Sheffield Hallam University

Acoustic Analysis and Unsupervised Clustering of Brazilian Portuguese Accents

DE LIMA, Thales Aguiar <http://orcid.org/0000-0002-1043-8685> and DA COSTA ABREU, Marjory <http://orcid.org/0000-0001-7461-7570>

Available from Sheffield Hallam University Research Archive (SHURA) at:

https://shura.shu.ac.uk/37914/

This document is the Published Version [VoR]

Citation:

DE LIMA, Thales Aguiar and DA COSTA ABREU, Marjory (2026). Acoustic Analysis and Unsupervised Clustering of Brazilian Portuguese Accents. Acoustics, 8 (3): 60. [Article]

Copyright and re-use policy

See http://shura.shu.ac.uk/information.html

Sheffield Hallam University Research Archive
http://shura.shu.ac.uk

acoustics

MDPI

Article

Acoustic Analysis and Unsupervised Clustering of Brazilian Portuguese Accents

Thales Aguiar de Lima [1,†] **and Márjory Cristiany da Costa Abreu** [2,*,†]

[1] Federal Institute of Education, Science and Technology of Rio Grande do Norte, Academic Directorate, Caicó 59300-000, Brazil; thales.aguiar@ifrn.edu.br
[2] Department of Computing, Sheffield Hallam University, Sheffield S11WB, UK
* Correspondence: m.da-costa-abreu@shu.ac.uk
† These authors contributed equally to this work.

Abstract

Speech is a fundamental component of human communication and has become increasingly important with the widespread adoption of voice-based technologies, text messaging systems, Chatbots, and Large Language Models (LLMs). While AI systems are going through exponential breakthroughs, they often prioritize the dominant variant of a language. Using the TEDx Talks Brazilian Accents (TTBAcc) dataset, this research provides an acoustic analysis of two Brazilian Portuguese accents: Nordestino and Paulistano and an unsupervised accent clustering to find distinctive characteristics and investigate factors that can potentially support the monitoring of emerging and endangered accents. The acoustic analysis compares rhotics and vowels produced by speakers of both accents. The analysis of rhotics provided limited evidence of meaningful difference between accents. In contrast, the vowels exhibit significant differences in the fundamental frequency (f_0), the first and second formant frequencies (F_1 and F_2) and intensity, suggesting that these acoustic features may be more effective to distinguish between the investigated accents. For accent clustering, f_0, length, volume, and formant frequencies (F_1 and F_2) are used as inputs. Among the evaluated approaches, K-Means, with $K = 38$ achieves the best performance, obtaining a Rand Score of 0.775 ± 0.028 and 0.524 ± 0.032 homogeneity, the best performing model. The optimal number of cluster exceeds the number of classified accents, which may indicate either an over-segmentation of data or a finer-grained phonetic variation captured in acoustic features. These findings highlight the potential of using acoustic features for investigating the variation in regional accents, and provides a basis for further research in monitoring emerging and endangered accents.

Keywords: brazilian portuguese; accents; clustering; acoustics; TTBAcc; machine learning

Check for updates

Academic Editor: Georgios P. Georgiou

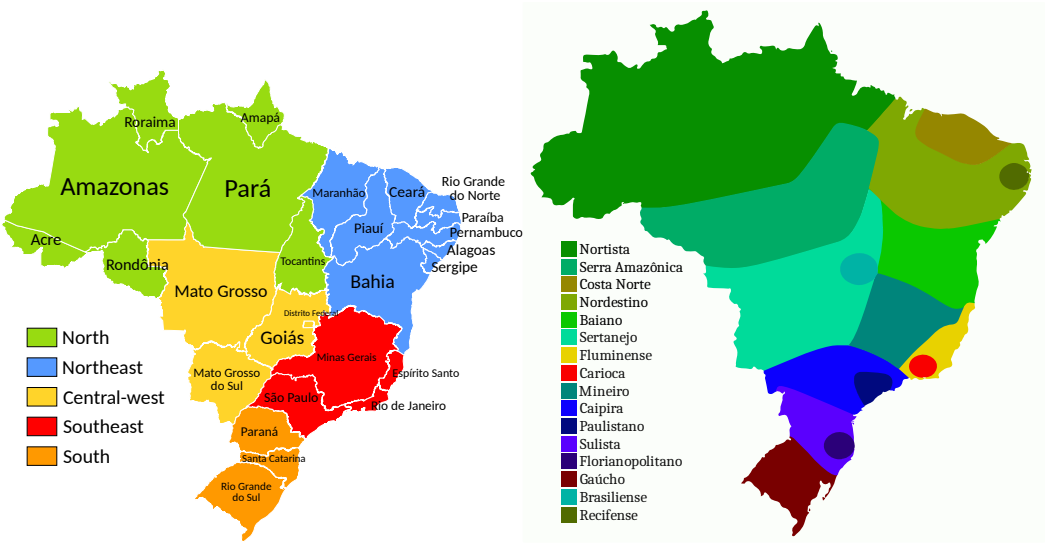
Received: 21 July 2026
Revised: 28 August 2026
Accepted: 29 August 2026
Published: 31 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

An accent can be defined as "Modo de articular os sons, palavras ou frases, peculiar de uma pessoa, de uma região etc." (i.e., way to produce the sounds, words or sentences, specific to a person, region etc.), while a dialect is a broader definition that includes pronunciation, vocabulary, and grammar [1]. The accents directly impacts on peoples communication and social interactions. They represent several cultural and social aspects of a person or population, making it a fundamental part of what defines them [2]. Thus, preserving and including such 'ways of producing' sounds in speech technologies is extremely important [3].

The importance of accents increases as they mark changes in languages, sometimes carrying the history of civilisations. For example, an Afro-Asiatic language called Coptic was found to be a modern version of Egyptian and was used to better understand missing sounds from hieroglyphics translations [4]. Similarly, the Brazilian Portuguese (PT-BR) has the influences of several cultures: Portuguese, African, Aboriginals (Indigenous people), Italian, and many others. Hence, preserving and understanding the accents, and the language itself, means to preserve the cultural aspects and the history of an ethnicity. Figure 1 presents a geopolitical map of Brazil alongside the linguistic boundaries of PT-BR dialects. The maps illustrate that political and linguistic boundaries do not necessarily coincide, although substantial overlap may occur in some regions. For example, the boundaries of the state of *Rio Grande do Sul* largely overlap with those of the *Gaúcho* accent. The linguistic boundaries shown in the figure are based on the findings of the *Atlas Linguístico do Brasil* [5], a project dedicated to documenting and characterizing the regional varieties of Brazilian Portuguese.



