RELATO REGISTRADO: PROTOCOLO

# TESTANDO A EFICIÊNCIA DE UM MÉTODO DE **SEGMENTAÇÃO PROSÓDICA AUTOMÁTICA** BASEADO EM APRENDIZADO DE MÁQUINA PARA O PORTUGUÊS BRASILEIRO

Caroline Adriane ALVES
Faculdade de Filosofia, Letras e Ciências Humanas – Universidade de São Paulo (USP)
São Paulo, SP, Brasil

Rian Pereira FERNANDES
Faculdade de Filosofia, Letras e Ciências Humanas – Universidade de São Paulo (USP)
São Paulo, SP, Brasil

Julio Cesar GALDINO
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
São Paulo, SP, Brasil

Giovana Meloni CRAVEIRO
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
São Paulo, SP, Brasil

Flaviane Romani Fernandes SVARTMAN
Faculdade de Filosofia, Letras e Ciências Humanas – Universidade de São Paulo (USP)
São Paulo, SP, Brasil

Sandra Maria ALUÍSIO
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
São Paulo, SP, Brasil

RESUMO

A fala natural é organizada em unidades compostas por segmentos cujas fronteiras são marcadas por elementos prosódicos. Métodos têm sido propostos



EDITORES
– Miguel Oliveira, Jr. (UFAL)
– Mailce Mota (UFSC)

AVALIADORES
– Plínio Barbosa (UNICAMP)
– Tommaso Raso (UFMG)

SOBRE OS AUTORES
- Caroline Adriane Alves
  Metodologia; Escrita – Esboço Original; Escrita – Revisão e Edição.
- Rian Pereira Fernandes
  Metodologia; Escrita – Esboço Original; Escrita – Revisão e Edição.
- Julio Cesar Galdino
  Escrita – Esboço Original; Escrita – Revisão e Edição.
- Giovana Meloni Craveiro
  Curadoria de Dados; Metodologia; Software; Escrita – Esboço Original; Escrita – Revisão e Edição.
- Flaviane Romani Fernandes Svartman
  Conceitualização; Metodologia; Supervisão; Escrita – Esboço Original; Escrita – Revisão e Edição.
- Sandra Maria Aluísio
  Conceitualização; Metodologia; Supervisão.

Recebido: 31/10/2025
Aceito: 10/02/2026
Publicado: 04/03/2026

COMO CITAR
ALVES, C.A.; FERNANDES, R.P.; GALDINO, J.C.; CRAVEIRO, G.M.; SVATMAN, F.R.F.; ALUÍSIO, S.M. (2026). Testando a eficiência de um método de segmentação prosódica automática baseado em aprendizado de máquina para o português brasileiro. *Cadernos de Linguística*, v. 7, n. 1, e912.

para identificar automaticamente tais unidades, levando em consideração os elementos prosódicos e visando aprimorar tarefas que envolvem a automatização da fala, como a conversão de texto em voz e a transcrição da fala. O método apresentado em Craveiro *et al*. (2025), que inclui um classificador treinado em português brasileiro (PB), revelou-se capaz de realizar previsões adequadas quanto à identificação dessas unidades de fala em uma amostra de dados dessa variedade do português. O presente trabalho pretende replicar esse método em uma nova amostra de dados do PB, a fim de verificar se os mesmos resultados são obtidos em uma nova condição. Espera-se que os resultados da nova amostra se aproximem daqueles alcançados no estudo original, por meio das mesmas métricas, ferramentas e técnicas estatísticas que foram empregadas.

PALAVRAS-CHAVE
Segmentação Prosódica Automática; Aprendizado de Máquina;
Português Brasileiro.

TITLE
TESTING THE EFFICIENCY OF A MACHINE LEARNING-BASED AUTOMATIC PROSODIC SEGMENTATION METHOD FOR BRAZILIAN PORTUGUESE

ABSTRACT
Natural speech is organized into units composed of segments whose boundaries are marked by prosodic elements. Methods have been proposed to automatically identify such units, taking into account prosodic elements and aiming to improve speech-related automatic tasks, such as text-to-speech conversion and speech transcription. The method proposed by Craveiro et al. (2025), which includes a classifier trained on Brazilian Portuguese (BP), has proven effective in predicting the identification of these speech units in a data sample from this Portuguese variety. This paper aims to replicate this method on a new BP data sample, to verify whether similar results can be obtained under new conditions. It is expected that the results from the new sample will approximate those achieved in the original study, based on the same metrics, tools, and statistical techniques employed.

KEYWORDS
Automatic Prosodic Segmentation; Machine Learning; Brazilian Portuguese.

# INTRODUÇÃO

A segmentação prosódica consiste no agrupamento da fala em unidades menores, com base em aspectos prosódicos que marcam as fronteiras dessas unidades, como a entoação, a intensidade e a duração. Tais unidades, que nem sempre correspondem a unidades morfossintáticas, ajudam a estruturar a fala e facilitar a compreensão da mensagem oralizada. Entre uma unidade e a seguinte, fronteiras prosódicas são inseridas. Há também estudos que fazem uma distinção entre enunciados de fronteiras terminais (TB - *terminal break*), que marcam sequências completas, ou seja, que comunicam a conclusão de uma ideia, constituindo a menor unidade de fala autônoma pragmaticamente, de enunciados de fronteiras não terminais (NTB - *non terminal break*), que sinalizam uma unidade não autônoma, cuja informação não é completada dentro da mesma unidade. A identificação dessas fronteiras baseia-se, sobretudo, na relevância perceptual (auditiva) das pistas prosódicas, como variações na frequência fundamental (F0), na duração do segmento e na presença de pausas (Serra, 2009; Raso; Teixeira; Barbosa, 2020), além da inspeção visual do sinal acústico.

