Simple and Fast Automatic Prosodic Segmentation of Brazilian Portuguese Spontaneous Speech

Giovana M. Craveiro
ICMC-USP, Brazil
giovana.meloni.craveiro@alumni.usp.br

Vinícius G. Santos
FFLCH-USP, Brazil
vinicius.santos@alumni.usp.br

Gabriel J. P. Dalalana
EESC-USP, Brazil
gabriel.jp.dalalana@usp.br

Flaviane R. F. Svartman
FFLCH-USP, Brazil
flavianesvartman@usp.br

Sandra M. Aluísio
ICMC-USP, Brazil
sandra@icmc.usp.br

Abstract

Detecting prosodic boundaries is a frequently studied task as it has a direct impact on automatic speech recognizers and synthesizers. For Brazilian Portuguese, this task has been mainly studied for the linguistic variety of Minas Gerais via supervised machine learning methods. As manually annotating a large corpus with prosodic boundaries is a costly task, this paper brings three main contributions: (1) a publicly available corpus, prosodically annotated automatically and manually revised; (2) the code of the heuristic method of Biron et al. (2021), that uses discontinuities in speech rates and silence pauses, adapted to segment Brazilian Portuguese spontaneous speech; and (3) the evaluation of the method in the scope of NURC-SP corpus, linguistic variety of São Paulo, which suggests that: (i) the method is more suitable for defining non-terminal boundaries than for defining terminal boundaries[1]; (ii) the method performs best by using all heuristics conjointly, but the silences' heuristics stands out; and (iii) there are no significant differences in performance among different speech genres (conversational or talks) but further analysis should be carried out. The pipeline created was intended to accelerate the manual revision of prosodic boundaries, and therefore, a simple and fast method was chosen as it does not require a training phase.

1 Introduction

Information in spoken language is transmitted through words associated with several non-segmental features (prosodic cues), such as pitch, volume, speech rate, rhythm, and timbre. Those speech chains bounded by prosodic cues can communicate coherent messages with a variety of linguistic functions that are expressed by different types of utterances (imperative, interrogative, assertive, or exclamatory). These prosodic groups are often called intonational phrases or *intonation units* (IUs) and although they are hard to define, one of their features is a well-defined ("single") pitch contour (Biron et al., 2021).

Detecting prosodic boundaries in natural languages is a frequently studied task in the speech processing literature (Wightman and Ostendorf, 1991; Ananthakrishnan and Narayanan, 2008; Huang et al., 2008; Jeon and Liu, 2009; Kocharov et al., 2017; Biron et al., 2021). This task remains an open problem due to multiple sources of variation in speech, including: speaker characteristics, such as age, gender, dialect variety; the recording environment, e.g., microphone used, room acoustics and noises; and production style, i.e., spontaneous vs. read speech, which are instances of the continuum unplanned-planned production style. This task has a direct impact on automatic speech recognizers (ASR) and speech synthesizers (TTS). For ASR, if the excerpt of speech used to train a model is based on IUs, the error rates for syllable, character, and word recognition are reduced (see Chen and Hasegawa-Johnson, 2004; Lin et al., 2019) and in the case of TTS, the adequate use of pause duration (for example), that are naturally used by human speakers, improves speech intelligibility, helping to capture the meaning of an excerpt of speech (Liu et al., 2022). It is expected that an effective automatic identification of prosodic boundaries will (i) facilitate linguistic studies on spontaneous speech, (ii) help to create more useful datasets to train ASR models and (iii) extend the power of speech-related applications working on spontaneous speech.

Automatic prosodic boundary recognition methods range from rule-based or heuristic systems (see, e.g., Biron et al., 2021) to supervised machine learning models using lexical and syntactic features that are combined with acoustic features (e.g. Kocharov

[1] Terminal boundaries mark the conclusion of the utterance. Non-terminal boundaries mark breaks of non-conclusive sequences of the utterance.

et al., 2017), generally applied to scripted speech, in which syntactic and prosodic conventions coincide, as disfluencies in this type of speech are rare. More recently, Roll et al. (2023) fine-tuned Whisper (Radford et al., 2023), a pretrained end-to-end ASR model, to segment spontaneous speech into IUs with great performance.

For American English, there are two resources frequently used in applications that consider prosodic boundaries: *Santa Barbara Corpus of Spoken American English* (SBC) (du Bois et al., 2000–2005) and the *Boston University Radio Speech Corpus* (BURSC) (Ostendorf et al., 1995). The first contains ≈20 hours of spontaneous speech of varying genres, transcribed and manually segmented into final and non-final IUs (du Bois et al., 1992), following the identification of a boundary. The second contains 10 hours of radio news, of which 3.5 hours are prosodically annotated according to the ToBI system (Beckman et al., 2005). For British English, the IViE Corpus[2] (Grabe et al., 2001) is a resource focusing on nine urban dialects of English spoken in the British Isles and is transcribed with an intonational phrase methodology — the IViE labeling system — adapted from the ToBI framework. It contains 36 hours of speech data and the speakers are male and female adolescents.

For Brazilian Portuguese (BP), the automatic detection of prosodic boundaries was explored within the scope of the C-ORAL-Brasil project[3], advancing studies in spontaneous speech by using phonetic-acoustic parameters and boundaries identified perceptually by trained annotators (Teixeira et al., 2018; Teixeira and Mittman, 2018; Raso et al., 2020). The studies use excerpts of male informal monological spontaneous speech (8 min 39 s of audio), from the annotated corpora *C-ORAL-Brasil I* and media and formal speech in natural context (8 min 29 s of audio), from *C-ORAL-Brasil II*, mainly of linguistic varieties of the Minas Gerais state (Raso and Mello, 2012; Mello et al., To appear).

