Capítulo 4

Medidas de duração

Após uma apresentação, na primeira seção, sobre o papel primordial da sílaba como unidade rítmica mínima, as próximas seções mostram como medir a duração de unidades prosódicas que vão do tamanho da sílaba até o do grupo respiratório, apontando o interesse da medida de cada unidade para a pesquisa experimental.

4.1 O Ancoramento do Ritmo na Sucessão Silábica

Desde os anos 1940 os foneticistas têm segmentado grupos acentuais tomando como limites o início da vogal (*vowel onset)*, fundamentando-se na experiência de que esse início é uma posição especialmente mais clara para o assinalamento da sílaba (basta o leitor ter em mente os grupos que começam com oclusivas e fricativas não vozeadas de baixa amplitude após pausa silenciosa para se conscientizar de que haveria impossibilidade de assinalar o inicio da sílaba fonológica no início com oclusiva - pois silêncio e intervalo de oclusão se confundem - ou uma grande imprecisão para marcar o início do grupo acentual pela dificuldade em saber onde começa a fricativa). Intuitivamente, percebiam que a sílaba é mais detectável na transição C-V. Para demonstrar esse aspecto, Dogil e Braun (1988) apresentaram evidências empíricas para a saliência do início da vogal na sílaba canônica (CV):

- Quando os sujeitos são solicitados a sincronizar tons puros regulares com sílabas que produzem em sequência, eles realizam a tarefa procurando alinhar uma região da sílaba chamada *perceptual*

center com a sequência de tons puros, e essa região se situa na vizinhança da transição C-V;

- Os parâmetros acústicos em torno das transições C-V e V-C em sílabas simétricas (e.g., /pap/, /bab/) não têm a mesma acurácia para assinalar características articulatórias: enquanto o início da vogal na transição C-V descreve de forma estável os traços do ponto de articulação da consoante precedente, os parâmetros em torno do final da vogal na transição V-C assinalam propriedades acústicas relevantes para a comunicação apenas em casos muito particulares;
- A articulação é melhor discriminada entre consoante e vogal na sequência CV do que na sequência VC, onde a coordenação gestual é mais imprecisa do que na sequência CV.

Confirmando também a estabilidade articulatória e, consequentemente acústica, da sílaba CV, Tuller e Kelso (1990, 1991) realizaram um experimento que mostrou que há mudança na coordenação entre os gestos laríngeo e supralaríngeo da consoante /p/ à medida em que as sílabas /ip/ e /pi/ são produzidas com taxa de elocução cada vez mais alta: a coordenação relativa entre os gestos na sílaba VC (/ip/) muda para a coordenação da sílaba CV (/pi/), enquanto a coordenação de gestos da sílaba CV se mantém estável. Sendo assim, por conta da estabilidade articulatória e acústica do início da vogal, sua sucessão age como ancoramento rítmico para a realização dos enunciados que requer, por economia, uma produção holística sob forma de um mecanismo oscilatório que garanta a rápida produção da sílaba.

De fato, estudos sobre a atividade temporal das redes neuronais apontam que a região do córtex motor que controla a fala é melhor descrita como oscilador neuronal que produz ciclos de impulsos elétricos na faixa de frequência 2 a 8 Hz que se refletirá na frequência natural de oscilação mandibular (POEPPEL; ASSANEO, 2020; DINGA et

al., 2017). Essa faixa coincide com a faixa de taxa de elocução em sílabas por segundo, como veremos na seção 4.6.

Pelo exposto, é mais condizente com a produção e a percepção da fala a delimitação da unidade prosódica mínima, a sílaba, por seus pontos de ancoramento sucessivo, os inícios de vogal. A unidade assim definida, a unidade VV[1], é uma sílaba fonética ancorada em seus limites pelos inícios (*onsets*) de duas vogais consecutivas na cadeia da fala, independentemente da presença ou não de pausa silenciosa entre esses dois inícios de vogal. A vantagem dessa unidade é sua eficiência em revelar a estruturação prosódica do enunciado, conforme amplamente detalhado num livro dedicado ao ritmo da fala (BARBOSA, 2006).

Mesmo que haja maior clareza na identificação do início de uma vogal pelo espectrograma de banda larga, a delimitação das fronteiras da unidade VV não é banal, por isso vamos mostrar em exemplos como fazê-la e como tomar determinadas decisões, especialmente quanto ao que deve ser segmentado e ao que deve ser etiquetado.

Tendo em vista que a sílaba é a unidade prosódica mínima e que, portanto, a adequada medida da duração de unidades do tamanho da sílaba tem consequência sobre as unidades maiores, começamos este capítulo guiando o leitor para medir a duração dessa unidade, não sem antes justificar seu papel para a constituição do ritmo da fala.

4.2 Medindo Durações de Unidades VV

Tomemos o enunciado “Em seguida apareceu um papagaio real que tinha fama de orador”, produzido por uma locutora universitária carioca de cerca de 25 anos na época da gravação. Ele foi retirado de um trecho de parágrafo lido que continua com o trecho “Subiu a

1 Observar bem que essa unidade é composta de uma única vogal, sendo que sua nomenclatura, VV, é apenas para lembrar que seus limites esquerdo e direito são inícios consecutivos de vogal.

tribuna de um poleiro de ouro", observação que terá sua importância explicada adiante. O trecho pode ser ouvido do repositório do livro pela etiqueta **EnunciadoLobatoCarioca** e a figura 4.1 mostra sua segmentação e etiquetagens iniciais. Por necessidade de discussão, mostra-se um trecho mais longo nas figuras, mas o ideal é segmentar as unidades VV pela imagem de espectrogramas entre 0,5 e 1 segundo, conforme ensinamos em outra obra (BARBOSA; MADUREIRA, 2015).

Para o trecho "Em seguida apareceu", é preciso inicialmente ouvir como a locutora pronunciou os segmentos acústicos e se guiar pelo espectrograma de banda larga para marcar os inícios de vocoides (vogais, ditongos, tritongos) pelo início do padrão de F2. É preciso observar no espectrograma os vocoides, de fato, pronunciados, os fenômenos de sândhi, de apagamento ou de epêntese. Nesse trecho inicial, a preposição "em" é pronunciada como ditongo nasalizado, a primeira vogal de "seguida" como [ɪ] e a última se funde por sândhi com a primeira vogal de "apareceu". Note o leitor que cada intervalo começa por um vocoide até o início do seguinte, assinalando dentro de cada um os símbolos dos segmentos nele contido.

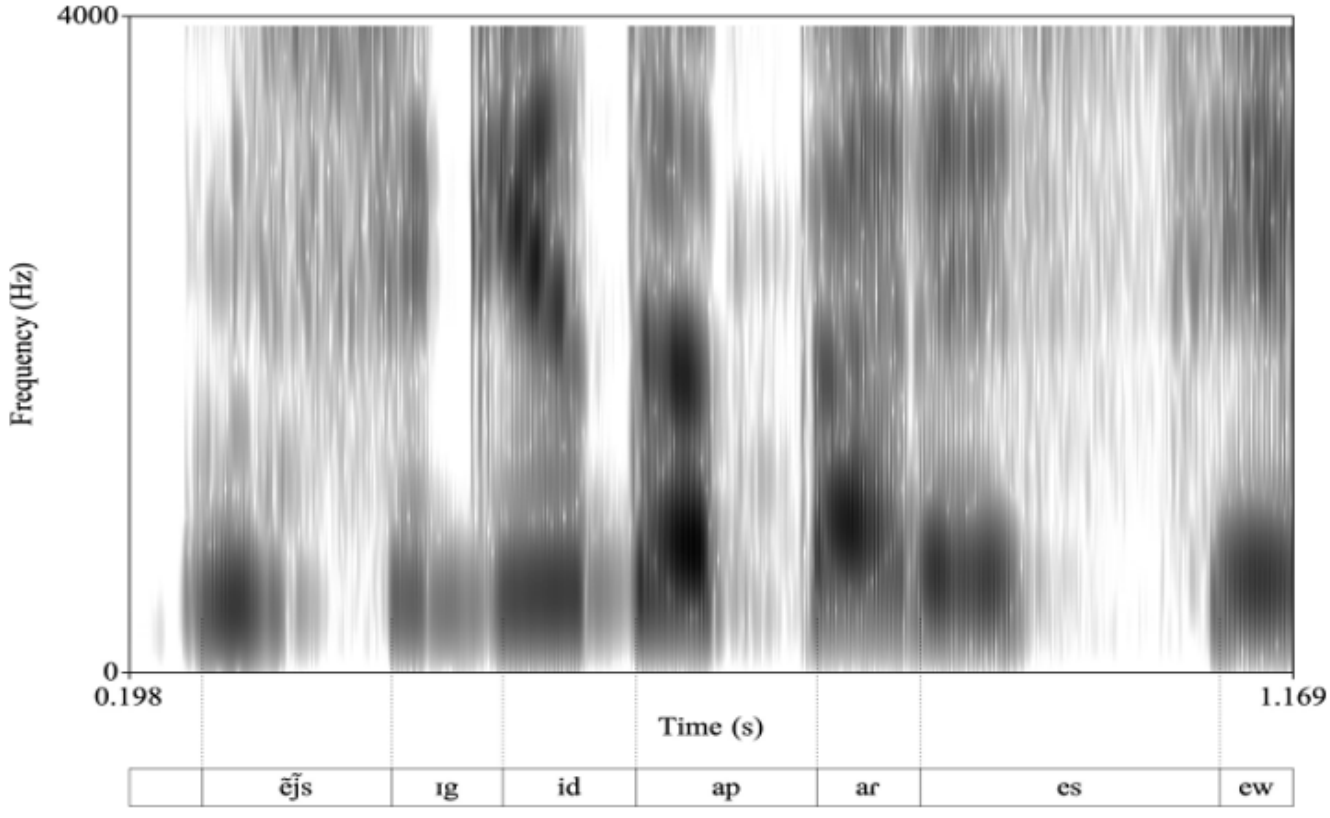

Figura 4.1 – Espectrograma de banda larga e camada de anotação para o trecho “Em seguida apareceu” do enunciado “Em seguida apareceu um papagaio real que tinha fama de orador”. Ver texto para saber como reproduzir a segmentação e etiquetagem.

Continuando o enunciado-exemplo, a figura 4.2 mostra a segmentação e etiquetagem do trecho “-ceu um papagaio real” que começa pelo ditongo [ew][2]. No espectrograma se vê que a vogal nasalizada do artigo “um” foi pronunciada de fato como vogal, não se integrando como semivogal ao ditongo precedente, por isso marcado como início de nova unidade VV. Essa mesma separação entre vogal e ditongo seguinte se vê ao final na pronúncia da palavra “real”, que inclui a consoante [k] da conjunção “que” pronunciada em seguida. No entanto, o final da palavra “papagaio” foi pronunciado como um vocoide que aparenta ser a sequência da vogal tônica, de uma vogal reduzida[3] e de uma semivogal [w]. Por conta da dificuldade de separação dos elementos desse trecho, o melhor é deixar tudo como única unidade

2 Da unidade VV precedente, [es], só aparece no espectrograma o final da [s]. fricativa

3 É vogal porque seu F2 tem um trecho horizontal, sendo a não horizontalidade a marca de movimento do corpo da língua, característica de uma aproximante.

VV etiquetando-a com os elementos que a constituem.

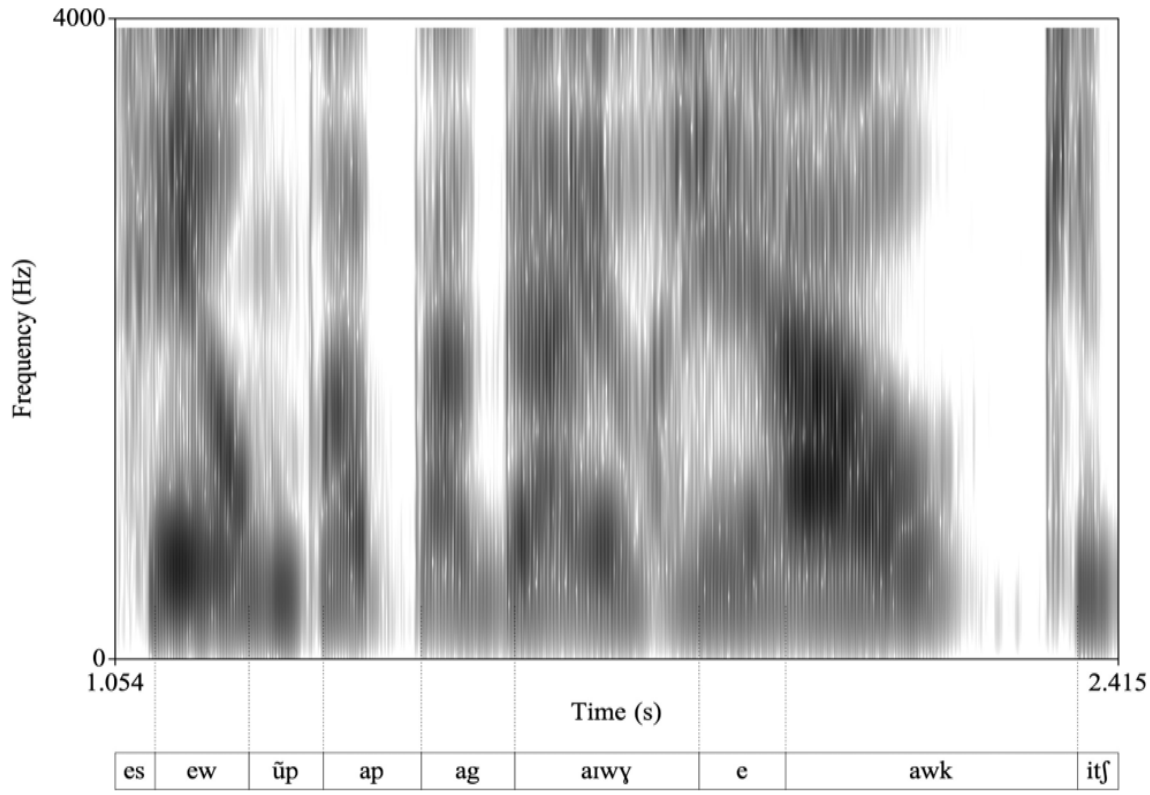

Figura 4.2 – Espectrograma de banda larga e camada de anotação para o trecho "-ceu um papagaio real" do enunciado "Em seguida apareceu um papagaio real que tinha fama de orador". Ver texto para saber como reproduzir a segmentação e etiquetagem.

Para concluir, observe o leitor a anotação do trecho final, "que tinha fama de orador", na Figura 4.3. Marcou-se a realização do /t/ de "tinha" como africada e, na mesma palavra, não houve produção da nasal palatal. Por isso, a vogal nasalizada tônica aparece sozinha no intervalo. Na sequência "de orador" a preposição se une à palavra hospedeira mudando para um ditongo crescente e a palavra final, com /r/ realizado como fricativa glotal não vozeada ([h]) tem o [s] na etiqueta da unidade VV([ohs]) por conta da palavra "subiu" que segue na leitura do parágrafo. É importante observar que, se não houvesse a continuação da leitura, a última unidade VV seria a correspondente ao intervalo etiquetado por [ad], que termina no início do [o] da sílaba final de "orador". O que começa pela vogal [o] não tem limite à direita nesse caso porque não tem vogal seguinte para assinalá-lo, portanto

não forma uma nova unidade VV.

As durações brutas (em milissegundos) das unidades VV de todo o enunciado da locutora carioca podem ser visualizadas na Figura 4.4. Ouvindo o áudio correspondente no repositório do livro, **EnunciadoLobatoCarioca**, as locuções destacadas pela locutora foram “em seguida”, “apareceu” e “real”, com “orador” terminando o enunciado com grande pausa silenciosa antes do próximo trecho de fala. As durações brutas refletem isso parcialmente, uma vez que as unidades VV de “em seguida” são das de menor duração. Por outro lado, embora as maiores durações sejam das tônicas de “apareceu”, “real” e “orador”, o início da palavra “fama” também é relativamente longo. A não correspondência estreita entre duração medida e percepção de funções de proeminência e fronteira prosódica só pode ser contornada com a normalização da duração.