(**a**) (**b**)

Figure 1. Geopolitical (**a**) and dialectal (**b**) limits of Brazil.

The outstanding features of an accent are reflected in its prosodic characteristics, particularly in acoustic parameters such as fundamental frequency, intensity, duration, and spectral distribution. However, listeners distinguish accents based on how speech is perceived and heard. Accent perception is therefore associated with perceived melody, loudness, duration, and timbre, which correspond directly—or, in some cases, indirectly—to these underlying acoustic parameters.

1.1. Literature Review

The use of signal processing techniques to describe, compare and measure the sound information in voice is called acoustic analysis [6]. To describe accents, it is common base the analysis on fundamental frequency, pitch, energy, and other speech parameters in the form of statistics. An acoustic analysis has been done for other languages [7], and for PT-BR as well [8]. There are a variety of accent datasets available [9], unfortunately, the majority covers only English accents, while low resource languages are rarely represented.

The novelty on this investigation is on the diversity of the dataset. This research uses a more general approach, relying on the dataset size and variety of accents, cultures and regions. Moreover, we also present an unsupervised methodology for clustering accents.

1.2. Paper Organisation

The next section describes the dataset transformations and any parameters used to extract or compare the accents. More specifically, this study starts with a comparison between Nordestino and Paulistano rhotics in Section 2.1. Another acoustic analysis is performed for the vowels in Section 2.2. Then, Section 2.3 describe the material and methods for the unsupervised clustering. The results are presented in the same order, in Section 3. Finally, there is a brief discussion in Section 4.

2. Materials and Methods

This section describes the methods and data transformations. It first presents an acoustic analysis of the rhotics, followed by the vowels, and subsequently details the configuration adopted for the unsupervised clustering procedure.

The TEDx Talks Brazilian Accents Dataset (TTBAcc) [10] is a recently introduced dataset for Brazilian Portuguese (PT-BR). It comprises 520 PT-BR speakers, including 251 males, from different regions of Brazil. The dataset contains approximately 109.8 h of speech and covers 77.8% of the country's territory, representing 21 of the 27 Brazilian states.

The dataset was automatically collected from Portuguese-language TEDx Talks available on YouTube. Speakers from other Portuguese varieties, such as European Portuguese and Mozambican Portuguese, are not considered in this study.

Based on the speaker's birthplace and the linguistic boundaries shown in Figure 1, a sample is classified as belonging to the Nordestino (NRD) accent when the speaker was born in one of the following states: Alagoas (AL), Pernambuco (PE), Paraíba (PB), Rio Grande do Norte (RN), Ceará (CE), or Bahia (BA). Among the speakers from these states, 70 have available phonetic data. Of these, only 45 have sufficient regional information for the present analysis. Therefore, this study considers a total of 45 speakers representing 8 Brazilian Portuguese accents.

The data used for investigation covers 4 out 5 regions of Brazil, the Center-west is the only one without speakers. These 45 speakers have 23 men and represent 8 accents, including NRD and PLST. Age is well distributed across many intervals. All this information is available in Table 1.

Table 1. Speaker demographics.

Variable	Category	N
Region	North	1
	Northeast	3
	Center-West	0
	Southeast	31
	South	10
Accent	PLST	19
	FLU	9
	GAU	6
	SUL	4
	MIN	3
	BAI	2
	NRD	1
	NRT	1
Gender	Male	23
	Female	22
Age	[0, 15]	2
	[16, 20]	2
	[21, 40]	21
	[41, 60]	13
	[61, 100]	7

2.1. Acoustic Analysis: Rhotics

The rhotics are the sounds of R, i.e., /ɾ/ (Go to https://bit.ly/3Wwn9f2, accessed on 3 November 2022, and listen to a sample /ɾ/ in a word.) and /ʁ/ (Go to https://bit.ly/3Wwn9f2, accessed on 3 November 2022, and listen to a sample /ʁ/ in a word.). These sounds are said to have a high variation across PT-BR accents and are very characteristic [2,11]. Their main characteristics in PT-BR are a vanishing /ɾ/ at the end of words and a louder /ʁ/ [2]. In this study, we restrict the analysis to the length in milliseconds and the distance between formants (gaps) 3, 4 and 5 [12]. The gaps will be referred to as ΔF_{43}, which means the distance between formants 4 and 3. To estimate the formants frequency, this study uses Linear Predictive Code (LPC). Figure 2 illustrates the features in the spectrogram. All features are extracted from the interval where the sound was produced.