The study reported in this paper was set out to accomplish three research objectives:

1. make publicly available the implementation of a simple rule-based method with three heuristics related to discontinuities in speech rate (DSRs) and silent pauses, which are prosodic acoustic cues marking prosodic boundaries, already evaluated for the English language (Biron et al., 2021). This method was adapted for BP using a forced aligner based on ASR, named UFPAlign (Batista et al., 2022). The code is available at `https://github.com/nilc-nlp/ProsSegue`;

2. evaluate the method in excerpts of the NURC-SP corpus, with ≈334 hours of transcribed speech, of which 19 hours were prosodically annotated in two types of IU boundaries (terminal and non-terminal), henceforth TB and NTB (Santos et al., 2022); different than Biron et al. (2021) that evaluates only IU terminal boundaries without specifying them; and

3. make publicly available a subcorpus of NURC-SP corpus, prosodically annotated with the method described in this paper and manually revised. The subcorpus is available at `http://tarsila.icmc.usp.br:8080/nurc/catna`.

NURC-SP (NURC-São Paulo)[4] recordings feature speakers with higher education; born and raised in the city; children of native Portuguese speakers; equally divided into men and women; and distributed into three age groups (25–35, 36–55, and 56 years onwards). The recordings were made in three situations, generating different discursive genres: lectures/classes in a formal context given by a speaker (EF); dialogues between documenters and a participant (DID); and dialogues between two participants mediated by documenters (D2). The version of NURC-SP used in this research is made up of 375 inquiries, some of which already had transcriptions — but, until then, not aligned with the audio — and the vast majority is composed of audio only. NURC-SP was divided into three work subcorpora: (i) the *Minimum Corpus* (MC) (21 recordings + transcriptions) used to evaluate automatic processing methods of the entire collection (Santos et al., 2022); (ii) the *Corpus of Non-Aligned Audios and Transcriptions* (CATNA) (26 recordings + transcriptions), which is the focus of this paper, as we are making this subcorpora publicly available; and (iii) *Audio Corpus* (328 recordings without transcription), which has been automatically transcribed by WhisperX (Bain et al., 2023) that provides fast automatic speech recognition (70x realtime with the large-v2 model

[2] `www.phon.ox.ac.uk/files/apps/old_IViE/`

[3] `www.c-oral-brasil.org/`

[4] `https://nurc.fflch.usp.br/`

of Whisper (Radford et al., 2023)) and speaker-aware transcripts, using the speaker diarization tool pyannote-audio[5].

2 The Heuristic-based Method to Detect Prosodic Boundaries

According to Biron et al. (2021), the lengthening of speech rate at the end of a unit together with the acceleration at its beginning, called discontinuities in speech rate (DSRs), is a prominent signal for identifying boundaries. Using two acoustic cues related to timing, DSRs and silent pauses, they proposed a heuristic method, using the output of an ASR system, to identify boundaries in spontaneous speech in English. The first heuristic made use of a threshold set to 88% of the largest difference in speech rate values of a single turn. Differences among consecutive speech rate measurements that were higher than this threshold were tagged as boundaries; the second heuristic set the threshold to 70% and was applied only to the resulting stretches that were longer than 3 seconds and contained more than 10 words; finally the third heuristic used silent pause durations longer than 0.3 seconds as a cue to indicate a boundary. To measure the speech rate values, an average of all non-silent phonemes inside a time window of 0.3 seconds is estimated for each word, starting at their beginning.

Biron et al. (2021) uses the Kaldi-based software Montreal Forced Aligner (MFA) Version 0.9.0 (McAuliffe et al., 2017) in order to obtain the timestamps of the beginning and ending of each phone present in the transcription. However, we opted for the Brazilian forced aligner UFPAlign (Batista et al., 2022), as it is also Kaldi-based and specifically adapted to Brazilian Portuguese. It is important to note that inquiries of NURC-SP vary, generally, from thirty minutes to one hour and thirty minutes (see Table 1), therefore, the original versions were split into files of ten minutes, along with their corresponding transcriptions, to be used in the forced aligner UFPAlign and merged back at the beginning of the prosodic segmentation method.

For our initial results, presented in this paper, we maintained the values of the six parameters used in Biron et al. (2021):

1. time window (window_size) used to measure the discontinuity rate: 300 ms (average word duration in English);
2. pause duration (silence_threshold) to determine a prosodic boundary: 300ms;
3. threshold (delta1) that determines the largest difference in speech rate values for the first heuristic: 88% ;
4. threshold (delta2) that determines the largest difference in speech rate values for the second heuristic: 70% ;
5. minimum number of words (interval_size) to determine any stretch between consecutive DSRs as eligible: 3;
6. minimum duration (min_words_h2) to determine any stretch between consecutive DSRs as eligible: 10 seconds.

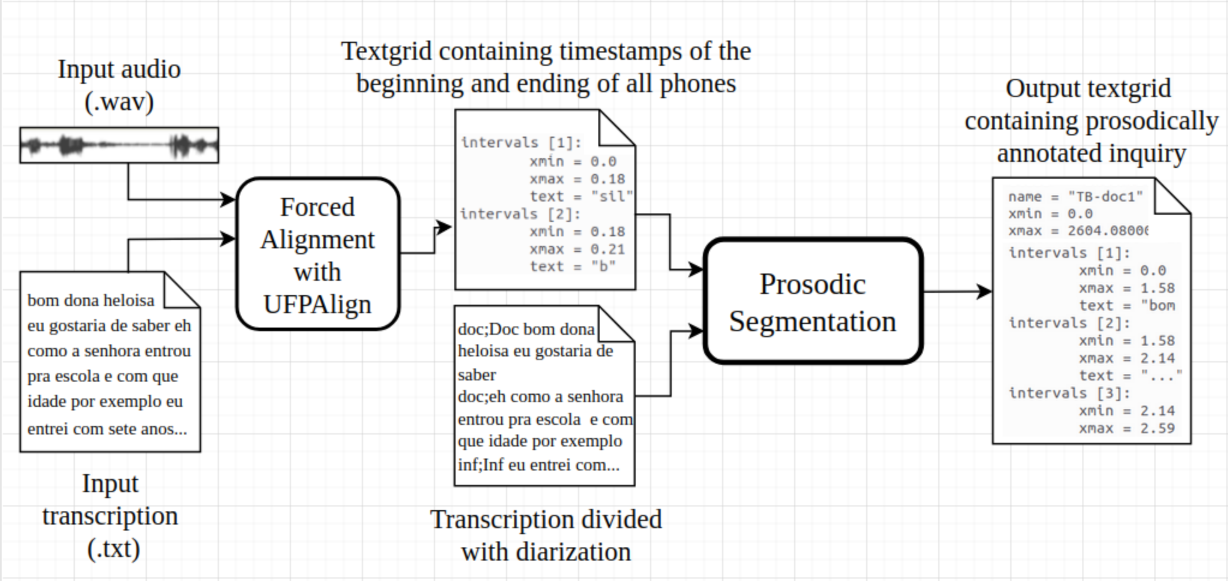
The final output is a Textgrid document composed of two layers for each speaker, one for terminal boundaries and one for non-terminal boundaries, each containing their speech divided by the identified boundaries (further details in Section 3.1). As the method is not yet adapted to estimate these two types of boundaries differently, these layers are identical for the same speaker. To evaluate our results (further details in Section 3), we experimented a hit threshold varying among 0.01, 0.1, 0.2, and the chosen value of 0.25 seconds, as its f1-score was better and was still beneath 0.3 seconds, our threshold for defining a silence boundary. Our complete pipeline can be seen in Figure 1.