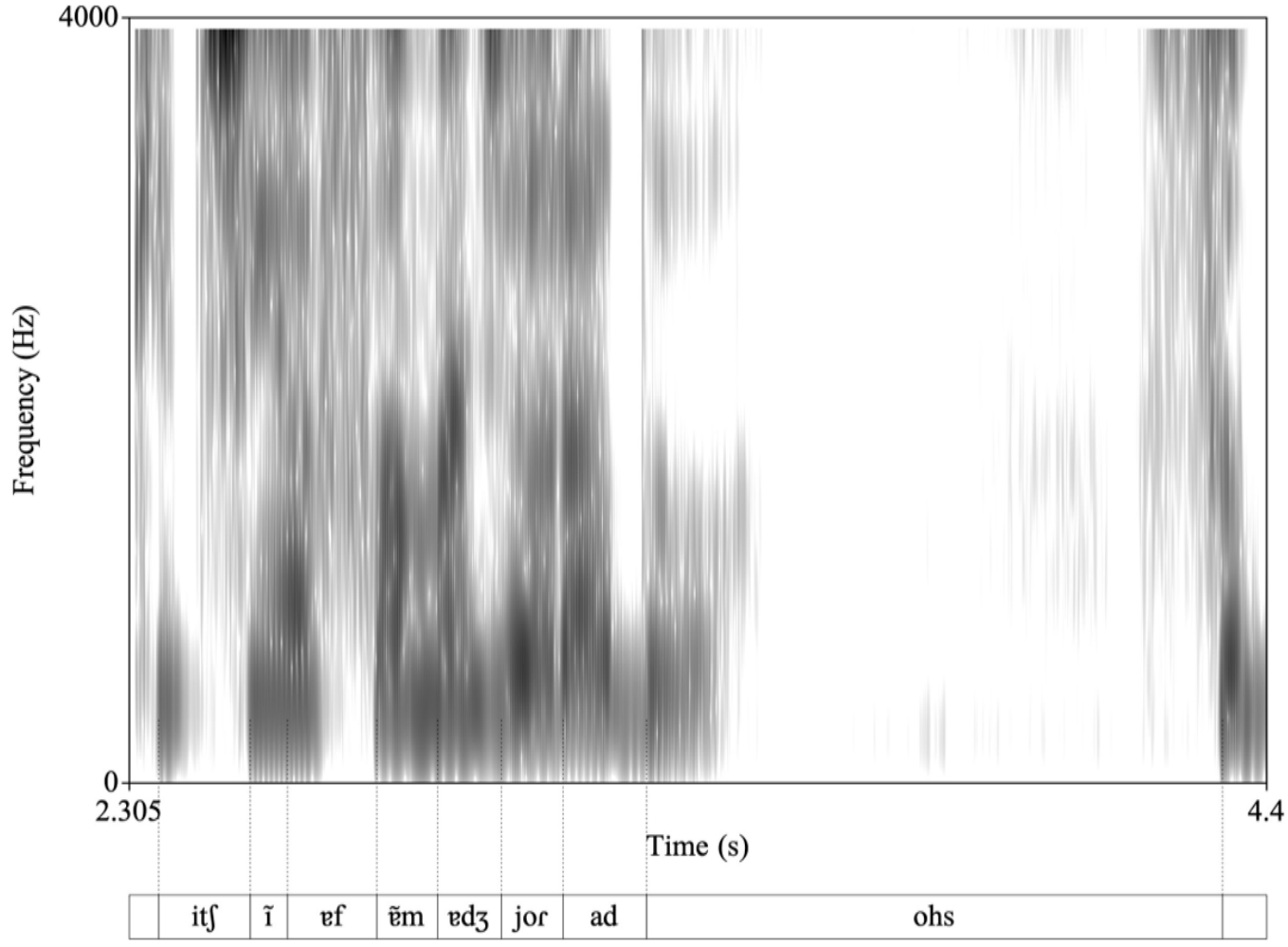

Figura 4.3 – Espectrograma de banda larga e camada de anotação para o trecho “que tinha fama de orador” do enunciado “Em seguida apareceu um papagaio real que tinha fama de orador”. Ver texto para saber como reproduzir a segmentação e etiquetagem.

Para melhor entender o papel da normalização, mostramos aqui as durações de unidade VV da leitura do mesmo trecho por outra locutora, dessa vez paulista. Pode-se ouvir no arquivo de áudio **EnunciadoLobatoPaulista** que essa locutora destaca mais as locuções “em seguida”e “real”, terminando com a palavra “orador”. Ela faz uma pausa silenciosa menor entre o fim do enunciado e o início do próximo, como pode ser visualizado na Figura 4.5[4]. Na locutora paulista, as unidades VV da locução “em seguida” têm durações maiores do que várias outras no trecho, correspondendo melhor à percepção. Embora “real” tenha duração bem maior que várias unidades, também o início

4 Para tornar mais clara a comparação entre as locutoras, juntamos unidades VV de uma delas para parear com as mesmas posições ao longo dos enunciados do trecho.

de “fama” é longo e isso não corresponde à percepção.

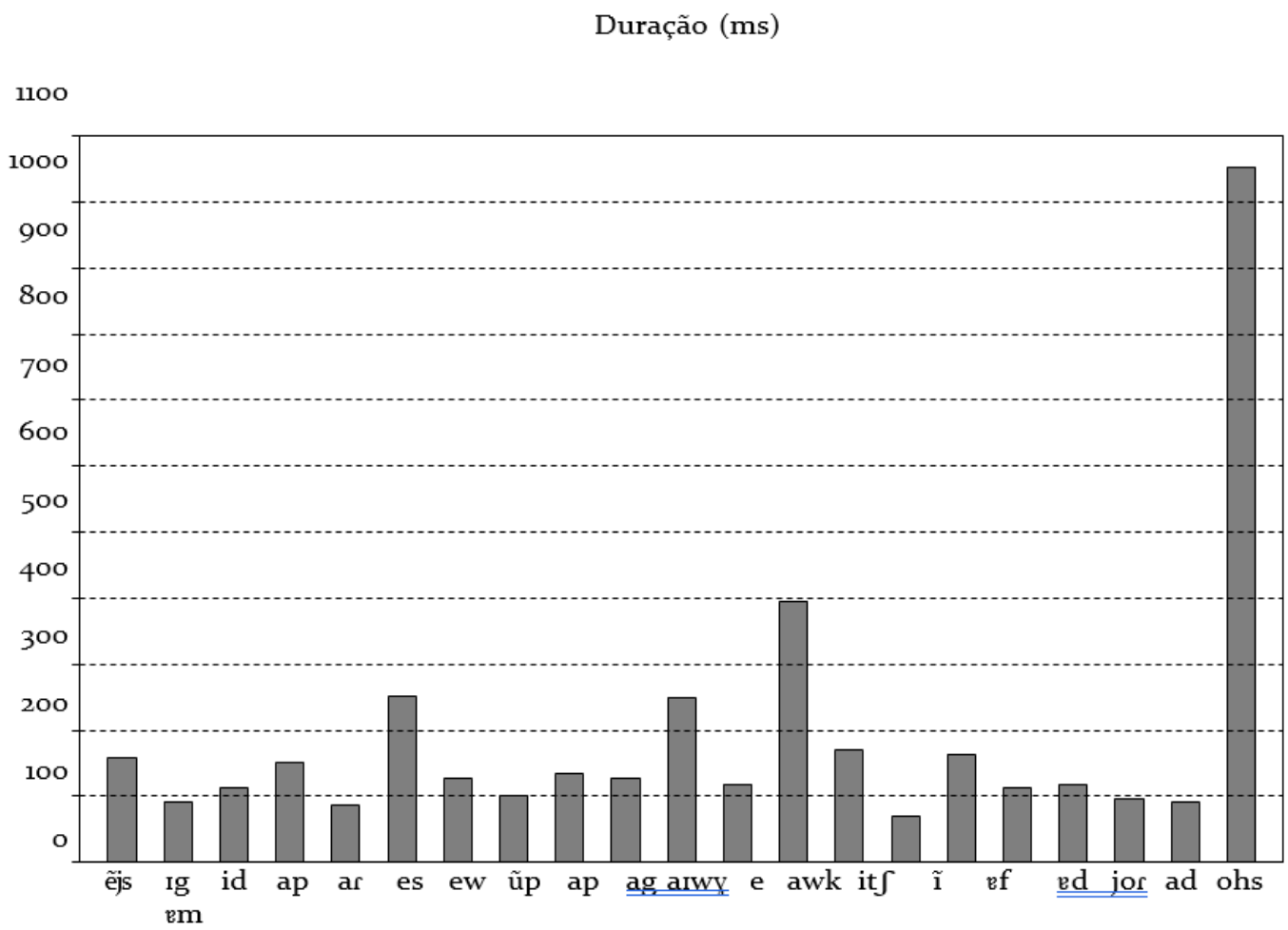

Figura 4.4 – Durações brutas, em milissegundos, do enunciado “Em seguida apareceu um papagaio real que tinha fama de orador” de locutora carioca.

A razão para a não correspondência estreita entre duração bruta de unidade VV e percepção dessa duração é o fato de que a percepção da duração requer a saliência da unidade em relação ao contexto fônico em sua vizinhança. Essa saliência se expressa por um afastamento da duração bruta em relação a uma expectativa sobre sua duração, internalizada pela experiência que temos em perceber a duração das unidades silábicas. Assim, a duração intrínseca de cada sílaba e de seus elementos constitutivos não chama a nossa atenção a não ser que difira de seu valor esperado. Por exemplo, um [s] é normalmente um som longo em relação a vários outros sons, e essa extensão temporal que lhe é própria se chama de duração intrínseca[5]. Mas sua duração

5 Duração esperada ou média são termos empíricos equivalentes.

realizada num enunciado específico só é percebida como relevante para a organização prosódica do enunciado quando está bem aquém ou bem além da duração intrínseca. O mesmo vale para uma unidade do tamanho da sílaba como a unidade VV ou uma sílaba fonológica. É por isso que, para se ter uma adequada avaliação da duração dessas unidades que reflita algo sobre sua percepção, é preciso normalizar a duração bruta. É o procedimento que descreveremos a seguir.

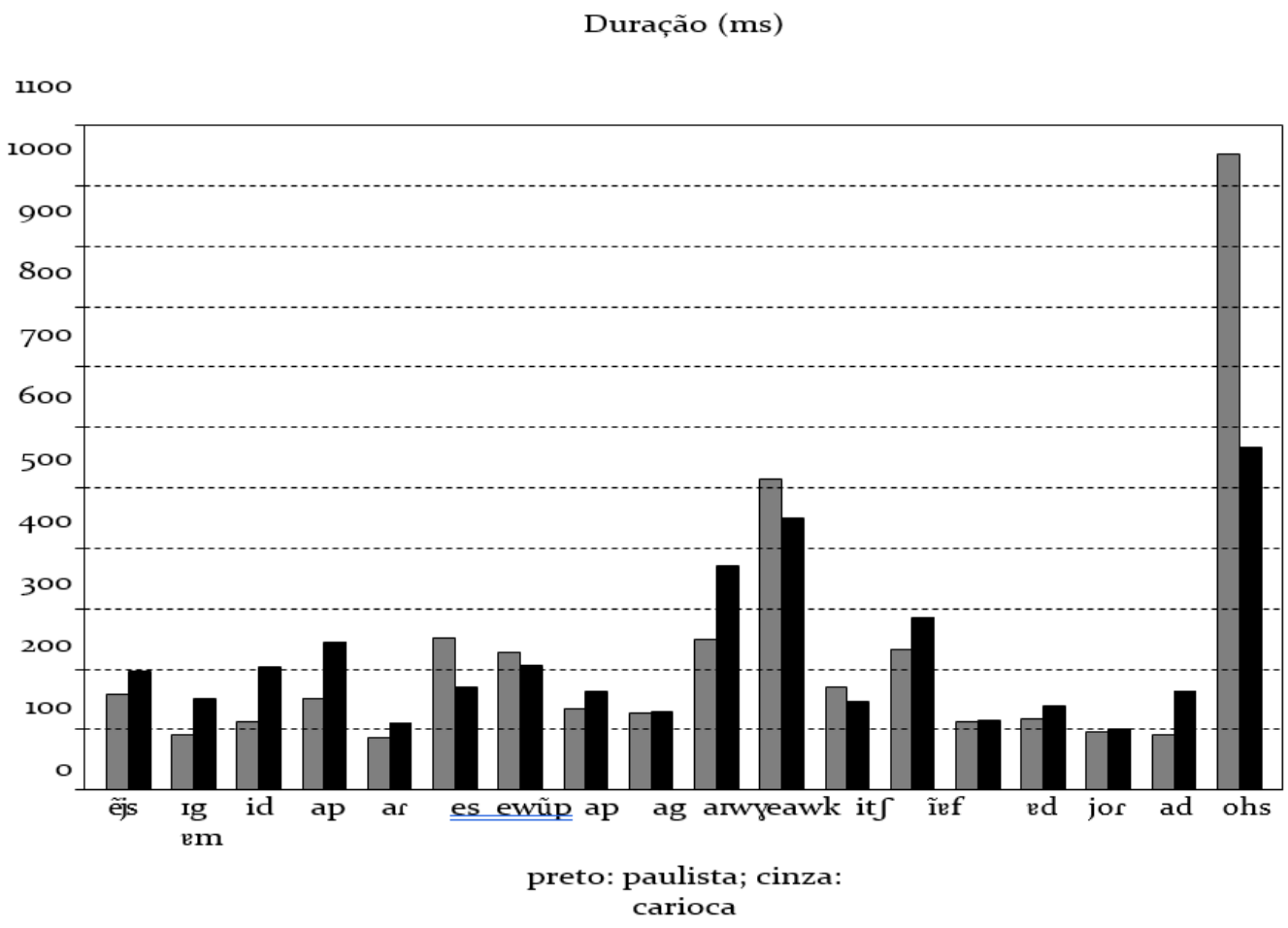

Figura 4.5 – Durações brutas, em milissegundos, do enunciado "Em seguida apareceu um papagaio real que tinha fama de orador" por locutoras carioca (cinza) e paulista (preta).

4.3 Normalização da Duração de Unidades VV

A normalização da duração de uma unidade do tamanho da sílaba, a menor unidade prosodicamente relevante, é fundamental para revelar o grau de saliência prosódica dessa unidade. Frequentemente

se normaliza essa duração dividindo-a pela duração de alguma unidade de referência mais extensa dentro da qual se encontra, como o enunciado ou a palavra. Entre essas duas referências, a divisão pela duração do enunciado é mais interessante por permitir comparar durações provindas de enunciados falados em taxas de elocução distintas, uma vez que se passa a trabalhar com a proporção que as durações dessas unidades ocupam no enunciado respectivo, independentemente de quanto duram em termos brutos. Mas esse procedimento não elimina o efeito da duração intrínseca, isto é, se uma unidade como [as] é longa porque contém dois segmentos longos do PB, continuará sendo proporcional- mente longa no enunciado. Por conta disso, o melhor procedimento de normalização é o que calcula o *z–score* da duração.

O cálculo do *z-score* é um procedimento de normalização básico em estatística e expressa o quanto um valor está afastado de uma média em unidades de desvio-padrão. Para tanto precisamos ter valores de média e desvio-padrão de duração de referência para todo tipo de unidade VV. Ora, como isso envolveria a gravação de um corpus de tamanho muito extenso, tendo em vista a combinatória de diferentes fones em cada unidade VV, fizemos o cálculo pela via da duração média e do desvio-padrão dos segmentos que compõem uma unidade VV, como já usado no trabalho de Campbell (1992). Dessa forma, precisamos apenas do inventário de realizações de segmentos do tamanho do fonema. É isso que expressa a equação 4.1.

$$z = \frac{dur - \sum_i \mu_i}{\sqrt{\sum_i var_i}} \qquad (4.1)$$

Nessa equação, *dur* é a duração bruta da unidade VV em milissegundos e o par de variáveis (μ_i, var_i) são a média e a variância de duração dos segmentos fônicos contidos na mesma unidade de um lo-

cutor do PB. Esses valores podem ser encontrados em Barbosa (2006, p. 489) para o PB, mas valores para outras línguas como inglês britânico, espanhol europeu, alemão padrão, francês padrão e português europeu estão disponíveis para rodar com o script que normaliza a du- ração, o *SG Detector*, disponível no endereço https://github.com/ pabarbosa/prosody-scripts.

O fato de usarmos para o procedimento de normalização um locutor da mesma língua, mas distinto daquele que fala não é um obstáculo, uma vez que se mostra que a mudança de locutor referência, aquele de que foram extraídas as médias e desvios-padrão das durações dos fones, não altera as posições em que se encontram nem os graus relativos dos picos locais de duração de unidade VV ao longo dos enunciados, como mostrou Vieira (2007, p. 81-85).