A segmentação prosódica é aplicada em uma variedade de áreas, incluindo sintetizadores de fala (TTS) e sistemas de reconhecimento de fala (ASR), além de análises linguísticas (Chen; Hasegawa-Johnson, 2004; Liu *et al*., 2022; Lin *et al*., 2019; Viola; Madureira, 2008). Muitos dos estudos que abordaram segmentação prosódica automática consideraram apenas corpora de fala controlada e lida. Nesses casos, fronteiras prosódicas e sintáticas coincidem, já que o falante segue a pontuação da escrita, baseada em fronteiras sintáticas marcadas pelas convenções da escrita, e consequentemente realiza, na fala, fronteiras prosódicas nas mesmas posições onde fronteiras sintáticas marcadas na escrita ocorrem. Contudo, estudos que abordam fala espontânea podem ter mais dificuldade em atingir bons resultados devido à presença de disfluências, e ao fato de serem raras em fala controlada (Biron *et al*., 2021), e as fronteiras prosódicas são menos claras, já que o falante formula o texto simultaneamente à produção da fala, frequentemente realizando fronteiras em momentos imprevisíveis, diferentemente do que ocorreria em uma tarefa de leitura de um texto previamente pontuado.

A tarefa de segmentação prosódica automática de fala espontânea é um desafio de longa data (Biron *et al*., 2021), que continua sendo tema relevante de estudos atuais, devido aos obstáculos mencionados e ainda não ultrapassados. As abordagens de segmentação prosódica automática incluem métodos baseados em heurísticas, aprendizado de máquina tradicional e aprendizado de máquina profundo. Há abordagens baseadas exclusivamente em sinais acústicos, outras que se baseiam também em informações lexicais e sintáticas, incluindo extensivas etapas de preparação, como anotação manual. Aqui apresentaremos uma revisão de literatura que cobre os oito estudos apresentados no trabalho de Craveiro *et al*. (2025), cuja abordagem estamos nos propondo a replicar. Tais estudos foram selecionados por terem sido desenvolvidos para o português ou por

terem alcançado resultados relevantes para o inglês através de diferentes tipos de abordagens, seja por meio da utilização de heurísticas, de aprendizado de máquina tradicional ou de aprendizado de máquina profundo.

Biron *et al*. (2021) detectaram fronteiras prosódicas em fala espontânea de inglês americano através de heurísticas baseadas na duração de pausas e descontinuidades de taxa de fala (SRDs). Os autores utilizaram o *Santa Barbara Corpus* (SBC), um corpus balanceado em gênero, composto por aproximadamente vinte horas de áudio, e também a ferramenta *Montreal Forced Aligner*[1] para realizar o alinhamento fonético forçado dos áudios, gerando previsões de início e final de cada fone. Com o uso dessas marcações, para cada palavra e a partir de seu início, a duração de todos os fones não silenciosos presentes em uma janela de 300 ms é extraída e, a partir desses valores brutos de duração, é calculada a média de duração, correspondente à taxa de elocução[2] daquela palavra. As SRDs são indicadas quando a diferença de taxa de elocução de palavras subsequentes excede determinado patamar: a primeira heurística utiliza como patamar 88% da maior diferença de valores de taxa de elocução em um turno e a segunda heurística utiliza como patamar 70% e só é aplicada a trechos resultantes da primeira heurística que sejam também maiores que 3 segundos e que contenham mais de 10 palavras. Esse estudo reportou uma medida f1[3] de 66% e comparou os resultados obtidos automaticamente com a anotação manual utilizada como referência, reportando similaridades nas características das fronteiras prosódicas, como a localização do valor mais alto da frequência fundamental (F0) da unidade entoacional[4] (UE) na segunda palavra da UE.[5]

1 https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

2 Embora os autores tenham calculado a taxa de elocução a partir da média de duração de fones, normalmente a taxa de elocução é calculada em sílabas por segundo ou palavras por minuto.

3 A métrica "precisão" avalia quantas vezes o modelo detectou a fronteira corretamente em relação a todas as vezes que indicou uma fronteira. Já a métrica "*recall*" mede quantas vezes o modelo detectou uma fronteira corretamente em relação a todas as vezes que deveria ter identificado uma fronteira, sem se importar com quantas vezes ele acertou ou errou. Em suma, se a precisão está focada em ser correta quando afirma que há uma fronteira, o recall está focado em não perder nenhuma fronteira correta. Por sua vez, a métrica "f1" combina precisão e recall por meio de uma média harmônica, proporcionando uma avaliação equilibrada do modelo. O valor de f1 varia de 0 a 100, pois é calculado por uma média harmônica. Assim, é sensível a baixos valores de precisão ou *recall*, exigindo que ambos sejam altos para um f1 elevado.

4 Os autores reportam que não há consenso a respeito da definição de unidade entoacional, mas sugerem que está relacionada à presença de um contorno de frequência e que as fronteiras prosódicas que separam tais unidades estão relacionadas a pausas, mudanças em valores de intensidade e frequência, desaceleração na taxa de elocução no final da unidade, bem como aceleração em seu início.

5 Conforme Biron *et al*. (2021, p. 10): *Given the mean duration of first words of IUs and the distribution of IU durations, the timing of peak pitch would typically correspond to the second word of the IU. [...] To summarize, although the boundary detection itself did not make use of pitch data in any way, and although the recordings varied in speakers, genre and communicative purpose, a consistent and clear pitch reset was observed. As expected, randomly segmenting speech into intervals of about one second (the mean duration of a phrase) and then averaging over them exhibited no such decline in pitch. We conclude that measurements of pitch reset and of pitch variability at the closure of phrases support the notion of similarity between the automatic and manual boundary detection.*

Em Kocharov, Kachkovskaia e Skrelin (2017), a predição de unidades segmentadas prosodicamente foi proposta com base na combinação de informação sintática e acústica e com o uso de um classificador *Random Forest*[6]. Essa abordagem assume que certas fronteiras entre palavras, como entre preposições e substantivos, são improváveis posições de fronteiras prosódicas, de modo que a sintaxe restringe potenciais localizações de fronteiras prosódicas. De fato, o estudo reportou que 97% das fronteiras prosódicas localizavam-se em posições sintaticamente possíveis, e que de 2% a 2,5% dos remanescentes 3% poderiam ser identificados por meio de regras específicas do idioma[7] abordado. Os autores utilizaram o *Boston University Radio Corpus* (BURNC), um corpus balanceado em gênero de cerca de três horas de fala espontânea em inglês americano. Na primeira etapa, os autores desenvolveram um sistema que prevê potenciais fronteiras prosódicas sempre que duas palavras adjacentes não estão sintaticamente conectadas. Para tal, utilizaram uma árvore de dependências[8] e adicionaram uma série de regras simples. Na segunda etapa, um classificador Random Forest determina quais potenciais fronteiras prosódicas previstas na primeira etapa são de fato fronteiras prosódicas, a partir de características acústicas. Tais características incluem declínio de contorno da frequência fundamental (F0), desaceleração na taxa de elocução ao fim do enunciado ou sintagma entoacional[9], intensidade[10] e pausas. O estudo obteve medida f1 de 76% e acurácia de 86,5%. Variações da F0, da taxa de elocução e da intensidade foram indícios acústicos fundamentais na predição de fronteiras de unidades segmentadas nesse trabalho. Os autores também reportaram que erros de *parsing*[11] devem ter sido responsáveis parcialmente pelos casos de erro, mas que não puderam calcular tal porcentagem já que o padrão ouro da anotação sintática do material utilizado ainda não estava disponível.