3 Experiments and Results

3.1 Dataset

Six inquiries were selected from the NURC-SP MC, two from each discourse genre, to carry out an acoustic analysis in order to select the study corpus of the segmentation method (see Table 1). Five inquiries were classified as good/clear audio quality and one inquiry as low audio. Figure 2 shows the multilevel transcription of NURC-SP MC using interval layers annotated in the speech analysis program Praat (Boersma and Weenink, 2023): (i) 2 layers (TB-, NTB-) in which the speech of each main speaker (-L1, -L2) and documenter (-Doc1, -Doc2) is segmented into prosodic units and transcribed according to standards adapted from the NURC project; (ii) 1 layer (LA) for transcribed and segmented speech from any random speaker; (iii) 1 layer for comments regarding the audio recording; (iv) 1 layer containing the normalized (-normal)

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


Input audio (.wav)
Textgrid containing timestamps of the beginning and ending of all phones
Output textgrid containing prosodically annotated inquiry
Forced Alignment with UFPAlign
Prosodic Segmentation
Input transcription (.txt)
Transcription divided with diarization

```
intervals [1]:
        xmin = 0.0
        xmax = 0.18
        text = "sil"
intervals [2]:
        xmin = 0.18
        xmax = 0.21
        text = "b"
```

bom dona heloisa
eu gostaria de saber eh
como a senhora entrou
pra escola e com que
idade por exemplo eu
entrei com sete anos...

doc;Doc bom dona
heloisa eu gostaria de
saber
doc;eh como a senhora
entrou pra escola e com
que idade por exemplo
inf;Inf eu entrei com...

```
name = "TB-doc1"
xmin = 0.0
xmax = 2604.0800(
intervals [1]:
        xmin = 0.0
        xmax = 1.58
        text = "bom
intervals [2]:
        xmin = 1.58
        xmax = 2.14
        text = "..."
intervals [3]:
        xmin = 2.14
        xmax = 2.59
```

Figure 1: An audio file (.wav) and its transcription are fed to the forced aligner, which outputs a .TextGrid document. Then, the resulting document, along with a .txt document that contains each sentence of the inquiry and its respective speaker ("speaker diarization"), is used as input to the method. The output of the pipeline is a textgrid with the prosodically segmented content of the inquiry.

version of the transcript of all TB and LA layers; and (v) 1 layer containing the punctuation (-point) that ends each TB (. ? ! . . .).

Appendix A presents the acoustic analysis and Section 3.2 presents the evaluation of the segmentation method adapted for BP.

3.2 Evaluation of the Segmentation Method Adapted for BP

Our evaluation dataset is composed of four inquiries and totals 4:47:18 h (see the inquiries in bold in Table 1; we calculated the number of filled pauses in four of these inquiries, using the following list: hum, uhum, éh, ah, ha, ahn, han, uhn, eh, ehn, hein, oh, hun).

Here, we use the same metrics to evaluate the boundary identification task reported in Biron et al. (2021) that are derived from the true positive (TP), false negative (FN), false positive (FP), and true negative (TN) values of the automated boundary detection method compared to the reference corpus. In our specific scenario, there are cases where the method creates a boundary that does not exist at the reference (FP), cases where the method does not identify a boundary that exists at the reference (FN), and cases where the boundary is placed at a similar timestamp for both documents (TP). Timestamps when neither the reference nor the method places a boundary (TN) can not be accounted for because the timeline is continuous. We also computed the metric SER (Slot Error Rate) that calculates the total number of wrong slots annotated by the method divided by the total number of slots annotated in the reference corpus that corresponds to the NIST SU error rate (Liu and Shriberg, 2007), and can have values greater than 100%. Therefore, here, precision (p), recall (r), F1-score (f1) and slot error rate (ser) are defined as: **p = TP/(TP+FP)**, **r = TP/(TP+FN)**, **f1 = 2*p*r/(p+r)** and **ser = (FP+FN)/(TP+FN)**. Table 2 illustrates our results.

Concerning our first research question — Is the heuristic method more suitable for segmenting TB or NTB? —, by looking at the f1-scores for all inquiries, we can see that the method performed better at identifying NTB (results varied from 33% to 50%) than at identifying TB (results ranged from 16% to 29%).

As for the second one — What is the best of the three heuristics for the boundary types TB and NTB (i.e., which one performs best for each type of boundary)? —, for all examples, the version that outperformed the others considered all heuristics. However, it should be noted that the silences' heuristics alone nearly achieved the same numbers in all cases (with a difference ranging from 0 to 3%). And only at inquiry SP_D2_360, heuristics 1 and 2 contributed more significantly, with a higher f1-score than the silences' heuristics at TB and values still significantly higher at NTB (ranging from 16% to 18%) than at the other inquiries (ranging

Discourse genre	Audio quality	Duration	Interviewee’s Gender	Voice of the speakers and external events	# Filled Pauses
SP_EF_153	+	01:11:11	M	very good audio	—
SP_EF_156	+	01:35:37	F	very good sound	73
SP_DID_242	+	00:44:08	F	clear audio	71
SP_DID_235	+	00:34:49	F	clear audio	—
SP_D2_255	+	01:24:01	M/M	clear sound	104
SP_D2_360	-	01:03:32	F/F	a little bit low audio	260
Total Duration		**06:33:18**			

Table 1: Six inquiries of the Minimum Corpus were used in the acoustic analysis. They are characterized by discourse genre, audio quality, duration, interviewees’ gender, a description related to both the voice of the speakers and external events, and number of filled pauses. Those four in bold were chosen to evaluate the speech segmentation method.