Após o cálculo do *z-score* das durações das unidades VV de um determinado excerto de fala, suaviza-se a sequência de valores para atenuar efeitos de implementação do acento lexical e salientar a extensão prosodicamente relevante da duração silábica. Para tanto, após um teste com médias móveis de 3 a 9 pontos, a média móvel[6] de 5 pontos foi aquela que mais correspondeu à percepção da duração da unidade VV como unidade proeminente ou marcadora de fronteira prosódica. A aplicacão dessa suavização por média móvel se dá pela equação 4.2 a partir da sequência de *z – scores* (z_i) obtida pela equação 4.1.

$$z^i_{suav.} = \frac{5.z^i + 3.z^{i-1} + 3.z^{i+1} + 1.z^{i-2} + 1.z^{i+2}}{13} \qquad (4.2)$$

6 A média móvel é um procedimento matemático que calcula da mesma forma uma média ponderada em cada posição de pontos de uma curva. Seu efeito é o de suavizar a curva, eliminando oscilações de pequena extensão.

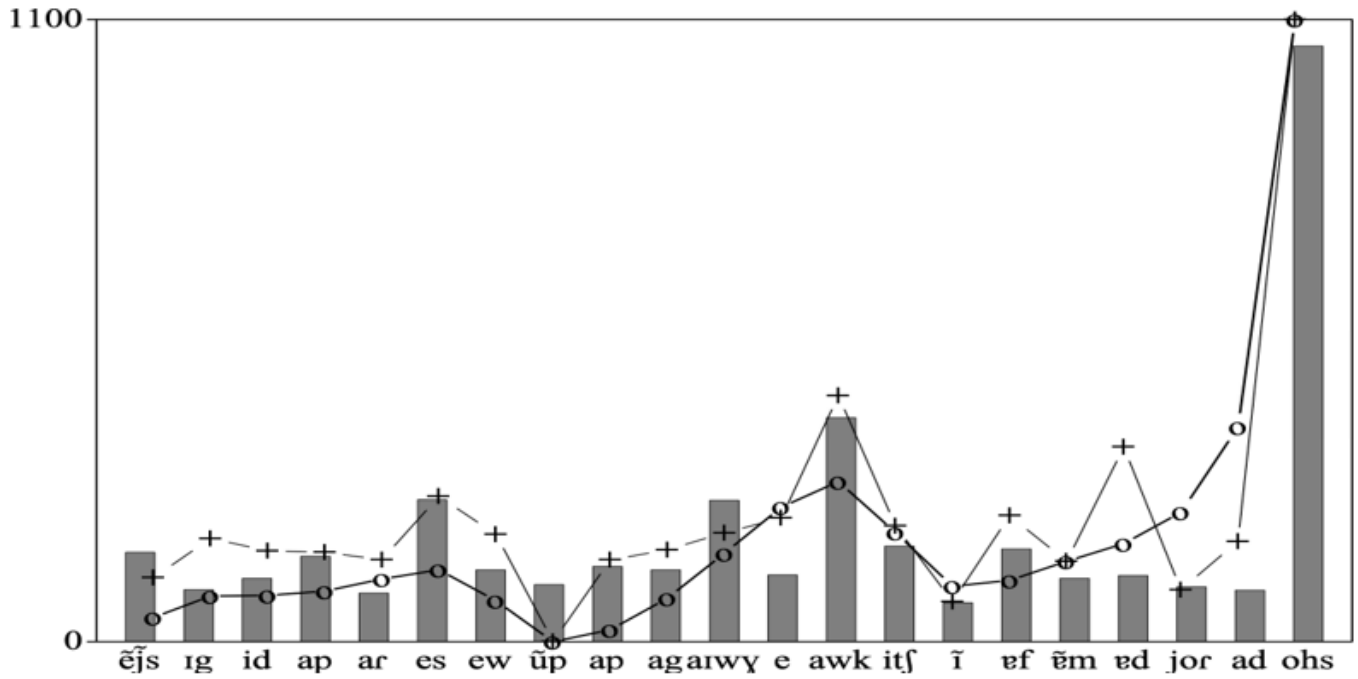

Figura 4.6 – Durações brutas em milissegundos (barra cinza), por *z- score* (pontos + conectados) e por *z- score* suavizado (pontos *o* conectados) do enunciado “Em seguida apareceu um papagaio real que tinha fama de orador” por locutora carioca.

O efeito do procedimento de normalização que termina com o contorno de *z-score* suavizado das durações das unidades VV pode ser visto na Figura 4.6 para a locutora carioca cuja duração bruta de unidades VV foi apresentada acima.

Conforme vimos acima, as locuções destacadas pela locutora carioca foram “em seguida”, “apareceu”, “real” e “orador”. São justamente as unidades VV dessas locuções que têm picos locais de *z-score* suavizado, nesta ordem decrescente de valor: “orador” ao final, “real”, logo antes da conjunção “que”, “apareceu” e “em seguida”. Observe que nesse contorno a palavra “fama” não é pico local como é na duração bruta. A palavra “papagaio”, que tem um pico local de duração bruta na unidade VV final, perde esse pico local já na primeira fase de normalização, o contorno de *z-score* antes da suavização por média móvel (marcado com o símbolo ’+’).

Outro interesse em se ter os valores normalizados da duração de uma unidade silábica como a unidade VV é a possibilidade de compa-

ração das diferentes formas de uso da duração silábica em diferentes locutores e estilos de elocução, como veremos na seção seguinte.

Embora a duração da unidade VV assim normalizada indique alongamento ou encurtamento (vide o mínimo do *z score* suavizado em [u˜p] que pode corresponder à realização tanto de proeminência quanto de marcação de fronteira prosódica), do ponto de vista prático, podemos considerar cada pico local de z-score suavizado como marcador da posição final de um grupo acentual, definido como sequência de unidades não proeminentes terminadas por uma unidade proeminente. Não obstante uma unidade VV antes de fronteira não ser necessariamente proeminente[7], tomar todo pico local de duração normalizada como fron teira à direita de grupo acentual tem a vantagem da automatização do procedimento sem grandes prejuízos para o conhecimento que se pode construir a respeito da duração de grupos acentuais, como mostraremos na seção 4.7.

4.4 Avaliando diferenças no ritmo da fala via duração

Há uma vantagem metodológica no uso dos picos locais de durações normalizadas de unidades VV para comparar a fala de diferentes locutores ou um mesmo locutor em diferentes estilos de elocução. O uso de um método para calcular a distância entre diferentes valores desses picos permite fornecer um índice de proximidade entre os ritmos das falas. Tomemos em primeiro lugar diferenças entre os estilos de elocução leitura (de história) e narração consecutiva (logo após ler a história) em homens e mulheres.

7 É o caso de trechos após a realização de um foco estreito, por exemplo, pois embora suas unidades VV não sejam proeminentes, precedem uma fronteira com realização de pausa silenciosa ou alongamento de sílaba final se a fala continua.

4.4.1 Distâncias de ritmo da fala

O corpus usado aqui é o corpus Belém, já mencionado anteriormente. Trechos de fala entre 10 e 20 segundos de 5 homens e 5 mulheres universitários e de idade entre 20 e 35 anos foram extraídos nos dois estilos, segmentados em unidades VV e devidamente etiquetados. Em seguida, utilizaram-se os procedimentos sucessivos de normalização e suavização descritos acima para gerar os valores de *z-score* suavizado. Para cada participante e estilo há, então, um conjunto de valores de *z-score* suavizado que assinalam o ritmo de cada um no respectivo estilo. Para calcular o quanto distam esses conjuntos de valores propusemos a equação 4.3 de distância entre distribuições em que $média_i$ e $média_j$ são as médias aritméticas dos valores de *z-score* suavizado dos conjuntos respectivos *i* e *j*, enquanto var_i e var_j são suas respectivas variâncias.

$$dist_{conjunto_i, conjunto_j} = \frac{|\text{média}_i - \text{média}_j|}{\sqrt{var_i + var_j}} \quad (4.3)$$

Homens: Geral: 0,25	**Leitura vs Narração**	**Mulheres:** Geral: 0,32	**Leitura vs Narração**
MT	0,27	AG	0,32
LA	0,47	RA	0,08
CA	0,2	NP	0,38
EM	0,4	GR	0,34
FA	0,38	DF	0,02

Tabela 4.1 – Distâncias entre amostras de *z-score* suavizado entre leitura e narração de locutores paulistas entre 20 e 35 anos.

A Tabela 4.1 mostra as distâncias dos valores de *z-score* suavizado entre leitura e narração para cada um dos participantes separados por

sexo. Observe que há locutores que não diferem muito ao ler e narrar, como as mulheres RA e DF[8]. De fato, ao escutar trechos dos dois estilos das duas locutoras, percebe-se que ambas são rápidas nos dois estilos. Essas distâncias entre estilos são em geral maiores do que aquelas entre locutores num mesmo estilo, como se vê pelos números nas tabelas 4.2 a 4.5.

Na Tabela 4.2 se vê pelas distâncias que as mulheres RA e GR diferem mais do que todas as outras ao lerem, uma hesitando e lendo mais lentamente que a outra[9]. A mesma locutora GR dista pouco de NP ao ler, como se percebe escutando trechos de leitura das duas[10], o que é assinalado pela distância 0,09. Locutoras próximas em seu ritmo de leitura, como essas duas, podem não o ser na narração, que é justamente o caso de DF e AG (com uma distância de 0,31, Tabela 4.3), como se depreende da escuta de trechos nos dois estilos para ambas as locutoras[11].

Mulheres - Leitura					
	AG	RA	NP	GR	DF
AG					
RA	0,13				
NP	0,17	0,27			
GR	0,28	0,38	0,09		
DF	0,03	0,16	0,16	0,27	

Tabela 4.2 – Distâncias entre amostras de *z-score* suavizado entre diferentes leituras de locutoras paulistas. Nesta e nas próximas tabelas o fundo verde aponta as maiores distâncias e o fundo amarelo, as menores distâncias.

8 Ouvir do repositório do livro os trechos BPDFREFE10 (leitura de DF) vs. BPDFSTFE01 (narração de DF) e BPRAREFE09 (leitura de RA) vs. BPRASTFE02 (narração de RA).

9 Ouvir do repositório do livro os trechos BPRAREFE09 (leitura de RA) vs. BPGRREFE09 (leitura de GR).

10 Ouvir do repositório do livro os trechos BPNPREFE05 (leitura de NP) vs.BPGRREFE09 (leitura de GR).

11 Ouvir do repositório do livro os trechos BPDFSTFE02 (narração de DF) vs. BPAGSTFE05 (narração de AG).

Mulheres - Narração					
	AG	RA	NP	GR	DF
AG					
RA	0,11				
NP	0,09	0,17			
GR	0,11	0,01	0,17		
DF	0,31	0,21	0,30	0,17	

Tabela 4.3 – Distâncias entre amostras de *z-score* suavizado entre diferentes narrações de locutoras paulistas.

O mesmo tipo de comportamento nos dois estilos têm os homens CA e FA, como se vê nas tabelas 4.4 e 4.5. Ao narrar, CA e FA são muito próximos no modo de pausar, alongar segmentos, por isso a distância de apenas 0, 01 entre eles[12].

Homens - Leitura					
	MT	LC	CA	EM	FA
MT					
LC	0,01				
CA	0,22	0,21			
EM	0,34	0,33	0,12		
FA	0,05	0,05	0,18	0,30	

Tabela 4.4 – Distâncias entre amostras de *z-score* suavizado entre diferentes leituras de locutores masculinos paulistas.

Homens - Narração					
	MT	LC	CA	EM	FA
MT					
LC	0,22				
CA	0,11	0,12			
EM	0,39	0,20	0,31		
FA	0,12	0,11	0,01	0,31	

Tabela 4.5 – Distâncias entre amostras de *z-score* suavizado entre diferentes narrações de locutores masculinos paulistas.

12 Ouvir do repositório do livro os trechos BPCASTMA06 (narração de CA) vs.BPFASTMA04 (narração de FA). Comparar com a leitura dos mesmos ouvindo os áudios **BPCAREMA01** (leitura de CA) e **BPFAREMA08** (leitura de FA), comdistância 0,18.

Essa técnica, como se vê, permite a quantificação das diferenças de emprego da duração silábica entre estilos de elocução e entre locutores distintos, fornecendo um meio de avaliar mudanças no ritmo da fala. Pode-se entrever aplicações para a detecção de mudanças prosódicas como as causadas por ansiedade e estresse e mudanças emocionais durante uma interação comunicativa, além de quaisquer outras mudanças comportamentais.

4.4.2 Hierarquia de proeminências e fronteiras prosódicas

O emprego da técnica de cálculo de distâncias entre os ritmos das falas que acabamos de ver considera as durações normalizadas das unidades VV de todo o trecho, sem distinção de saliência acústica. De fato, proceder assim é fundamental para considerar todos os aspectos rítmicos dos trechos de fala sendo comparados. Mas é também possível investigar a realização de graus distintos na realização das funções de proeminência e de marcação de fronteira prosódica levando-se em conta as durações normalizadas apenas nos seus pontos de máximo.

Os histogramas que seguem consideram apenas valores de duração normalizadas nesses pontos de máximo para três locutores paulistas do corpus Belém, tanto para a leitura quanto para a narração. Neles se podem ver indícios de mais de uma moda, apontando para a possibilidade de amostras de populações estatísticas distintas que corresponderiam a níveis distintos da implementação das duas funções prosódicas mencionadas acima.

Embora sugira agrupamentos distintos, a inferência de qual são os grupos estatisticamente distintos se dá por meio de técnicas estatísticas de classificação e agrupamento. Para os exemplos aqui empregamos a técnica de k-médias. Essa técnica descobre os agrupamentos

distintos de um conjunto de dados, desde que se informe previamente quantos grupos serão discriminados. O algoritmo é feito de tal forma que os dois primeiros valores mais próximos constituem um grupo e, à medida que se analisa um novo valor compara-se esse com o primeiro agrupamento constituído e se avalia a distância para ver se pertence a esse agrupamento ou faz parte de novo agrupamento e assim iterativamente. Com isso se descobrem os valores que pertencem ao número de distribuições imposto de antemão.

Os histogramas da locutora LC superpostos na Figura 4.7 sugerem cerca de cinco grupos para ambos os estilos. Usando a técnica k-médias com esse número de grupos, obtemos os seguintes intervalos para ambos os estilos: o primeiro grupo com *z-score* suavizado inferior a 3, 5, o segundo com valores de 3, 5 a 9, 0, o terceiro de 9, 0 a 18, 0 , o quarto de 18, 0 a 29, 0 e o último superior a esse último número.

Cabe ao pesquisador associar esses grupos a uma função prosódica específica. Por exemplo, o grupo com os menores valores de máximos de duração normalizada está associado a fronteiras de enunciado dentro de um mesmo subtópico. Os grupos de valores intermediários assinalam fronteiras entre tópicos ou subtópicos distintos e os maiores valores estão frequentemente associados a hesitações e ao macroplanejamento, no caso da narração.

Observando a figura 4.8 referente à locutora AV, utilizamos a técnica de k-médias com cinco grupos e obtivemos os seguintes intervalos de valores para cada um dos grupos considerando ambos os estilos jun- tos: o primeiro grupo com *z–score* suavizado inferior a 2, 8, o segundo com valores de 2, 8 a 7, 0, o terceiro de 7, 0 a 13, 5, o quarto de 13, 5 a 24, 0 e o último superior a esse último número, o que não é muito distinto da locutora LC.

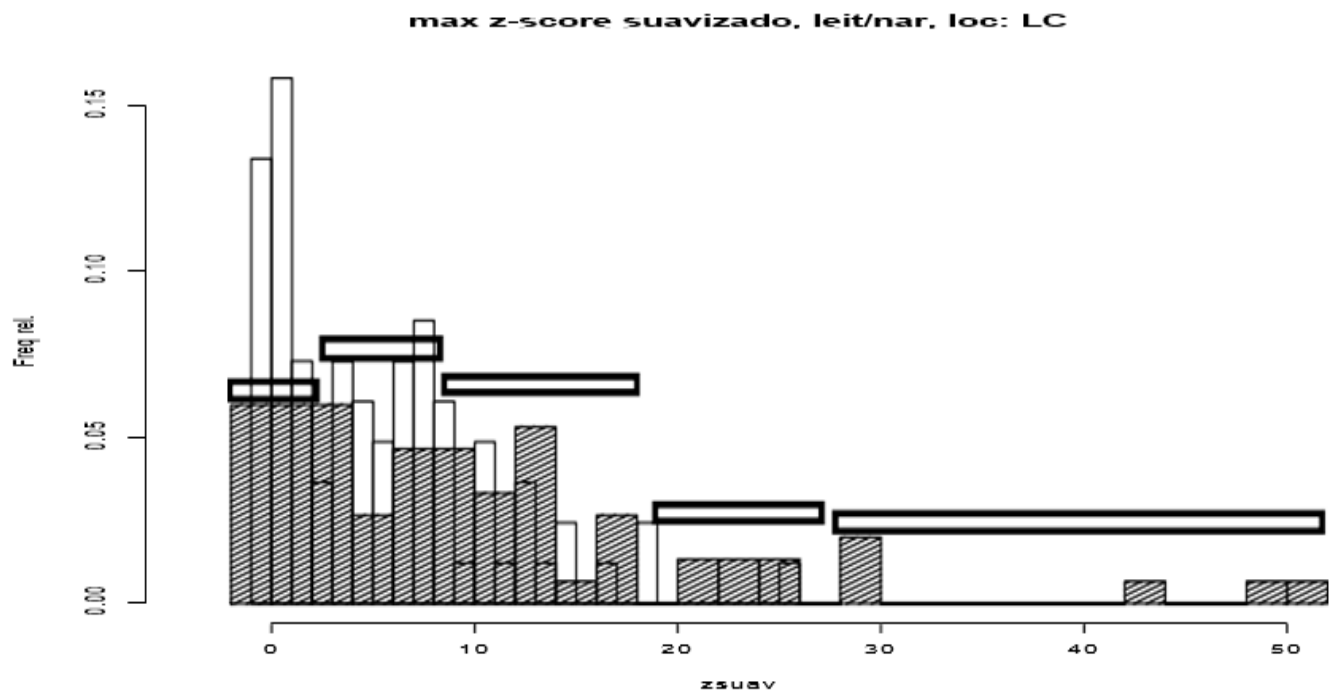

Figura 4.7 – Histogramas superpostos dos picos de z-score suavizados de leitura (barras claras) e narração (barras hachuradas) da locutora paulista LC.