$$k = 2 + \frac{F_s}{1000} \tag{1}$$

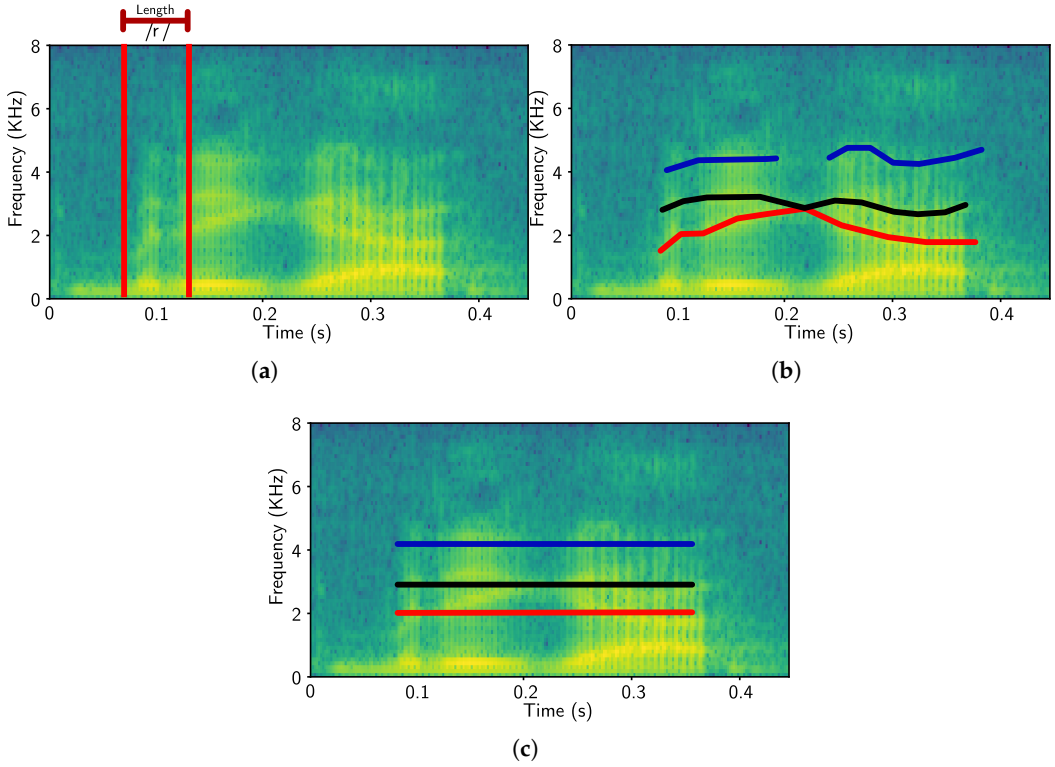


Figure 2. Length L (**a**), formants contour (**b**) and static formants frequency (**c**) for /ɾ/ in "Obrigado" (i.e., Thank you). The third, fourth and fifth formants are red, black and blue, respectively.

Figure 3 illustrates the process of extracting the length L and formants from a spectrogram. First, compute the L of the rhotic, then the signal passes through a Hamming Window of size L and is emphasised with a high-pass filter with a gain of 0.63. The spectrum is obtained with an NFFT and window length of 128 and 32 samples of stride, formants are obtained by computing the LPC residuals with positive imaginary parts, and filter order is adjusted with Equation (1), and only roots between 90 Hz and 400 Hz are considered.

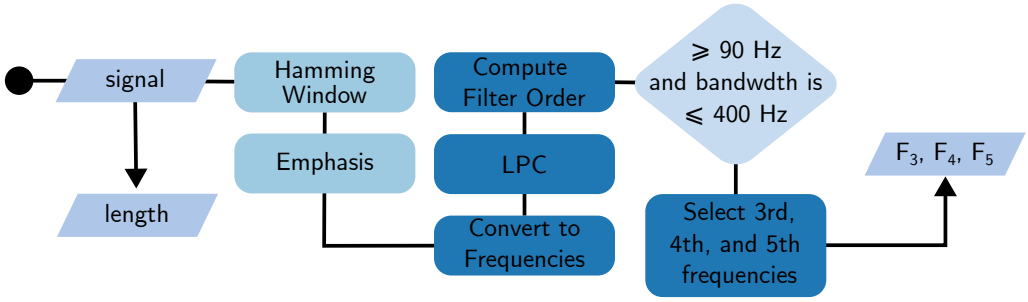


Figure 3. Rhotics feature extraction.

2.2. Acoustic Analysis: Vowels

This section examines differences in the production of oral and nasal vowels between the NRD and PLST accents. Nasal vowels are represented by a ~ over the vowel symbol (e.g., a). As with rhotics, vowel characteristics can provide useful cues for distinguishing between accents [11]. To account for differences in the number of observations across vowel categories, this study employs bootstrapping with 5000 resamples, a 95% confidence interval (CI), the median as the statistic, and the percentile resampling strategy.

Among the vowels analyzed, / *i*/ is the least represented, with 3948 samples, whereas / *o*/ is the most frequent, with 42,604 samples. The distribution of vowel occurrences by speaker gender is shown in Figure 4.

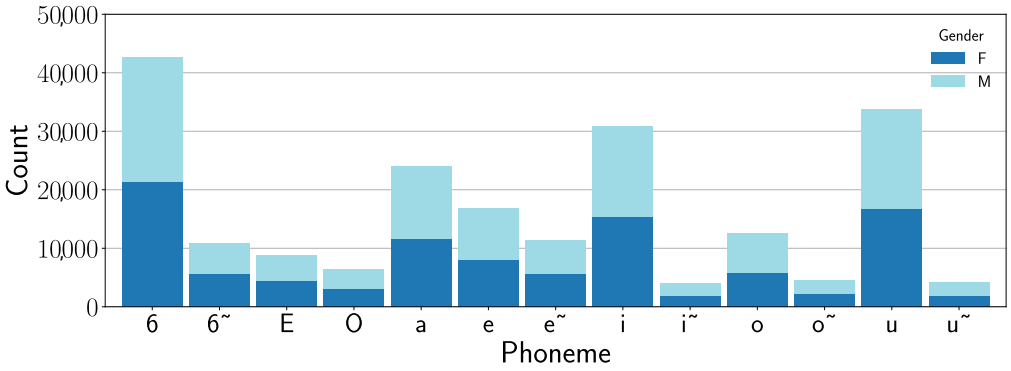


Figure 4. Vowels frequency from speakers between 20 and 35 years.

To compare accents, this study selected the volume, duration *L*, fundamental frequency, and formants [13,14]. The extraction process is depicted in Figure 5. The signal is first windowed with a hamming window of size *L*, then it is emphasised with a high-pass filter using a gain of 0.63. The intensity of the phoneme is calculated as the mean dB of its spectrogram, which is computed with a hamming window, a NFFT and window length of 128, and 32 samples of stride. The resulting spectrum is converted to dB using the maximum amplitude of each frame as reference. Finally, the formants are computed as the roots from the LPC residuals with positive imaginary parts. To adjust the filter order, this study used Equation (1) which solves to 18 since all recordings are sampled at 16 kHz. Results with frequency over 90 Hz and bandwidths below 400 Hz are selected as the formant estimation. This process is illustrated in Figure 5.