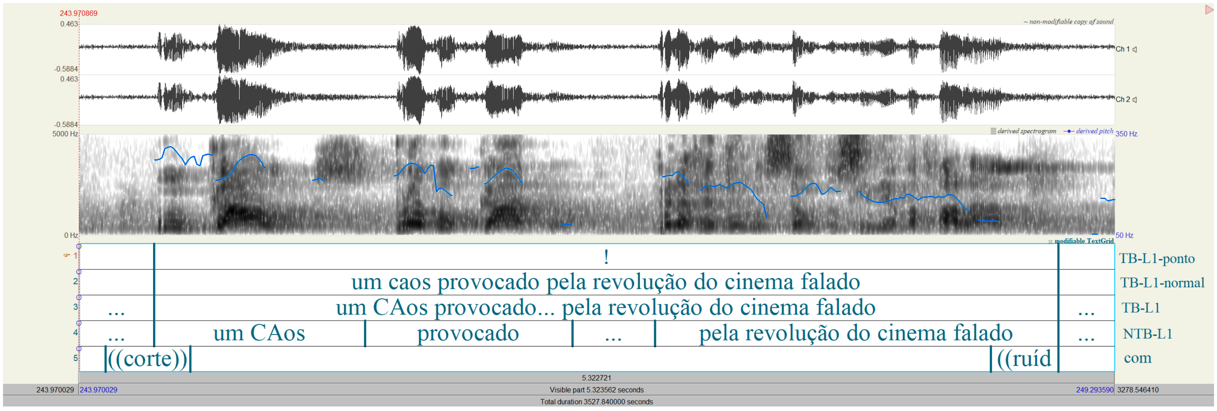


Figure 2: Excerpt from the SP_EF_153 inquiry with five layers annotated in Praat: the first is used to indicate the punctuation that ends each TB (. ? ! . . .), the second contains the normalized excerpt, without the annotation used for transcription in the NURC project, the next two for each speaker that appears in the inquiry (TB-L1, NTB-L1) and the last one for comments on the audio recording (com).

	SP_EF_156								**SP_DID_242**							
	TB				**NTB**				**TB**				**NTB**			
H	sil	h1	h1 + h2	all	sil	h1	h1 + h2	all	sil	h1	h1 + h2	all	sil	h1	h1 + h2	all
f1	**0.18**	0.0	0.01	**0.18**	0.4	0.0	0.03	**0.41**	**0.29**	0.02	0.05	**0.29**	0.49	0.01	0.05	**0.5**
p	0.12	**0.14**	0.04	0.12	**0.48**	0.43	0.38	0.47	**0.23**	0.14	0.14	0.22	**0.71**	0.27	0.36	0.68
r	**0.38**	0.0	0.01	**0.38**	0.34	0.0	0.01	**0.36**	0.4	0.01	0.03	**0.41**	0.38	0.0	0.02	**0.39**
ser	3.41	**1.01**	1.16	3.55	1.03	**1.0**	1.01	01.04	1.91	**1.04**	1.15	02.03	**0.78**	1.01	1.02	0.79
mf1								0.295								0.395
	SP_D2_255								**SP_D2_360**							
	TB				**NTB**				**TB**				**NTB**			
H	sil	h1	h1 + h2	all	sil	h1	h1 + h2	all	sil	h1	h1 + h2	all	sil	h1	h1 + h2	all
f1	**0.16**	0.02	0.05	**0.16**	0.32	0.02	0.04	**0.33**	0.17	0.19	0.18	**0.2**	0.4	0.16	0.18	**0.42**
p	**0.11**	0.08	0.08	**0.11**	0.4	0.19	0.24	**0.39**	0.13	**0.2**	0.17	0.14	**0.5**	0.34	0.32	0.43
r	0.3	0.01	0.03	**0.32**	0.27	0.01	0.02	**0.28**	0.24	0.17	0.19	**0.37**	0.33	0.11	0.13	**0.42**
ser	3.11	**1.15**	1.35	3.31	1.14	**1.03**	1.05	1.17	2.32	**1.52**	1.71	2.92	**1.01**	1.1	1.14	1.14
mf1								0.245								0.31

Table 2: Overall results of the adapted method for BP. We also show an ablation study to measure the impact of the three heuristics in the adapted method, in row H: silence pauses (sil), heuristic 1 (h1), and heuristic 2 (h2), all show results for the three heuristics combined. mf1 stands for macro-f1, i.e. arithmetic mean over harmonic means. The macro-f1 measure of our dataset is 0.31125.

from 0 to 5%).

With respect to speech genre (our third question — Which speech genre has the best segmentation performance (EF/D2/DID)? —, the best results were achieved with SP_DID_242 with a macro-f1 score of 39.5%, which might suggest that, for this method, dialogues between documenters and a participant are the most adequate speech genre among the ones tried. However, with only four inquiries analyzed, it is hard to draw any conclusions. To support that argument, inquiries of type D2 were not adjacently ranked, and their difference of 6.5% is relatively close to the difference of 15% among the highest and lowest macro-f1 scores obtained.

Regarding the number of filled pauses in each inquiry, there is no direct correlation to the impact on the macro-f1 measure, as the second best value of macro-f1 is related to SP_D2_360 (31%) which has the largest number of filled pauses (260) (see Table 1). But we cannot be sure that filled pauses are not affecting all the inquiries as they appear more in conversation inquiries (D2 and DID) and less in classes and talks, but in all the inquiries of NURC-SP MC.