Na figura 4.9, referente ao locutor FA, podem-se ver de três a quatro agrupamentos nas duas distribuições de leitura e narração. Usando a técnica das k-médias especificando quatro grupos, obtivemos os seguintes intervalos para ambos os estilos: o primeiro grupo com *z-score* suavizado inferior a 2, 5, o segundo com valores de 2, 5 a 9, 5, o terceiro de 9, 5 a 22, 0 e o último superior a esse último número. O alongamento das unidades VV em fronteira é menor neste locutor, um professor do ensino médio com grande experiência na exposição das matérias. Isso faz com que hesite menos e organize melhor seus tópicos e subtópicos na narração.

É notório observar como o primeiro agrupamento tem *z-score* suavizado na vizinhança de 2, 5 para os três locutores, limite inferior que serviu no trabalho de Barbosa (2020) para a detecção automática de fronteira prosódica correspondente a um enunciado ou a uma unidade entoacional inferior (fronteira não terminal).

A análise da composição das amostras de valores de *z-score*

suavizado fornece uma riqueza de detalhes sobre a forma como se organiza ritmicamente a cadeia de fala. O resultado da aplicação de uma técnica estatística de análise por agrupamentos para os picos de *z–score* suavizado sugere que essa organização é feita em níveis hierárquicos distintos. No entanto, essa descoberta não impede a quantificação das distâncias rítmicas entre trechos de fala de diferentes estilos de elocução e entre locutores, pois a forma de hierarquizar diferentes constituintes prosódicos também é parte da variação entre estilos e entre locutores. Um aspecto importante dessa organização é a realização de pausas silenciosas e preenchidas durante a enunciação. Medir sua taxa de produção e sua duração fornece pistas importantes para quantificar diferenças no ritmo da fala.

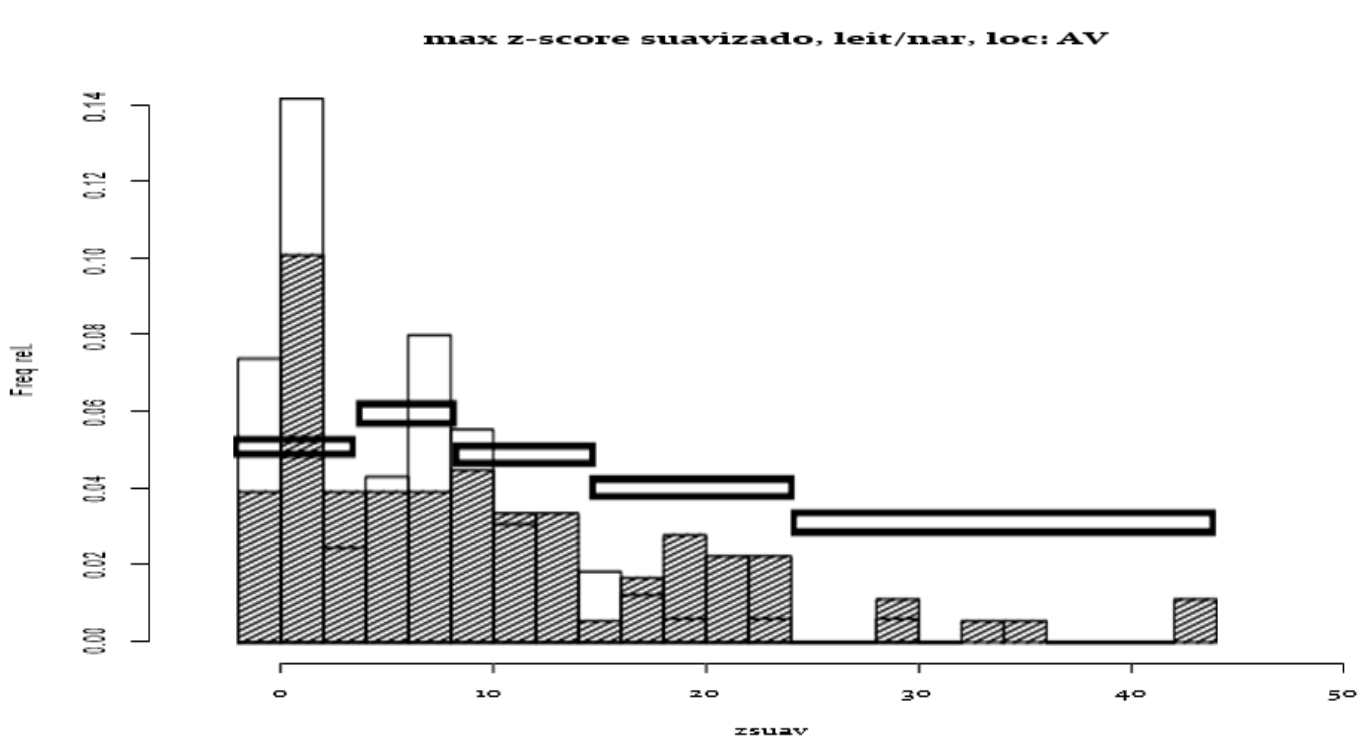

Figura 4.8 – Histogramas superpostos dos picos de *z-score* suavizados de leitura (barras claras) e narração (barras hachuradas) da locutora paulista AV.

4.5 Medindo durações de pausas silenciosas e preenchidas

A pausa é uma quebra momentânea no curso da enunciação que tem por finalidades tanto organizar em partes menores aquilo que se diz, função da pausa não hesitativa, quanto ganhar tempo para planejar o que ainda se dirá, função da pausa hesitativa. A pausa não hesitativa pode ser uma pausa silenciosa[13] ou um alongamento de vogal ou consoante para marcar uma fronteira prosódica no enunciado, como em "Manuel tinha entrado para o mosteiro há quase um a:no /, mas ainda não se acostumara àquela maneira de viver". com o sinal ":" indicando alongamento do /a/ e a barra (/) indicando uma pausa silenciosa. Já a pausa hesitativa é composta de material sonoro e por isso mesmo também é chamada de pausa preenchida. Ela pode ser realizada por trechos sonoros não lexicais como "uhm", "ahn" ou trechos sonoros lexicais como "né", "e:", "quer dizer", desde que esteja associada a uma organização do pensamento. Aqui incluiremos alongamentos em fim de sílaba que cumprem essa função de "ganhar tempo" na classe das pausas preenchidas. Para uma classificação semelhante, ver a tese de Rose (1998).

13 Estritamente falando pode haver inspiração ou expiração audível, como será apresentado na seção 4.9.

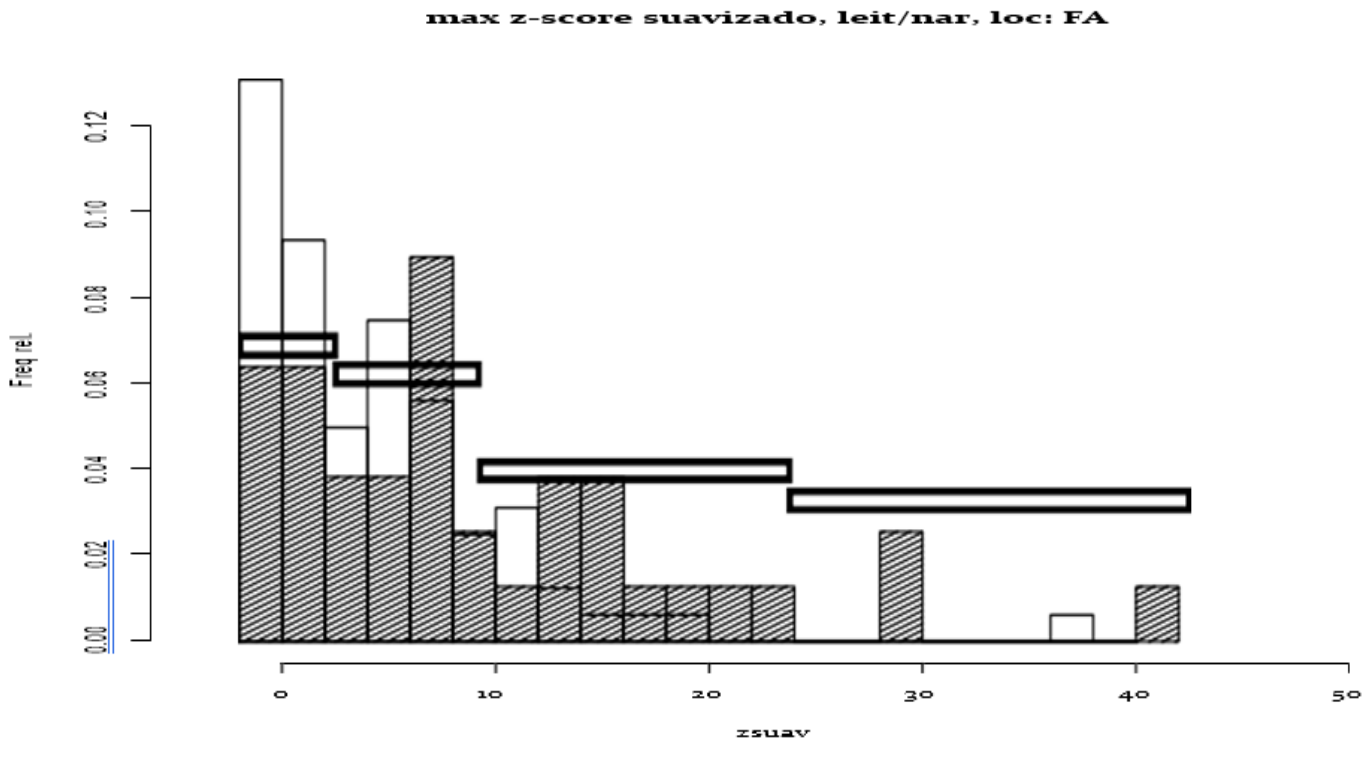

Figura 4.9 – Histogramas superpostos dos picos de *z-score* suavizados de leitura (barras claras) e narração (barras hachuradas) do locutor paulista FA.

Para ilustrar como medir e como analisar as durações de pausas silenciosas e preenchidas, utilizamos dados de dois participantes que não eram irmãos, extraídos do corpus da tese de Cavalcanti (2021), que contou com entrevistas por telefone entre gêmeos univitelinos, todos do Estado de Alagoas e do sexo masculino com idades entre 19 e 35 anos com pelo menos o Ensino Fundamental completo. A gravação de cada um deles foi feita com microfones de lapela, não passando assim pelo filtro telefônico. A conversa entre os gêmeos, que visou em sua tese a aplicação forense, tem a vantagem de se obter longos trechos de fala por conta da familiaridade entre os interlocutores. Para essa análise e para possibilitar revelar uma diferença maior entre locutores, tomamos trechos da conversa de um dos dois locutores gêmeos extraídos de dois diálogos dos quais segmentamos mais de 40 pausas silenciosas e preenchidas de cada um para análise neste livro. A Figura 4.10 mostra como fizemos a marcação da vogal da pausa preenchida, indicando a sílaba em que foi produzida (como em “que”, “e”) e a pausa silenciosa,

com a etiqueta “PS’ ’ na camada inferior.

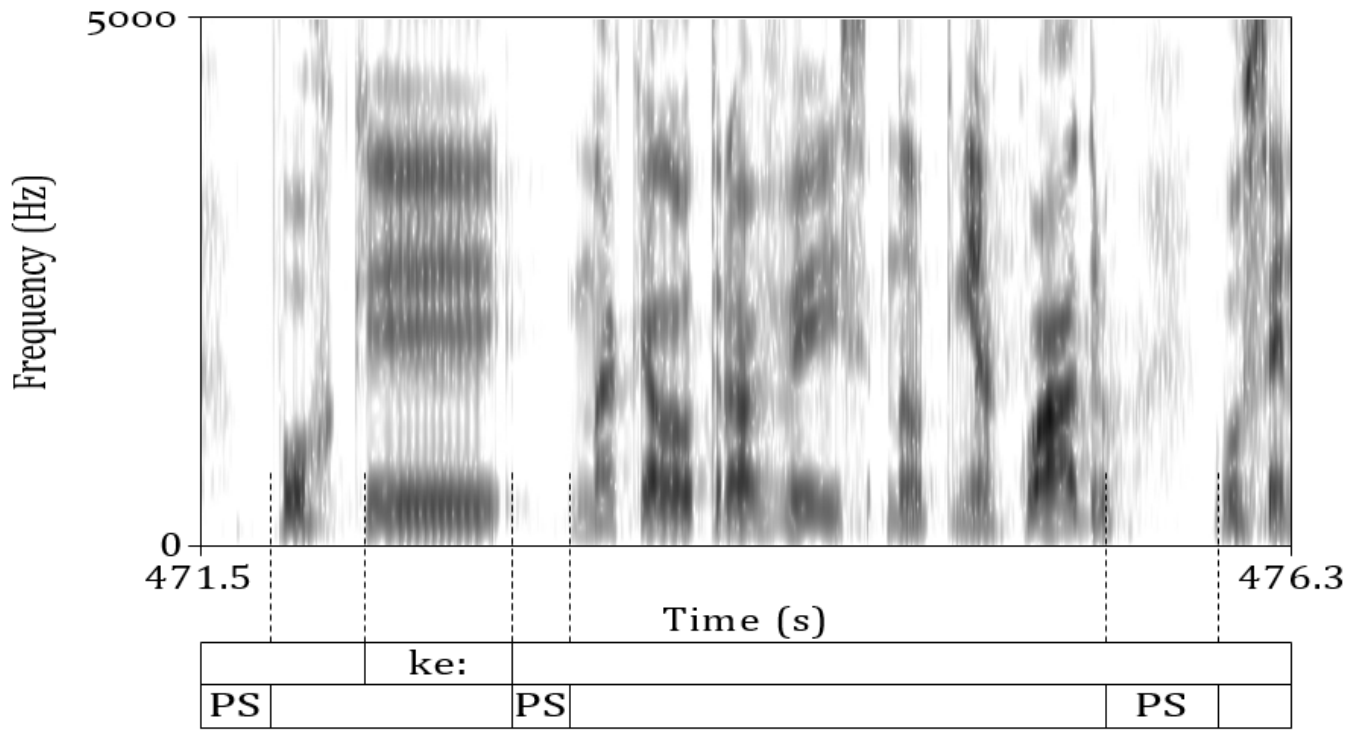

Figura 4.10 – Espectrograma de banda larga e segmentação de pausas preenchida (acima, indicando o segmento produzido) e silenciosa (abaixo, identificada por PS).

A partir dessa segmentação, tabelamos as durações em milissegundos dos dois tipos de pausa para cada locutor, bem como o tempo transcorrido entre o início da produção da pausa precedente e a pausa corrente em segundos, independentemente do tipo de pausa. Esse tempo entre pausas permite calcular a taxa de produção de pausas em cada locutor, possibilitando o exame de eventuais diferenças quanto a essa variável. Observe na Figura 4.11 o histograma das durações de pausas preenchidas e silenciosas do locutor DV.