6 Algoritmo de aprendizado de máquina que combina várias árvores de decisão, de maneira aleatória, formando o que pode ser considerado uma floresta, em que cada árvore é utilizada para a escolha do resultado final.

7 Os autores afirmam que para cada idioma há um conjunto de dez a vinte regras que dizem respeito a parênteses, nomes compostos, combinações de verbos com preposições ou advérbios e que funcionam como expressões idiomáticas, entre outros.

8 Uma árvore de dependências representa a estrutura sintática de uma frase. As arestas representam as relações de dependência entre as palavras, representadas pelos nós. Em geral, o nó raiz é o verbo principal da frase.

9 Kocharov, Kachkovskaia e Skrelin (2017, p. 2) definem sintagma entoacional nos termos de Ladd (1986): *As discussed by Ladd (1986, p. 311), IP in its traditional sense has the following main properties: (i) they are the largest phonological chunk into which utterances are divided, extending from one phonetically definable boundary to the next; (ii) they are a specifiable intonational structure, including—in most versions of the theory—a single most prominent point (primary stress, tonic, nucleus); (iii) the are phonological units which are nevertheless assumed, ideally, to match up in poorly understood way with elements of syntactic or discourse-level structure.*

10 A intensidade é estimada da seguinte maneira no trabalho dos autores: *We estimate this reset using amplitude values: for each clitic group, it is the difference between its mean amplitude values and that of the following clitic group. Amplitude was calculated as absolute value of speech signal* (Kocharov; Kachkovskaia; Skrelin, 2017, p. 4).

11 *Parsing,* em PLN, é um processo de análise gramatical de uma cadeia de texto para determinar sua estrutura sintática, ou seja, identificar suas partes constituintes, como sujeito, verbo, objeto, advérbios etc. No estudo, uma ferramenta automática de *parsing* foi utilizada, o que pode ocasionar erros que são propagados para as etapas seguintes da tarefa.

Roll, Graham e Todd (2023) introduziram o método PSST, que faz um *fine-tuning* (ajuste) do modelo *Whisper*de 764M[12] de parâmetros para segmentar a fala, integrando informações prosódicas e sintáticas, funcionando também como uma ferramenta de transcrição. Eles utilizaram o *Santa Barbara Corpus*, já mencionado anteriormente, e revisaram manualmente as transcrições, preservando pausas preenchidas e disfluências, mas removendo tokens indesejados, como, por exemplo, tokens compostos por sons de respiração e risadas. Os autores testaram a influência das informações sintáticas e probabilidades léxicas/sintáticas para a segmentação através de duas outras versões do modelo, uma com a sintaxe mascarada e outra sem as informações acústicas, a qual partiu diretamente dos tokens de texto transcritos pelo *Whisper*. A versão com a sintaxe mascarada foi construída através da substituição de todos os tokens por um token comum, preservando somente as informações acústicas e marcações de fronteiras. O modelo que obteve melhor desempenho foi o que combinava informações acústicas e sintáticas, alcançando 96% de acurácia e 87% de medida f1. O método é semissupervisionado e não requer extensivas anotações ou recursos computacionais, tornando-o prático para aplicações de processamento de linguagem natural (PLN), e foi disponibilizado[13] pelos autores.

Teixeira (2022) desenvolveu um classificador de análise discriminante linear (LDA) aplicado à fala espontânea em PB, baseado em parâmetros acústicos. Os dados utilizados no estudo consistiram em gravações de áudio de aproximadamente um minuto, extraídas dos corpora C-ORAL BRASIL I e II, compostos por 7 amostras de cada corpora, representando fala espontânea monológica informal, fala midiática e fala formal em contexto natural em 14 amostras (denominadas aqui amostra I e amostra II), totalizando 17 minutos de fala masculina com limites prosódicos anotados por especialistas. Foram extraídas 111 características fonético-acústicas, por meio do *script* Praat, ao longo do sinal de fala para todas as unidades V-V em uma janela centrada em todos os limites entre palavras fonológicas. As características extraídas compreenderam 5 grupos de medidas: 1) Velocidade e ritmo da fala; 2) Duração normalizada; 3) Frequência fundamental; 4) Intensidade; 5) Pausa silenciosa (presença/ausência e duração). As posições em que pelo menos 50% dos anotadores indicaram um limite do mesmo tipo foram consideradas limites. Vários modelos foram treinados para identificar limites terminais (LTs), e limites não terminais (LNTs)[14]: (i) o modelo TB-b1, com pausa e F0 como parâmetros principais, foi treinado na amostra I (balanceada), e o teste

12 Um modelo *transformer* de reconhecimento de fala estado da arte, proposto por Radford *et al*. (2023), que é pré-treinado em milhares de horas de fala, capaz de realizar eficientemente uma variedade de tarefas relacionadas à fala e de generalizar para muitos conjuntos de dados e domínios. Os autores disponibilizam modelos treinados com diferentes quantidades de dados; o modelo de 764 M de parâmetros é considerado de tamanho médio.

13 https://github.com/Nathan-Roll1/PSST

14 Os modelos que identificam limites terminais usam a sigla TB em seus nomes, enquanto aqueles que identificam limites não terminais usam a sigla NTB.

na amostra II teve uma acurácia de 76,3% para LTs; (ii) o modelo TB-b2 foi treinado na amostra II (balanceada), e o teste na amostra I teve uma acurácia de 80,8% para LTs; (iii) o modelo NTB-b1 foi treinado na amostra I e testado na amostra II, tendo uma acurácia de 71,2% para LNTs; (iv) o modelo NTB-b2 foi treinado na amostra II e testado na amostra I, tendo uma acurácia de 75,6% para LNTs. As pausas, o reset de F0 e a inclinação média de F0 em unidades próximas ao final das palavras fonológicas foram as principais características relacionadas à identificação dos limites terminais, enquanto a pausa, a taxa de articulação e a duração do segmento padrão foram as principais características para a identificação dos limites não terminais.

