Diversity in Data for Speech Processing in Brazilian Portuguese

Giovana Meloni Craveiro[1][0009−0008−8509−219X] and Julio Cesar Galdino[1][0000−0001−6378−4648]

ICMC-USP, Brazil
giovana.meloni.craveiro@alumni.usp.br
juliogaldino@usp.br

Abstract. Striving to attend AI ethical guidelines is essential when developing and testing AI systems in order to ensure safe and trustworthy applications. However, these guidelines can be too general. The analysis presented here concerns the ethical principle of diversity, by discussing its application to the field of speech processing, using the task of prosodic segmentation of spontaneous speech as a case study. Particularly, it covers the relevance of including diversity of speaker's profiles and regional variants in data used for training and developing AI applications, in the context of Brazilian Portuguese (BP). The contributions brought by this study are: (i) a discussion of the application of the diversity principle in the context of corpora for speech applications, considering some relevant aspects and the process we formulated to select a diverse sample of speakers to compose our corpus; (ii) a literature review of the current scenario of available corpora for the task of prosodic segmentation of spontaneous speech in BP, focused on the diversity of the data; (iii) a publicly available speech corpus[1] containing 2hs32min15s of spontaneous speech audios in BP, their revised transcriptions with automatic prosodic segmentation annotation, elaborated to comprise diversity of age, gender, and accents (17 Brazilian states).

Keywords: Speech processing · Diversity · Prosodic Segmentation.

1 Introduction

Great effort has been directed towards regulation of Artificial Intelligence internationally, in order to promote safe and trustworthy AI applications. In march 2024, the EU Artificial Intelligence Act was approved by the European Parliament. In Brazil, Bill Project 2338/2023, inspired on the European AI Act, is currently under consideration and should be voted before the end of the year. The High-Level Expert Group On Artificial Intelligence, set up by the European Commission, also makes available an Assessment List for Trustworthy Artificial Intelligence (ALTAI) [2]. This list contains a series of questions to guide

[1] The corpus is publicly available in our Github repository https://github.com/nilc-nlp/MuPe-Diversidades/ under the CC BY-NC-ND 4.0 license.

developers in order to ensure their AI systems adhere to the seven required ethic principles as proposed by the EU: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being, and accountability. However, these are general guidelines that can be applied to all kinds of AI systems, and it could be challenging to effectively know how to apply such orientations when developing and testing a specific tool.

This article, guided on the referred assessment list, intends to elaborate mainly on the ethical principle of diversity inside the scope of speech processing. The idea behind it is to bring to light some important factors that one should consider when choosing data for developing and testing speech applications. To do so, along with the discussion, we present a use case of applying such considerations when creating a corpus proposed for the task of prosodic segmentation of spontaneous speech in Brazilian Portuguese (BP). As Brazil comprehends great diversity of linguistic variants, the principle of diversity, which emphasizes the relevance of developing applications that work for every individual, despite of their differences, can be widely explored. The contributions brought by this study are: (i) a discussion of the application of the diversity principle in the context of corpora for speech applications, considering some relevant aspects and the process we formulated to select a diverse sample of speakers to compose our corpus; (ii) a literature review of the current scenario of available corpora for the task of prosodic segmentation of spontaneous speech in Brazilian Portuguese (BP), focused on the diversity of the data; (iii) MuPe-Diversidades, a publicly available spontaneous speech corpus in BP, with automatic prosodic segmentation annotation. It contains 2hs32min15s and is more diverse (balanced in gender, comprehending accents of 17 Brazilian states, various education levels and ages) than those currently available with prosodic segmentation annotation. We present the discussion of some AI ethic principles and guidelines when applied to this scenario at Section 2, the task of prosodic segmentation at Section 2.1, the current scenario of corpora available today for prosodic segmentation of spontaneous speech in BP at Section 3, the corpus we generated during this study at Section 4, the process we formulated to make it as diverse as possible at Section 4.1 and its statistics at Section 4.2.

2 Background

One of the principles covered at ALTAI [2] is named “Diversity, non-discrimination and fairness” and alludes to the risk that AI systems could exacerbate prejudice unintentionally, by making use of incomplete or biased data. Additionally, it sustains that the systems should be accessible to all people, regardless of their age, gender, abilities or characteristics. One of the proposed questions at the assessment list is "Did you consider diversity and representativeness of end-users and/or subjects in the data?". In the context of speech processing, in order to guarantee that inclusion, several perspectives must be accounted for. For instance, it is important to consider the diversity of segments (consonants and vowels) and suprasegmentals, such as tone, speech rate, etc. (which can be prosodic in nature). The diversity in how these speech units can be expressed by different people justifies careful consideration when dealing with speech processing applications. Some of those units and variations are explained below.

Speech contains specific features related to the production of vowels and consonants. Vocalic segments, for example, can be classified based on their height (open, closed) and may have secondary properties, such as undergoing nasalization [11]. The proposal of [28] presents seven oral vowels: high front [i], high back [u], mid-high front [e], mid-high back [o], mid-low front [ɛ], mid-low back [ɔ], and central [a]. Pretonic vowels, that is, those occurring before the stressed syllable, represented by "e" and "o", create a distinction between speech patterns in the North/Northeast of Brazil, which typically exhibit an open production, while speech in the Central/West/Southeast/South regions tend to have a closed realization [15].