Observe que, em geral, as pausas silenciosas têm uma gama de variação maior do que a das pausas preenchidas, exibindo valores bem mais longos. Em DV, as pausas preenchidas têm intervalo de confiança a 95%[14] de 213 a 896 ms, enquanto as silenciosas, de 214 a 1451

14 O intervalo de confiança revela em que faixa a grande maioria dos valores está concentrada. Quando é a 95% significa que 95% dos valores estão nesse intervalo.

ms. Observa-se assim que a diferença entre os tipos de pausa consiste na possibilidade de fazer uma pausa mais longa pelo uso do silêncio. A mediana das durações de pausas preenchidas é de 333 ms enquanto para a pausa silenciosa é de 603 ms, praticamente o dobro.

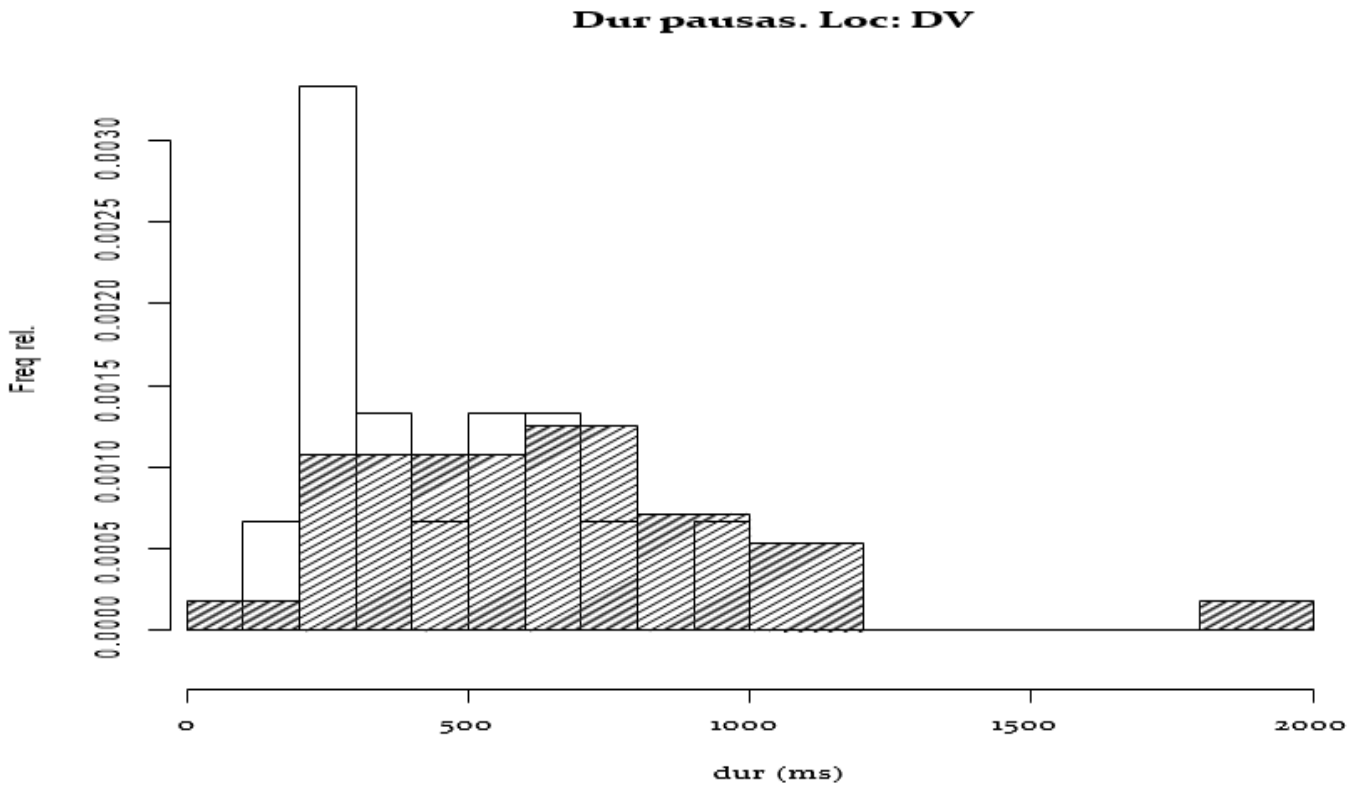

Figura 4.11 – Histogramas superpostos das durações das pausas preenchidas (retângulos claros) e silenciosas na fala do locutor DV em milissegundos.

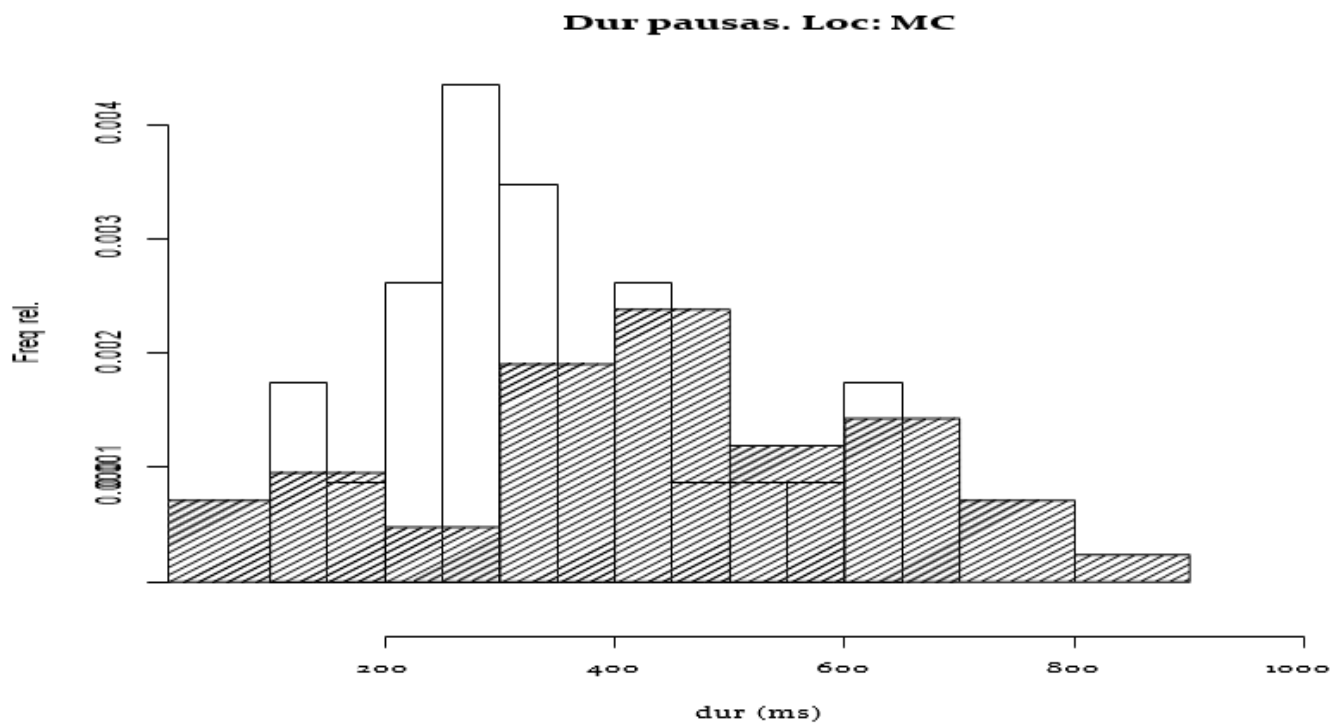

Figura 4.12 – Histogramas superpostos das durações das pausas preenchidas (retângulos claros) e silenciosas na fala do locutor MC em milissegundos.

Comparando com a fala do locutor MC, cujos histogramas de duração de pausas podem ser vistos na Figura 4.12, se vê claramente que suas pausas silenciosas também têm uma gama de variação maior do que a das pausas preenchidas. Na fala deste locutor, as pausas preenchidas têm intervalo de confiança a 95% de 126 a 604 ms, enquanto as silenciosas, de 76 a 792 ms. Observe-se que é possível ter valores bem baixos de pausas silenciosas, normalmente logo depois de uma pausa preenchida (vide Figura 4.13 para o participante MC com pausa silenciosa de 136 ms após uma pausa preenchida), estando associado ou não a um fenômeno precedente de laringalização. Em MC, a mediana das durações de pausas preenchidas é de 306 ms enquanto para a pausa silenciosa é de 449 ms. Observe que MC faz pausas silenciosas mais curtas que DV (teste de Wilcoxon com W = 226 e valor p = 0,01), enquanto a duração média da pausa preenchida não é significativamente distinta entre os dois participantes. Essa não distinção pode estar relacionada aos limites de alongamento sonoro, algo que não se dá ao fazer um silêncio, que estaria mais relacionado ao tempo para

preparar o próximo trecho de fala. Observe-se na Figura 4.14 a longa duração de pausa silenciosa que é usada por DV para reiniciar sua fala.

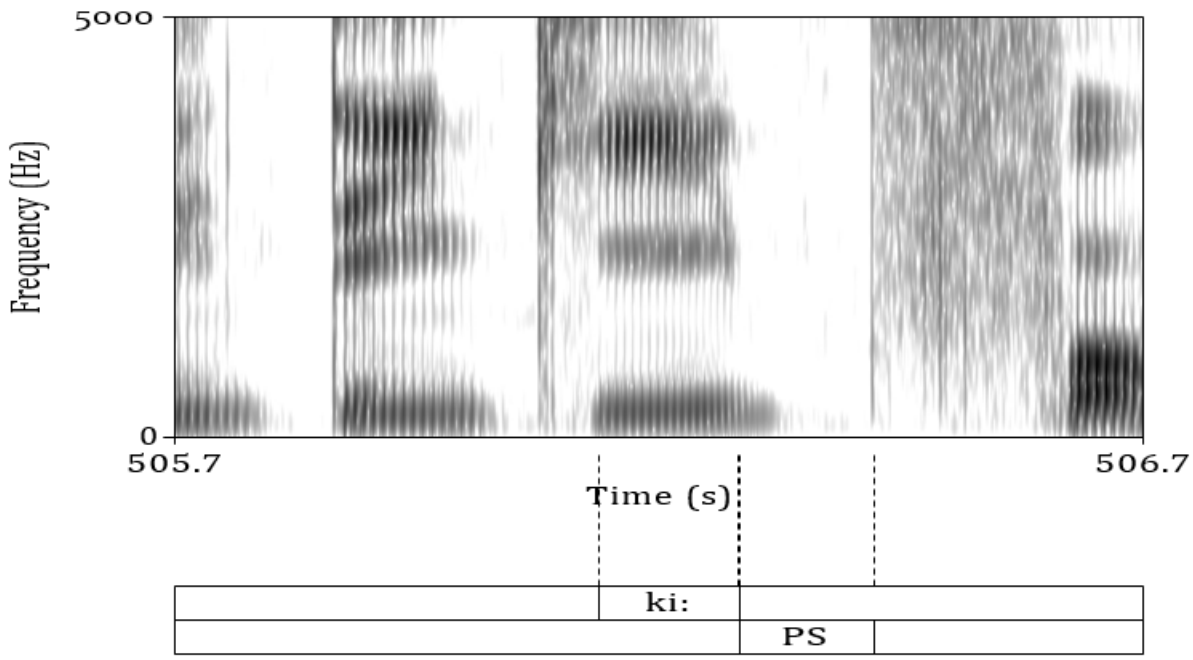

Figura 4.13 – Espectrograma de banda larga e segmentação de pausa silenciosa de duração 136 ms na fala do locutor MC.

Além da extensão das pausas silenciosas diferir entre os dois locutores, há diferenças na variabilidade das durações e na taxa de produção de pausas. De fato, o coeficiente de variação[15] do participante DV é de 50% para duração de pausa preenchida e de 58% para duração de pausa silenciosa, contra 42% e 47% respectivamente para pausa preenchida e silenciosa em MC. Assim, além de produzir em média pausas silenciosas mais curtas, MC varia menos, sendo, portanto, mais regular na produção dos dois tipos de pausa.

15 O coeficiente de variação é a razão entre o desvio-padrão e a média, medindo de forma relativa a variabilidade de uma amostra de dados.

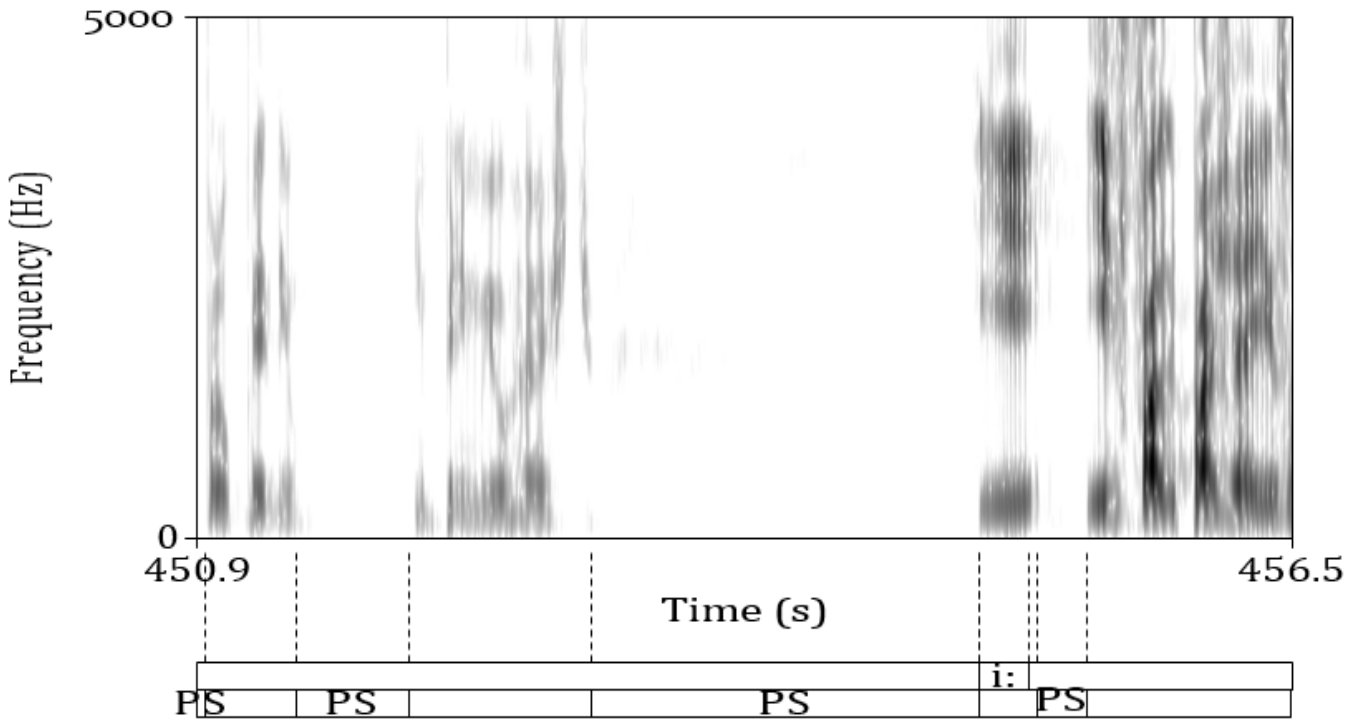

Figura 4.14 – Espectrograma de banda larga e segmentação de pausa silenciosa de duração 2000 ms na fala do locutor DV no trecho central da figura.

No que diz respeito à recorrência da produção de pausas, MC demora mais tempo a produzir uma pausa, com mediana de 1,69 s (35 pausas por minuto) contra 1,06 s (56 pausas por minuto) em DV. Quanto à variabilidade dessa produção, no entanto, ela é maior na fala de MC: 77% de coeficiente de variação contra 64% na fala de DV. Observe que, em seu conjunto, os descritores estatísticos aqui mostrados para os dois locutores permitem inferir comportamentos distintos num contexto de uma conversa telefônica. Em nenhum dos casos as pausas consideradas envolveram pausas entre turnos, foram sempre no interior de um trecho monológico. É evidente que os resultados aqui mostrados têm implicações forenses, uma vez que assinalam a possibilidade de reconhecer a "assinatura" vocal de uma pessoa pela forma como pausa.

Além da análise feita até aqui considerando o conjunto de pausas de cada tipo como um todo, é possível também examiná-las por sua função, que está associada a diferentes durações, como se depreende dos histogramas mostrados acima que, no geral, parecem apontar para

três agrupamentos possíveis. Utilizando a técnica das k-médias para as durações de pausas dos dois tipos, encontramos para DV um grupo de durações abaixo de 450 ms, outro entre esse valor e 800 ms e o último acima desse valor. Para MC os agrupamentos são as durações abaixo de 300 ms no primeiro grupo, o segundo entre esse valor e 550 ms e o último acima desse valor. Em ambos os participantes, as pausas de durações menores estão relacionadas a rápidas reformulações do que se diz, as de duração intermediária a algum tipo de microplanejamento do discurso e as maiores a uma mudança relacionada a macroplanejamento, para usar termos da pesquisa de Levelt (1989).