Hoi, Sun e Im (2022) propuseram um método baseado na detecção de pausas utilizando espectrogramas e uma rede neural convolucional (CNN)[15]. Os autores extraíram 15.000 sentenças[16] do *website* de notícias RTP[17], totalizando 33 horas de fala lida em português europeu, sem balanceamento de gênero. O método detectou se pausas maiores ou com exatamente 250 ms marcavam fronteiras terminais ou não terminais. Janelas de áudio (100 ms antes da pausa + 300 ms depois da pausa) foram classificadas com uma CNN de três camadas. Sem o uso de alinhamento fonético forçado ou de características semânticas, o modelo atingiu 95,6% de acurácia. Apesar de ser eficiente e funcionar para qualquer idioma, o método só lida com fronteiras baseadas em pausas, de modo que não identifica fronteiras prosódicas sem esse aspecto, o que nos parece ser mais relevante para fala lida do que para fala espontânea.

Craveiro *et al*. (2024) adaptaram a metodologia descrita em Biron *et al*. (2021) para o português brasileiro, utilizando o alinhador fonético forçado *UFPAlign* (Batista; Dias; Neto, 2022), elaborado para o PB. Por trabalharem com áudios longos (30-90 minutos), foi necessária a segmentação dos áudios em trechos de 10 minutos para o alinhamento. A metodologia utilizada levou em consideração as mesmas heurísticas de Biron *et al*. (2021), que utilizaram janelas de áudio de 300 ms para detectar pausas e descontinuidades de taxa de elocução. Os autores aplicaram a metodologia a trechos do corpus NURC-SP, os quais contêm fala espontânea de dois homens e quatro mulheres, todos advindos de São Paulo e com educação de nível superior, totalizando aproximadamente 5 horas. A medida f1 reportada é de 31% com margem de acerto de 0,25 segundos, considerando uma média de resultados para fronteiras terminais e não terminais. A margem de acerto considera em qual segundo a fronteira prosódica de referência está localizada e define a previsão do segmentador

15 Uma CNN é um algoritmo de aprendizado profundo que aprende como atribuir importância a determinados aspectos de uma imagem para diferenciá-la de outra.

16 Embora "sentença" se refira a uma unidade da escrita, segundo os autores, eles utilizam esse termo ao referirem a enunciado, visando manter consistência quanto ao uso da mesma terminologia de outros estudos: "*... we adopt the notion of utterance in this research, a stream of speech bounded by pauses or constituting a single semantic unit [...]. However, to remain consistent with the use of terminology in other studies, we use the term "sentence" to represent what we mean by "utterance", even though we have a clear understanding of the difference in linguistic meaning between them.* (Hoi; Sun; Im, 2022, p.1-2).

17 https://www.rtp.pt/noticias/

como acerto, somente se o tempo de previsão da fronteira estiver a uma distância de tempo anterior ou posterior de, no máximo, 0,25 segundos da fronteira de referência. O código[18] da metodologia apresentada está disponível publicamente.

O estudo de Craveiro *et al*. (2025), que pretendemos replicar neste trabalho, inspira-se no trabalho de Ananthakrishnan e Narayanan (2008), no qual os autores exploraram três classificadores de aprendizado de máquina: um LDA (*Linear Discriminant Analysis*), um GMM (*Gaussian Mixture Model*), e uma rede neural, baseando sua abordagem de identificação de fronteiras de frases prosódicas[19] em informações acústicas, mas também na combinação de informações sintáticas e lexicais. Os dados utilizados foram parte do *Boston University Radio Speech Corpus*, totalizando aproximadamente três horas de fala espontânea em inglês americano, balanceada em termos de gênero dos falantes. Os autores extraíram as seguintes informações acústicas de cada sílaba: (i) duração de pausas imediatamente após as sílabas; (ii) duração da vogal núcleo; (iii) diferença da F0 mínima e máxima; (iv) diferença da F0 mínima e média da sílaba; (v) diferença da F0 média e máxima da sílaba; (vi) diferença da energia mínima e média da sílaba; (vii) diferença da energia média e máxima da sílaba; (viii) diferença da energia máxima e mínima; e (ix) diferença da média de F0 da sílaba e da média de F0 do enunciado falado (*spoken utterance* nas palavras dos autores). Com seu classificador baseado em rede neural, obtiveram acurácia de 91,6% com a abordagem que utiliza informações sintáticas e acústicas, e 89,9% com a abordagem que se vale apenas de informações acústicas. Contudo, o parâmetro relativo à diferença da F0 média da sílaba e da F0 média do enunciado depende de conhecimento prévio das posições de fronteiras prosódicas para ser utilizado, já que para calcular a média de F0 de um enunciado é preciso saber quais sílabas pertencem a esse enunciado. Tal conjunto de parâmetros é necessário tanto para treino quanto para teste, já que o classificador depende dessas informações acústicas dos dados de teste para realizar as previsões de posição das fronteiras prosódicas.

Por fim, Craveiro *et al*. (2025) basearam-se na extração das mesmas informações acústicas de Ananthakrishnan e Narayanan (2008), mas disponibilizando também um modelo que considera apenas as oito informações que não requerem uma anotação prévia das fronteiras prosódicas. As autoras adaptaram a abordagem para a identificação somente de fronteiras prosódicas terminais em PB, utilizando o alinhador fonético forçado *UFPAlign* e o corpus MuPe-Diversidades[20] (Craveiro;

18 https://github.com/nilc-nlp/ProsSegue/tree/main/baseline%20approach

19 Termo utilizado pelos autores, cuja função é agrupar um conjunto de unidades semânticas presentes em um enunciado. Tais unidades são divididas em dois tipos de categorias: fronteira fraca de sintagma intermediário e fronteira "completa" de sintagma entoacional. De acordo com (Ananthakrishnan *et al*., 2008, p. 217): "*Prosodic phrase boundaries serve to group together semantic units in the utterance. These are divided in two coarse categories, weak intermediate phrase boundaries and full intonational phrase boundaries*".

