RELATÓRIO

ANÁLISE COMPUTACIONAL

O que o *text as data* *não diz*

Uma análise multimodal do confronto: Aécio Neves × Renan Calheiros no Senado Federal

VINCIUS SANTOS

88%	36%	34:1
dos momentos classificados pelo texto como “defesa argumentativa”	de amplitude na intensidade vocal entre o mínimo e o máximo	de variação na velocidade dos gestos ao longo do trecho

Corpus · Senado Federal · 137,1 s · 49 segmentos

Vídeo analisado · youtube.com/watch?v=2YXbSc88HRs

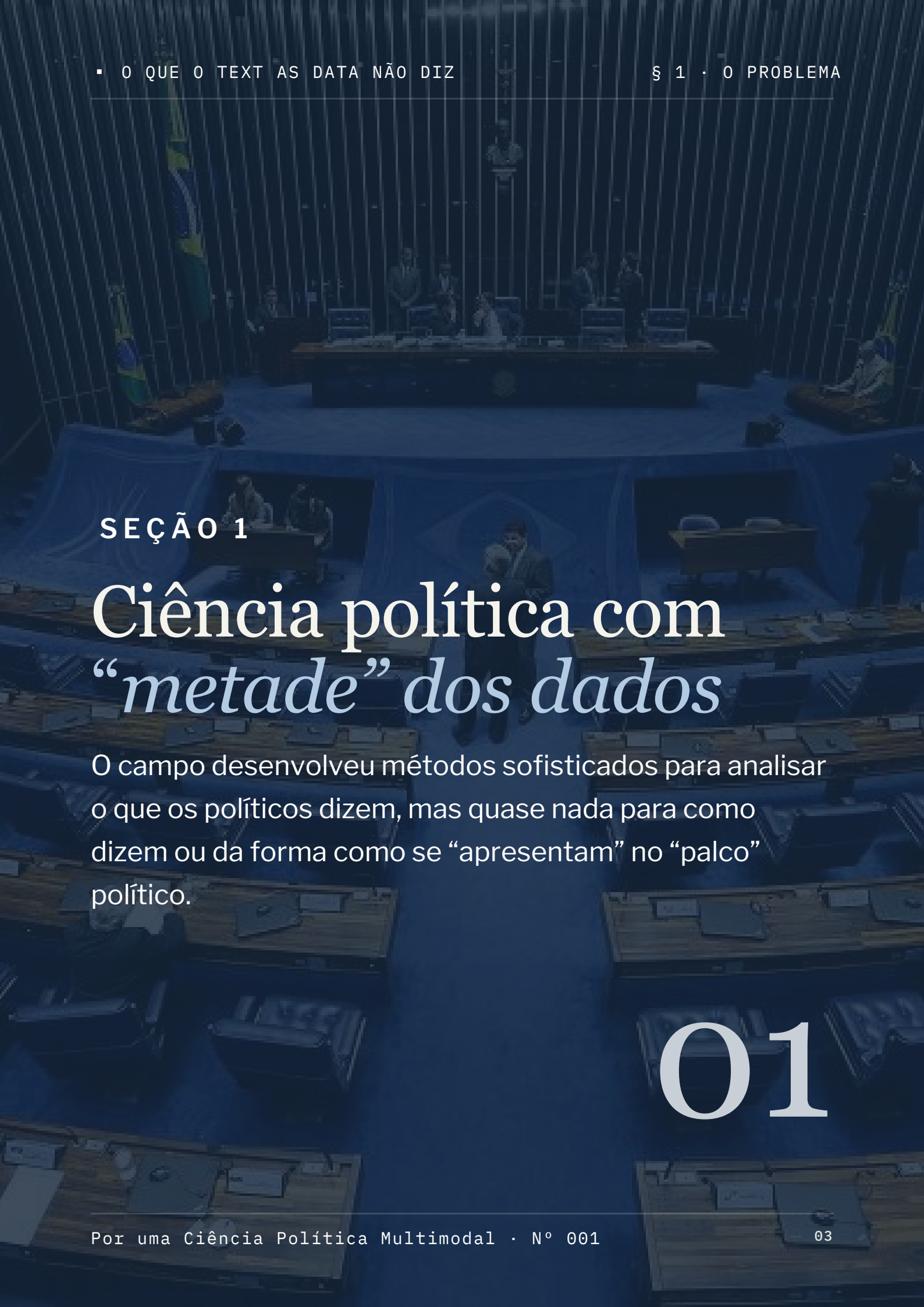
RESUMO EXECUTIVO

A análise textual oferece uma *representação parcial* de eventos cuja natureza *excede* o que *pode ser expresso em palavras*

Este relatório analisa um debate político brasileiro usando **três canais de informação simultâneos**: texto transcrito, sinal acústico e movimentação corporal. A partir de 49 segmentos extraídos de um confronto entre Aécio Neves e Renan Calheiros no Senado Federal, mostramos que a análise textual, método recorrente na ciência política (com tendencias crescentes em abordagens quantitativas), classifica **88% dos momentos** como "defesa argumentativa e resposta formal", produzindo uma imagem essencialmente plana do evento.

Porém, as modalidades vocal e gestual revelam uma realidade diferente: a intensidade vocal varia **36%** entre o mínimo e o máximo do debate; a velocidade dos gestos oscila numa proporção de **34 para 1**; e o segmento de maior intensidade acústica do corpus é classificado pelo texto como defensivo, não como confrontacional.

> A análise computacional multimodal não substitui a interpretação política: ela **recupera a dimensão do debate que o texto sistematicamente apaga**.



SEÇÃO 1

Ciência política com *"metade" dos dados*

O campo desenvolveu métodos sofisticados para analisar o que os políticos dizem, mas quase nada para como dizem ou da forma como se "apresentam" no "palco" político.

01

SETEMBRO DE 1960

A mesma conversa, dois resultados dependendo do o canal monitorado.

Em setembro de 1960, pela primeira vez na história americana, um debate presidencial foi transmitido simultaneamente em rádio e televisão. Os ouvintes de rádio, que processavam apenas o conteúdo verbal, consideraram **Richard Nixon** o vencedor.

Os telespectadores que viam o suor, a rigidez corporal e o contraste visual entre os dois candidatos declararam **John Kennedy** vencedor por margem expressiva. A mesma conversa. Dois resultados opostos. Dependendo de qual canal você monitorava.