It is important to note that all the results reported in Table 2 use the transcriptions provided by the original NURC-SP project. Therefore, we performed an evaluation to measure the impact of using the revised transcription with the support of the software tool Praat in the pipeline of Figure 1. We reran the pipeline for the inquiry SP_DID_242. Our findings exhibited a change within the range of 0-2%, with an updated macro F1-score of 41% for SP_DID_242.

When dealing with boundaries identified by more than one heuristic, Biron et al. (2021) attributes the hits to the DSRs, rather than to the silences' heuristic. In our ablation study, each heuristic's performance was calculated separately and there may be overlaps among the boundaries covered. Therefore, on Table 2, it can be seen that the summation of the value obtained using each heuristic separately does not necessarily equal the value obtained using all of them conjointly.

4 Discussion

4.1 Related Work on Automatic Detection of Prosodic Boundaries

Table 3 presents six studies that have developed boundary detection methods, and compares their methodologies and results. Three of them deal with the Portuguese language (Brazilian and European) and three with the English language. With regard to datasets, only the BP one is small (≈17 min) compared to the others which are longer than four hours. All the datasets but one (the dataset that was crawled from the site of RTP[6]) are resources frequently used in applications that consider prosodic boundaries. Three of them are annotated with TB and NTB boundary types, although in one of them (Hoi et al., 2022), the terms used are *sentences* and *phrases*, respectively. This dataset annotated with labels of sentences and phrases is balanced, being composed of 7.500 sentences and 7.500 phrases for training, and 200 samples of each for testing. The model proposed by Hoi et al. (2022) was set to identify if a silent pause indicates a terminal or non-terminal boundary but uses the spectrogram of speech as a feature in order to recognize and segment sentences/phrases. There are three studies that deal with only one type of boundary (IU). While the method presented in Kocharov et al. (2017) was initially developed for processing Russian speech, here we only show results for English speech to facilitate the comparison among studies, notwithstanding the fact that the methods were not applied to the same dataset.

Table 3 summarizes evaluation metrics of previous boundary identification methods for spontaneous speech. It is important to note that Raso et al. (2020) and Biron et al. (2021) remove IUs composed of filled pauses from the evaluation. There is no information about the treatment of filled pauses in the other three studies described in this section. Our work was evaluated with filled pauses and this choice was due to the important discursive roles that these elements play. Filled pauses are typical manifestations of oral speech planning and can play the role of discursive markers with an interactional and cohesive function of the spoken text.

Preserving filled pauses may be one of the causes for the discrepancy between our results and the results of Biron et al. (2021). Another one could be the different average length of IUs between languages (English and Portuguese) as we have not yet customized the parameters used in the method for our corpus. Finally, we selected a challenging corpus (see details in Section 4.2), created in the 1980s when acoustic tools were not available to aid annotators in audio transcriptions.

Raso et al. (2020) reports a lower performance

[6] www.rtp.pt/noticias/

Source	Dataset	Lang.	Training	Features	Boundary Types	F1-score/Accuracy
This work	Part of the NURC-SP MC (∼5hrs)	BP	No	DSR and Silent Pause	TB NTB	31%/—
Raso et al. (2020)	C-Oral-Brasil I C-Oral-Brasil II (∼17 min)	BP	Yes, LDA algorithm	Speech Rate, Duration, f0, Intensity, Pause	TB NTB	68%/—
Hoi et al. (2022)	RTP (∼33 hrs)	EP	Yes, CNN API of keras Library	Spectrogram	TB NTB	—/95%
Biron et al. (2021)	SBC (∼20 hrs)	EN	No	DSR and Silent Pause	IU	66%/—
Kocharov et al. (2017)	BURSC (∼10 hrs)	EN	Yes. Two-stage procedure combines syntax and acoustics	Pause, PBL, Df0C	IU	76.2/—%
Roll et al. (2023)	SBC (∼20 hrs) IViE (∼36 hrs)	EN	Whisper was fine-tuned to annotate IU	—	IU	87%/96% (SBC) 73%/93% (IViE)

Table 3: Segmentation Methods and Corpora containing spontaneous speech used in the previous boundary identification methods for spontaneous speech. TB stands for Terminal Boundary, NTB stands for Non-Terminal Boundary. DSR stands for Discontinuities in Speech Rate. PBL stands for pre-boundary lengthening and Df0C stands for declination of f0 contour.

of the classifier of NTB (54.5% F1) than the TB classifier (81.5% F1). The main features responsible for the performance of TB were pause and f0, while for NTB these features were pause, f0, and speech rate. In our evaluation, we found the inverse: our best results came from the detection of NTB labels. Kocharov et al. (2017) proposes a two-stage procedure that combines syntax and acoustics, using a rule-based system over a dependency tree followed by a Random Forest classifier based on acoustic features. Their results, F1 of 76%, show 10% of improvement over the heuristic-based method of Biron et al. (2021) although the methods were evaluated in different corpora. It is amazing how the best results of the methods compared here (Roll et al., 2023) are obtained with a simple fine-tuning of Whisper for the task of detecting prosodic boundaries. The authors justify the reasons for this performance showing that ASR Whisper captures, in its model, the prosodic characterization to segment speech in IUs, in addition to the task for which it was modeled, which is automatic transcription of speech.

4.2 Error Analysis of the Automatic Segmentation

Through error analysis, we aimed to verify whether the automatic segmentation method impacts positively or negatively on the annotation process. To this end, we measured the time required to annotate an inquiry — namely, SP_D2_012 — in two situations: (i) from the final output generated by the method and (ii) manually, that is, without the help of the method.

In order to prepare the textgrid for evaluation, we added an interval tier to the SP_D2_012 textgrid (generated by the method), dividing it into 300-second chunks. We selected two subsequent excerpts in the initial, medial, and final positions of the file; then, one excerpt of each pair was annotated from the method output and the other was manually annotated[7]. The intervals were adjusted to match the beginning and end of a complete TB. We then copied the timestamped tier to another textgrid to be used in the manual annotation process.

The annotation was carried out by one of the authors, an expert in prosodic annotations.