Tendo tirado lições da investigação das pausas para a pesquisa prosódica, convém examinar a questão das taxas de elocução e articulação, não apenas como medi-las, mas também como essas duas medidas podem revelar diferenças rítmicas eventuais entre indivíduos, estilos e comportamentos linguageiros.

4.6 Medindo taxas de elocução e de articulação

A Figura 4.15 ilustra um trecho da fala do locutor alagoano MC ao conversar por telefone com seu irmão gêmeo. A última camada é aquela que segmenta as unidades VV que, como vimos, são sílabas fonéticas que nos permitem calcular a taxa de elocução. No exemplo que consideramos aqui, tomamos um trecho de cerca de 32 s, pois autores como Arantes, Eriksson e Lima (2018) mostraram que é preciso cerca de 15 segundos para a estabilização da taxa de elocução, por isso, o trecho que escolhemos dura mais do que esse limiar.

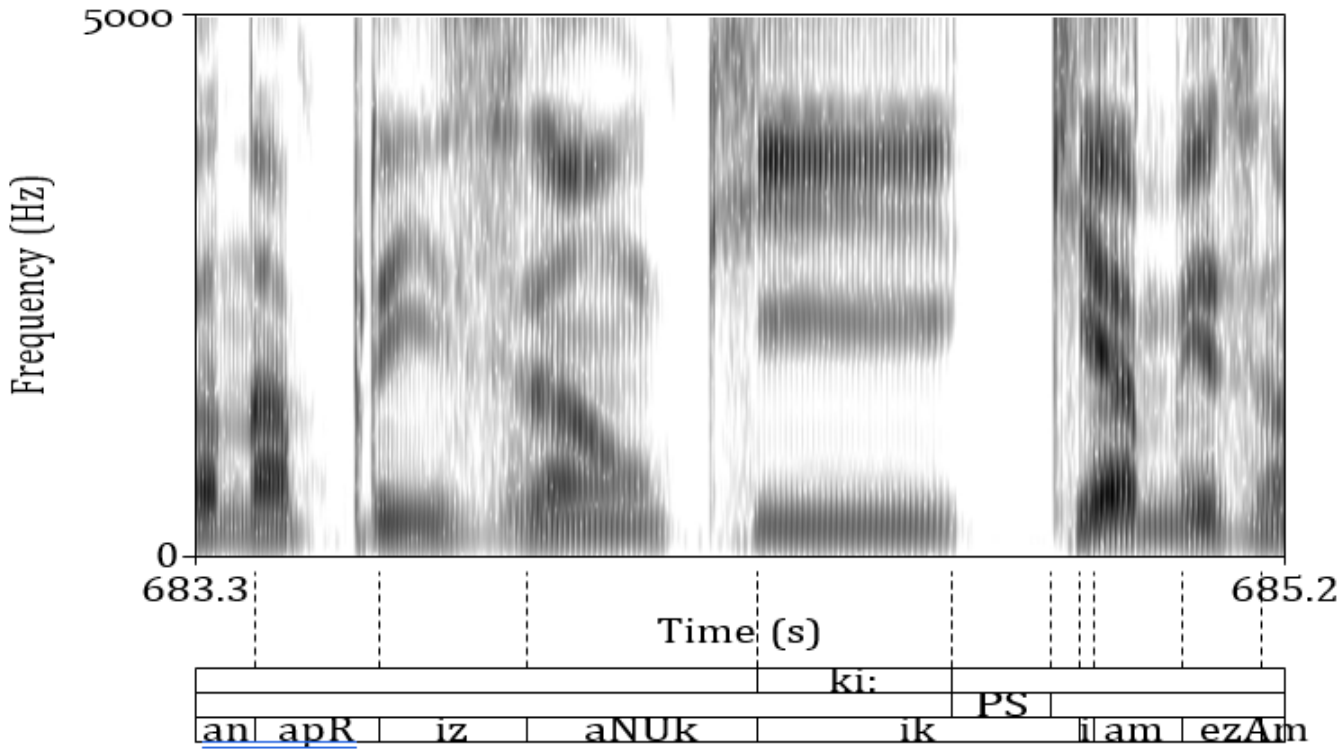

Figura 4.15 – Espectrograma de banda larga e segmentação de pausa preenchida (acima), pausa silenciosa (meio) e unidades VV (abaixo). Trecho da fala de MC, “na prisão que a mesa”.

Para calcular a taxa de elocução precisamos saber apenas duas coisas: quantas sílabas foram pronunciadas no trecho e qual a duração desse trecho, incluindo qualquer tipo de pausa. Procedendo assim para o excerto selecionado de amostra de fala de MC, temos a duração de 32,3 segundos com o número de 152 unidades VV. Dividindo o último número pelo primeiro temos a taxa de elocução de 4,7 unidades VV (sílabas fonéticas) por segundo. A taxa de articulação, por sua vez, pressupõe a retirada, do cálculo da duração do trecho, a soma do total de durações de pausas silenciosas, apenas essas, uma vez que também há som nas pausas preenchidas. A duração total de pausas silenciosas nesse trecho é de 3,33 segundos e, portanto, a duração apenas de trecho sonoro é de 32, 3 3, 33 = 28, 97 segundos, sendo a taxa de articulação a razão do número de 152 sílabas fonéticas pelo valor de trecho sonoro, o que resulta em 5,2 sílabas fonéticas por segundo.

Calculando essas mesmas medidas para o locutor DV encontramos os seguintes valores: 34,2 segundos de duração para 109

unidades VV e uma duração total de pausa silenciosa de 11,1 segundos. Isso dá 3,2 sílabas fonéticas por segundo de taxa de elocução e 4,7 sílabas fonéticas por segundo de taxa de articulação. Vê-se que a diferença maior entre os dois locutores é quanto à taxa de elocução, por conta da produção de pausas silenciosas mais longas em DV, como vimos na seção anterior.

A significância quanto à diferença entre as taxas de elocução pode ser avaliada comparando as distribuições das durações das unidades VV de cada participante, uma vez que a duração média da unidade VV é o inverso dessa taxa[16]. Utilizando o teste de Wilcoxon para comparar as médias de duração da unidade VV nos excertos dos dois participantes, confirma-se que a diferença é significativa (W = 6512,5, com valor p = 0,0018). Somente depois deste teste podemos então dizer que, nos excertos respectivos, MC fala mais rapidamente que DV (respectivamente 4,7 e 3,2 sílabas fonéticas por segundo).

No intuito de explorar ao máximo as diferenças rítmicas entre dois excertos quaisquer de fala, examinemos as diferenças nas distribuições dos grupos acentuais, tanto sua duração quanto o número de unidades VV que contêm. Com isso, podemos examinar questões de variabilidade e centralidade dessas durações em diversas situações, como entre locutores num mesmo estilo de elocução, entre dois estilos de elocução, duas atitudes ou mesmo entre duas emoções diferentes, bastando que se escolham os dados de cada distribuição.

4.7 Medindo durações de grupos acentuais

Como mostramos em outro lugar (BARBOSA, 2019) e assinalamos acima, em PB e em línguas que nesse domínio têm proeminência à direita, o grupo acentual é uma unidade que termina com uma sílaba proeminente sendo as sílabas à esquerda não proe-

16 Isto é, taxa de elocução = 1/(duração média unidade VV).

minentes. Vimos na seção 4.3 que o procedimento de normalização das durações de unidades VV permite associar os picos de *z-score* suavizados com posições proeminentes. Mostramos num trabalho anterior que a duração normalizada que corresponde a esses picos (BARBOSA, 2010), que ocorrem em uma determinada palavra que contém a unidade VV saliente acusticamente, têm uma correlação com a proporção de percepção de uma palavra como proeminente por ouvintes que varia entre 61 e 90%. O fato de não haver correspondência perfeita entre percepção de proeminência e picos de duração normalizada se dá por dois motivos.

O primeiro motivo da não correspondência entre percepção e produção tem a ver com o chamado limiar de percepção de alguma grandeza acústica. Para que percebamos que a duração silábica marca uma proeminência ou assinala uma fronteira prosódica é preciso que ela exceda um determinado valor em relação ao contexto fonético que seja capaz de atrair a atenção de nosso sistema cognitivo. Esse valor é chamado de limiar de percepção. Não é um valor fixo, mas depende do contexto, por isso é difícil de ser estimado. Mas podemos adotar como regra inicial que o *z-score* de um pico local de duração da unidade VV deve ser pelo menos acima de 1, 5 da média dos valores fora da condição de pico local.

O outro motivo da não correspondência entre percepção e produção é o fato de que percebemos numa unidade linguística mais do que a sua duração, mas também parâmetros melódicos, intensivos e a qualidade da vogal, por exemplo. Sendo assim, podemos dizer que uma palavra é proeminente por conta de um acento de *pitch* sem ter a duração maior do que a vizinhança.

Tendo feito as ressalvas acima, de um lado, as que requerem a investigação prosódica completa dos parâmetros que assinalam proeminência e, de outro lado, a correspondência em sua maioria dos picos locais de duração normalizada de unidade VV com proeminências, é possível assumir que esses picos marcam a proeminência e que, por-

tanto, terminam um grupo acentual. A vantagem dessa assunção é a automatização do procedimento de detecção de grupos acentuais.

De fato, há alguns anos implementamos o script *SGDetector* para o Praat, que realiza a normalização das durações de unidades VV previamente segmentadas e etiquetadas, gerando assim os valores de *z-scores* suavizados e a identificação dos máximos locais que são os picos de duração que assinalam a fronteira à direita do grupo acentual. O script também gera um arquivo com a duração e o número de unidades VV em cada grupo. Essa riqueza de informação serve para avaliar também diferenças rítmicas entre trechos de fala. As aplicações são as mesmas mencionadas anteriormente, a de avaliar a distância rítmica entre locutores e entre estilos de elocução. O script requer apenas a camada de anotação do Praat, o objeto TextGrid, bem como uma tabela de referência de médias e desvios-padrão da duração de fones da língua, que é fornecida juntamente com o script e disponível para o PB, o português europeu, o alemão, o espanhol, o francês, o sueco e o inglês britânico, conforme explicado em seu repositório em https://github.com/pabarbosa/prosody-scripts.

O exame da extensão dos grupos acentuais complementa o observado nos histogramas de picos de durações normalizadas que vimos na seção 4.4.2 para os mesmos locutores, ocasião em que se observou que há valores mais extremos de *z-score* na narração e, portanto, da duração de unidades VV salientes, o que contribui para grupos acentuais mais longos nesse estilo de elocução. Pelos diagramas de blocos da Figura 4.16 é possível ver claramente, para os locutores FA (homem) e LC (mulher), uma mediana de duração maior dos grupos acentuais na narração. A diferença tomando-se os três locutores é significativa por um teste de Wilcoxon ($W = 28718$, com valor $p = 0{,}003$) com valores de mediana de 1681 ms no estilo narração e 1439 ms no estilo leitura, 242 ms a menos. O intervalo de confiança a 95% vai de 538 a

3088 ms na leitura e de 514 a 3960 ms na narração. Observar que esse limite superior em torno de 3 s na leitura corresponde ao tempo da leitura de um verso alexandrino. Assim, a poesia exploraria os limites da extensão de um grupo acentual. Um resultado semelhante para o caso do hemistíquio no verso alexandrino e sua relação com o número de sílabas fonéticas mediano é apontado adiante.

Quanto à variabilidade da duração do grupo acentual, somente encontram-se diferenças para LC entre seus dois estilos 684 (RE) e 1084 ms (NR), com p = 0, 06 em teste de permutação para comparação pareada de variâncias.

Quanto ao número de unidades VV por grupo acentual, se vê na Figura 4.17 que a maior diferença entre esses números para os dois estilos ocorre para AV e FA, em que há maior número de unidades na leitura: medianas de 5 (FA) e 6 (AV) unidades VV na narração comparado a 6 (FA) e 7 (AV) unidades VV na leitura. Essa mediana em torno de seis sílabas fonéticas corresponde a um hemistíquio, a metade de um verso alexandrino, ponto em que se costuma fazer uma pausa ao se declamar e, portanto, fronteira de grupo acentual, numa declamação em que as duas únicas proeminências são a palavra final dos primeiro e segundo hemistíquios. O intervalo de confiança a 95% na leitura para os três locutores é de 2 a 11 unidades VV na narração e de 3 a 11 unidades VV na leitura. Não há diferença alguma quanto à variabilidade, nem entre estilos nem entre locutores.

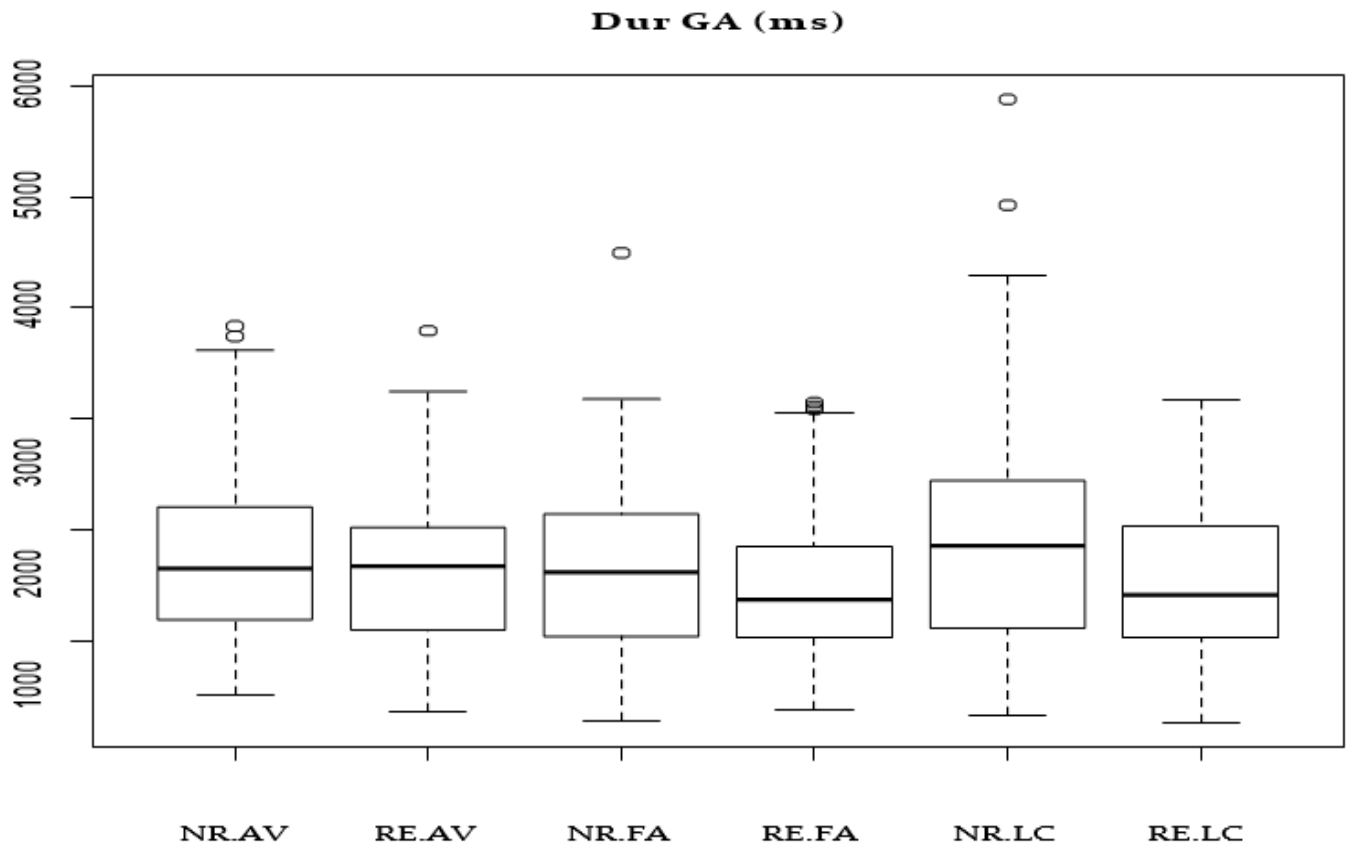

Figura 4.16 – Diagramas de blocos da duração dos grupos acentuais em milissegundos de três locutores paulistas (AV, LC, FA) nos estilos lido (RE) e narrado (NR).