> *Sessenta e seis anos depois, a ciência política ainda analisa debates políticos como se fossem rádio.*

O campo desenvolveu métodos sofisticados para o texto: análise de sentimento, modelos de tópicos, embeddings semânticos, classificadores de frame. **Sabe-se muito sobre o que os políticos dizem.** Sabe-se pouco, sistematicamente, sobre como dizem — com que intensidade, com que postura, com que cadência gestual.

O QUE DIZEM × COMO DIZEM

Essa lacuna não é irrelevante.

Estudos de psicologia social demonstram que juízos de competência e liderança formados a partir de expressões **não-verbais** predizem resultados eleitorais com precisão comparável ou superior à análise de conteúdo verbal.

O debate político é, antes de mais nada, um evento multimodal. Analisá-lo por um único canal é uma escolha metodológica que deveria ser **justificada, não assumida**.

Este relatório apresenta uma pipeline de análise computacional que integra três modalidades em paralelo e aplica essa metodologia a um caso real: o confronto entre Aécio Neves e Renan Calheiros no Senado Federal brasileiro.

> Se o **único canal** monitorado é o **verbal**, a conclusão inevitável é a de que pouco de relevante acontece. **As modalidades vocal e gestual contam outra história.**

SEÇÃO 2

O confronto: Aécio × Renan

Um trecho de 137,1 segundos do canal oficial do Senado Federal, transcrito e rastreado quadro a quadro.

02

O MATERIAL ANALISADO

137,1 segundos extraídos do Senado Federal.

O corpus desta análise consiste em um segmento de debate político extraído do canal oficial do Senado Federal no YouTube. O trecho tem **137,1 segundos de duração** (2,3 minutos) e foi transcrito, analisado acusticamente e rastreado por *pose estimation*, resultando em **49 segmentos temporais alinhados** com duração média de 2,8 segundos cada.

137,1s	49	2,8s
Duração do trecho (2,3 min)	Segmentos temporais alinhados	Duração média por segmento

AÉCIO X RENAN

UMA ESCOLHA DELIBERADA

Onde a contenção verbal encontra a expressividade do corpo.

A escolha deste corpus é deliberada. Trata-se de um evento claramente confrontacional, dois senadores com posições institucionais antagônicas em sessão aberta, mas onde o **protocolo formal do Senado** cria pressão constante para que o discurso verbal permaneça dentro de certos limites retóricos.

É exatamente nessa tensão — entre a contenção discursiva exigida pelo ambiente institucional e a expressividade não-verbal que escapa aos rituais de fala: que a análise multimodal tem mais a dizer.

Em termos técnicos: o corpus apresenta **baixa variância semântica no texto** e **alta variância nas modalidades acústica e gestual**. Isso o torna um caso de teste ideal para a hipótese central deste trabalho.

> Texto e sinal físico carregam informações **parcialmente independentes**.

Frame Fonte: youtube.com/watch?v=2YXbSc88HRs

SEÇÃO 3

Quatro camadas de *análise*

Pipeline construída em Python e R, com processamento modular, replicável em novos corpora sem reconfiguração manual.

03

A PIPELINE EM QUATRO ETAPAS

Toda etapa parte das mesmas janelas temporais da transcrição.

A pipeline foi construída em Python e R, com processamento modular por debate, permitindo replicação em novos corpora sem reconfiguração manual. Os timecodes do Whisper definem as janelas temporais; tudo o que vem depois é agregado dentro delas e integrado num único dataframe.

3.1 Texto	**Whisper** Transcrição automática com timecodes de início e fim por segmento: a espinha dorsal do alinhamento.
3.2 Voz	**librosa — envelope acústico · F0** Intensidade vocal (onset strength) e pitch fundamental por janelas de alta resolução.
3.3 Gesto	**MediaPipe — Google · pose 3D** 33 landmarks corporais por quadro a 25 fps: velocidade, expansividade e inclinação.
3.4 · Semântica	**CLAP — LAION · zero-shot** Classificação semântico-acústica do sinal por similaridade entre áudio e texto.

A PIPELINE EM QUATRO ETAPAS

3.1 Transcrição automática — Whisper

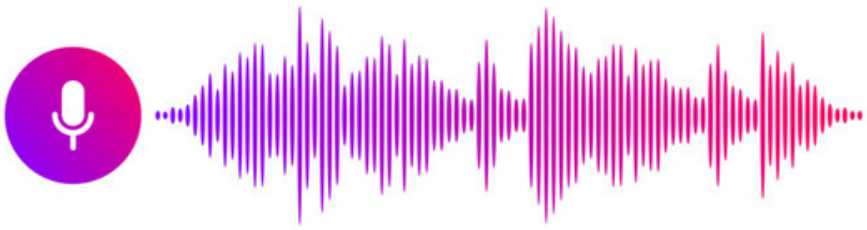
O áudio bruto foi extraído via yt-dlp e transcrito com o modelo Whisper (OpenAI, 2022), que produz não apenas o texto da fala mas também os **timecodes** de início e fim de cada segmento. Esses timecodes são a espinha dorsal do alinhamento: todo o processamento subsequente parte das janelas temporais definidas pela transcrição.

O Whisper opera por reconhecimento acústico neural treinado em **680 mil horas** de áudio multilíngue e produz segmentações mais naturais do que os sistemas clássicos de diarização, ele segmenta por unidade semântica, não apenas por silêncio. Para o debate Aécio X Renan, o modelo base foi utilizado com língua fixada em português.

3.2 Análise acústica — librosa

Paralelamente à transcrição, o mesmo arquivo de áudio foi processado com a biblioteca librosa para extração do **envelope de intensidade vocal** (onset strength) e do pitch fundamental (F0) por janelas de 23 ms com hop de 11 ms. Essa granularidade produz séries temporais de alta resolução, depois agregadas por segmento Whisper.