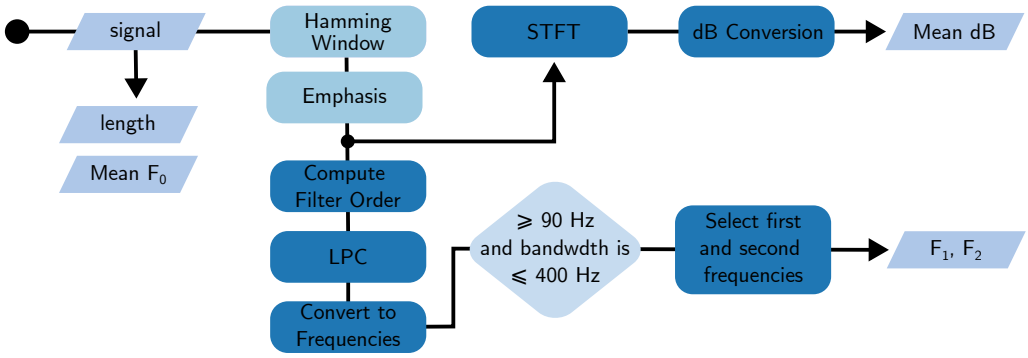


Figure 5. Acoustics extraction process for vowels.

2.3. Accent Clustering

This section takes advantage of the vowel information from the acoustic analysis to cluster the PT-BR accents. The dataset is preprocessed, and three unsupervised clustering algorithms are used: K-Means, Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), and Gaussian Mixture Model (GMM), all from the SkLearn library. The decision of the algorithms is based on their popularity for unsupervised problems.

The vowel features are the mean fundamental frequency (f_0), their duration in milliseconds, the mean volume in decibels, and the first and second formants (F_1 and F_2L). The data was computed using 1,215,659 vowel from 45 speakers from 4 regions of Brazil and 8 accents. We include every vowel in the speaker features, which makes each of them to 65 features (13 times 5). This data is summarized in Table 2.

Table 2. Accent Clustering features description. Statistics calculated using 1,215,659 vowel samples.

Feature	Mean ± Std	Min	25%	50%	75%	Max
$\overline{f_0}$ (Hz)	217.0 ± 133.5	80.6	162.0	203.8	249.8	3480.9
Length (ms)	82.2 ± 57.2	9.0	40.0	70.0	100.0	1970.0
Volume (dB)	−45.6 ± 6.2	−70.8	−49.8	−45.7	−41.4	−15.7
F_1 (Hz)	417.5 ± 153.7	90.0	308.8	386.8	496.7	3209.4
F_2 (Hz)	1107.0 ± 526.5	164.1	698.2	963.1	1435.1	4273.5

To use these features for clustering, the samples are grouped by speaker, then we calculate the mean of each feature. Figure 6 shows a data projection in two dimensions obtained with t-SNE algorithm. A few clusters can be identified, but the small number of speakers for the complexity of the problem make the data appear very noisy.

This study uses the Rand score and homogeneity to evaluate the quality of the resulting clusters. To obtain a more reliable estimate of model performance on unseen samples, 5-fold cross-validation was employed. Due to the relatively small size of the dataset, splitting the data into separate training and test sets would substantially reduce the amount of data available for model development. Cross-validation provides a more robust approximation of the model's expected performance on unseen data while allowing all samples to contribute to both training and evaluation across different folds. The choice of five folds also provides a balance between retaining a larger proportion of the data for training and obtaining a reliable estimate of model performance.

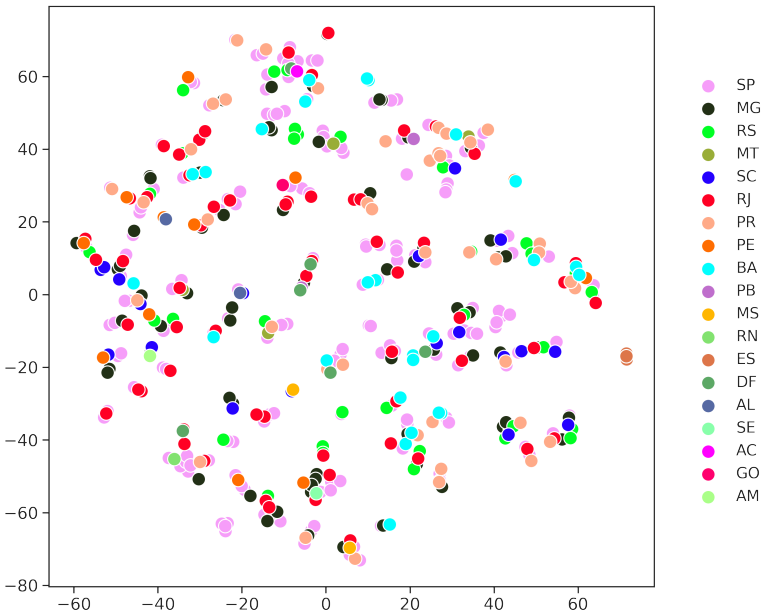


Figure 6. Dataset in two-dimensional projection from t-SNE.

2.4. AI Statement

During the preparation of this manuscript, the authors used Generative to review the grammar and spelling (at some section, not all), adjust BibTeX entries, and improve the readability of tables. After using this tool, the authors reviewed the generated content as needed.

3. Results

3.1. The Rhotics

The distributions of the accents for each rhotic are similar, and mostly skewed to the left due to high frequency of outliers in the opposite direction, as illustrated in Figure 7.