The consonantal system of Brazilian Portuguese can be identified by place and manner of articulation [13]. One of the secondary articulations of consonants that stands out in terms of linguistic variety is the process of palatalization. In this phenomenon, the tongue changes direction towards the hard palate, resulting in the sound of [s], for example, being produced as what we know as a wheezing in some regions of the country (fe[s]ta x fe[ʃ]ta; me[z]mo x me[ʒ]mo) [6]. This change occurs in a coda position, which is a final syllable constituent attached to a CV (consonant-vowel) unit [3]. Apart from [s], Brazilian Portuguese exhibits a wide variety in the production of [r] in coda, with occurrences of simple tap [ɾ], vibrant [r], retroflex [ɻ], velar fricatives [x] and glottal [h] [16].

Additionally, regardless of language, speakers commonly utter pauses at grammatical junctures. The regularity and duration of pauses vary according to speech rate (how fast one is talking), text genre, emotional state, stylistic intention, age and experience of the speaker [31]. Speech rate and fundamental frequency (F0), that is, the most common frequency or pitch that a person uses when speaking, also vary according to the characteristics of the speaker and spoken text, such as text genre, age and experience.

2.1 Prosodic Segmentation

The task of prosodic segmentation consists of relying on prosodic markers to divide an audio of speech into intonational units (IUs) [21]. There is no precise definition of IUs but they necessarily consist of a grouping of words delimited by prosodic cues, which often include a well-defined pitch contour [4]. IUs can also be split considering terminal boundaries (TB) and non-terminal boundaries (NTB), which establish concluded utterances, and breaks in non-concluded utterances, respectively [25]. Computationally, prosodic segmentation has been explored through diverse approaches, which consider specific sets of features that could be of acoustic or syntactic nature. Such features include change in intensity [14, 20], pauses [14, 20, 4], F0 [14, 20], and speech rate [14, 20, 4].

2.2 Speaker's Profiles

Each individual has a unique speaking style, which can vary in many traits. However, despite individual differences, in Brazilian Portuguese, a study that measured F0 yielded an average fundamental frequency of 105 Hz for men, 213 Hz for women, 290 Hz for children before puberty, and 440 Hz for newborns [24]. Additionally, the study by Reubold, Harrington and Kleber [22] shows a longitudinal analysis of the extent to which age affects F0 and formant frequencies. It presents the analysis of recordings of two speakers over a 50 year period. For the female speaker studied, a decrease of F0 and F1 (first formant in vowels) is observed as she gets older.

Furthermore, there are geographical and social factors that also highly impact speech style. For example, depending on the level of education, speakers may or may not adhere to the standard language norm [26], such as the syncope of proparoxytone words ("abóbora" - "abóbra") [5] and the consonant substitution ("planta" - "pranta") [9], which are more frequent in the speech of people with lower levels of education [26].

2.3 Regional Linguistic Variants

There are several characteristics that differentiate accents among different regions in Brazil. Not only each state has a unique accent, but a single city among a state could have its own specific accent as well. The Atlas Linguístico do Brasil (ALiB) [17] presents a set of speech characteristics that occur in the capitals of Brazil. ALiB demonstrates that there is preference for the realization of open pretonic vowels, as in "t[é]l[é]fone" and "c[ó]p[é]rar", in the North and Northeast regions of Brazil, while in other regions closed pretonic vowels prevail, as in "t[ê]l[ê]fone" and "c[ô]p[ê]rar". Tonic vowels are usually nasalized throughout the country, when the following segment is also nasal, but there are dialects in which pretonic vowels can be nasalized j[ã]nela or open j[a]nela [11].

Among consonantal segments, the process of palatalization of /S/ in coda position presents differences among regions. While it is alveolar at most regions, it tends to be produced with a wheezing, characteristic of the palatalization, by speakers from Rio de Janeiro, Recife and Northern states [7].

This process can also occur with the consonants /T/ and /D/, assimilating the characteristic of the following high vowel /i/, as in "tia", which can be pronounced with or without a wheezing [‘tʃia], [‘dʒia] [12]. In these consonants, three variants can exist, with higher rates of alveopalatal affricates [tʃ, dʒ] in Rio de Janeiro and lower rates in Recife, as well as other variations of the phenomenon, involving alveolar occlusives [t, d] and alveolar affricates [ts, ds] [1]. A review of this phenomenon can be verified in [29].

In Brazil, there is also a wide variety of /R/ sounds, especially at the end of syllables. ALiB shows that in the capitals, there is more realization of the glottal fricative in the North and Northeast of the country, and a higher index of velar fricative in Rio de Janeiro and Espírito Santo. The Atlas indicates that the tap is prominent in São Paulo, Curitiba, and Porto Alegre, while the retroflex has higher indices in Mato Grosso and Mato Grosso do Sul.

Furthermore, [27] shows that there are two melodic patterns when Brazilians ask a certain type of question, dividing the country into two: the North and Northeast regions follow an ascending pattern, while the Central-West, Southeast, and South regions exhibit a different pattern. There are also descriptions related to the intonation of declarative sentences, characterizing the intonational variety of Brazilian capitals.

3 Literature Review on Prosodic Segmentation

The available corpora for the task of prosodic segmentation of spontaneous speech in Brazilian Portuguese are NURC-SP [23], NURC-Recife [18], C-ORAL BRASIL I and II. While the annotation at NURC-SP and C-ORAL BRASIL I and II follow the linguistic theory presented at Raso, Teixeira and Barbosa's work [20], which considers terminal and non-terminal intonational units, the annotation at NURC-Recife relied on annotators intuitively segmenting the audios into IUs. It presented a high annotator agreement rate (Fleiss' kappa > 0.7) and each sample was revised thrice by other annotators [18].