20 https://github.com/nilc-nlp/MuPe-Diversidades/

Galdino, 2024), apresentando também uma avaliação de viés, considerando o perfil de falante, definido por gênero, região de origem, idade e nível de escolaridade. O corpus MuPe-Diversidades permite tal avaliação, pois contém fala espontânea de um conjunto de 30 falantes, balanceados em termos de estado de origem (17 estados brasileiros estão contidos) e gênero, e englobando diferentes idades e níveis de escolaridade, totalizando aproximadamente duas horas e meia de áudio. Na avaliação de viés, as autoras compararam a performance do modelo para cada grupo de falantes separados por aspecto de perfil de falante, verificando se ele é igualmente eficaz para cada grupo. Através de uma validação cruzada (K-fold=5), Craveiro *et al*. (2025) testaram sete classificadores, optando por um *Random Forest*. As autoras reportaram: medida f1 de 77%, medida f1 binária de 55% e acurácia de 97%, além de um total de apenas 8,2g de emissão de carbono, medido com a biblioteca de python codecarbon[21]. Durante os últimos anos, houve expressiva popularização de métodos de aprendizado de máquina profundo e modelos de linguagem, os quais exigem poder computacional massivo e, consequentemente, geram significativo impacto ambiental (Bender *et al*., 2021; Ferraro *et al*., 2024). Por isso, Craveiro *et al*. (2025) enfatizam que têm a preocupação de trabalhar com um modelo energeticamente eficiente. As autoras reportaram, também, que a avaliação de vieses teve resultados inconclusivos. Os modelos elaborados no estudo e o código[22] do método estão disponíveis publicamente.

O objetivo do trabalho que iremos desenvolver e ao qual se refere este relato registrado é replicar a abordagem descrita em Craveiro *et al.* (2025), que se baseia exclusivamente em informações acústicas, utiliza aprendizado de máquina tradicional e considera fala espontânea. Esse estudo inova ao treinar um modelo de segmentação prosódica automática para o PB baseado em duas horas e meia de fala por pessoas de perfis relativamente diversos, levando em conta variáveis como gênero, idade, nível de escolaridade e região de origem dos falantes. Ademais, Craveiro *et al*. (2025) foi selecionado para replicação pois o código do classificador e os modelos treinados estão disponíveis publicamente, permitindo sua reprodução. O estudo que será por nós realizado tem a intenção de avaliar a robustez da abordagem proposta em Craveiro *et al*. (2025), através de sua aplicação em um novo conjunto de dados, o corpus NURC-CM (Santos *et al*., 2022), especificado na seção "1. Métodos". Esse corpus foi selecionado pela significativa quantidade de horas com anotação de segmentação prosódica (17 h 35 min 19 s), apesar de as gravações terem sido feitas na década de 70, implicando baixa qualidade em alguns dos áudios. Considerando esse intuito, temos a seguinte pergunta de pesquisa: resultados semelhantes quanto à segmentação automática das unidades de fala são obtidos com o mesmo segmentador em amostras de dados de corpus de fala diferentes do português brasileiro?

21 https://codecarbon.io/
22 www.github.com/nilc-nlp/ProsSegue

# 1. MÉTODOS

A fim de testar a eficiência do segmentador prosódico automático em um corpus diferente daquele utilizado no estudo original, usaremos os dados do Corpus Mínimo CORAA NURC-SP. Esse corpus é um subcorpus do NURC-SP, sendo um recurso em português brasileiro que fornece:

- 21 arquivos de áudio (*.wav*, 2 canais, 16 bits, 48 kHz), totalizando 17 h 35 min 19 s, 155.394 palavras, com seis monólogos, classificados como elocuções formais (EF) (4 h 28 min 52 s, 29.607 palavras), seis diálogos entre dois informantes (D2) (6 h 55 min 07 s, 71.350 palavras) e nove diálogos entre informante e entrevistador (DID) (6 h 11 min 20 s, 54.437 palavras).
- Arquivos texto alinhados à fala (*.textgrid*, UTF-8), contendo as seguintes camadas de intervalos anotadas no software Praat (Boersma; Weenink, 2025), conforme ilustrado na Figura 1[23]:


**Figura 1.** Excerto do inquérito SP_EF_153 com cinco camadas anotadas no *Praat*. **Fonte**: Adaptado de Santos *et al.* (2022).

- 2 camadas (TB-, NTB-) nas quais a fala de cada locutor (-L1, -L2) e documentador (-DOC1, -DOC2) é segmentada em unidades prosódicas e transcrita de acordo com as normas do Projeto NURC.
- 1 camada (LA) para a fala transcrita e segmentada de locutores eventuais[24].
- 2 camadas para comentários: uma voltada a observações gerais sobre o áudio (com), como qualidade sonora, presença de ruídos ou trechos inaudíveis; e outra destinada a anotações

23 Na Figura 1, não estão representadas a camada LA, uma vez que não houve locutor eventual nesse áudio, nem a camada de comentários com anotações temporárias (com-anotadores), já excluída na etapa do trabalho em que a imagem foi gerada.

24 Participantes que não integram o conjunto principal de locutores da gravação (L1, L2) nem exercem o papel de documentador, mas que podem, ocasionalmente, intervir de forma pontual e não sistemática na gravação, geralmente por meio de breves comentários.

temporárias dos anotadores (com-anotadores), utilizadas para o registro de dúvidas, decisões analíticas provisórias e observações metodológicas, sendo excluídas ao final da anotação.

- 1 camada contendo a versão normalizada (-normal)[25] da transcrição de todas as camadas TB e LA[26].
- 1 camada contendo a pontuação (-ponto) que marca a fronteira final de cada TB.

- Arquivo de metadados (*.csv*) associados a cada inquérito, contendo informações relativas ao inquérito (ID, duração e qualidade do áudio, data e tema da gravação) e aos principais locutores (ID, sexo, idade, faixa etária, estado civil, ocupação e locais de origem do participante e de seus pais).