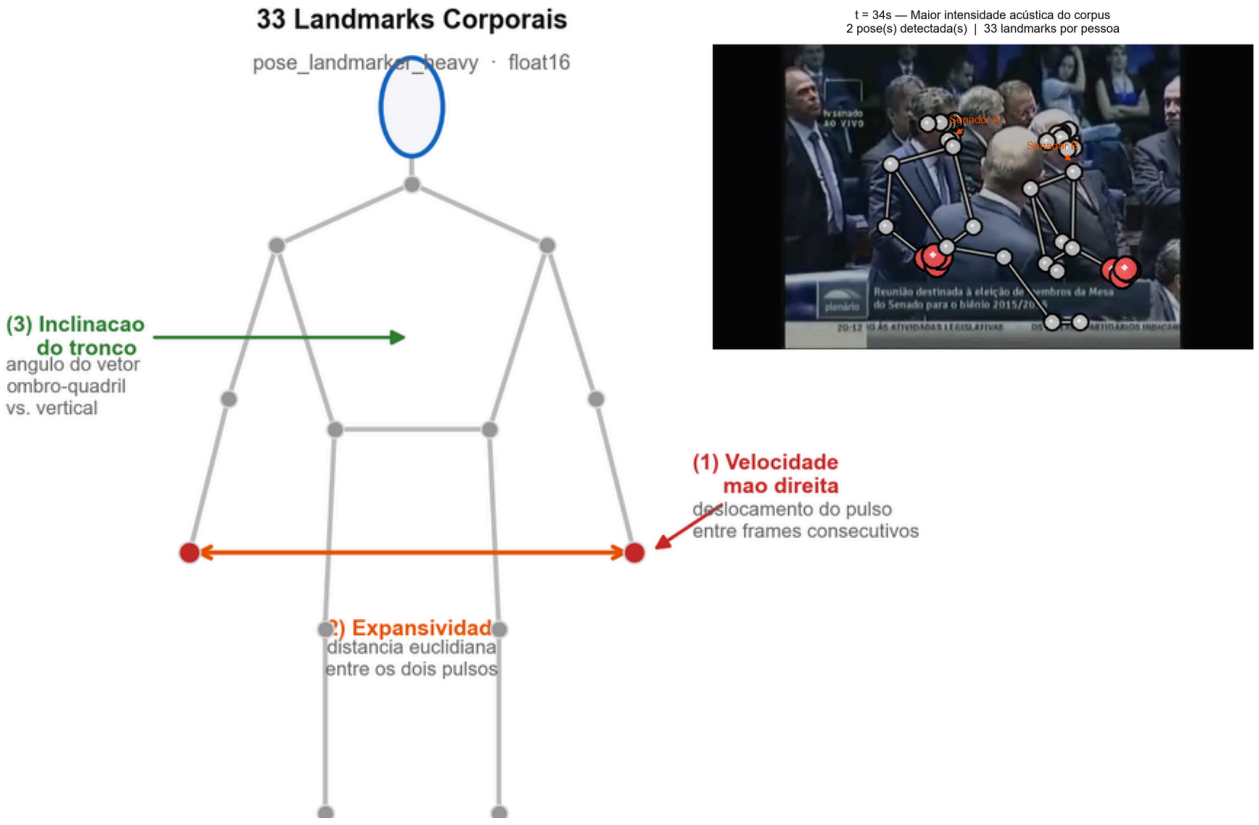
O envelope de onset strength captura o ataque acústico — picos de energia que correspondem a ênfases, acelerações e inflexões vocais. É uma proxy robusta de *arousal* acústico: não informa o conteúdo da fala, mas **informa sua força**.

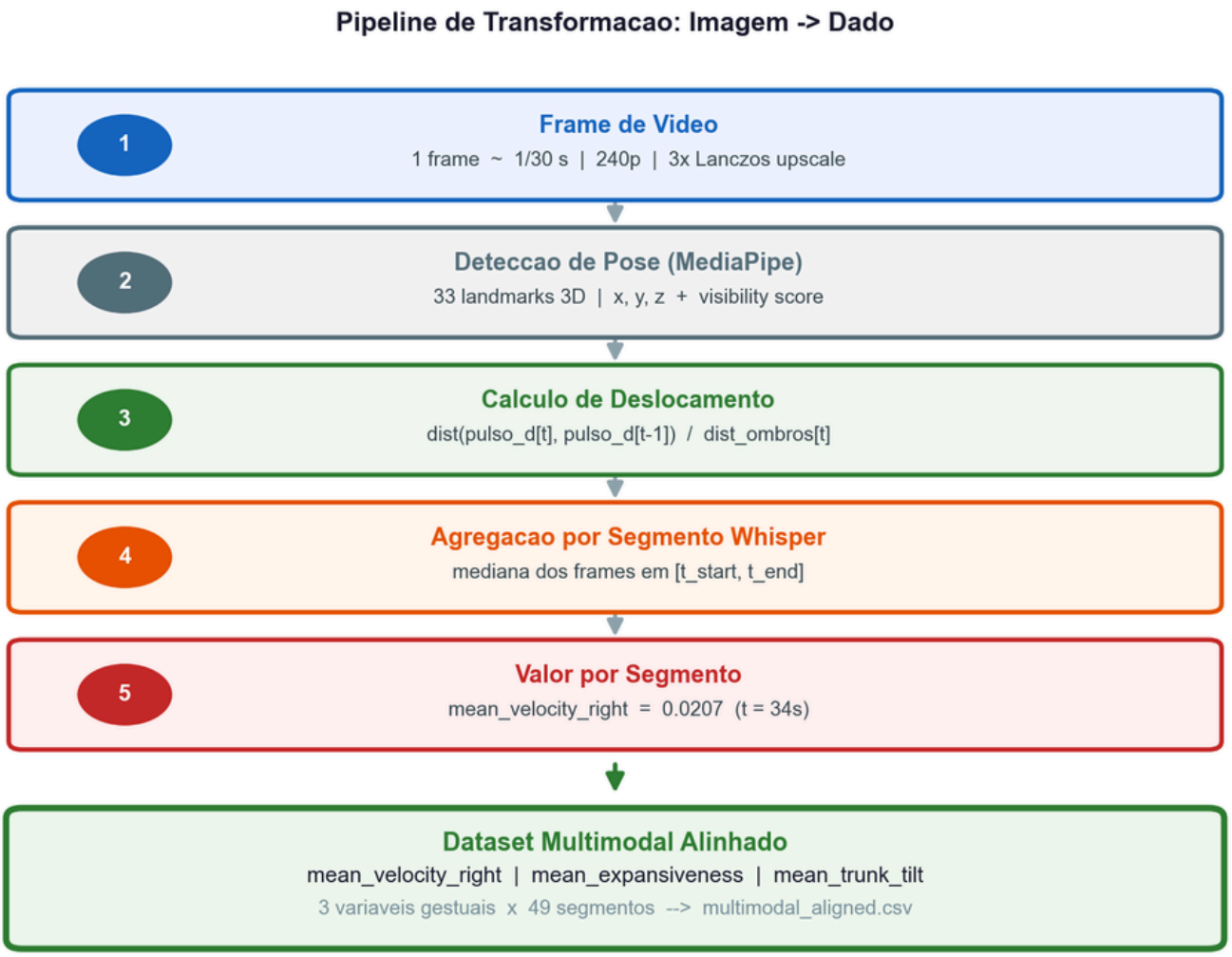


3.3. “Strike a pose” (Rastreamento) — MediaPipe

A dimensão gestual foi extraída com MediaPipe Pose (Google), que estima **33 landmarks** do corpo em cada quadro a 25 fps. Da nuvem de pontos 3D calculamos três indicadores por frame, agregados por mediana dentro de cada janela de segmento:

01	**Velocidade da mão** Deslocamento euclidiano do pulso direito entre frames, normalizado pela distância entre ombros.
02	**Expansividade** Distância tridimensional entre os dois pulsos, capturando a abertura do gesto.
03	**Inclinação do tronco** Ângulo do vetor ombro-ombro em relação ao plano vertical.





3.4. Classificação semântica zero-shot — CLAP

O modelo CLAP (Contrastive Language-Audio Pretraining, LAION 2023) opera num espaço de embeddings conjunto áudio-texto: a similaridade coseno entre o áudio de cada segmento e seis descrições fornece uma pontuação de relevância **sem treinamento supervisionado**. O resultado não analisa o texto transcrito: analisa o sinal sonoro, sensível à prosódia, ao timbre e à intensidade.

Confronto verbal e acusação política	Defesa argumentativa e resposta formal
Discurso técnico e argumentação jurídica	Fala emocional com indignação ou ênfase
Tom conciliatório e retórica persuasiva	Interrupção e pressão sobre o interlocutor

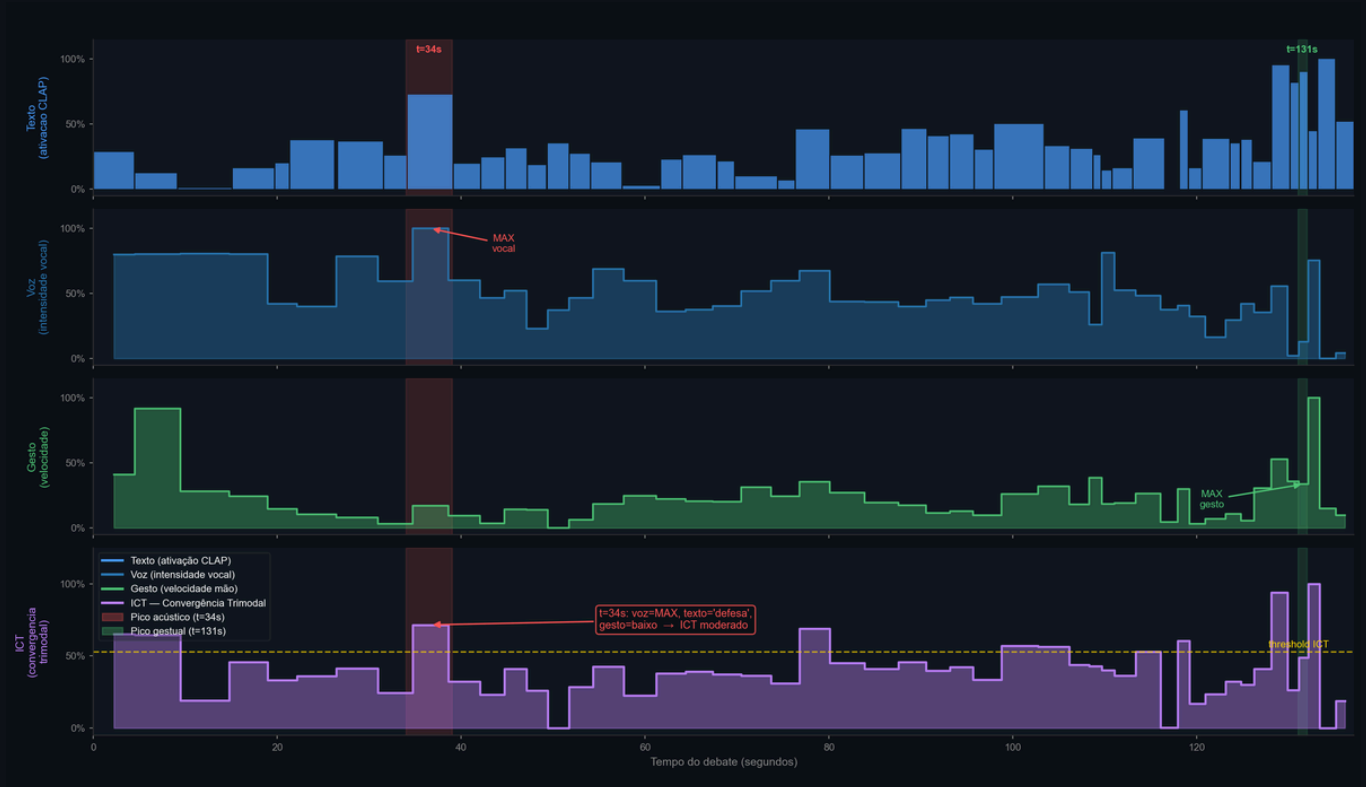
3.5 — ALINHAMENTO TEMPORAL

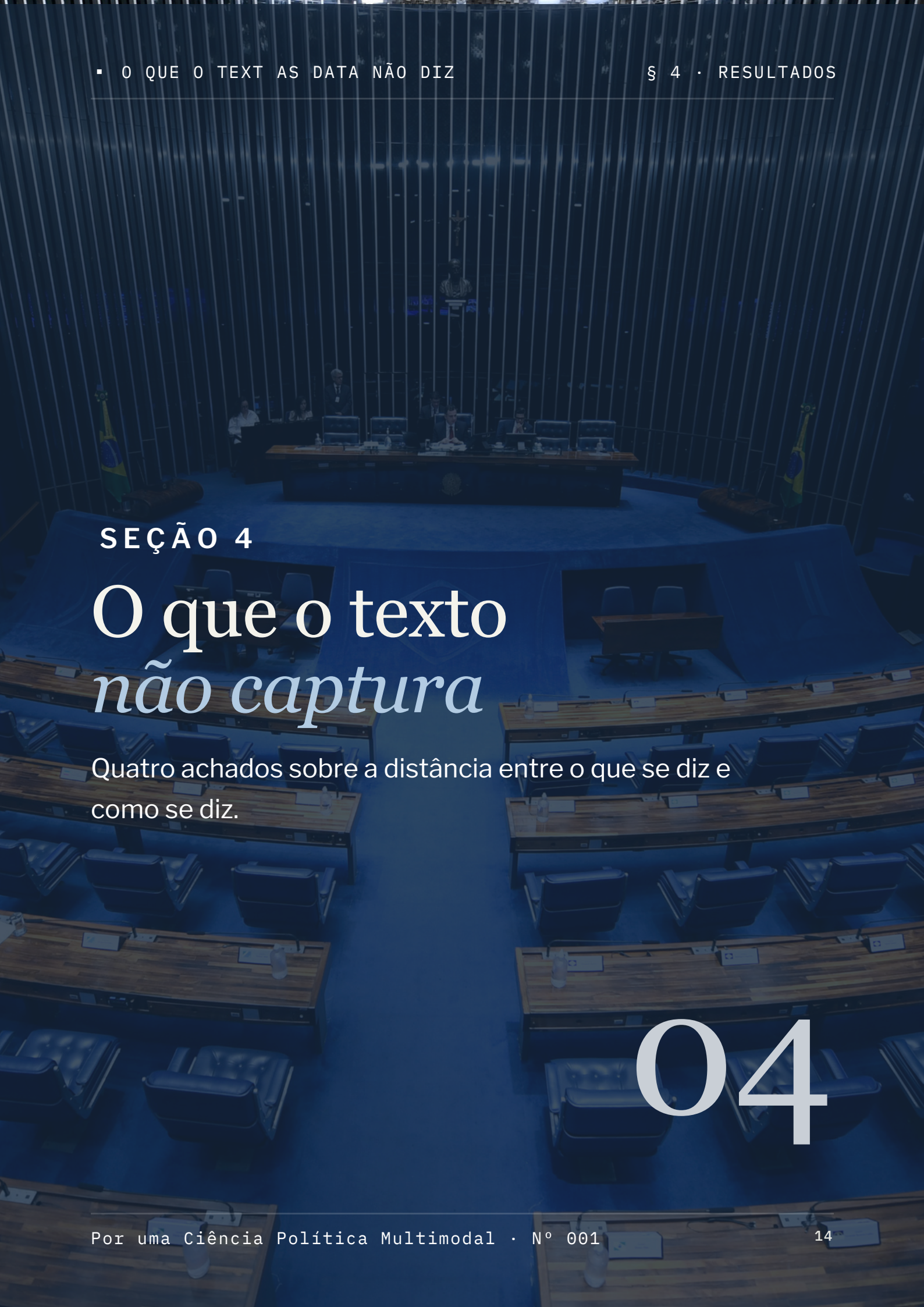
Índice de Convergência Trimodal

As três séries temporais — texto, voz e gesto — são integradas num dataframe de **49 linhas × 31 colunas**, onde cada linha é um segmento Whisper com todos os indicadores alinhados. Adicionalmente, calculamos o Índice de Convergência Trimodal (ICT) como a média geométrica das três modalidades normalizadas:

$$\text{ICT} = \sqrt[3]{\text{audio}_n \times \text{gesture}_n \times \text{text}_n}$$

A média geométrica penaliza qualquer desalinhamento: se uma modalidade está próxima de zero, o ICT colapsa independentemente das outras. Momentos de ICT alto são aqueles em que texto, voz e gesto convergem simultaneamente para alta ativação: os **picos expressivos multimodais** do debate.





SEÇÃO 4

O que o texto *não captura*

Quatro achados sobre a distância entre o que se diz e como se diz.

04

RESULTADOS

4.1 O que o texto captura e o que não captura

A classificação CLAP dos 49 segmentos produz a seguinte distribuição. **88% dos momentos** deste debate são classificados como “defesa argumentativa”: para uma análise convencional baseada em conteúdo, o debate parece um exercício de diplomacia formal — duas posições reafirmadas educadamente.

CATEGORIA SEMÂNTICA (CLAP)	N	%
Defesa argumentativa e resposta formal	43	**88%**
Confronto verbal direto e acusação política	5	**10%**
Fala emocional com indignação ou ênfase	1	**2%**

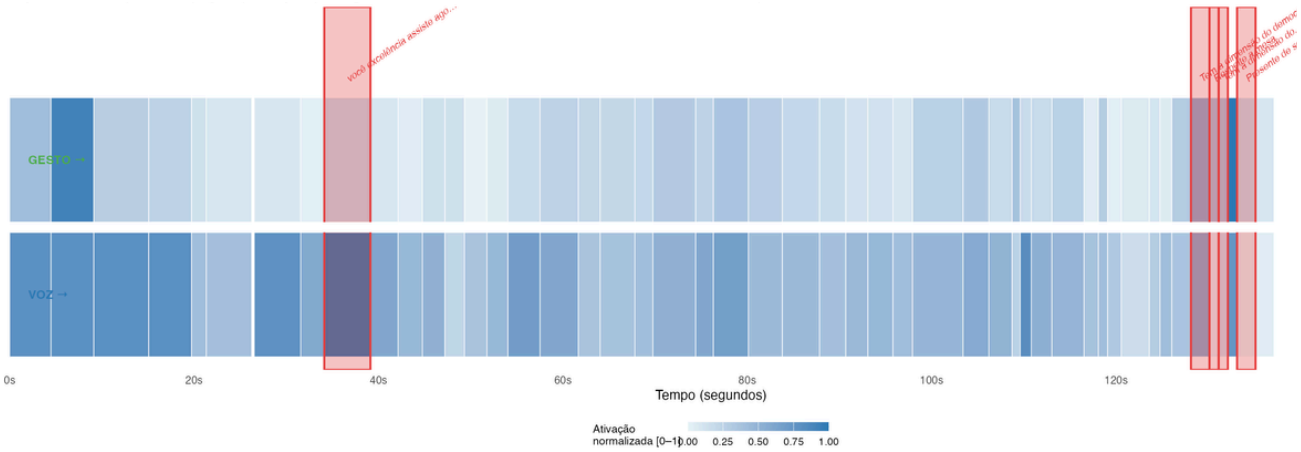


Figura 1 · faixa superior — A categorização CLAP produz uma imagem quase monocromática.