In Raso, Teixeira, and Barbosa's work [20, 30], the dataset used was composed of a sample of monological audio extracted from C-ORAL-BRASIL I and II, then annotated in terms of prosodic segmentation considering TBs and NTBs. It totals approximately 17 minutes of spontaneous speech, about 1 minute from each of the 14 speakers. There are 4 speakers from Minas Gerais (Sete Lagoas, Rio Pomba, Rio Espera, Belo Horizonte), 2 from the city Rio de Janeiro, 1 from Pará (Belém), 1 from Santa Catarina (Florianópolis) and 2 from São Paulo (Diadema, São Paulo). The remaining 4 are of unknown origin. All speakers are male and there is no information about their age and level of education. The samples comprise three text genres: informal and formal speech in natural context and television speech, and all were classified with high acoustic quality.

NURC-SP comprises solely the linguistic variant of the capital of São Paulo, but contains around 44 hours of audios, with revised transcription and manual annotation of prosodic segmentation with TBs and NTBs. NURC-Recife comprises solely the linguistic variant of Recife and contains around 300 hours, of which some feature prepared speech. Both NURC-SP and NURC-Recife feature speakers with higher education, with age ranging from 25 to over 56 years old, comprise audios of varying acoustic quality and are equally divided between men and women [8, 10].

None of these corpora contain information about the specific speech phenomena and accent actually perceived in its speakers or information about whether they migrated from their place of birth, and only C-ORAL BRASIL is not publicly available.

Table 1 exhibits characteristics of existing corpora that contain spontaneous speech in BP with prosodic segmentation annotation, including the one we created during this study in an effort to embrace more Brazilian linguistic variants

(MuPe-Diversidades), which is described at Section 4. The speech variety of 8 states of Brazil and Distrito Federal are not comprehended by any of the corpora. Particularly, the majority of states not comprehended are from the North region of Brazil (Amazonas, Roraima, Acre, Amapá, Tocantins) while two Northeastern states (Maranhão and Rio Grande do Norte) and two Midwestern (Mato Grosso and Distrito Federal) also lack representation.

Table 1: Statistics of existing corpora of spontaneous speech in BP with prosodic segmentation annotation. In total, 17 Brazilian states are comprehended and 9 are lacking. Note: Dur: duration, Bal: balanced, unk: unknown, NE: No education, IES: Incomplete Elementary School, CES: Complete Elementary School, TE: technical education, IB: Incomplete Bachelor's degree, CB: Complete Bachelor's degree, M: Master's degree.

Corpus	**States**	**Gender**	**Age**	**Education**	**Dur**
NURC-SP	SP	Bal	25-56+	Higher Education	∼44hs
NURC-RECIFE	PE	Bal	25-56+	Higher Education	∼300hs
C-ORAL-BRASIL I,II	MG, RJ, PA, SP, SC	Male	unk	unk	∼17min
MuPe-Diversidades	**AL, BA, CE, ES, GO, MG, MS, PA, PB, PE, PI, PR, RJ, RO, RS, SE, SP**	**Bal**	**20-91**	**Varied NE, IES, CES, TE, IB, CB, M**	**∼2,5hs**

4 MuPe-Diversidades

Project TaRSila is working on several corpora of annotated audio [2], one of which consists of a collection of 289 anonymized life stories shared by São Paulo's Museu da Pessoa[3] (MuPe) in a partnership with ICMC-USP and the Federal University of Goiás (UFG) and totals 324.09 valid hours of audio automatically transcribed with WhisperX [19], an ASR model trained on large datasets that provides accurate automatic transcriptions using the large-v2 model of Whisper, and diarized via *pyannote-audio*[4], an open-source Python toolkit for speaker diarization. The automatic transcriptions were reviewed by ten students from the course of linguistics at the Federal university of Alagoas (UFAL) and at the Faculty of Philosophy, Languages and Human Sciences of the University of São Paulo (FFLCH-USP). This dataset is named CORAA MuPe and is not publicly available yet, but it will soon be released.

[2] https://sites.google.com/view/tarsila-c4ai/coraa-versions

[3] https://museudapessoa.org/

[4] https://github.com/pyannote/pyannote-audio

CORAA MuPe includes some information about each speaker, but, in many cases, some of the categories were incomplete. Education level was one of those cases: there are 39 speakers with a bachelor's degree, and from 1 to 7 speakers that varied among incomplete middle school, complete middle school, incomplete high school, complete high school, incomplete bachelor's degree, master's degree, doctorate, technical education. They were all born among 1899 and 1991. For the majority of speakers, there is information about the state and country where they were born. 17 states of Brazil, and a few countries are mentioned. Additionally, around 46.3% (133) of speakers were labeled as women and the remaining 53.7% (154) were labeled as men. Although the dataset is quite balanced in gender, there is great discrepancy in the perspective of place of origin. Graph 1 exhibits the distribution of speakers by state of birth. There are 15 speakers labeled with "-", who were born in different countries, but live in Brazil and speak Portuguese.

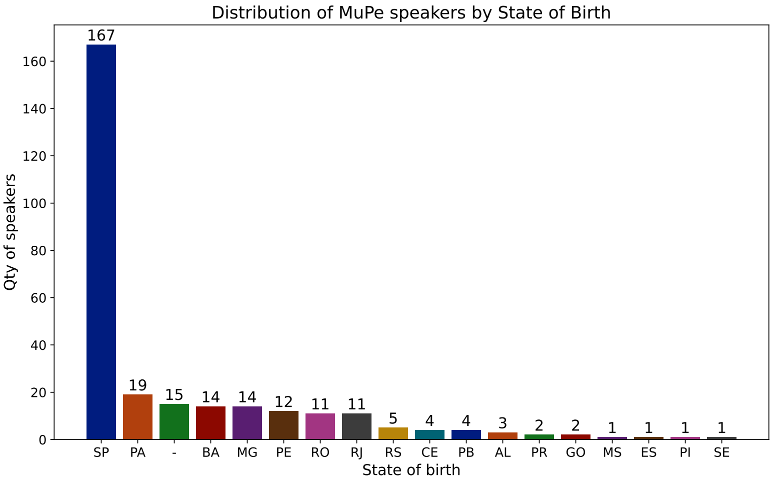


Fig. 1: Distribution of speakers from CORAA MuPe by state of birth.