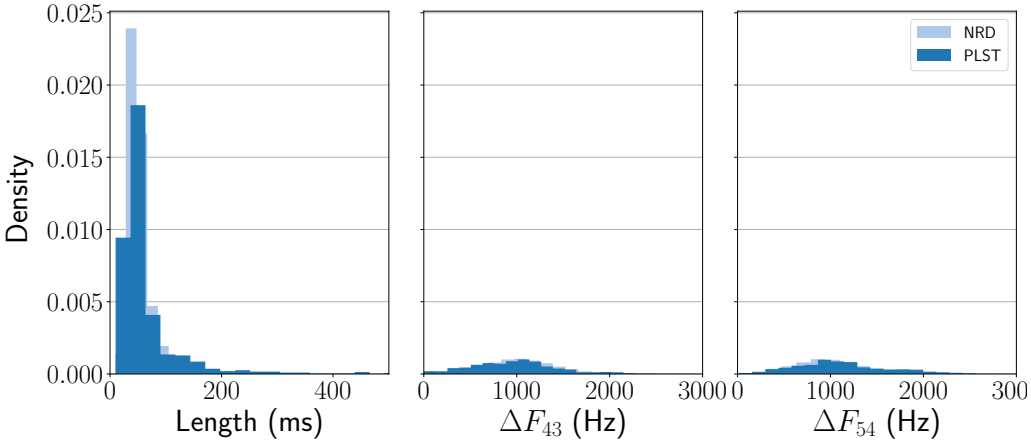


Figure 7. Features distribution for PT-BR rhotics. Length in ms with mean 64.1, $\sigma = 61.0$ for NRD, and 54.5, $\sigma = 37.7$ for PLST. ΔF_{43} in Hz with mean 984.9, $\sigma = 479.7$ for NRD, and 1029.2, $\sigma = 415.2$ for PLST. ΔF_{54} in Hz with mean 1129.5, $\sigma = 523.6$ for NRD, and 1028.0, $\sigma = 469.6$ of PLST.

To assess the statistical significance of each feature, this study employed the Kruskal–Wallis test and bootstrapping. The bootstrap analysis was performed using 5000 resamples, a 95% confidence interval (CI), the median as the statistic, and the *percentile* resampling method. The results are summarized in Table 3.

Table 3. Kruskall-Wallis *p*-value, NRD Bootstrapping CI, and PLST medians for rhotics. Signifficant differences are in bold.

	Feature	Kruskall-Wallis	Bootstrap	Median PLST
/ɾ/	Len	0.67	(40 to 40)	40
	ΔF_{43}	**<0.05**	(947.4 to 1024.4)	1022.9
	ΔF_{54}	**<0.05**	(1047.3 to 1132.5)	978.6
/ʁ/	Len	**0.05**	(70 to 90)	60.0
	ΔF_{43}	0.	(891.2 to 1116.0)	963.2
	ΔF_{54}	0.66	(947.5 to 1091.2)	975.1

Using the TTBAcc and the results shown in the Table 3, there are three main indications. First, considering /ɾ/, NRD has a significant difference for both ΔF_{43} and ΔF_{54}. The Kruskall-Wallis *p*-value and CIs indicate such difference for the ΔF_{54}. Therefore, NRD has a smaller distance between formants 5 and 4, indicating a faster hit of the tongue into the palate for this accent. Second, despite the statistically significant *p*-value observed for ΔF_{43}, the median PLST value falls within the bootstrap confidence interval, *suggesting limited evidence of a meaningful difference*. Finally, results for /ʁ/ indicates that *NRD* pronounces /ʁ/ faster than PLST.

3.2. The Vowels

This section describes notable differences between NRD and PLST vowels production. Figure 8 compares their formants, showing several shifts in both F_1 and F_2.

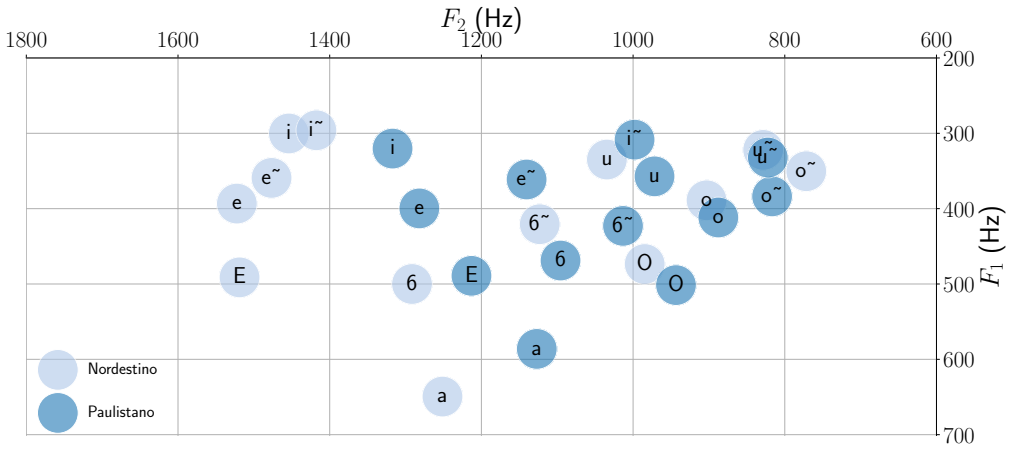


Figure 8. Vowel plane, comparing NRD to PLST.

The formant distributions appear to be similar across the two accents, as illustrated in Figure 9. To assess potential differences, a bias-corrected and accelerated bootstrap analysis (SciPy stats module.) was performed for the NRD F_1 values using 5000 resamples, a 95% confidence interval, and the median as the statistic. The resulting confidence interval ranged from 377.41 to 384.39 Hz, with a standard deviation of 1.80 Hz. The PLST median was 385.18 Hz, falling outside the NRD confidence interval, suggesting a significant difference between the accents for this measure.

The same procedure was applied to F_2 and intensity (dB), with the results summarized in Table 4. The PLST median for F_2 falls outside the NRD confidence interval, indicating a difference between the accents. In particular, PLST speakers tend to produce vowels with an F_2 approximately 300 Hz lower than that of NRD speakers. In contrast, the PLST F_1 median is close to the upper bound of the NRD confidence interval, and the observed difference does not provide sufficient evidence of a significant difference for this feature. Overall,

PLST speakers tend to produce vowels with lower intensity, with a median approximately 4.4 dB lower than that of NRD speakers.