Esse resultado não é uma falha do método: é precisamente o que o texto diz. A gramática do discurso parlamentar é estruturalmente defensiva. **O texto está certo; o desafio é que ele está incompleto.**

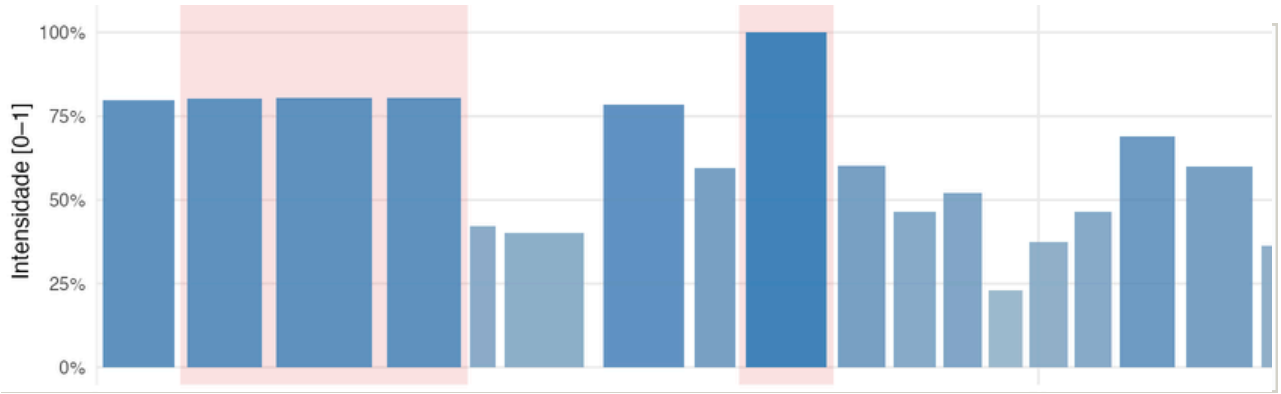
RESULTADOS

4.2 O que a voz revela

A intensidade vocal média ao longo do debate é **2,10 unidades** (escala de onset strength librosa), com desvio padrão de 0,16. A amplitude total, entre o momento mais silencioso e o mais intenso: corresponde a **36% da média**. Cinco segmentos ultrapassam o limiar de alta intensidade.

O achado mais contundente está no segmento **t = 34–39 s**: intensidade vocal de 2,493: o valor mais alto de todo o corpus. O texto desse segmento? "Você excelência assiste agora à apresentação de uma chapa...". O CLAP o classifica como "defesa argumentativa e resposta formal". A voz o classifica como o momento de maior energia acústica do debate.

> *"Você excelência assiste agora à apresentação de uma chapa..."*
>
> t = 34–39 s · intensidade 2,493 (máx. do corpus) · CLAP: "defesa argumentativa"



Still · t = 34–39 s: o momento de maior energia acústica do debate, classificado pelo texto como rotina defensiva.

Há também uma tendência temporal significativa: a correlação entre tempo e intensidade vocal é **r = −0,565**. O debate começa com alta intensidade vocal e decresce progressivamente. A voz esquenta no início e esfria — e o texto não registra essa dinâmica.

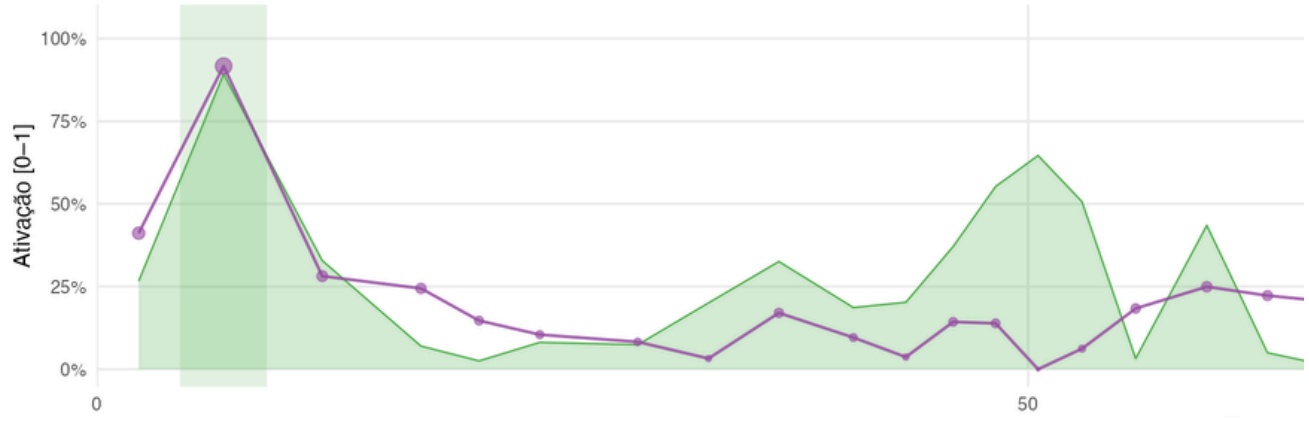
RESULTADOS

4.3. O que o gesto mostra

A velocidade da mão direita apresenta variação dramaticamente maior do que o áudio: entre o segmento menos ativo (v = 0,003) e o mais ativo (v = 0,106), há uma diferença de **34 vezes**. Enquanto a voz é um sinal relativamente contínuo, o gesto é intermitente e explosivo.

O mais expressivo ocorre em **t = 131–132 s**, com velocidade gestual normalizada de 1,0: o máximo do corpus. O texto transcrito desse segmento é uma única palavra:

> *"Senador."*
>
> `t = 131–132 s · velocidade gestual normalizada = 1,0 (máx. do corpus)`



Still · t = 131–132 s - o máximo gestual de todo o debate ocorre sobre uma única palavra. Velocidade gestual (Roxo) e expansividade das mãos (verde)

O máximo gestual do debate inteiro ocorre no momento em que o texto está no seu mínimo. A correlação temporal do gesto é essencialmente nula (**r = 0,052**), em contraste com a queda vocal: voz e gesto seguem trajetórias temporais independentes. A correlação entre intensidade vocal e velocidade gestual é **r = 0,303** (p = 0,034) — significativa, mas fraca. **Nenhuma é redundante da outra.**

4.4 A “impressão digital” de cada tipo de momento

A análise de perfil multimodal em coordenadas polares representa cada tipo de momento como um polígono sobre quatro dimensões: intensidade vocal, velocidade gestual, expansividade das mãos e ativação CLAP. **Três perfis emergem com geometrias distintas:** não apenas tamanhos distintos.