Aiming at increasing diversity of Brazilian regional linguistic variants and speaker's profiles in speech corpora, particularly for the task of prosodic segmentation, we have carefully created a selection of samples from CORAA MuPe, named MuPe-Diversidades, and are making it publicly available at github [5]. Our dataset contains a sample of approximately 10 minutes of speech from each state, with their respective transcriptions with prosodic segmentation annotation. In its current state, version 0, each sample consists of the audio, its revised transcription with automatic prosodic segmentation annotation considering TBs and NTBs, conducted according to the methodology described at our previous work

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

[10]. However, during the manual review of the prosodic segmentation, we noticed that automatic cuts at every 30 seconds, which were part of the automatic transcription process carried with WhisperX, jeopardize the quality of the prosodic segmentation process, as some prosodic information is lost at every cut. Thus, a second version, named version 1.0, containing the samples of audios without cuts and their respective transcriptions with manually reviewed prosodic segmentation annotation will be available soon. Furthermore, the anonymization process consists of replacing names with their first letter, such as replacing "Ana" with "[A]" or "Zé Luís" with "[ZL]" in transcriptions, and silencing the respective parts in audio.

4.1 Methodology for creating a diverse sample, MuPe-Diversidades

Regarding the selection, it would be ideal to include a balanced number of people to be representative of each unique combination of features (gender, age, location, education level), but there are so many variables and often limited resources to do so that it may only be feasible to formulate strategies to include as much diversity as possible. An interesting strategy regarding age is to create different age groups. NURC-SP creates three (I:25-35, II:35-55, III:56+) [23], which we use at Table 2 to classify our speakers by age group. In our case, we could only afford to include 2 speakers of each state, so we opted for aiming at including the oldest and youngest speakers of those states, one of each gender, from different cities, and could not consider education as a criteria due to the lack of data. It would be ideal to choose only speakers that did not migrate, but in many cases, it was not possible. 5 minutes of audio were extracted from each selected speaker, except in cases of states with only one speaker, in which around 10 minutes were extracted. Additionally, considering that other speech datasets for BP target the linguistic variety of capitals of Brazil, such as NURC-SP and NURC-Recife, here we prioritized cities other than the capitals. For each state, whenever possible, the rules to select a speaker were as follows, in this exact order:

1. Select the speaker with the earliest year of birth, as long as their city of origin was not the capital of their state;
2. Select the speaker with the latest year of birth, as long as they are from other gender, other city and not from the capital;
3. In case the speaker migrated to another state and presented some characteristics of the accent of the state to which they moved, while did not present some characteristics of the accent of their state of origin, according to ALiB, the next youngest/oldest speaker who did not migrate, migrated at an older age, or/and did present such characteristic was preferred (this substitution occurred only for RJ1 regarding the [ʃ]);

The steps above were elaborated considering the context of our corpus, which has a very limited number of speakers for many states and contains fewer people from age groups I and II. It would also be relevant, due to differences in F0, to include teenager's and children's speech samples, but CORAA MuPe did not contain any speakers below the age of 18.

4.2 Statistics

The resulting corpus contains 14 (47%) male speakers and 16 (53%) female speakers, ranging from 20 to 91 years old, born in 17 distinct states of Brazil. Table 2 presents information about the duration of each audio sample, the gender, city, state, level of education (information which was collected through hearing the interviews), year of birth of the respective speaker, and whether they migrated or not to another state, at which approximate year, if such information was available. There is one speaker who did not remember her age, as affirmed in her interview. For her case, there is the indication of the year of the interview. For the others, the age was estimated based on the year of their interview. Metadata indicates that at least 67% of speakers migrated from their state of origin, many of them at a young age. That could mean they could have lost some of the characteristics of their original accents.

It is important to remember that one or two speakers are not enough to reliably represent all characteristics of the accent of a region, and that their accent may be influenced by other factors, such as migration. Also, each person presents individual speech characteristics, other than their accent, and having few speakers could make it harder for an ML training to generalize and identify correctly the elements of a given accent. Furthermore, we could only include 17 states of Brazil.

To present greater transparency of which phonetic phenomenon were included in our corpus, we also classified some of the phonetic occurrences explained in Section 2. They were observed in the audio samples, with the help of Praat. Table 3 presents such characteristics, including classification of the enunciation of /R/, enunciation of /S/ with or without wheezing, at the end or in the middle of words, open or closed pretonic vowels, and palatalization of /T/ and /D/. For the latter one, we considered only two categories, either palatalized or non palatalized /T/ and /D/, we make no distinction among [ts, ds] and [tʃ, dʒ]. For letter /R/, we consider tap [ɾ], retroflex[ɻ], vibrant [r] and fricative, making no distinction among glottal [h] and velar [x] fricatives. The audio samples of the speakers were fully heard and the categories attributed to them were the ones that were recognized at least twice, and appear at the table in order of apparent frequency from left to right. The production of pretonic vowels was particularly hard to classify because their realization can differ depending on the word, so speakers that realized an open pretonic vowel at least once in the audio were attributed with "open".