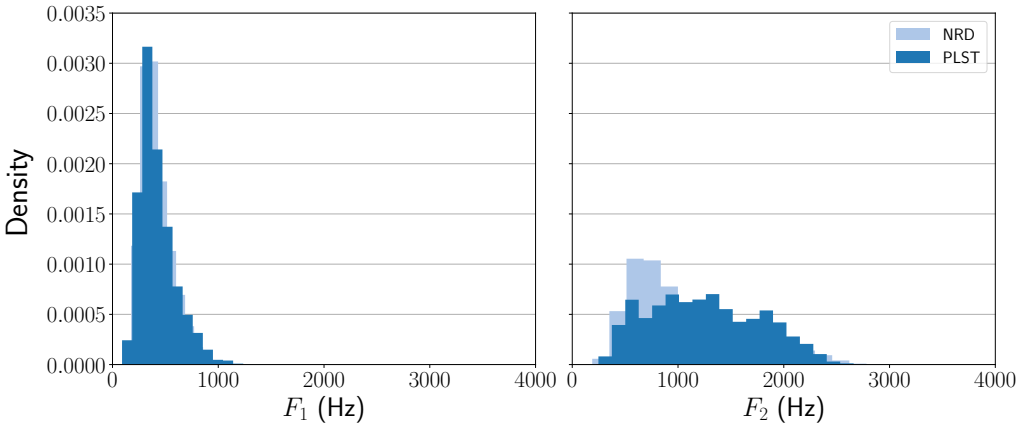


Figure 9. NRD and PLST formants distributions. NRD F_1 with mean 423.3, $\sigma = 178.9$ and PLST with mean 413.6, $\sigma = 152.1$. NRD F_2 with mean 12.2, $\sigma = 520.5$ and PLST with mean 1060.6, $\sigma = 516.1$.

Table 4. Bootstrapping results for NRD.

Feature	Median (PLST)	Confidence Interval	σ
F_1	385.18	(377.41 to 384.39)	1.80
F_2	914.47	(1205.80 to 1238.15)	8.32
dB	−47.00	(−42.60 to −42.32)	0.07

To apply the tests, each feature was bootstrapped and subsequently combined across vowels. The bootstrapping settings were preserved, except that the sampling method was replaced with basic. This was necessary due to low variance, which is worsened when isolated per vowel. That smaller variation also causes a "Degenerate Distribution" (A degenerate distribution is a probability distribution with support only at a single point.) issue with the bias-corrected and accelerated method. The confidence intervals, standard deviation, and Mdn. of PLST for each vowel is available at Tables 5 and 6.

The PLST /ẽ/ is often 10 ms (12.5%) longer than NRD. This small difference, and the other vowels medians falling within the confidence intervals, is likely due to the tool used for extracting the phonemes: the Montreal Forced Aligner. Since it was the same tool and model across all phonemes and accents, they will tend to have a similar or fixed length. In contrast, the other features are much less dependent on the tool, being more related to the speaker characteristics.

The fundamental frequency has demonstrated its ability to distinguish between speakers in several previous studies [15]. In this study, except for /ĩ/ and /õ/, all PLST f_0 medians lie outside the confidence intervals. For these vowels, PLST exhibits higher f_0 values than NRD, with increases ranging from 3.4% to 16.8% on average.

Formant frequencies are another set of features that distinguish the two accents. The PLST median falls outside the NRD confidence interval for every oral vowel except /ɔ/. Tables 5 and 6 also indicate differences in the first formant (F_1). However, no differences in the F_1 medians are observed for any of the nasal vowels.

Table 5. Vowel length, fundamental frequency (f_0), and first formant (F_1) statistics for NRD and PLST. Values are reported as bootstrap confidence intervals (CI), standard deviation (σ), and PLST median (Mdn.). Values in bold are significant differences between the two accents.

	Length			f_0			F_1		
Vowel	**CI (NRD)**	**σ (NRD)**	**Mdn. (PLST)**	**CI (NRD)**	**σ (NRD)**	**Mdn. (PLST)**	**CI (NRD)**	**σ (NRD)**	**Mdn. (PLST)**
/i/	(60.0–60.0)	1.23	60.0	(172.8–172.8)	1.28	**200.9**	(286.6–294.3)	2.01	**304.5**
/u/	(40.0–50.0)	2.92	50.0	(179.9–185.6)	1.45	**199.7**	(318.9–331.9)	3.21	**347.6**
/e/	(60.0–60.0)	0.45	60.0	(181.8–188.8)	1.96	**203.2**	(378.8–390.5)	3.11	**396.1**
/o/	(50.0–60.0)	4.76	60.0	(176.9–184.5)	1.92	**198.4**	(373.3–388.2)	3.67	**406.9**
/ɛ/	(70.0–80.0)	3.00	80.0	(169.6–182.0)	3.09	**192.9**	(493.4–513.9)	5.56	**487.8**
/ɐ/	(60.0–60.0)	0.00	60.0	(176.9–182.9)	1.60	**193.8**	(487.7–508.3)	5.42	**457.2**
/ɔ/	(70.0–90.0)	4.26	90.0	(172.4–183.0)	2.85	**189.3**	(472.3–501.1)	8.17	501.8
/a/	(84.9–90.0)	1.58	90.0	(166.1–172.8)	1.63	**188.8**	(669.5–694.2)	6.80	**592.2**
/ĩ/	(70.0–90.0)	5.27	80.0	(182.9–196.3)	3.67	198.0	(289.4–311.7)	5.58	297.5
/ẽ/	(80.0–80.0)	1.57	**90.0**	(179.9–187.2)	2.03	**195.4**	(346.8–363.2)	3.82	346.6
/õ/	(80.0–100.0)	5.82	80.0	(184.4–200.5)	4.16	202.3	(322.9–365.8)	8.34	370.2
/ũ/	(80.0–100.0)	4.52	80.0	(182.0–195.9)	3.81	**205.1**	(300.7–325.8)	7.00	315.1
/ɐ̃/	(70.0–80.0)	4.25	80.0	(187.1–196.9)	2.56	**205.9**	(420.4–438.0)	4.51	419.6

Note. CI = bootstrap confidence interval; σ = standard deviation; Mdn. = median. Oral vowels are shown above the horizontal rule and nasal vowels below.

Table 6. Second formant (F_2) and intensity statistics for NRD and PLST. Values are reported as bootstrap confidence intervals (CI), standard deviation (σ), and PLST median (Mdn.). Values in bold are significant differences between the two accents.