Complementando a técnica de cálculo de distância rítmica entre locutores, feita ao nível da unidade VV, o exame dos grupos acentuais que acabamos de fazer promove uma compreensão de que o estilo narrativo tem sílabas fonéticas mais longas, grupos acentuais mais extensos temporalmente, mas muito pouco a mais em ternos de número dessas sílabas. Grande parte desse alongamento está relacionado ao planejamento do discurso que conta também com a presença de trechos sonoros hesitativos, que são pausas preenchidas, como vimos acima.

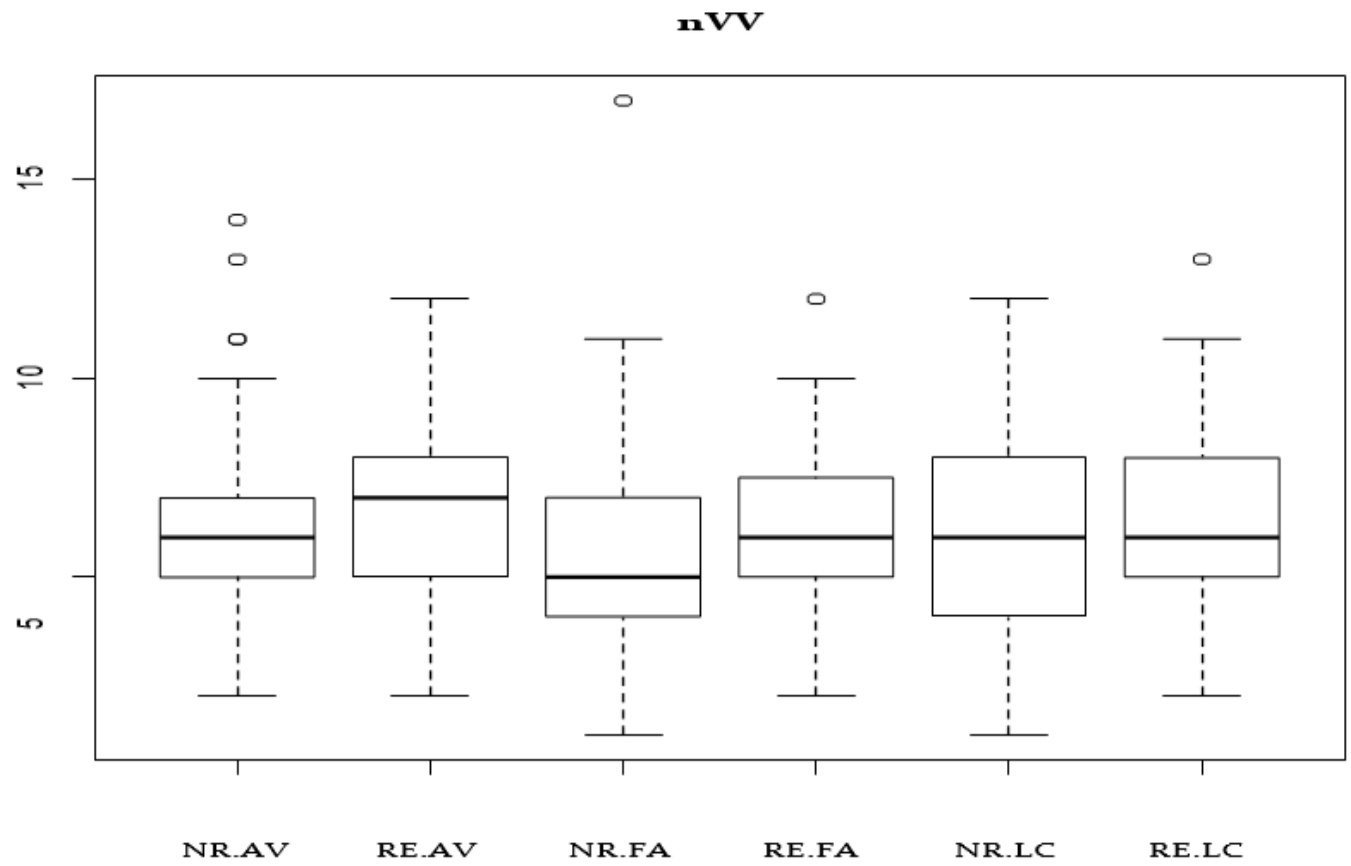

Figura 4.17 – Diagramas de blocos do número de unidades VV nos grupos acentuais de três locutores paulistas (AV, LC, FA) nos estilos lido (RE) e narrado (NR).

4.8 Medindo durações de eventos de natureza dialógica

A teoria da Língua em Ato, formulada por Cresti (2000), avalia as ilocuções por seu perfil prosódico, que vai determinar sua função no enunciado. A teoria propõe, a partir da pesquisa em corpora do italiano, mas corroborado pela pesquisa em corpora do PB pelos trabalhos de Raso (2012) e Raso e Mello (2012), seis unidades dialógicas que se distinguem das unidades encontradas em monólogos por não serem composicionais sintaticamente com o resto do enunciado nem contribuí- rem para a interpretação de seu significado. Por essa definição negativa, essas unidades são aquelas que nas demais abordagens se chamam de marcadores discursivos.

O trabalho de Gobbo (2019) examina, do ponto de vista da análise

prosódico-acústica embasada estatisticamente, essas seis unidades em um corpus de fala espontânea do PB mineiro. Dentre as seis unidades que ele investigou, vamos ilustrar aqui o incipitário, o conativo, o alocutivo e o fático. O incipitário é a unidade dialógica que marca o início de um turno e normalmente tem uma duração curta em relação ao contexto imediato, um valor elevado de Fo e uma maior intensidade (GOBBO, 2019, p. 11). Essas mesmas características são encontradas no conativo, mas sem um padrão claro para o perfil melódico, unidade usada para encorajar o interlocutor. O alocutivo interpela o interlocutor sendo de baixa intensidade e normalmente ao final do enunciado com perfil melódico baixo e nivelado. Já o fático é a unidade dialógica de menor duração, usada para assinalar o interlocutor que está sendo ouvido, mantendo o canal de comunicação aberto. Exemplos dessas unidades serão mostradas a seguir para apontar a dificuldade de sua segmentação.

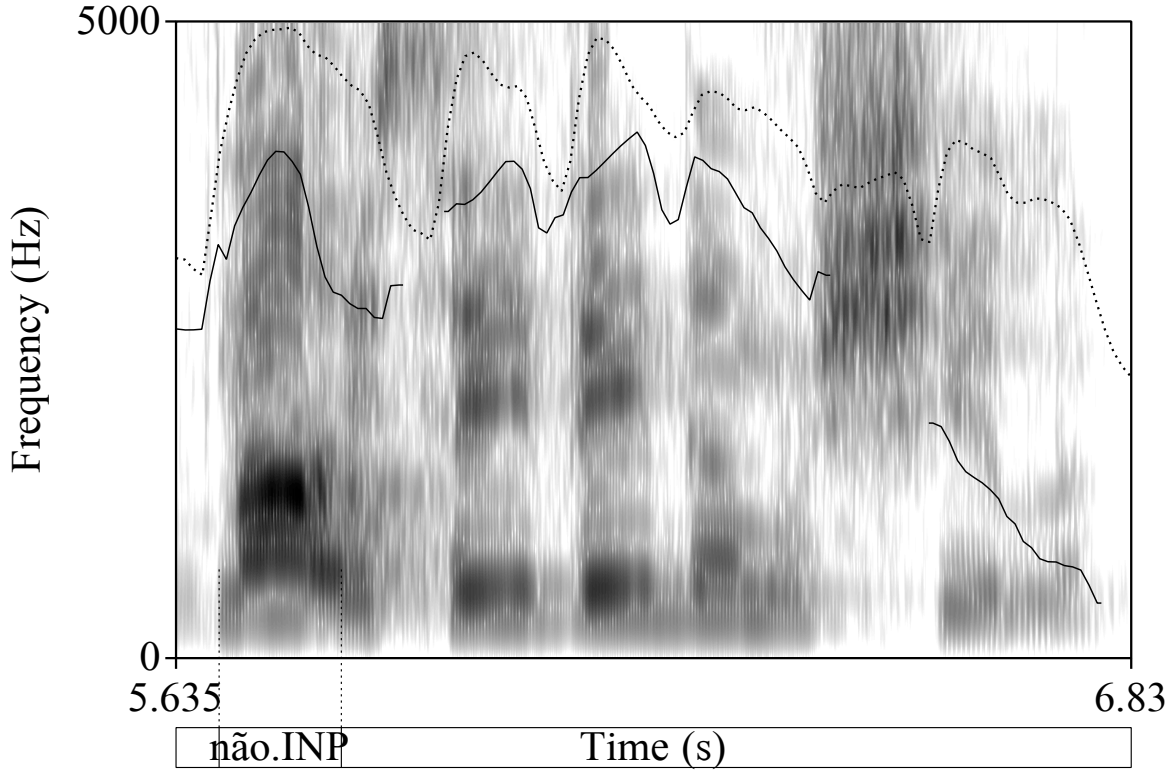

Figura 4.18 – Espectrograma de banda larga e curvas de Fo (cheia) e intensidade (pontilhada) do trecho “não, isso aí veio da mochila” do locutor 1 tendo sido segmentado o incipitário “não”.

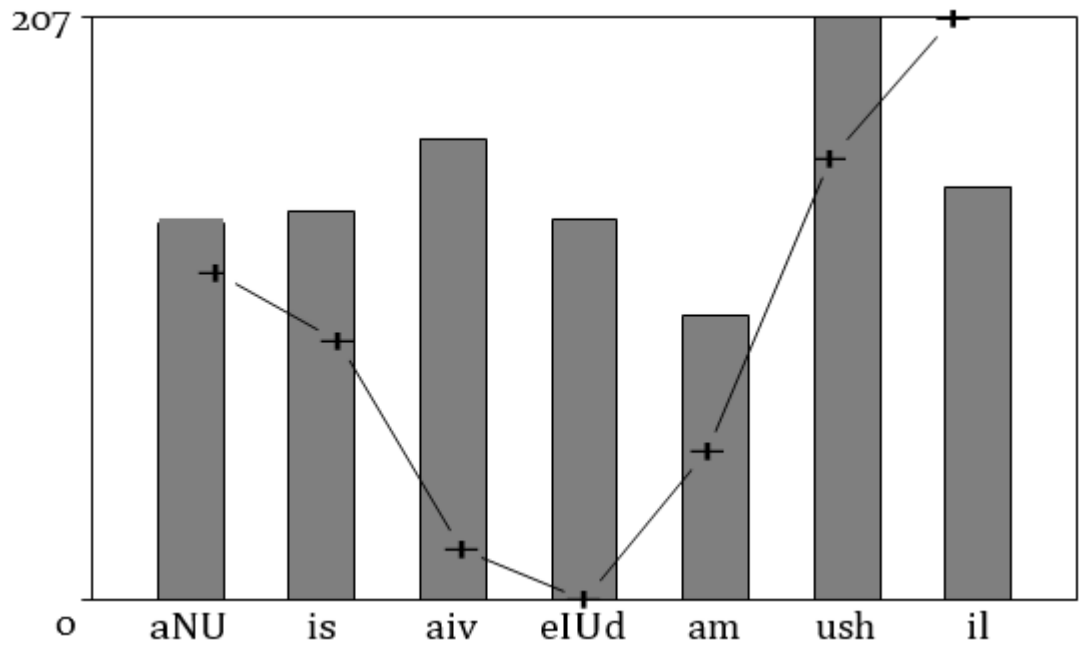

Figura 4.19 – Valores de duração bruta (ms) e *z-score* suavizado das unidades VV do trecho “não, isso aí veio da mochila” do locutor 1. A primeira unidade VV corresponde à rima de “não”.

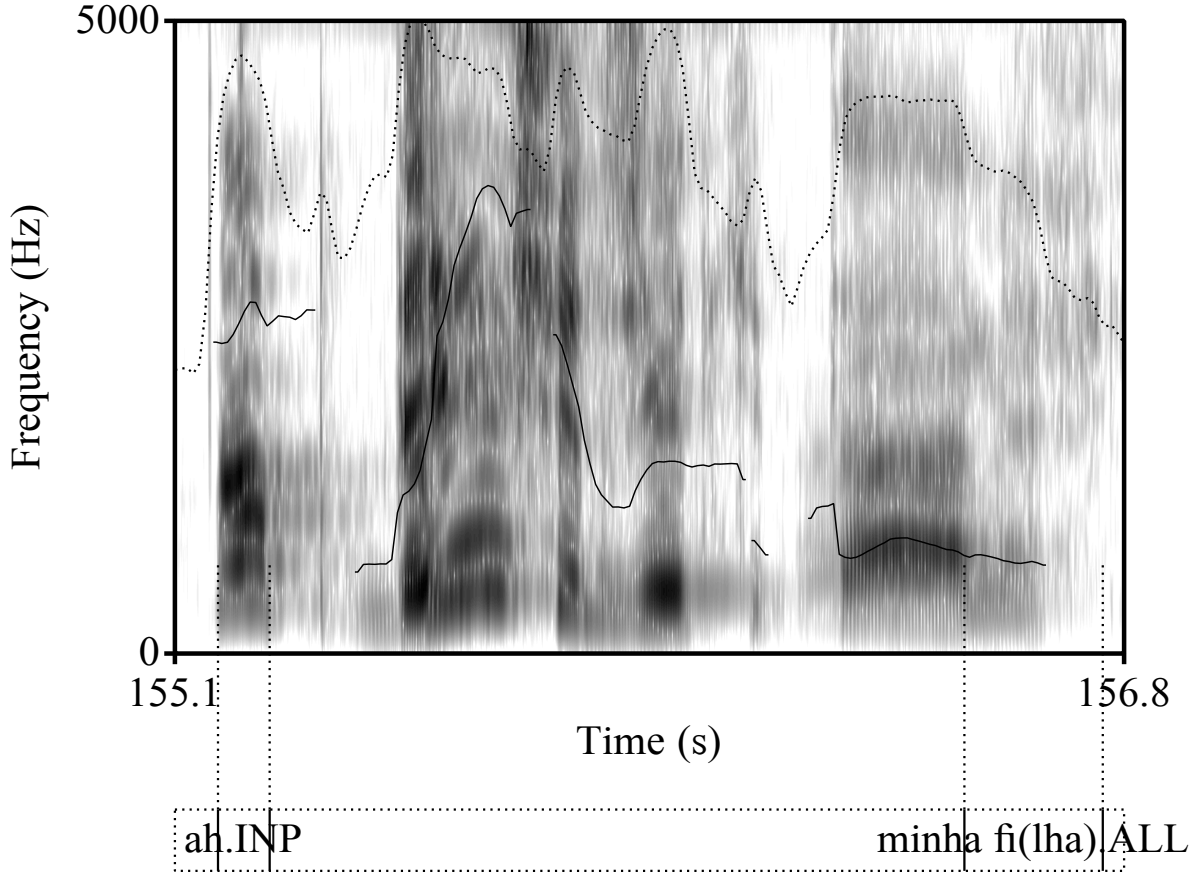

Figura 4.20 – Espectrograma de banda larga e curvas de Fo (cheia) e intensidade (pontilhada) do trecho “ah! deixa do jeito que tá, minha fi(lha).” do locutor 1 tendo sido segmentados o incipitário “ah” e o alocutivo “minha fi(lha)”.

Os trechos que seguem foram extraídos do corpus C-ORAL-Brasil (RASO; MELLO, 2012), através do endereço http://www.c-oral-brasil.org/. O primeiro é um diálogo entre dois estudantes de pós-graduação mineiros que falam sobre o empacotamento de material de gravação nas dependências da UFMG. Os interlocutores são um homem (locutor 1) e uma mulher (locutor 2). Falam durante cerca de 7,5 minutos e produzem no total 243 enunciados e 32 unidades dialógicas, sendo 21 dos tipos incipitário, alocutivo e conativo.