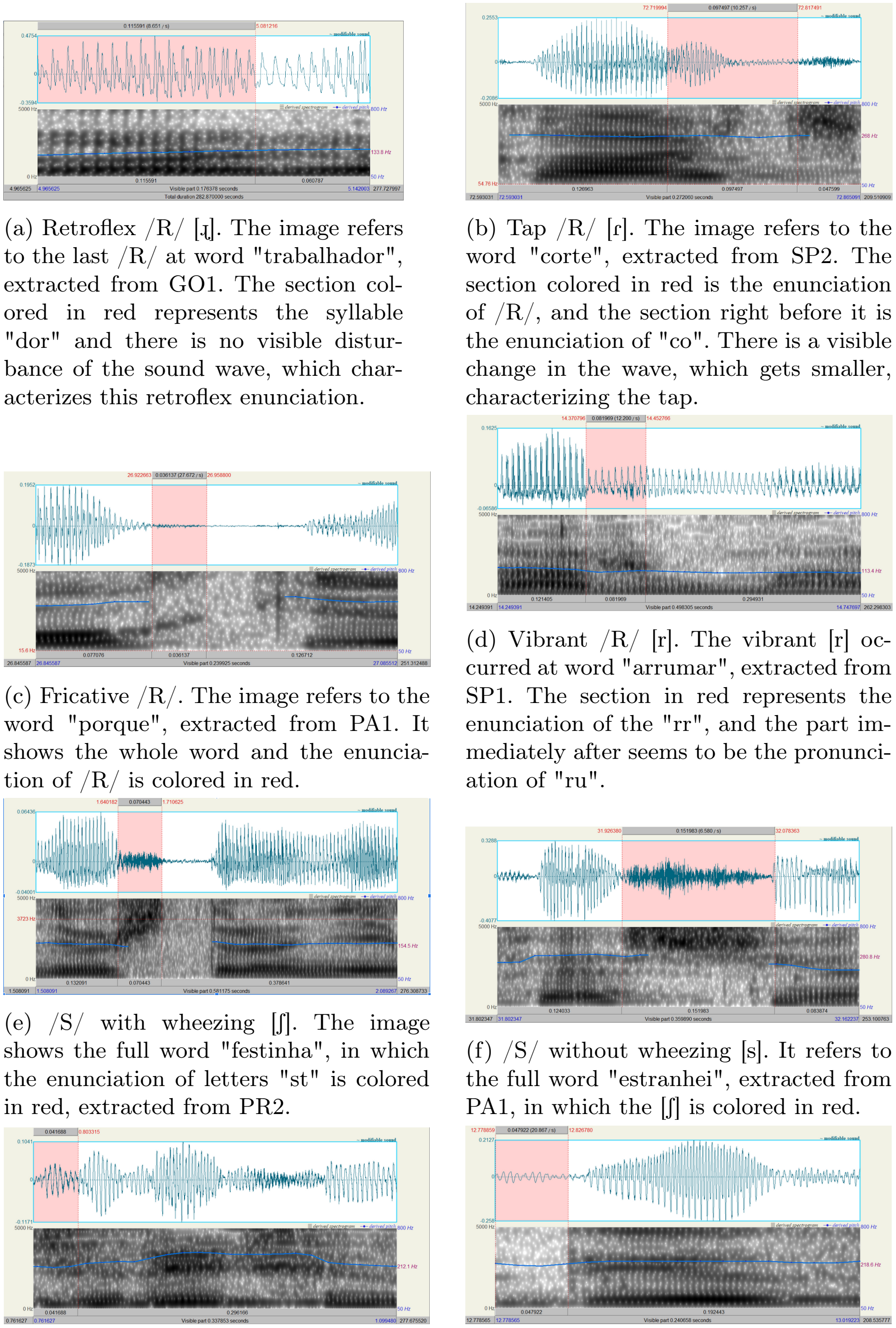
According to ALiB, open pronunciation of pretonic vowels should prevail in North and Northeastern States, but 10 speakers from those regions realized closed pronunciation, opposed to 6 from the same region who produced open pretonic vowels. No clear relation can be established among migration of such speakers, as the majority of them migrated, regardless of the pronunciation of pretonic vowels. The variation of pronunciation of the /S/ seems more well represented, as most speakers from Pará, Rondônia, Rio de Janeiro and Pernambuco realize the wheezing, as do some speakers from other states, mainly Northeastern. However, it is crucial to remember that ALiB's conclusions rely on the accent of

Table 2: Information about each speaker of MuPe-Diversidades, which is identified by an ID. The table contains information about gender (G) (M:male, F:female), year of birth (YOB), estimated age at the time of the recording, stratified age group (S) (I:25-35, II:35-55, III:56+), city and state of origin, education level (NE: No education, IES: Incomplete Elementary School, CES: Complete Elementary School, TE: technical education, IB: Incomplete Bachelor's degree, CB: Complete Bachelor's degree, M: Master's degree), approximate year and destination of migration to another state if it occurred (Y), or whether it didn't (No), and the duration of the audio sample included at the corpus (Dur.). Given the nature of this data, which was collected through the interview, some information about migration is lacking, which is marked with "unk" for "unknown". The corpus totals 2hs32min15s.