	F_2			Intensity (dB)		
Vowel	**CI (NRD)**	**σ (NRD)**	**Mdn. (PLST)**	**CI (NRD)**	**σ (NRD)**	**Mdn. (PLST)**
/i/	(1745.8–1826.1)	20.5	**1024.8**	(−42.5–−41.7)	0.22	−**46.1**
/u/	(1000.1–1034.4)	9.47	**878.9**	(−45.6–−44.9)	0.18	−**49.6**
/e/	(1743.5–1783.2)	10.1	**1306.8**	(−42.4–−41.7)	0.20	−**45.9**
/o/	(887.4–930.7)	11.2	**825.4**	(−47.1–−46.0)	0.29	−**50.1**
/ɛ/	(1687.0–1749.9)	15.9	**1041.2**	(−40.1–−38.2)	0.48	−**45.1**
/ɐ/	(1326.2–1360.5)	9.00	**970.6**	(−40.9–−40.2)	0.18	−**45.9**
/ɔ/	(951.4–1001.0)	13.01	**933.9**	(−44.6–−43.3)	0.38	−**48.6**
/a/	(1287.0–1315.0)	7.24	**1079.7**	(−41.2–−40.3)	0.26	−**45.0**
/ĩ/	(1012.1–1841.5)	276.16	**656.1**	(−43.0–−40.0)	0.77	−**48.7**
/ẽ/	(1807.8–1889.9)	21.4	**691.2**	(−39.8–−38.9)	0.23	−**46.0**
/õ/	(676.1–789.0)	32.6	744.7	(−47.3–−45.7)	0.36	−**51.4**
/ũ/	(729.3–824.2)	19.4	727.3	(−45.9–−44.3)	0.49	−**40.5**
/ɐ̃/	(1107.3–1112.2)	12.3	**888.8**	(−41.7–−40.7)	0.24	−**46.2**

Note. CI = bootstrap confidence interval; σ = standard deviation; Mdn. = median. Oral vowels are shown above the horizontal rule and nasal vowels below.

The second formant (F_2) exhibits a substantially larger difference, with an average median difference of 384.5 Hz, making it a more prominent distinguishing feature than fundamental frequency and F_1. Differences in F_2 are observed for all oral vowels and for three of the five nasal vowels. Similarly, the PLST vowel intensity is consistently outside the NRD confidence intervals. On average, PLST speakers produce vowels with a median intensity approximately 3.62 dB lower than that of NRD speakers.

The data presented in this and the previous sections provide evidence of differences between the NRD and PLST accents in both vowel and consonant features. The corresponding results are summarized in Table 4. When considering the combined set of features, the first formant (F_1) falls within the confidence interval when the standard deviation is taken into account, whereas the remaining features show significant differences between the accent medians.

In the following analysis, the features extracted from rhotics and vowels are combined to evaluate the effectiveness of clustering techniques in distinguishing between the NRD and PLST accents.

3.3. Accent Clustering

This section leverages the vowel information from the acoustic analysis to cluster the PT-BR accents. The hyperparameters and results are described in Tables 7–9.

Table 7. Unsupervised accent clustering results for BIRCH. Values are reported as mean ± standard deviation (σ) over the evaluated runs. Bold indicates the best configuration.

Branching	Threshold	Rand Score	Homogeneity
50	0.25	**0.588 ± 0.030**	**0.078 ± 0.013**
	0.50	0.588 ± 0.030	0.078 ± 0.013
	0.75	0.588 ± 0.030	0.078 ± 0.013
	1.50	0.588 ± 0.030	0.078 ± 0.013
150	0.25	0.588 ± 0.030	0.078 ± 0.013
	0.50	0.588 ± 0.030	0.078 ± 0.013
	0.75	0.588 ± 0.030	0.078 ± 0.013
	1.50	0.588 ± 0.030	0.078 ± 0.013
250	0.25	0.588 ± 0.030	0.078 ± 0.013
	0.50	0.588 ± 0.030	0.078 ± 0.013
	0.75	0.588 ± 0.030	0.078 ± 0.013
	1.50	0.588 ± 0.030	0.078 ± 0.013

Table 8. Unsupervised accent clustering results for Gaussian Mixture Model (GMM). Values are reported as mean ± standard deviation (σ) over the evaluated runs. Bold indicates the best configuration.

Covariance	# Components	Rand Score	Homogeneity
Diag	8	0.717 ± 0.024	0.209 ± 0.014
	14	0.747 ± 0.028	0.306 ± 0.027
	20	0.757 ± 0.028	0.371 ± 0.020
	26	0.766 ± 0.026	0.416 ± 0.040
	32	0.770 ± 0.023	0.478 ± 0.030
	38	0.771 ± 0.025	0.497 ± 0.033
Spherical	8	0.710 ± 0.028	0.212 ± 0.027
	14	0.746 ± 0.027	0.305 ± 0.036
	20	0.760 ± 0.026	0.398 ± 0.016
	26	0.771 ± 0.027	0.463 ± 0.017
	32	0.773 ± 0.027	0.488 ± 0.025
	38	**0.774 ± 0.026**	**0.520 ± 0.029**

Table 9. Unsupervised accent clustering results for K-Means. Values are reported as mean ± standard deviation (σ) over the evaluated runs. Bold indicates the best configuration.

Tolerance	# Clusters	Rand Score	Homogeneity
0.0001	8	0.712 ± 0.028	0.197 ± 0.018
	14	0.744 ± 0.030	0.306 ± 0.019
	20	0.760 ± 0.024	0.387 ± 0.032
	26	0.770 ± 0.028	0.446 ± 0.015
	32	0.773 ± 0.026	0.491 ± 0.029
	38	0.774 ± 0.028	0.520 ± 0.034
0.00001	8	0.708 ± 0.026	0.185 ± 0.022
	14	0.751 ± 0.029	0.325 ± 0.020
	20	0.760 ± 0.027	0.380 ± 0.027
	26	0.768 ± 0.026	0.441 ± 0.038
	32	0.771 ± 0.027	0.467 ± 0.028
	38	**0.775 ± 0.028**	**0.524 ± 0.032**

4. Discussion and Conclusions

The results indicate that K-Means with $K = 38$ provides the most effective unsupervised clustering configuration among the models evaluated. Although the GMM with 38 components achieves comparable performance, the difference between the two approaches is marginal. Additionally, it is important to highlight that increasing the number of clusters also increases the homogeneity score. Therefore, the choice is also based on the rand score results.

Consequently, the simpler K-Means algorithm is preferred, as it provides similar clustering performance with lower model complexity and fewer assumptions regarding the underlying data distribution. Furthermore, both K-Means and GMM attain their optimal performance with 38 clusters/components, which may indicate an over-segmentation of the data. This result suggests that the acoustic features may capture finer-grained phonetic variation than the 16 accent categories identified by the Atlas Linguístico do Brasil.