For the *manual annotation process* (without the method), it was necessary *(i)* to create tiers for annotation (TB, NTB, comments), *(ii)* to copy the text from an external textfile (the diarized transcription) into the tiers, audio-aligning it according to the TB and NTB concepts, *(iii)* and to review the transcription, according to the annotation standards adopted for CATNA[8]. As for the *annotation process using the method output*, since the tiers (TB, NTB, comments) were already created and the text was already aligned and segmented, it was only necessary *(i)* to adjust the text-to-audio alignment according to the division into TBs and NTBs and

[7] The selection of excerpts at relatively distributed points in the inquiry was designed to reduce possible differences between more complex and less complex transcription parts, whether due to automatic segmentation or to the dialogue dynamics itself.

[8] CATNA's annotation standards — a simplified version of those used in MC (see Santos et al., 2022) — are as follows: **(a)** transcription for words is based on written BP standards; **(b)** no punctuation mark or any special character; **(c)** lowercase letter only; **(d)** numbers are written in full; **(e)** phatic expressions are always written; **(f)** empty parentheses for incomprehensible words; **(g)** single parentheses for hypotheses of what was heard; **(h)** laughs are transcribed as a tag *((risos))* and segmented as a separate NTB; **(i)** acronyms are expanded for their forms of pronunciation, and the tag *((sigla))* is set in the comments tier; **(j)** proper names are extended (e.g., M. → *Maria*), and the tag *((name))* is set in the comments tier.

(ii) to review the transcription.

We present the annotation time measurements for each excerpt in Table 4. In short, the data show that the manual annotation was relatively faster, with a difference of -1h37min, even though the annotation speeds between the revision methods are similar.

Interestingly, regardless of the position of the excerpt (initial, medial, final) or the nature of the review (based on the method or completely manual), we noticed that all six excerpts are balanced in terms of duration, the number of characters, and the total number of IU boundaries, be it before or after the review (see Table 5). We therefore believe that these factors had a similar impact on the time taken to annotate all the excerpts.

On the other hand, the text-to-audio misalignment seen throughout the inquiry seems to be crucial for the annotation slowdown. The initial 82% of the first excerpt of the inquiry is relatively well aligned (i.e., much of the text corresponds to the audio recording); after that, the match is lost, meaning that none of the text contained in an interval from the second and third excerpts matches the recording to which it was forced-aligned. Because of this, text from later intervals had to be moved to the preceding ones, slowing down the annotation process.

During the transcription review, the following adjustments had to be made: *(a)* space insertion between words (casovocê → *caso você*); *(b)* spelling correction and adequacy to writing standards (musica → *música*, pro → *para o*); *(c)* word correction (fachoto → *pacheco*), *(d)* extra or missing words/phrases adjustment ("jornal informar o artigo" → "*jornal informativo*", "eu pela manhã" → "*eu começo pela manhã*"). Thus, in addition to low audio quality and overlapping voices, the transcription used as input for the forced aligner may have contributed to the misalignment we have noted, especially in the cases specified in (c) and (d).

Therefore, the misalignment negatively affects the phones' timestamps to be used in the automatic segmentation method and, consequently, the insertion of DSR-based prosodic boundaries. All these factors lead us to the need to create a human-reviewed version of the CATNA transcription files in order to provide a transcription that is faithful to the audio recordings and suitable for training future natural language processing systems. Despite the evaluation results, we believe that the prosodic segmentation method presented here has the potential to assist in the segmentation of other corpora (provided that an adequate transcription is guaranteed as input for the forced aligner), as well as to assist annotators less experienced in prosodic annotation.

5 Concluding Remarks and Future Work

The relevance of a prosodically processed and annotated BP corpus lies in the fact that the delimitation of prosodic boundaries improves the performance of natural language processing systems and is input for automatic punctuation prediction, such as the Whisper ASR does. Manually annotating a large corpus with prosodic boundaries is a costly task, therefore, to have a baseline method available, as the one made available in this work, can help to foster this research area. Furthermore, it is possible to use the corpus, also made available, as a reference set for training ASRs and, thus, leveraging the development of BP speech processing methods and enabling new linguistic studies. Regarding our results, our f1-macro reaches 31%, significantly lower than Biron et al.'s (2021) performance of 66% (see Table 3). We suspect that is due to three reasons. The first one is that we did not remove the filled pauses from the corpus, as was part of Biron et al.'s (2021) pre-processing. The second reason is that Biron et al. (2021) is adapted to English and for our initial results, we applied the method to our corpus without customizing the six parameters (see Section 2) to BP. The third is due to a few challenges of the NURC-SP corpus: (1) "overlapping speakers' voices" present in inquiries of types D2 and DID, (2) low audio quality in some of the inquiries, which impacts even manual transcription, causing several annotations of "incomprehension of words or segments" and "hypothesis of what was heard" (Gris et al., 2022), (3) the transcriptions of the corpus were carried in the 1980s, when acoustic tools were not available to support the annotators, who had to rely solely on auditory perception.

Regarding future work, we foresee two lines of research. In the first one, we intend to perform hyperparameter tuning for Portuguese, using the complete Minimum Corpus of NURC-SP and techniques such as grid search or random search (e.g., GridSearchCV and RandomizedSearchCV (Pedregosa et al., 2011)). The second is inspired by the best results that can be seen in Table 3, obtained using Whisper's fine-tuning at Roll et al. (2023). We intend to study the correlation between

Excerpt	Revision from the method			Manual revision		
	Duration (s)	Annotation time spent (h:m:s)	Annotation speed	Duration (s)	Annotation time spent (h:m:s)	Annotation speed
Initial	296.6	2:03:48	25	304	1:43:43	20.5
Medial	300.6	2:33:35	30.7	294	1:28:25	18
Final	285.5	2:00:35	25.3	310.2	1:48:31	21
	882.8	6:37:57	27	908.2	5:00:39	19.9

Table 4: Duration, annotation time spent, and annotation speed (= ratio of *annotation time* to *duration*) for the SP_D2_012 inquiry excerpts.