ID	**G**	**State**	**City**	**YOB/Age/S**	**Ed.**	**Migration**	**Dur.**
RS1	M	RS	Rio Grande	1952/55 yrs/II	IB	Y (RJ, SP; '72)	4m39s
RS2	F	RS	Bagé	1987/22yrs/I	IB	Y (SC, SP; '07)	4m39s
PR1	M	PR	Rolândia	1946/63yrs/III	TE	Y (SP; '59)	4m41s
PR2	F	PR	Curitiba	1969/41yrs/II	CB	Y (SP; '89)	4m45s
GO1	M	GO	Catalão	1937/73yrs/III	CES	Y (SP; '70)	4m43s
GO2	M	GO	Goiânia	1950/57yrs/III	CB	Y (SP; '73)	4m21s
MS1	F	MS	Cassilândia	1965/42yrs/II	CB	Y (SP; '71)	8m49s
RO1	F	RO	Mutum-Paraná	unk/2010/III	NE	No	4m04s
RO2	M	RO	Jaci-Paraná	1943/67yrs/III	NE	No	4m45s
PA1	F	PA	Juruti	1927/83yrs/III	IES	No	4m38s
PA2	M	PA	Castanhal	1962/48yrs/II	CES	No	4m40s
PI1	F	PI	Teresina	1965/45yrs/II	CB	Y (MA,SP; '78, '83)	4m44s
SE1	F	SE	Simão Dias	1981/27yrs/I	TE	Y (SP; unk)	4m22s
AL1	F	AL	Cacimbinhas	1940/68yrs/III	NE	Y (SP; unk)	4m26s
AL2	M	AL	Santana de Ipanema	1942/68yrs/III	IES	Y (SP; '71)	4m44s
PB1	F	PB	Serra Branca	1945/63yrs/III	NE	Y (SP; '92)	4m04s
PB2	M	PB	Imaculada	1967/40yrs/II	NE	Y (SP; '84)	4m35s
CE1	M	CE	Tauá	1971/37yrs/II	IES	Y (SP; unk)	4m47s
CE2	F	CE	Várzea Alegre	1952/56yrs/III	IES	Y (DF, SP; '72)	4m20s
BA1	F	BA	Abaíra	1925/84yrs/III	NE	Y (SP; unk)	4m49s
BA2	M	BA	Lençóis	1978/30yrs/I	IES	Y (SP; '00)	4m42s
ES1	F	ES	Vitória	1965/44yrs/II	IB	unk	9m31s
RJ1	F	RJ	Rio de Janeiro	1936/71yrs/III	IES	Y (SP; unk)	4m13s
RJ2	M	RJ	Duque de Caxias	1965/44yrs/II	TE	Y (EUA; '87)	4m37s
MG1	M	MG	Minas Novas	1943/66yrs/III	NE	No	4m22s
MG2	F	MG	Ibiá	1968/40yrs/II	CB	unk	4m27s
PE1	M	PE	Caruaru	1917/91yrs/III	TE	Y (RJ; unk)	3m37s
PE2	F	PE	Manari	1987/20yrs/I	IB	unk	3m42s
SP1	M	SP	Catanduva	1922/88yrs/III	NE	No (SPcity - unk)	4m37s
SP2	F	SP	Mogi das Cruzes	1981/27yrs/I	M	unk	4m42s

Table 3: Speech phenomena perceived in each speaker from MuPe-Diversidades, identified by their IDs. Here, [s] indicates no wheezing, and [ʃ] indicates wheezing at /S/, "mid" stands for a mid-word occurrence and "end" indicates occurrences at the end of the word. "NP" stands for "non palatal", "P" stands for "palatal", "f" stands for "fricative", "r" stands for "retroflex", "t" stands for "tap", "v" stands for "vibrant". The final rows indicate the frequency of the phenomena throughout the corpus considering how many audios of the corpus contain each pronunciation style.

ID	**/R/**	**[s] x [ʃ] end**	**[s] x [ʃ] mid**	**/T/ /D/ P x NP**	**pretonic vowels**
RS1	t	[s]	[s]	P	closed
RS2	t	[s]	[s]	P	closed
PR1	t, f, r	[s]	[s]	P	closed
PR2	t	[s]	[s]	P	closed
GO1	f, t, r	[s]	[s]	P	closed
GO2	f, r	[ʃ]/[s]	[s]	P	closed
MS1	t, r	[s]	[s]	P	closed
RO1	f	[s]	[s]	NP	closed
RO2	f	[ʃ]/[s]	[ʃ]	P	closed/open
PA1	f	[ʃ]	[ʃ]	NP/P	closed
PA2	f	[ʃ]/[s]	[ʃ]/[s]/[r]	P	closed
PI1	f, t	[s]	[ʃ]/[s]	P	closed
SE1	f	[s]	[ʃ]/[s]	NP/P	closed
AL1	r, t	[s]	[ʃ]	NP/P	closed
AL2	f, t	[s]	[s]	NP/P	closed
PB1	f, t	[s]	[ʃ]/[s]	NP	open
PB2	f, t, r	[s]	[s]/[r]	NP/P	open
CE1	f, t	[s]	[s]	P/NP	closed
CE2	f	[s]	[s]/[ʃ]/[r]	NP/P	open
BA1	f	[s]	[s]	NP/P	open
BA2	f	[s]	[s]	NP/P	open
ES1	f	[s]	[s]	P	closed
RJ1	f	[ʃ]/[s]	[ʃ]/[s]/[r]	P	closed
RJ2	f	[ʃ]	[ʃ]/[s]	P	closed
MG1	f	[s]	[s]/[ʃ]	P/NP	closed
MG2	f, t, r	[s]	[s]	P	closed
PE1	t, f	[s]/[ʃ]	[ʃ]/[s]	P/NP	closed
PE2	f	[ʃ]/[s]	[s]	NP	open
SP1	v, t	[s]	[s]	NP/P	closed
SP2	r	[s]	[s]	P	closed
%	t - 57%; f - 77%;	[s] 93%	[s] 90%	P - 90%	closed - 80%
	r - 20%; v - 3%	[ʃ] 27%	[ʃ] 12%; [r] 13%	NP - 50%	open - 20%



(a) Retroflex /R/ [ɻ]. The image refers to the last /R/ at word "trabalhador", extracted from GO1. The section colored in red represents the syllable "dor" and there is no visible disturbance of the sound wave, which characterizes this retroflex enunciation.