These findings indicate a higher degree of variation in the acoustic realisation of vowels and rhotics in Brazilian Portuguese (PT-BR) than might be expected from the accent descriptions provided by [5]. This discrepancy may reflect differences in the objectives and methodologies of the two approaches. While the Atlas Linguístico do Brasil focuses primarily on geographic and linguistic variation, the present study relies on acoustic characteristics extracted from spontaneous speech. Consequently, the clusters identified here may capture finer-grained phonetic variation that does not necessarily correspond directly to traditionally defined geographic accents. A more systematic exploration of the TTBAcc dataset, including a more balanced distribution of speakers and regions, could further clarify these patterns and potentially lead to more consistent clustering results.

Overall, the findings suggest that emerging and endangered accents can be investigated by combining acoustic analysis with unsupervised clustering techniques. The proposed approach provides a data-driven alternative for identifying patterns of accent variation without requiring predefined accent labels. Nevertheless, several limitations remain. In future work, the authors intend to increase the number of speakers and apply speaker anonymization techniques to reduce the possibility that speaker-specific characteristics are captured by the clustering process and subsequently interpreted as accent-related variation. This is particularly important because acoustic measures may reflect individual speaker characteristics, such as vocal tract differences, speaking style, or recording conditions, in addition to regional accent.

Although the results are promising, the methodology can also be further refined. The combination of acoustic analysis and accent clustering proved useful, but the overall process is relatively complex and involves several stages in which methodological choices may influence the final clusters. Future studies could therefore investigate a simplified and more robust version of the proposed pipeline, including a more consistent statistical analysis approach using a Linear Mixed-Effects Model, normalization procedures, and validation strategies. Such improvements could make the approach easier to reproduce and facilitate comparisons across datasets. Furthermore, incorporating evaluation metrics can contribute to a more comprehensive comparison between the methods, primarily by reducing the dependency of homogeneity score on the number of clusters.

Ultimately, this approach could contribute to the documentation and monitoring of the evolution of PT-BR and other languages. By combining acoustic measurements with unsupervised learning, it may be possible to identify subtle changes in pronunciation and detect emerging patterns of linguistic variation that are not yet represented in traditional dialectological classifications.

Author Contributions: Conceptualization, T.A.d.L. and M.C.d.C.A.; Methodology, T.A.d.L. and M.C.d.C.A.; Software, T.A.d.L.; Validation, T.A.d.L. and M.C.d.C.A.; Formal Analysis, T.A.d.L.; Investigation, T.A.d.L.; Resources, T.A.d.L. and M.C.d.C.A.; Data Curation, T.A.d.L.; Writing—Original Draft Preparation, T.A.d.L.; writing—Review and Editing, T.A.d.L. and M.C.d.C.A.; Visualization, T.A.d.L.; Supervision, M.C.d.C.A.; Project Administration, M.C.d.C.A.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Michaelis. Definição de "Sotaque". In *Dicionário Brasileiro da Língua Portuguesa*; Editora Melhoramentos: Sao Paulo, Brazil, 2022.
2. Callou, D. *Iniciação à Fonética e à Fonologia*; Jorge Zahar: Rio de Janeiro, Brazil, 1990.
3. Veliche, I.E.; Fung, P. Improving Fairness and Robustness in End-to-End Speech Recognition Through Unsupervised Clustering. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5. [CrossRef]
4. Layton, B. *Coptic in 20 Lessons*; Peeters Publishers: Leuven, Belgium, 2007; p. 210.
5. Cardoso, S.A.M.; Mota, J.A.; Aguilera, V.d.A.; Aragão, M.d.S.S.d.; Isquerdo, A.N.; Altenhofen, C.V.; Razky, A.; Margotti, F.W. *Atlas Linguístico do Brasil*; EDUEL: Delhi, India, 2014; Volume 1.
6. Hegde, S.; Shetty, S.; Rai, S.; Dodderi, T. A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders. *J. Voice* **2019**, *33*, 947.e11–947.e33. [CrossRef] [PubMed]
7. Labov, W. Contraction, Deletion, and Inherent Variability of the English Copula. *Language* **1969**, *45*, 715. [CrossRef]
8. dos Santos, P.F. Da Região da Costa Verde ao Noroeste Fluminense: A Prosódia dos Enunciados Interrogativos Totais do Rio de Janeiro. Master's Thesis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, 2016.
9. Salifu, A.; Mensah, H.N.; Tchao, E.T.; Ibrahim, A.M.; Acheampong, F.A.; Kponyo, J.J.; Agbemenu, A.S. A systematic review of accent classification techniques and datasets for inclusive speech recognition. *Int. J. Data Sci. Anal.* **2026**, *21*, 12. [CrossRef]
10. Aguiar, T.; Da Costa-Abreu, M. Automatic Collection of Transcribed Speech for Low Resources Languages. In Proceedings of the 2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS), Guayaquil, Ecuador, 4–7 July 2023; pp. 1–6. [CrossRef]
11. Ramsammy, M.; Raposo de Medeiros, B. Rhotic Variation in Brazilian Portuguese. *Languages* **2024**, *9*, 364. [CrossRef]
12. University of Manitoba. Identifying Sounds in Spectrograms. 2005. Available online: https://home.cc.umanitoba.ca/~krussll/phonetics/acoustic/spectrogram-sounds.html (accessed on 29 November 2022).
13. de Lima, M.F.B.; de Camargo, Z.A.; Ferreira, L.P.; Madureira, S. Qualidade vocal e formantes das vogais de falantes adultos da cidade de João Pessoa. *Rev. CEFAC* **2007**, *9*, 99–109. [CrossRef]

14. dos Santos, G.B. Análise Fonético-Acústica das Vogais Orais e Nasais do Português: Brasil e Portugal. Ph.D. Thesis, Universidade Federal de Goiás, Goiânia, Brazil, 2013.
15. Fatima, N.; Zheng, T.F. Short Utterance Speaker Recognition A research Agenda. In Proceedings of the International Conference on Systems and Informatics, Yantai, China, 19–20 May 2012; ICSI; pp. 1746–1750. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.