE, E.; NUDELL, J.; QUARTEY, M.; MENGESHA, Z.; TOUPS, C.; RICKFORD, J. R.; JURAFSKY, D.; GOEL, S. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, v. 117, n. 14, p. 7684–7689, 2020. DOI: <https://doi.org/10.1073/pnas.1915768117>.
- KRIVOKAPIC, J. Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics*, v. 35, n. 2, p. 162–179, 2007. DOI: <https://doi.org/10.1016/j.wocn.2006.04.001>.
- KUMAR, Y.; SINGH, N. A comprehensive view of automatic speech recognition system – a systematic literature review. In: INTERNATIONAL CONFERENCE ON AUTOMATION, COMPUTATIONAL AND TECHNOLOGY MANAGEMENT (ICACTM), 2019, London. IEEE, 2019. p. 168–173. DOI: <https://doi.org/10.1109/ICACTM.2019.8776714>.
- KUZNETSOVA, A.; BROCKHOFF, P. B.; CHRISTENSEN, R. H. B. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, v. 82, n. 13, p. 1–26, 2017. DOI: <https://doi.org/10.18637/jss.v082.i13>. Accessed on: 3 April 2026.
- LADEFOGED, P. *Phonetic data analysis: an introduction to fieldwork and instrumental techniques*. Malden: Blackwell, 2003.
- LENTH, R. V. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 2.0.3. 2024. Available at: <https://CRAN.R-project.org/package=emmeans>. Accessed on: 3 April 2026.
- MARKL, N. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT 2022), 5., 2022, Seoul. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, 2022. p. 521–534. DOI: <https://doi.org/10.1145/3531146.3533117>.
- MCKINNEY, W. Data structures for statistical computing in Python. In: PYTHON IN SCIENCE CONFERENCE, 9., 2010, Austin. *Proceedings...* Austin: SciPy, 2010. p. 56–61. DOI: <https://doi.org/10.25080/Majora-92bf1922-00a>.
- MEHRALIAN, P.; VAN HAMME, H. Leveraging geographic metadata for dialect-aware speech recognition. In: *Proceedings of International Speech Communication Association (INTERSPEECH 2025)*. Rotterdam, The Netherlands: International Speech Communication Association (ISCA), 2025. p. 1153–1157. DOI: <https://doi.org/10.21437/Interspeech.2025-1839>.
- MEIRELES, A.; GAMBARINI, V. Rhythm typology of Brazilian Portuguese dialects. In: 6th International Conference on Speech Prosody, Shanghai, China. International Speech Communication Association (ISCA), 2012. *Proceedings of Speech Prosody 2012*, p. 474–477.
- MIRMAN, D. *Growth curve analysis and visualization using R*. Boca Raton: CRC Press, 2014.
- MOKSONY, F. Small is beautiful: the use and interpretation of R² in social research. *Szociológiai Szemle*, p. 130–138, 1999.
- MOLNAR, C. *Interpretable machine learning*. 2. ed. [S. l.]: Lulu.com, 2022. Available at: <https://christophm.github.io/interpretable-ml-book/>. Accessed on: 2 April 2026.
- MOYER, A. *Foreign accent: the phenomenon of non-native speech*. Cambridge: Cambridge University Press, 2013.
- NAKAGAWA, S.; SCHIELZETH, H. A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, v. 4, n. 2, p. 133–142, 2013. DOI: <https://doi.org/10.1111/j.2041-210x.2012.00261.x>.
- NIELSEN, A. *Análise prática de séries temporais: predição com estatística e aprendizado de máquina*. Tradução de Cibelle Ravaglia. Rio de Janeiro: Alta Books, 2021.
- ORTEGA-LLEBARIA, M.; SILVA JR., L.; NAGAO, J. Macro and micro-rhythm in L2 English: exploration and refinement of measures. In: SKARNITZL, R.; VOLÍN, J. (ed.). *Proceedings of the 20th International Congress of Phonetic Sciences*, 2023. p. 1582–1586. Available at: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/830.pdf. Accessed on: 22 March 2026.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Available at: <https://jmlr.org/papers/v12/pedregosa11a.html>. Accessed on: 2 July 2026.

R CORE TEAM. R: *A language and environment for statistical computing*. Computer software. Vienna: R Foundation for Statistical Computing, 2026. Available at: <https://www.R-project.org/>. Accessed on: 25 January 2026.