O conjunto de dados está disponível publicamente no repositório Portulan Clarin[27], sob a licença CC BY-NC-ND 4.0. O corpus compreende pelo menos 55 falantes distintos: sendo 27 informantes principais, 23 documentadores e 5 falantes eventuais. Entre os locutores principais, há 14 homens e 13 mulheres, com idades variando de 25 a 85 anos (média = 44; desvio padrão = 16,8). Todos são naturais da cidade de São Paulo, com exceção de dois participantes que nasceram em outras cidades e se mudaram para São Paulo ainda jovens. Esses falantes pertencem a diversas áreas profissionais, incluindo: advocacia, biblioteconomia, docência, economia, engenharia, estatística, jornalismo, nutrição, odontologia, pedagogia, psicologia, publicidade e vendas.

O conjunto de dados contém gravações datadas de dezembro de 1971 a maio de 1977. Além de aulas e palestras gravadas (sobre língua, cinema, estética, desenvolvimento intelectual, dinheiro, arte pré-histórica), o corpus contém conversas sobre uma ampla variedade de tópicos, como família, saúde, alimentação, tempo, vestuário, profissão, educação, religião, transporte e viagens, entretenimento, cinema, televisão, rádio e teatro, telecomunicações, finanças, casa, terreno, vegetais, agricultura, animais e gado.

As gravações originais foram capturadas com gravadores de rolo, como *National* RQ 501s, *Sony Tapecorder* TC-105 e AKAI 707, e ocorreram em diferentes locais. Assim, os arquivos de áudio digitalizados atuais possuem diferentes níveis de inteligibilidade como resultado do equipamento de gravação utilizado, do ambiente de gravação ou da deterioração das fitas de gravação. No arquivo de metadados, são fornecidos comentários sobre (i) o volume do áudio percebido pelos comentaristas e (ii) a qualidade das gravações em relação à voz dos locutores e à presença de

25 Nas camadas TB e NTB, há marcações de pontuação indicando, por exemplo, silêncios marcados por "..." ou risadas marcadas por "((risos))". Nas camadas normalizadas, não há esse tipo de marcação.

26 Não há camadas normalizadas referentes às NTBs.

27 https://hdl.handle.net/21.11129/0000-000F-73CA-C

eventos externos[28] (como chiado, ruído de fundo ou música, interferência aleatória dos locutores), consistindo em descrições positivas (bom, muito bom, audível, claro) e negativas (baixo, muito baixo, grave, ruidoso). Assim, há 10 arquivos de áudio com avaliação positiva, 6 com avaliação negativa e 5 com avaliação mista. É interessante incluir os áudios com avaliação negativa, pois áudios gravados em condições cotidianas podem não refletir alta qualidade, e desse modo é possível também avaliar a performance do modelo em condições não ideais de qualidade de áudio.

No estudo, utilizaremos apenas as camadas de TB normalizadas de cada falante. Antes de iniciar os experimentos no modelo de segmentação prosódica automática, os dados serão pré-processados, tendo em vista os procedimentos relatados no estudo original. Converteremos os áudios em 16 kHz, em sinal monofônico e no formato *.wav*. No experimento original, os dados do MuPe-Diversidades tinham exemplos de cinco a dez minutos. Para a replicabilidade a ser realizada em nosso estudo, adaptações serão necessárias, uma vez que o NURC-Corpus Mínimo possui áudios mais longos, mas faremos uso do mesmo alinhador fonético utilizado no estudo original, que não processa áudios muito longos. Visto que para uma boa generalização do classificador, o ideal é utilizar a maior quantidade de dados disponíveis para treinamento, o que também contribui com a avaliação de robustez do método, optamos por processar o corpus inteiro. Assim, verificaremos se o modelo suporta processar os áudios com duração de aproximadamente 20 minutos. Desse modo, dividiremos os inquéritos do NURC-Corpus Mínimo nessa estimativa. As tentativas podem levar a possíveis limitações do alinhador fonético, o que pode resultar em cortes do corpus em trechos ainda menores, de 5 a 10 minutos. Os cortes serão feitos manualmente, de forma que não haja interrupções inadequadas de fala, como, por exemplo, no meio das palavras.

Cada uma das transcrições dos áudios precisará conter palavras separadas por um único espaço, sem sobreposições e sinais de pontuação. Para esta normalização, será empregado o mesmo *script*[29] utilizado no estudo original. Cada transcrição será alinhada manualmente com os arquivos de áudio do NURC-Corpus Mínimo dividido em trechos menores.

Após o pré-processamento, os dados estarão prontos para serem testados no modelo. O experimento será realizado em três etapas: (i) alinhamento fonético com *UFPAlign*; (ii) extração de informações prosódicas; (iii) segmentação prosódica automática do áudio com o classificador de Craveiro et al. 2025, já treinado em PB. Na primeira etapa, o alinhamento fonético vai marcar o tempo inicial e final de cada fone, sílaba e palavra dos dados. Na segunda, as informações prosódicas ((i) duração de pausas imediatamente após as sílabas; (ii) duração da vogal núcleo; (iii) diferença da F0 mínima e máxima; (iv) diferença da F0 mínima e média da sílaba; (v) diferença da F0 média e máxima da sílaba; (vi) diferença da energia mínima e média da sílaba; (vii) diferença da energia média

28 No arquivo de metadados não há informação sobre razão sinal-ruído.
29 https://github.com/nilc-nlp/ProsSegue/blob/main/utils/textgridToCleanTxt.py

e máxima da sílaba; (viii) diferença da energia máxima e mínima; e (ix) diferença da média de F0 da sílaba e da média de F0 do enunciado falado) serão extraídas com o auxílio da biblioteca *Parselmouth*[30]. Finalmente, a segmentação prosódica automática será executada, usando o modelo *Random Forest* treinado no estudo original. Assim como no estudo original, o produto do método empregado será um *textgrid* com diferentes intervalos, segmentado em unidades, separadas prosodicamente pelo classificador.

As ferramentas que serão utilizadas incluem: *UFPAlign* para o alinhamento fonético forçado; *Python* (com bibliotecas *sklearn*, *tgt*, *parselmouth*, pandas, *scipy*, entre outras) para a segmentação e extração de informações acústicas; e Praat para análise linguística qualitativa dos resultados. Usaremos exatamente as mesmas técnicas estatísticas do trabalho original (Craveiro *et al*. 2025), pois o estudo que será replicado é muito recente.