(b) Tap /R/ [ɾ]. The image refers to the word "corte", extracted from SP2. The section colored in red is the enunciation of /R/, and the section right before it is the enunciation of "co". There is a visible change in the wave, which gets smaller, characterizing the tap.

(c) Fricative /R/. The image refers to the word "porque", extracted from PA1. It shows the whole word and the enunciation of /R/ is colored in red.

(d) Vibrant /R/ [r]. The vibrant [r] occurred at word "arrumar", extracted from SP1. The section in red represents the enunciation of the "rr", and the part immediately after seems to be the pronunciation of "ru".

(e) /S/ with wheezing [ʃ]. The image shows the full word "festinha", in which the enunciation of letters "st" is colored in red, extracted from PR2.

(f) /S/ without wheezing [s]. It refers to the full word "estranhei", extracted from PA1, in which the [ʃ] is colored in red.

(g) /D/ with palatalization [d] at word "direitinho", extracted from RS2.

(h) /D/ without palatalization at word "dias", extracted from PE2.

Fig. 2: Screenshots of productions of specific phonemes in samples extracted from MuPe-Diversidades, analyzed at Praat.

the capitals, while this corpus features mainly speakers from other cities of the state, which could have distinct accent characteristics.

Concerning the /R/, while, in our corpus, tap occurs in all south states and in São Paulo, it also occurs in many other states, which could be explained by the migration of many speakers to São Paulo. Many speakers across states pronounce a fricative /R/, composing the most common pronunciation of the /R/ in the corpus, occurring in 77% of the audios.

At figure 2, we include some screenshots of audio samples of certain segments observed with Praat to exemplify what we considered to be members of each category. They contain the respective oscillograms and spectrograms of the sound. As for the difference between open and closed pretonic vowels, we extracted the values of formants 1 and 2 of two samples to exemplify. The sample of the open pretonic vowel of the word "português" from PE2 showed F1 702.52Hz and F2 1246.58hZ at 10.57s, while the sample of the closed pretonic vowel of the word "coordenador" from RS2 showed F1 467.72Hz and F2 1067.18Hz at 19.22s.

5 Conclusions and Future Work

As discussed in Sections 2.2 and 2.3, in the context of speech processing, it is important to consider covering different speaker's profiles, including level of education, age, gender, place of origin and accent, in order to attend the AI ethical principle of diversity, and thus, promote technology that is equally efficient to all kinds of speakers.

Moreover, the transparency principle at ALTAI [2] includes the question "Did you communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/or error rates?". Thus, it is also sensible to consider testing the applications to different linguistic variants and speaker's profiles, and analyzing its performance for each case. As it is reasonable to argue that reporting those characteristics about speech corpora is relevant for ensuring transparency and allowing developers to take such limitations into consideration.

To what concerns Mupe-Diversidades, it is far from being complete and representative of all kinds of speaker's profiles and regional variants of BP. It lacks material from people under 18, which are known to present different ranges of F0, speakers of 9 Brazilian states, speakers with different combinations of age, gender, place of origin and education level, as well as many other BP speakers who may utter phenomena that are not covered at this small sample. It also comprehends only one text genre: interviews about life-stories. However, the corpus also represents the inclusion of 13 regional variants of Brazilian states that were not yet covered and the inclusion of 2hs32min15s of audio with prosodic segmentation annotation to the collection of corpora available today for prosodic segmentation of spontaneous speech in BP. We hope this work encourages researchers to reflect on these issues, contribute to the creation of additional resources for BP, and prioritize the inclusion of greater diversity in future research.

Acknowledgments. This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. This project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law No. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published Residence in TIC 13, DOU 01245.010222/2022-44. Special thanks to Caroline Alves, responsible for the manual review of the prosodic segmentation annotation. Thanks are also addressed to the Academic Excellence Program (PROEX) of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES).