RYBKA, J.; JANICKI, A. Comparison of speaker dependent and speaker independent emotion recognition. *International Journal of Applied Mathematics and Computer Science*, v. 23, n. 4, p. 797–808, 2013. DOI: <https://doi.org/10.2478/amcs-2013-0060>.

SASSE, L.; NICOLAISEN-SOBESKY, E.; DUKART, J.; EICKHOFF, S. B.; GÖTZ, M.; HAMDAN, S.; KOMEYER, V.; KULKARNI, A.; LAHNAKOSKI, J. M.; LOVE, B. C.; RAIMONDO, F.; PATIL, K. R. Overview of leakage scenarios in supervised machine learning. *Journal of Big Data*, v. 12, n. 1, p. 135, 2025. DOI: <https://doi.org/10.1186/s40537-025-01193-8>.

SHMUELI, G. To explain or to predict? *Statistical Science*, v. 25, n. 3, p. 289–310, 2010. DOI: <https://doi.org/10.1214/10-STS330>.

SILVA, C.; GOMES, F.; SANTOS, J.; SANTOS, C. Uma análise comparativa das técnicas de machine learning: regressão logística, árvores de decisão, Random Forest e SVM. *Apoena*, v. 7, p. 501–511, 2023. Available at: <https://publicacoes.unijorge.com.br/apoena/article/view/183>. Accessed on: 15 January 2026.

SILVA JR., L. SupplementaryMaterials: Tools for Speech Science Research. *Zenodo*, 2026. DOI: <https://doi.org/10.5281/zenodo.19421930>.

SILVA JR., L.; BARBOSA, P. A. Foreign accent and forensic speaker identification in voice lineups: the influence of acoustic features based on prosody. *Journal of Visualized Experiments*, n. 211, p. 1–21, 2024. DOI: <https://doi.org/10.3791/66313>.

SILVA JR., L.; BARBOSA, P. A. Voice disguise and foreign accent: prosodic aspects of English produced by Brazilian Portuguese speakers. *Journal of Experimental Phonetics*, v. 32, p. 195–226, 2023. DOI: <https://doi.org/10.1344/efe-2023-32-195-226>.

SILVA JR., L.; BARBOSA, P. A. Foreign accent and L2 speech rhythm of English: a pilot study based on metric and prosodic parameters. In: *COLÓQUIO BRASILEIRO DE PROSÓDIA DA FALA, 2., 2022, Belo Horizonte. Proceedings of the 2nd Brazilian Prosody Conference*. v. 1. 2022. p. 41–50. Available at: http://www.periodicos.letras.ufmg.br/index.php/anais_coloquio/article/view/34734 Accessed on: 15 January 2026.

SILVA JR., L.; BARBOSA, P. A. Speech rhythm of English as L2: an investigation of prosodic variables on the production of Brazilian Portuguese speakers. *Journal of Speech Sciences*, v. 8, n. 2, p. 37–57, 2019. DOI: <https://doi.org/10.20396/joss.v8i2.14996>.

SILVA JR., L.; BARBOSA, P.; MONTE, J. PROSODY: A Semi-Automated Protocol for Experimental Prosody Research. In: *Proceedings of Speech Prosody 2024*, Leiden, The Netherlands: International Speech Communication Association (ISCA), 2024. p. 1245–1249. DOI: <https://doi.org/10.21437/SpeechProsody.2024-251>.

SILVA JR., L.; SILVA, J.; MEER, P. Prosodic aspects of Brazilian L2 English: A comparison of duration-based rhythm and F0 measures with American English, Indian English, and Brazilian Portuguese. In: *Proceedings of Speech Prosody 2024*, Leiden, The Netherlands: International Speech Communication Association (ISCA), 2024. v. 7. p. 101–105. DOI: <https://doi.org/10.21437/SpeechProsody.2024-21>.

SILVEIRA, G.; ALBERT, A.; GRICE, M. Probing prosodic differences between two regional varieties of Brazilian Portuguese. In: *Proceedings of Interspeech 2025*. Rotterdam, The Netherlands: International Speech Communication Association (ISCA), 2025. p. 2935–2939. DOI: <https://doi.org/10.21437/Interspeech.2025-1863>.