Em relação aos critérios de inclusão e exclusão, todos os áudios que passarem automaticamente na fase de pré-processamento e de processamento do modelo serão analisados. Caso os dados não sejam processados pelo modelo em alguma das duas etapas, serão excluídos e listados no artigo final.[31]

Quanto à análise dos resultados obtidos, utilizaremos as mesmas métricas do estudo original, considerando a segmentação do corpus inteiro e também os resultados obtidos em grupos específicos divididos por idade e gênero. Avaliaremos a relevância estatística desses resultados e a relevância das informações prosódicas utilizadas pelo classificador. Nessa replicação, apenas escolaridade e região de origem, duas das variáveis de análise do estudo original não serão contempladas, uma vez que a nova amostra (NURC-CM) não apresenta a possibilidade de verificar esses vieses, porque os dados são apenas de uma região de origem (SP) e relativos a falantes com grau de escolaridade superior completo. Será também realizada a análise linguística qualitativa, com base em inspeção de parâmetros acústicos, dos erros da segmentação automática aplicada no presente estudo, de forma comparativa aos erros obtidos no estudo de Craveiro *et al.* (2025)[32].

O experimento será aplicado de forma automatizada. O método analítico previsto irá averiguar a relevância estatística com *one-way* ANOVA da biblioteca *SciPy* e analisar, de forma comparativa aos resultados observados no corpus original, os resultados das medidas f1 binária e f1 macro com

30 https://parselmouth.readthedocs.io/en/latest/api/parselmouth.html

31 Caso o alinhamento fonético forçado ou a extração de features não sejam concluídos com sucesso, é impossível utilizar o classificador para a previsão das fronteiras prosódicas, já que não teremos as informações prosódicas de cada sílaba, seja por não saber quando cada sílaba se inicia e se finaliza, ou por não termos conseguido extrair suas informações acústicas.

32 Ressaltamos que, devido a critérios de delimitação, Craveiro *et al.* (2025) não fizeram uma análise linguística qualitativa de seus resultados. Diante disso, propomos, neste trabalho, realizar essa análise qualitativa dos resultados obtidos previamente pelas autoras, a fim de compará-los aos novos resultados alcançados, o que nos permite testar, assim, por outros critérios (qualitativos), a replicabilidade do método apresentado.

*sklearn.metrics* obtidos no novo corpus. Enquanto a f1 binária[33] considera apenas os valores de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos para a classe TB, a f1 macro considera uma média entre os valores da classe TB e da classe NB, uma classe secundária que indica todas as sílabas que não estão imediatamente anteriores a fronteiras. Utilizaremos o código aberto disponibilizado pelas autoras para esses cálculos. Uma reprodução bem-sucedida ocorrerá caso:

(i) a replicação do método e geração de arquivos segmentados prosodicamente seja concluída com sucesso, ou seja, sem erros de algoritmo que inviabilizem a conclusão do processo; e

(ii) com alcance de resultados que atinjam padrões quantitativos semelhantes aos observados no corpus original, incluindo f1 binária acima de 50% e f1 macro acima de 70%.

Em relação ao cronograma aproximado de atividades, pretendemos iniciar os experimentos imediatamente após a aprovação desta fase inicial do relato registrado. O pré-processamento, o processamento e a análise dos dados serão executados em aproximadamente 8 semanas. A escrita final será concluída dentro de 4 semanas. Assim, a finalização do relatório está prevista para cerca de 3 meses após o recebimento da aprovação deste presente estágio 1.

## INFORMAÇÕES COMPLEMENTARES

### CONFLITO DE INTERESSE

Os autores declaram que não possuem interesses financeiros ou relações pessoais que possam ter influenciado o trabalho relatado neste artigo.

### DECLARAÇÃO DE DISPONIBILIDADE DE DADOS

O compartilhamento de dados não é aplicável a este artigo, pois nenhum dado novo foi criado ou analisado neste estudo.

### DECLARAÇÃO DE USO DE IA

Os autores declaram que nenhuma ferramenta de IA foi utilizada na criação deste manuscrito nem em qualquer aspecto dos trabalhos realizados cujo resultado será reportado no manuscrito.

33 Conforme descrito anteriormente, a fórmula para o cálculo de medida f1 se dá multiplicando a precisão pela revocação, e multiplicando esse resultado por 2: 2*precisão*revocação).

## AVALIAÇÃO E RESPOSTA DOS AUTORES

Avaliação: https://doi.org/10.25189/2675-4916.2026.V7.N1.ID912.R

Resposta dos Autores: https://doi.org/10.25189/2675-4916.2026.V7.N1.ID912.A

## REFERÊNCIAS

ANANTHAKRISHNAN, S.; NARAYANAN, S. S. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 16, n. 1, p. 216-228, 2008. DOI: http://dx.doi.org/10.1109/TASL.2007.907570. Acesso em: 27 out. 2025.

BATISTA, C.; DIAS, A. L.; NETO, N. Free resources for forced phonetic alignment in Brazilian Portuguese based on Kaldi toolkit. *EURASIP Journal on Advances in Signal Processing*, v. 1, n. 11, 2022. DOI: https://doi.org/10.1186/s13634-022-00844-9. Acesso em: 27 out. 2025.

BENDER, E. M.; GEBRU, T.; MCMILLAN-MAJOR, A.; SHMITCHELL, S. On the dangers of stochastic parrots: Can language models be too big? *In*: 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY. *Proceedings of [...]*. Online, 2021. p. 610-623. DOI: https://doi.org/10.1145/3442188.3445922. Acesso em: 8 fev. 2026.

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*, v. 16, n. 5, p. 1-21, 2021. DOI: https://doi.org/10.1371/journal.pone.0250969. Acesso em: 27 out. 2025.

BOERSMA, P.; WEENINK, D. *Praat*: doing phonetics by computer [Computer program]. Version 2025: University of Amsterdam, 2025. Disponível em: https://www.fon.hum.uva.nl/praat/. Acesso em: 27 out. 2025.

CHEN, K.; HASEGAWA-JOHNSON, M. A. How prosody improves word recognition. *In:* ISCA INTERNATIONAL CONFERENCE ON SPEECH PROSODY 2004. *Proceedings of [...]*. Nara, Japan, 2004. p. 583-586. DOI: http://dx.doi.org/10.21437/SpeechProsody.2004-134. Acesso em: 27 out. 2025.