Excerpt	Characters				Boundaries (TB,NTB)			
	Revision from the method		Manual revision		Revision from the method		Manual revision	
	Original	Reviewed	Original	Reviewed	Original	Reviewed	Original	Reviewed
Inital	4004	4121	4781	4963	508	455	472	477
Medial	4778	4953	4383	4618	456	547	496	480
Final	5025	5185	5497	5839	332	439	526	618
	13807	14259	14661	15420	1296	1441	1494	1575
	Incr. = 452 (3.3%)		Incr. = 759 (5.2%)		Incr. = 145 (11.2%)		Incr. = 81 (5.4%)	

Table 5: Number of characters and TB/NTB boundaries before and after human review on SP_D2_012 inquiry excerpts. The number of characters includes spaces. *Original* stands for the original transcription (whose source is the diarization textfile). *Incr.* stands for the increase over the reviewed version.

punctuations provided by Whisper and the prosodic boundaries of our method presented in this paper. For this study, we intend to transcribe the evaluation dataset with the ASR Whisper in order to compare the boundaries of both.

Acknowledgements

First of all, we would like to thank the annotators of the TaRSila project who were tireless in reviewing the automatic transcriptions, training and testing the models for various speech processing systems. This work was carried out at the Artificial Intelligence Center (C4AI-USP), with support from the São Paulo Research Foundation (FAPESP grant nº 2019/07665-4) and IBM Corporation. We also thank the support of the Center of Excellence in Artificial Intelligence (CEIA) funded by the Goiás State Foundation (FAPEG grant no. 201910267000527), the São Paulo University Support Foundation (FUSP) and the National Council for Scientific and Technological Development (CNPq-PQ scholarship, process 304961/2021-3). This project was also supported by the Ministry of Science, Technology and Innovation, with resources from Law nº 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published Residência no TIC 13, DOU 01245.010222/2022-44.

References

Zrar Kh. Abdul and Abdulbasit K. Al-Talabani. 2022. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158.

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

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, pages 4489–4493.

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

Mary E. Beckman, Julia Hirschberg, and Stefanie Shattuck-Hufnagel. 2005. The original ToBI system and the evolution of the ToBI framework. In Sun-Ah Jun, editor, *Prosodic typology: the phonology of intonation and phrasing*, pages 9–54. Oxford University Press, Oxford.

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

Paul Boersma and David Weenink. 2023. Praat: doing phonetics by computer [Computer program]. Version 6.3.10.

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

Rodrigo Colnago Contreras, Monique Simplicio Viana, Everthon Silva Fonseca, Francisco Lledo Dos Santos, Rodrigo Bruno Zanin, and Rodrigo Capobianco Guido. 2023. An experimental analysis on multicepstral projection representation strategies for dysphonia detection. *Sensors*, 23(11):5196.

John W. du Bois, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000–2005. *Santa Barbara corpus of spoken American English. Parts 1–4*. Linguistic Data Consortium, Philadelphia.

John W. du Bois, Susanna Cumming, Stephan Schvetze-Coburn, and Danae Paolino. 1992. *Discourse transcription*, volume 4 of *Santa Barbara Papers In Linguistics*. Department of Linguistics, University of California, Santa Barbara.

J.I. Godino-Llorente and P. Gomez-Vilda. 2004. Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Transactions on Biomedical Engineering*, 51(2):380–384.

E. Grabe, Brechtje Post, and F. Nolan. 2001. Modelling intonational variation in english. the ivie system. *Proceedings of Prosody 2000*.

Lucas Gris, Arnaldo Candido Junior, Vinícius Santos, Bruno Dias, Marli Leite, Flaviane Svartman, and Sandra Aluísio. 2022. Bringing nurc/sp to digital life: the role of open-source automatic speech recognition models. In *Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional*, pages 330–341, Porto Alegre, RS, Brasil. SBC.

Ling He, Margaret Lech, Namunu C Maddage, and Nicholas Allen. 2009. Stress detection using speech spectrograms and sigma-pi neuron units. In *2009 Fifth International Conference on Natural Computation*, volume 2, pages 260–264. IEEE.

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

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

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

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

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

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

Yang Liu and Elizabeth Shriberg. 2007. Comparing evaluation metrics for sentence boundary detection. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–185–IV–188.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Heliana Mello, Tommaso Raso, and Lúcia de Almeida Ferrari. To appear. C-ORAL–Brasil II: Corpus de referência do português brasileiro falado informal.

Mari Ostendorf, Patti Price, and Stefanie Shattuck-Hufnagel. 1995. The Boston University Radio news corpus.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Lawrence Rabiner and Ronald Schafer. 2010. *Theory and applications of digital speech processing*. Prentice Hall Press.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR.

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

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

Nathan Roll, Calbert Graham, and Simon Todd. 2023. Psst! prosodic speech segmentation with transformers.

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

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

Bárbara Helohá Falcão Teixeira and Maryualê Malvessi Mittman. 2018. Acoustic models for the automatic identification of prosodic boundaries in spontaneous speech. *Revista de Estudos da Linguagem*, 26(4):1455–1488.

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

Mohammed Zakariah, Yousef Ajmi Alotaibi, Yanhui Guo, Kiet Tran-Trung, Mohammad Mamun Elahi, et al. 2022. An analytical study of speech pathology detection based on mfcc and deep neural networks. *Computational and Mathematical Methods in Medicine*, 2022.

A Acoustic Analysis of the Sampling from Minimum Corpus

Mel scale spectrograms, also known as Mel spectrograms, constitute an extension of traditional spectrograms in which the frequency scale is transformed to the Mel scale, approximating the way the human ear perceives sounds. This makes Mel spectrograms particularly useful for tasks where frequency discrimination is critical, such as identifying phonemes in speech recognition, separating sound sources in noisy environments, and analyzing melodic features in music (Rabiner and Schafer, 2010; Zakariah et al., 2022). Bark scale spectrograms represent a sophisticated approach to analyze audio signals, offering a perspective that comes even closer to human auditory perception (Rabiner and Schafer, 2010; He et al., 2009). The Bark scale is designed to map frequencies in terms of the 25 critical bands of hearing, taking into account how the human ear perceives different frequencies at different sound intensity levels.