Figura 2 · coordenadas polares — Cada tipo de momento tem uma assinatura própria.

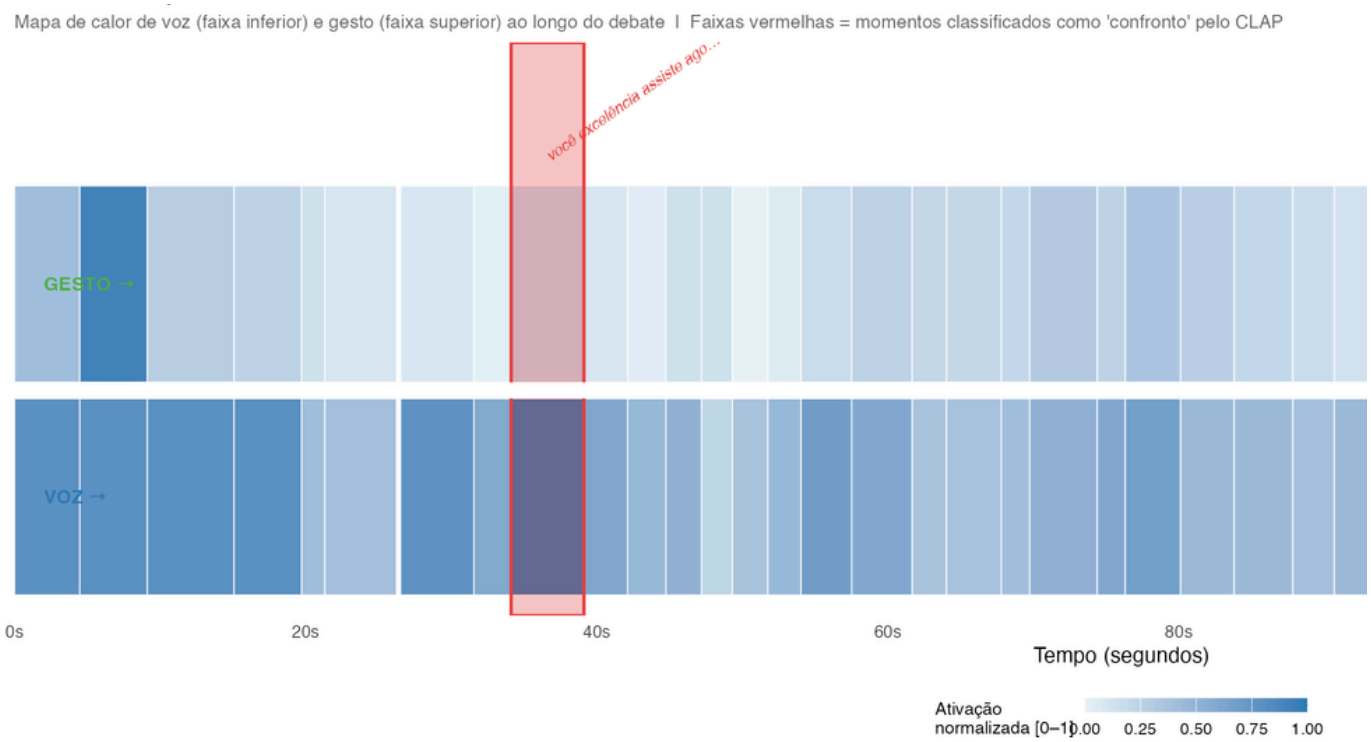
Pico de intensidade: polígono assimétrico com maior extensão em intensidade vocal; a voz domina e o gesto é secundário.

Confronto verbal: expandido em velocidade gestual: contra o esperado, os momentos de confronto não são os de maior intensidade vocal absoluta: o gesto é o canal dominante.

Momento típico: polígono pequeno e uniforme: a “defesa formal” de 88% do debate é, multimodalmente, de baixa ativação.

4.5 A conversa que o texto não transcreveu

O mapa de calor de ativação vocal e gestual ao longo dos 137 segundos revela a estrutura temporal do debate. As faixas vermelhas marcam os cinco momentos classificados como confronto.