CRAVEIRO, G. M.; ALVES, C. A.; SVARTMAN, F. R. F.; ALUÍSIO, S. M. Machine Learning Classifiers with Acoustic Features for Prosodic Segmentation in Brazilian Portuguese: A Comprehensive Evaluation. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 2025, Fortaleza/CE. *Anais [...]*. Porto Alegre: Sociedade Brasileira de Computação, 2025. p. 113-124. DOI https://doi.org/10.5753/stil.2025.37818. Acesso em: 27 out. 2025.

CRAVEIRO, G. M.; GALDINO, J. C. Diversity in Data for Speech Processing in Brazilian Portuguese. *In*: PAES A.; VERRI, F. A. N. (eds.) *Intelligent Systems. BRACIS 2024*. Lecture Notes in Computer Science, v. 15415. Springer, Cham. DOI: https://doi.org/10.1007/978-3-031-79038-6_9. Acesso em: 31 out. 2025.

CRAVEIRO, G. M.; SANTOS, V. G.; DALALANA, G. J. P.; SVARTMAN, F. R. F.; ALUÍSIO, S. M. Simple and fast automatic prosodic segmentation of Brazilian Portuguese spontaneous speech. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE, Santiago de Compostela. Proceedings of the 16th International Conference on Computational Processing of Portuguese - v. 1. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, p. 32-44, 2024. Disponível em: https://aclanthology.org/2024.propor-1.4/. Acesso em: 27 out. 2025.

FERRARO, V. R.; GULLO, G.; DA SILVA COSTA, D.; MOURA, P. N. D. S. Aprendizagem Profunda e Inteligência Artificial Verde: Caminhos para um Futuro mais Sustentável. *In*: WORKSHOP DE COMPUTAÇÃO APLICADA À GESTÃO DO MEIO AMBIENTE E RECURSOS NATURAIS (WCAMA). SBC, 2024. p. 159-168. DOI: https://doi.org/10.5753/wcama.2024.3033. Acesso em: 8 fev. 2025.

HOI, L. M.; SUN, Y.; IM, S. K. An automatic speech segmentation algorithm of Portuguese based on spectrogram windowing. *In*: IEEE WORLD AI IOT CONGRESS (AIIoT), 2022, Seattle. 2022 IEEE World AI IoT Congress (AIIoT): IEEE, p. 290-295, 2022. DOI: http://dx.doi.org/10.1109/AIIoT54504.2022.9817299. Acesso em: 27 out. 2025.

KOCHAROV, D.; KACHKOVSKAIA, T.; SKRELIN, P. Eliciting Meaningful Units from Speech. *In*: INTERSPEECH, p. 2128-2132, 2017. DOI: http://dx.doi.org/10.21437/Interspeech.2017-855. Acesso em: 27 out. 2025.

LIN, C.-H.; YOU, C.-L.; CHIANG, C.-Y.; WANG, Y.-R.; CHEN, S.-H. Hierarchical prosody modeling for Mandarin spontaneous speech. *The Journal of the Acoustical Society of America*, v. 145, n. 4, p. 2576-2596, 2019. DOI https://doi.org/10.1121/1.5099263. Acesso em: 27 out. 2025.

LIU, S.; NAKAJIMA, Y.; CHEN, L.; ARNDT, S.; KAKIZOE, M.; ELLIOTT, M. A.; REMIJN, G. B. How pause duration influences impressions of English speech: Comparison between native and non-native speakers. *Frontiers in Psychology*, v. 13, 2022. DOI https://doi.org/10.3389/fpsyg.2022.778018. Acesso em: 27 out. 2025.

RADFORD, A.; KIM, J. W.; XU, T.; BROCKMAN, G.; MCLEAVEY, C.; SUTSKEVER, I. Robust speech recognition via large-scale weak supervision. In: *INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML)*, 40., 2023, Honolulu. Proceedings of the 40th International Conference on Machine Learning (ICML 2023). Honolulu, Hawaii, USA: JMLR. org, 2023. Artigo n. 1182, p. 28492-28518. Disponível em: https://proceedings.mlr.press/v202/radford23a.html. Acesso em: 8 fev. 2026.

RASO, T.; TEIXEIRA, B.; BARBOSA, P. Modelling automatic detection of prosodic boundaries for Brazilian Portuguese spontaneous speech. *Journal of Speech Sciences*, v. 9, p. 105-128, 2020. DOI: http://dx.doi.org/10.20396/joss.v9i00.14957. Acesso em: 27 out. 2025.

ROLL, N.; GRAHAM, C.; TODD, S. Psst! prosodic speech segmentation with transformers. In: *CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING (CoNLL)*, Singapore. Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL). Singapore: Association for Computational Linguistics, p. 476-487, 2023. DOI https://doi.org/10.18653/v1/2023.conll-1.31. Acesso em: 27 out. 2025.

SANTOS, V. G.; ALVES, C. A.; CARLOTTO, B. B.; DIAS, B. A. P.; GRIS, L. R. S.; IZAIAS, R. L.; MORAIS, M. L. A.; OLIVEIRA, P. M.; SICOLI, R.; SVARTMAN, F. R. F.; LEITE, M. Q.; ALUÍSIO, S. M. CORAA NURC-SP Minimal Corpus: a manually annotated corpus of Brazilian Portuguese spontaneous speech. In: IBER SPEECH, 2022. *Proceedings of IberSPEECH 2022*. p. 161-165, 2022. DOI: https://doi.org/10.21437/IberSPEECH.2022-33. Acesso em: 29 out. 2025.

SERRA, C. R. *Realização e percepção de fronteiras prosódicas no português do Brasil*: fala espontânea e leitura. 2009. Tese (Doutorado em Linguística) - Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2009.

TEIXEIRA, B. H. F. Detecção automática de fronteiras prosódicas na fala espontânea. 2022. Tese (Doutorado em Estudos Linguísticos) - Universidade Federal de Minas Gerais, Minas Gerais, 2022. Disponível em: https://hdl.handle.net/1843/47273. Acesso em 5 fev. 2026.

VIOLA, I. C.; MADUREIRA, S. The roles of pause in speech expression. In: SPEECH PROSODY, Campinas. Speech Prosody, p. 721-724, 2008. DOI http://dx.doi.org/10.21437/SpeechProsody.2008-160. Acesso em: 27 out. 2025.