TATMAN, R. Gender and dialect bias in YouTube's automatic captions. In: *ACL WORKSHOP ON ETHICS IN NATURAL LANGUAGE PROCESSING (EthNLP)*, 1., 2017, Valencia. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics, 2017. p. 53–59. DOI: <https://doi.org/10.18653/v1/W17-1606>.

TATMAN, R.; KASTEN, C. Effects of talker dialect, gender & race on accuracy of Bing Speech and YouTube automatic captions. In: INTERSPEECH 2017, 2017, Stockholm. *Proceedings of Interspeech 2017*. [S. l.]: International Speech Communication Association, 2017. p. 934–938. DOI: <https://doi.org/10.21437/Interspeech.2017-1746>.

TEIXEIRA, L. A. A multidimensional analysis of English-L2 rhythm development. *Ilha do Desterro*, v. 76, n. 3, p. 223–250, 2023. DOI: <https://doi.org/10.5007/2175-8026.2023.e94742>.

THOMAS, E. R. *Sociophonetics: An introduction*. London: Palgrave Macmillan, 2011.

TORGBI, M.; CLAYMAN, A.; SPEIGHT, J.; MADABUSHI, H. Adapting Whisper for regional dialects: Enhancing public services for vulnerable populations in the United Kingdom. In: SCHERRER, Y.; JAUHIAINEN, T.; LJUBEŠIĆ, N.; NAKOV, P.; TIEDEMANN, J.; ZAMPIERI, M. (ed.). *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2025)*. New York: Association for Computational Linguistics, 2025. p. 29–38. DOI: <https://doi.org/10.48550/arXiv.2501.08502>.

VAN ROSSUM, G.; DRAKE JR., F. L. *Python 3 reference manual*. Scotts Valley: CreateSpace, 2009.

WASSINK, A. B.; GANSEN, K.; BARTHOLOMEW, A. Uneven success: Automatic speech recognition and ethnicity-related dialects. *Speech Communication*, v. 140, p. 50–70, 2022. DOI: <https://doi.org/10.1016/j.specom.2022.03.009>.

WIELING, M.; TIEDE, M.; TOMASCHEK, F. Quantitative identification of dialect-specific articulatory settings. *Journal of the Acoustical Society of America*, v. 142, n. 1, p. EL28–EL34, 2017. DOI: <https://doi.org/10.1121/1.4990951>.

WHITE, L.; MATTYS, S. L. Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, v. 35, n. 4, p. 501–522, 2007. DOI: <https://doi.org/10.1016/j.wocn.2007.02.003>.

WICKHAM, H.; AVERICK, M.; BRYAN, J.; CHANG, W.; D'AGOSTINO MCGOWAN, L.; FRANÇOIS, R.; GROLEMUND, G.; HAYES, A.; HENRY, L.; HESTER, J.; KUHN, M.; PEDERSEN, T. L.; MILLER, E.; BACHE, S. M.; MÜLLER, K.; OOMS, J.; ROBINSON, D.; SEIDEL, D. P.; SPINU, V.; TAKAHASHI, K.; VAUGHAN, D.; WILKE, C.; WOO, K.; YUTANI, H. Welcome to the tidyverse. *Journal of Open Source Software*, v. 4, n. 43, p. 1686, 2019. DOI: <https://doi.org/10.21105/joss.01686>. Accessed on: 14 March 2026.

YAEGER-DROR, M.; FAGYAL, Z. Analyzing prosody: Best practices for the analysis of prosody. In: DI PAOLO, M.; YAEGER-DROR, M. (ed.). *Sociophonetics: a student's guide*. London: Routledge, p. 119–130, 2011.

YARKONI, T.; WESTFALL, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, v. 12, n. 6, p. 1100–1122, 2017. DOI: <https://doi.org/10.1177/1745691617693393>.

ZHANG, C.; JEPSON, K.; LOHFINK, G.; ARVANITI, A. Comparing acoustic analyses of speech data collected remotely. *The Journal of the Acoustical Society of America*, v. 149, n. 6, p. 3910–3916, 2021. DOI: <https://doi.org/10.1121/10.0005132>.

ZHANG, Z. Speech feature selection and emotion recognition based on weighted binary cuckoo search. *Alexandria Engineering Journal*, v. 60, p. 1499–1507, 2021. DOI: <https://doi.org/10.1016/j.aej.2020.11.004>.