Both Mel scale and Bark scale spectrograms address the challenge of representing the spectral characteristics of an audio signal in a more meaningful way than a simple Fourier Transform. Their main differences lie in the details of the mapping scale: Mel scale spectrograms map frequencies in terms of the Mel scale, which is designed to approximate how the human ear perceives frequency differences. This makes them especially effective in tasks such as speech and music recognition (Rabiner and Schafer, 2010), where frequency discrimination is critical. Conversely, Bark scale spectrograms take into account the critical hearing bands and the variation of auditory perception with the level of sound intensity, resulting in an even more accurate representation of human perception. Therefore, Bark scale spectrogram was chosen in this work to present an acoustic analysis. Here, we analyzed the acoustics of the six audio sampling from the Minimum Corpus in order to choose one of each type (EF, D2, DID) to pursue the segmentation analysis (see Figure 3).

Considering the acoustics involved in the EF situation, we can notice that, as expected, there is a concentration of signal energy in low frequencies, particularly in those frequencies that are responsible for the physical human way of speaking. Furthermore, due to the formal/illustrative nature of the EF class, we can also notice a more continuous dialogue, without major discontinuities in the spectrograms. Continuing with the D2 case study, we can now infer, based on the spectrograms, two particular situations:

- A more intense dialogue in the *SP_D2_255* example, evidenced by the high distribution of energy within the entire conversation, with some “negative” spikes caused by the mediator; and
- A calmer example in *SP_D2_360*, with the energy concentrated in low frequencies, below 2048 Hz. We can also mention the low general amplitude of the signal caused by some effect during audio recording.

Moving on to the case of the last conversation (DID), we can deduce the more abrupt peaks and discontinuities compared to the EF and D2 scenarios, highlighting intervals of thought between the questions/inferences raised by the interviewee’s response time. To have a more quantitative way of describing the above statements, **the speed and acceleration of the signal** were calculated, represented by Δ and Δ^2 extracted by Mel Frequency Cepstral Coefficients (MFCCs) (Abdul and Al-Talabani, 2022; Godino-Llorente and Gomez-Vilda, 2004). It is worth mentioning that the adopted number of MFCC coefficients is 13, representing an average between the lower and upper limits that generally define the number of MFCCs to be extracted. A more in-depth study on MFCCs and other forms of application involving cepstral coefficients can be found in Contreras et al. (2023). That said, the values Δ and Δ^2 are shown in the Table 6.

Discourse genre	Δ	Δ^2
EF	-13.594	-23.953
D2	4.039	-64.476
DID	43.985	14.921

Table 6: Table of Average Speed (Delta) and Acceleration (Delta-Delta) for Each Conversation Class of the Minimum Corpus.

As expected, the dynamics of the signal recorded for EF conversation presents negative values for speed and acceleration, a behavior that emphasizes the **continuous speech with low frequencies expected in classrooms/speeches**. Note: here, the negative represents that the sporadic peaks that the speaker applies in the recording are immediately followed by a slowdown in intonation, i.e., high frequencies to low frequencies, to resume the "normal" mode of speech. For the $D2$ and DID speech types (case studies), we can note that: for the first, a positive speed indicates that speech occurs with quick responses, and negative acceleration also indicates that the conversational flow presents abrupt changes between speakers; for the latter, a positive Δ and Δ^2 shows that, even with the presence of considerable discontinuities generated by the speaker thinking about his response to speeches, we have direct conversational behavior that flows optimally within the scope of the speech interview.

Therefore, considering the differences between *SP_D2_255* and *SP_D2_360*, we decided to bring both to the segmentation analysis shown in Section 3.2.

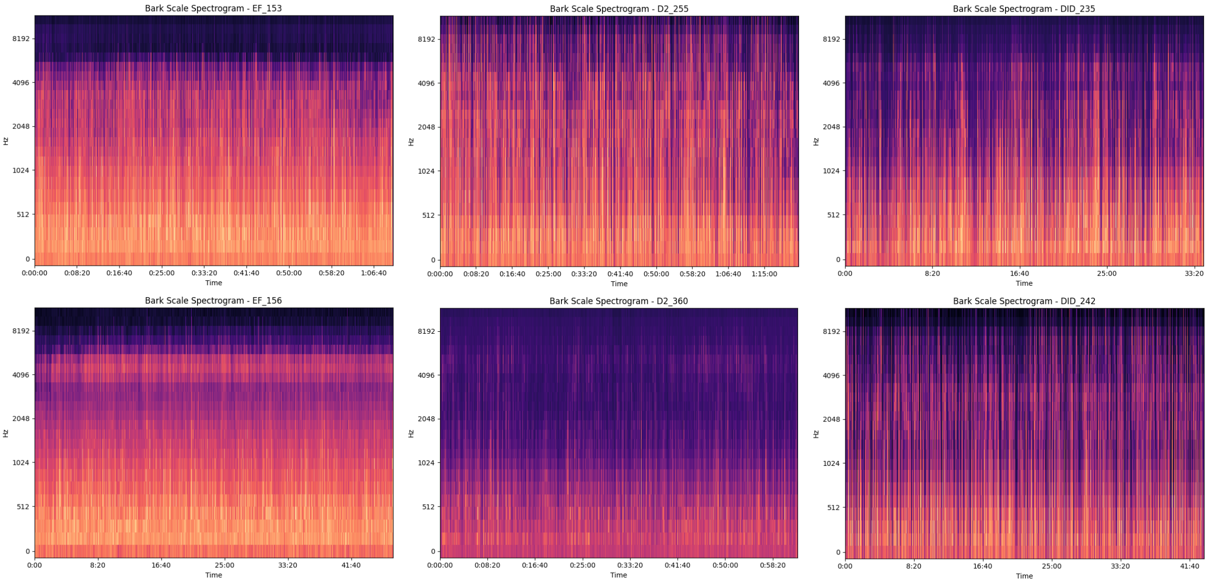


Figure 3: Bark scale spectrograms for the six inquiries selected from the NURC-SP Minimum Corpus: SP_EF_153, SP_EF_156, SP_D2_255, SP_D2_360, SP_DID_235, and SP_DID_242, respectively. Here, warmer colors, such as yellow and red, indicate greater energy intensity (range 0 dB to -40 dB), while cooler colors, such as blue and purple, indicate lower energy intensity (range -40 dB to -80 dB).