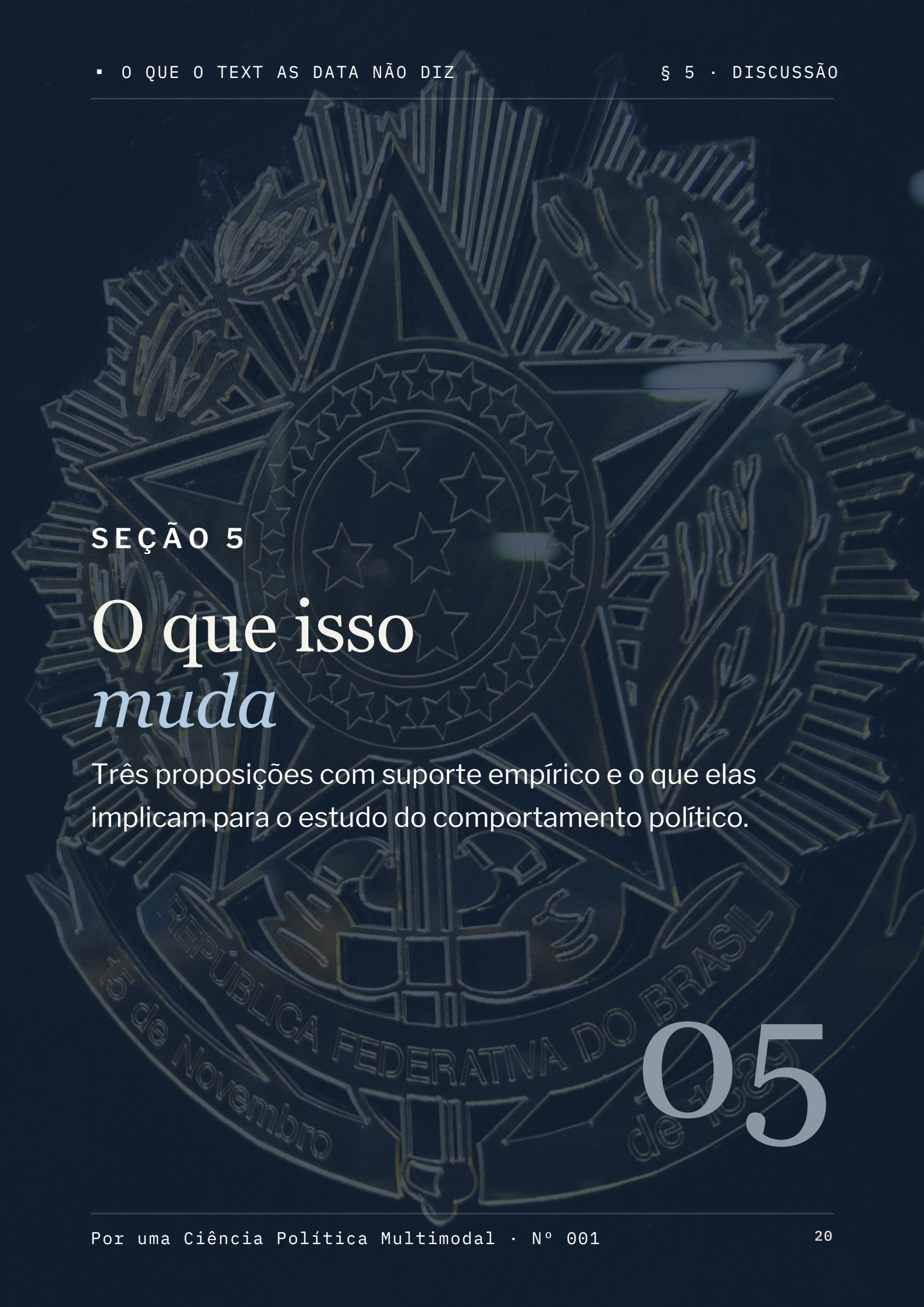


Ativação gestual (faixa superior) e vocal (faixa inferior); faixas vermelhas marcam confrontos.

Figura 3 · mapa de calor — Distribuição assimétrica dos momentos de confronto e progressão temporal das duas modalidades.

Há um confronto isolado em **t = 34 s** (alto índice vocal) e um cluster de quatro confrontos nos últimos sete segundos do trecho (**t = 128–135 s**). O debate tem uma estrutura: começa tenso, esfria no meio e explode no final.

> O ACHADO
>
> Essa narrativa de **escalada → resfriamento → re-escalada** não aparece em nenhuma análise de conteúdo textual dos mesmos segmentos.



SEÇÃO 5

O que isso *muda*

Três proposições com suporte empírico e o que elas implicam para o estudo do comportamento político.

05

DISCUSSÃO

Três proposições com suporte empírico.

P1 Texto e sinal físico são canais parcialmente independentes

A correlação áudio × gesto (r = 0,303) e a ausência de correlação entre classificação textual e picos vocais confirmam que as modalidades não são redundantes. Analisar apenas o texto não é uma simplificação do sinal completo — é a **eliminação de dois canais com variância própria**.

P2 O ambiente institucional suprime o texto, não o corpo

A gramática parlamentar força a "defesa argumentativa" como modo dominante (88%). Mas voz e gesto não seguem o mesmo protocolo: o momento de maior intensidade acústica é classificado textualmente como defensivo. **O corpo é menos treinável do que a linguagem formal** — com implicações para o estudo do comportamento político em tribunais, parlamentos e pronunciamentos preparados.

P3 O tempo importa de formas diferentes para cada modalidade

A voz esfria progressivamente (r = −0,565); o gesto permanece estável (r = 0,052). Um debate tem uma **dramaturgia própria** que se manifesta de modo distinto em cada canal. Análises que colapsam o evento num único vetor textual perdem essa estrutura temporal multimodal.

CONCLUSÃO

O texto diz o que tem de dizer, mas o debate é mais do que o texto.

Este relatório demonstrou, com dados empíricos de um evento político real, que a análise textual produz uma representação **sistematicamente empobrecida** do debate político. Não porque seja errada, o texto diz o que diz, mas porque o debate é um fenômeno multimodal e o texto é um único canal.

A contribuição metodológica desta pipeline não é substituir a análise de conteúdo. É **adicionar dois canais** que estavam sendo descartados por razões históricas , dificuldade de coleta, não teóricas. Com vídeo disponível em escala e modelos pré-treinados para áudio e pose, o custo marginal de processar as três modalidades é hoje comparável ao de processar apenas o texto.

> *O momento em que "Senador." — uma única palavra — corresponde ao pico gestual de um debate inteiro não é uma curiosidade. É um dado. E dado que não aparece em nenhuma análise de texto.*

NOTAS TÉCNICAS

Especificações da pipeline.

Ambiente computacional

Python 3.12 + R 4.x. Processamento modular por debate via variáveis de ambiente (DEBATE_ID, DEBATE_PROCESSED_DIR).

Transcrição

OpenAI Whisper base — 39M parâmetros, saída com timecodes por segmento. Idioma: português (pt-BR).

Análise acústica

librosa 0.10 — onset strength com Mel-spectrograma (n_mels=128, hop_length=512 a 22050 Hz). F0 por autocorrelação vetorizada.

Pose estimation

MediaPipe Pose (complexidade 1) — 33 keypoints 3D por frame a 25 fps. Velocidade por deslocamento euclidiano normalizado pela distância interombros.

Classificação semântica

CLAP laion/clap-htsat-unfused (512-d), similaridade coseno áudio × texto. Zero-shot, sem fine-tuning. Inferência em CPU.

Alinhamento

Médias por janela Whisper. ICT = média geométrica das três modalidades normalizadas [0,1] por min-max within-debate. Threshold de pico: mediana + 1,5 × MAD.

Corpus e fonte

debate_001 · Senado Federal · N = 49 · 137,1 s. Vídeo: youtube.com/watch?v=2YXbSc88HRs. Dados: multimodal_aligned.csv.

VINCIUS SANTOS

É **doutor em Ciência Política** pela UFMG e especialista na interseção entre **tecnologia, inteligência artificial e políticas públicas**, com foco no desenvolvimento de **Civic Technology e soluções em GovTechs.**

Seu trabalho combina **métodos computacionais, análise de sistemas complexos e governança de IA** para fortalecer e fomentar inovação cívica.

No **Senado Federal,** atuou como **assessor parlamentar e cientista de dados**, contribuindo com **análises e notas técnicas** para o **debate legislativo** sobre a **regulação de inteligência artificial**.

https://vsntos.github.io/site/

SÉRIE · Nº 001

Por uma Ciência Política *Multimodal*

Primeiro de uma série sobre análise multimodal de comunicação política: Texto, voz e gesto.