References

1. Abaurre, M.B.M., Pagotto, E.G.: Palatalização das oclusivas dentais no português do brasil. Gramática do português falado **8**, 557–601 (2002)
2. Ala-Pietilä, P., Bonnet, Y., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Bauer, W., Bouarfa, L., Chatila, R., Coeckelbergh, M., Dignum, V., et al.: The assessment list for trustworthy artificial intelligence (ALTAI). European Commission (2020)
3. Athayde, M.d.L., Baesso, J.S., Dias, R.F., Giacchini, V., Mezzomo, C.L.: O papel das variáveis extralinguísticas idade e sexo no desenvolvimento da coda silábica. Revista da Sociedade Brasileira de Fonoaudiologia **14**, 293–299 (2009)
4. Biron, T., Baum, D., Freche, D., Matalon, N., Ehrmann, N., Weinreb, E., Biron, D., Moses, E.: Automatic detection of prosodic boundaries in spontaneous speech. PLoS ONE **16**(5), 1–21 (May 2021). https://doi.org/10.1371/journal.pone.0250969
5. Bisol, L., Brescancini, C.: Fonologia e variação: recortes do português brasileiro. Edipucrs (2002)
6. Brescancini, C.R.: A palatalização em coda em florianópolis-sc: variáveis sociais. Working Papers em Linguística (Online) (2015)
7. Callou, D., Brandão, S.: O processo de palatalização no português do brasil. Linguística **18**, 57–73 (2006)
8. Candido Junior, A., Casanova, E., Soares, A., de Oliveira, F.S., Oliveira, L., Junior, R.C.F., da Silva, D.P.P., Fayet, F.G., Carlotto, B.B., Gris, L.R.S., et al.: Coraa asr: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese. Language Resources and Evaluation **57**(3), 1139–1171 (2023)
9. Costa, L.T.d.: Estudo do rotacismo: variação entre consoantes líquidas. Master's thesis, Universidade Federal do Rio Grande do Sul (2006)
10. Craveiro, G.M., Santos, V.G.d., Dalalana, G.J.P., Svartman, F.R.F., Aluísio, S.M.: Simple and fast automatic prosodic segmentation of brazilian portuguese spontaneous speech. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese (2024)
11. Cristófaro Silva, T.: Fonética e fonologia do português: roteiro de estudos e guia/exercícios. são paulo. Contexto (2015)
12. Freitag, R.M.K., Souza, G.G.A.: O caráter gradiente vs. discreto na palatalização de oclusivas em sergipe. Tabuleiro de Letras **10**(2), 78–89 (2016)
13. Kent, R.D., Read, C.: Análise acústica da fala. Cortez Editora (2015)
14. Kocharov, D., Kachkovskaia, T., Skrelin, P.: Eliciting meaningful units from speech. In: Proc. Interspeech 2017. pp. 2128–2132 (2017)
15. Leite, Y.F.: Como falam os brasileiros. Editora Schwarcz-Companhia das Letras (2002)
16. Lima, M.M.d.O.: As consoantes róticas no português brasileiro com notas sobre as róticas das variedades de Goiânia, Goiatuba e Uberlândia. Master's thesis, Universidade de Brasília (2013)

17. Mota, J.A., Ribeiro, S.S.C., de Oliveira, J.M.: Atlas Linguístico Do Brasil: Comentários às Cartas Linguísticas 1-V. 3. Universidade Estadual de Londrina. Editora (2023)
18. Oliveira, M.J.: Nurc digital: um protocolo para a digitalização, anotação, arquivamento e disseminação do material do projeto da norma urbana linguística culta (nurc). CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos **3(2)**, 149–174 (2016)
19. Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of the 40th International Conference on Machine Learning, vol. 202, pp. 28492–28518. PMLR (Jul 2023), https://proceedings.mlr.press/v202/radford23a.html
20. Raso, T., Teixeira, B., Barbosa, P.: Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech. Journal of Speech Sciences **9**, 105–128 (September 2020)
21. Reed, B.S.: Analysing conversation: An introduction to prosody. Bloomsbury Publishing (2017)
22. Reubold, U., Harrington, J., Kleber, F.: Vocal aging effects on f0 and the first formant: A longitudinal analysis in adult speakers. Speech communication **52**(7-8), 638–651 (2010)
23. Rodrigues, A.C., Macedo, A.A., Candido Jr, A., Svartman, F.R., Craveiro, G.M., Leite, M.Q., Aluísio, S., Santos, V.G., Garcia, V.M.: Portal nurc-sp: Design, development, and speech processing corpora resources to support the public dissemination of portuguese spoken language. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese. pp. 187–195 (2024)
24. Russo, I., Behlau, M.: Percepção da fala: análise acústica do português brasileiro. Lovise, São Paulo (1993)
25. Santos, V.G., Alves, C.A., Carlotto, B.B., Dias, B.A.P., Gris, L.R.S., de Lima Izaias, R., de Morais, M.L.A., de Oliveira, P.M., Sicoli, R., Fernandes-Svartman, F.R., Leite, M.Q., Aluísio, S.M.: CORAA NURC-SP Minimal Corpus: a manually annotated corpus of Brazilian Portuguese spontaneous speech. In: Proc. IberSPEECH 2022. pp. 161–165 (2022). https://doi.org/10.21437/IberSPEECH.2022-33
26. Schwindt, L.C.d.S., Quadros, E.S.d., Toledo, E.E., Gonzalez, C.A.: A influência da variável escolaridade em fenômenos fonológicos variáveis: efeitos retroalimentadores da escrita. Revista virtual de estudos da linguagem-ReVEL. Novo Hamburgo, RS. Vol. 5, n. 9 (ago. 2007), 12 f. (2007)
27. da Silva, J.C.B.: A prosódia regional em enunciados interrogativos espontâneos do português do brasil. Revista Gatilho **13** (2011)
28. Silva, T.C.: O método das vogais cardeais e as vogais do português brasileiro. Revista de Estudos da Linguagem **8**(2), 127–153 (1999)
29. Silva, T.C., Barboza, C., Guimarães, D., Nascimento, K.: Revisitando a palatalização no português brasileiro. Revista de Estudos da Linguagem **20**(2), 59–89 (2012)
30. Teixeira, B., Raso, T., Barbosa, P.A.: Detecção automática de fronteiras prosódicas entre unidades entonacionais. Gradus: Revista Brasileira de Fonologia de Laboratório (2020)
31. Vaissière, J.: Language-independent prosodic features. In: Cutler, A., Ladd, D.R. (eds.) Prosody: Models and Measurements, pp. 53–66. Springer Berlin Heidelberg, Berlin, Heidelberg (1983)