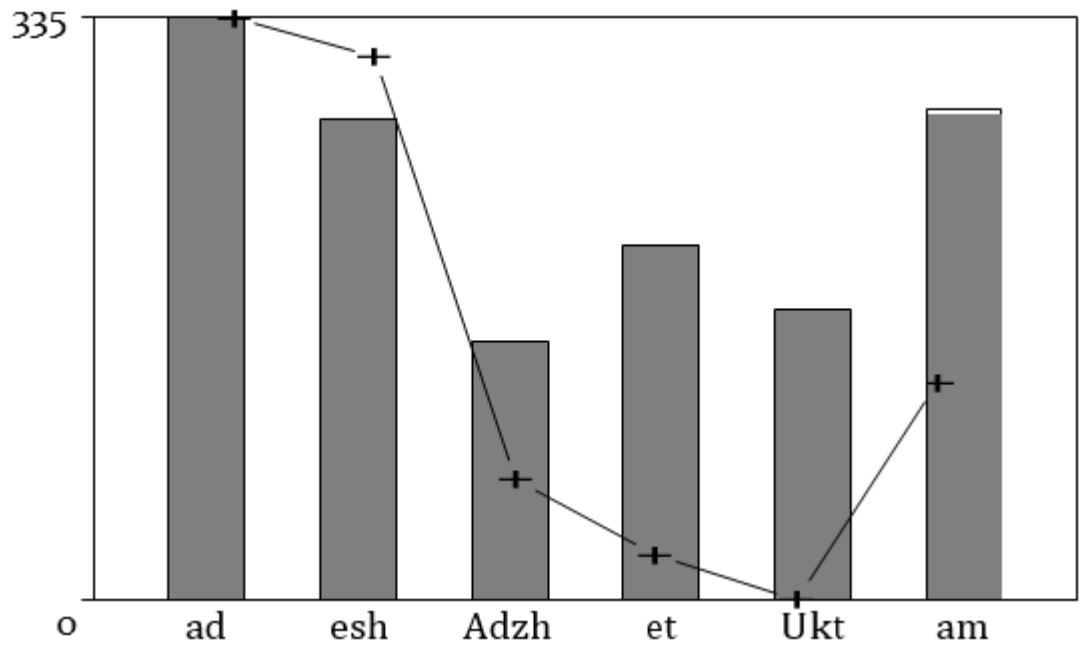

Figura 4.21 – Valores de duração bruta (ms) e *z-score* suavizado das unidades VV do trecho "ah! deixa do jeito que tá, minha fi(lha)." do locutor 1.

No trecho ilustrado na Figura 4.18 com espectrograma de banda larga, curva de F0 e de intensidade correspondente ao enunciado "não, isso aí veio da mochila" do locutor 1, ilustramos o incipitário "não", com duração de 153 ms e F0 médio de 242 Hz. Essa duração é compatível com as das sílabas seguintes (média de 184 ms), o valor médio de F0 é superior aos que seguem como se vê pela curva F0 descendente e a intensidade é maior do que a do restante do enunciado. Por conta da intensidade maior, sua delimitação não apresenta maior dificul-

dade. Do ponto de vista dos parâmetros prosódicos, é importante ter em mente que a relação de seus valores com o contexto imediato é importante para entender a função da unidade do ponto de vista pragmático. O trecho pode ser ouvido em **NaoLoc1INP**.

A relação da duração da unidade dialógica com a vizinhança fonética pode ser vista na Figura 4.19 tanto para a duração bruta quanto para a normalizada. Fica claro que o incipitário "não" é um pico local e portanto constitui um grupo acentual de uma unidade. Sua duração é intermediária às demais do grupo acentual seguinte.

No trecho ilustrado na Figura 4.20, com espectrograma de banda larga, curva de F0 e de intensidade correspondente ao enunciado "ah! deixa do jeito que tá, minha fi(lha)" do locutor 1, ilustramos o incipitário "ah" e o alocutivo "minha fi(lha)". O primeiro dura 93 ms com Fo médio bem elevada, de 348 Hz, compatível com as características descritas para essa unidade dialógica. Já o alocutivo, embora o trecho dure cerca de 250 ms, a duração média das unidades VV é pouco maior de 80 ms. O valor baixo e nivelado de Fo e menor de intensidade é compatível com sua descrição prosódica. A delimitação do segmento acústico não é simples e parece terminar sem que a última sílaba se pronuncie e com grau elevado de ensurdecimento. Recomendamos por isso a escuta atenta do trecho em **AhFiLoc1INPALL**.

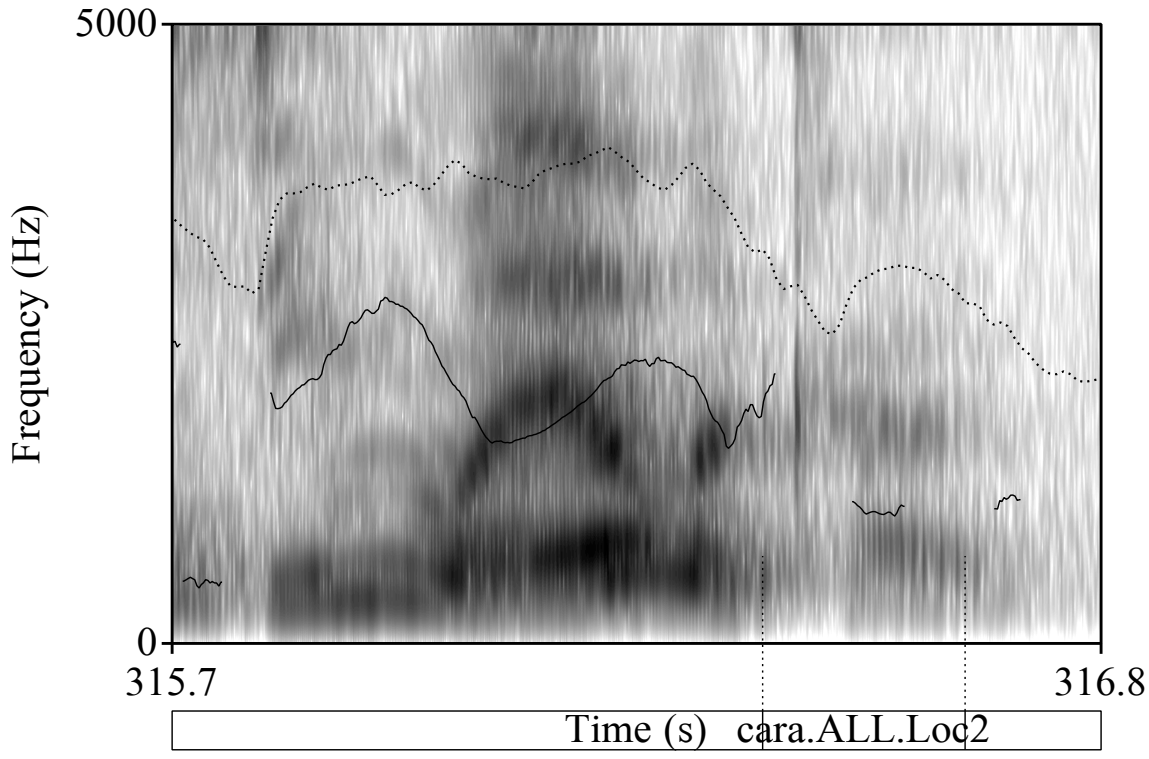

Figura 4.22 – Espectrograma de banda larga e curvas de Fo (cheia) e intensidade (pontilhada) do trecho “cê tá igualzinho ela, cara” do locutor 2 tendo sido segmentado o alocutivo “cara”.

A relação da duração da unidade dialógica com a vizinhança fonética pode ser vista na Figura 4.21 tanto para a duração bruta quanto para a normalizada. Também nesse caso o incipitário “ah” constitui um grupo acentual de uma unidade e todo o restante, comparado a essa unidade, tem duração normalizada menor. Ele se destaca em duração e também, como se viu na Figura 4.20, em F0 e intensidade. O trecho de “minha filha” só pôde ser medido até o [m] por falta de realização plena da vogal [i], por isso não se pode comentar nada a respeito de seu papel quanto à duração normalizada.

No trecho ilustrado na Figura 4.22, correspondente ao enunciado “cê tá igualzinho ela, cara” pela locutora 2, ilustramos o alocutivo “cara”, cuja duracão é de 232 ms (média de 116 ms por sílaba). Trata-se de um trecho ruidoso com Fo e intensidade bem mais baixas do que no trecho precedente. A delimitação do segmento acústico, que pode ser ouvido em **CaraLoc2ALL**, também não é simples.

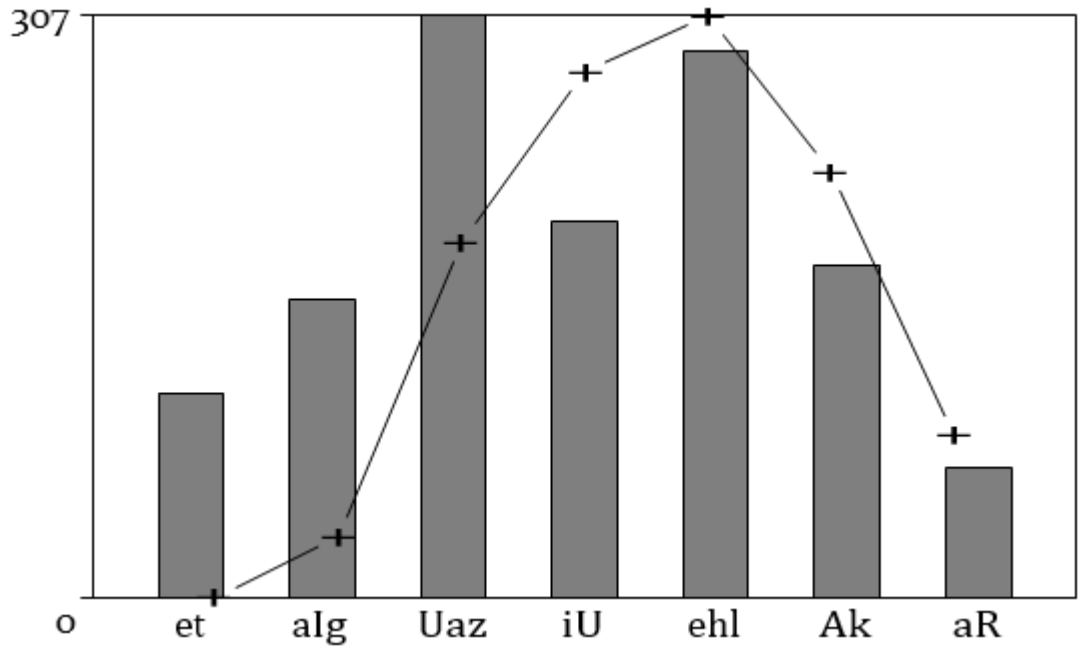

Figura 4.23 – Valores de duração bruta (ms) e *z-score* suavizado das unidades VV do trecho "cê tá igualzinho ela, cara" do locutor 2.

Para esse trecho, a relação de duração do alocutivo "cara" com a vizinhança fonética pode ser vista na Figura 4.23. O primeiro grupo acentual se encerra na tônica de "ela" e o alocutivo vem encerrar o grupo acentual final com uma duração normalizada que é fruto de uma diminuição progressiva desde o acento frasal em "ela".

O interesse dessas ilustrações em contexto dialógico é mostrar o cuidado que se deve ter em sua mensuração, não apenas em considerar valores relativos à vizinhança fonética, como também ter em consideração as unidades que, de fato, podem ser delimitadas com segurança e ter seus valores de F0 calculados com precisão, como mostraremos no próximo capítulo. Em todo diálogo em que existe uma certa familiaridade entre os interlocutores, e é o caso aqui, há muitos casos de superposição de fala que, sem microfones que cancelem completamente a fala do outro, devem ser descartados para análise acústica. O estudo da superposição de fala em si é relevante para uma compreensão das instâncias dialógicas, mas requer um equipamento que permita a separação das falas de cada interlocutor. Para um estudo acústico da

superposição ler o trabalho de Valle-Barbosa (2013). Além de superposições de fala, uma série de outros eventos sonoros ocorrem, especialmente num contexto dialógico, como tosses, risos e ruídos de inalação ou expiração, entre outros eventos.

4.9 Medindo durações de eventos sonoros não linguísticos

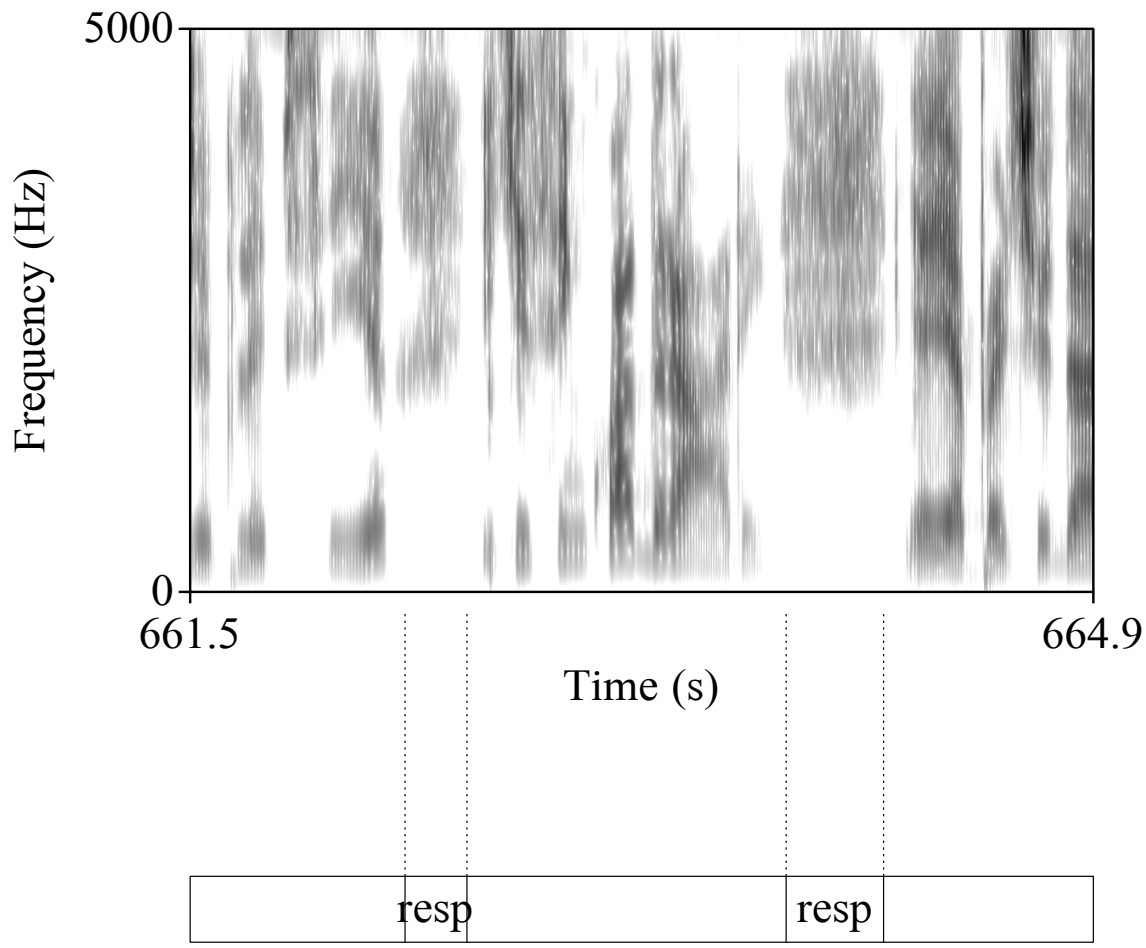

Figura 4.24 – Trechos com RRA em locutora feminina paulista de cercade 25 anos (MI da tabela 4.6). O primeiro ruído tem menor duração eé menos intenso porque marca uma fronteira prosódica mais fraca do que aquela em que a locutora produz o segundo ruído.

Eventos sonoros não linguísticos, também chamados de vocalizações não verbais (VNV), são produções de um indivíduo ao longo de sua fala e que são mais numerosos em diálogos e conversas. Exem-

plos dessas VNV são inalações e expirações audíveis, também chamadas de ruídos respiratórios audíveis (RRA), tosses, risos, risadas, gargalhadas, sopros, suspiros, bocejos, estalos de língua e lábios, puxadas de ar fortes com o nariz. Pela análise de seis corpora de conversação, Trouvain e Truong (2012) evidenciaram que, dentre esses, os RRA e os risos/risadas são de longe os mais frequentes. A importância desses eventos reside no fato de que podem revelar informações a respeito de níveis linguísticos, paralinguísticos e extralinguísticos no discurso, como a segmentação prosódica, carga cognitiva, estado afetivo e identidade do locutor (TROUVAIN, 2014).

De fato, Grosjean e Collins (1979) mostram que, tanto na fala lida como na espontânea, ruídos de inalação são encontrados durante pausas em fronteiras prosódicas fortes. Pausas que incluem esses ruídos são mais longas, como ilustrado na Figura 4.24 nos dados de entrevistas informais em PB. O primeiro ruído assinala a fronteira entre os trechos “e aí o que foi sugerido pelo estatístico é assim: a gente vai convidar todas as crianças do ambulatório que tiverem a ressonância que mostra o tempo de epilepsia” e o complemento “e que se enquadrarem”, enquanto o segundo ruído ocorre antes de um novo tema do relato, que começa por “e aí, a partir daí a gente fez por tipo de coleta”, portanto durante um intervalo que marca uma fronteira prosódica forte.

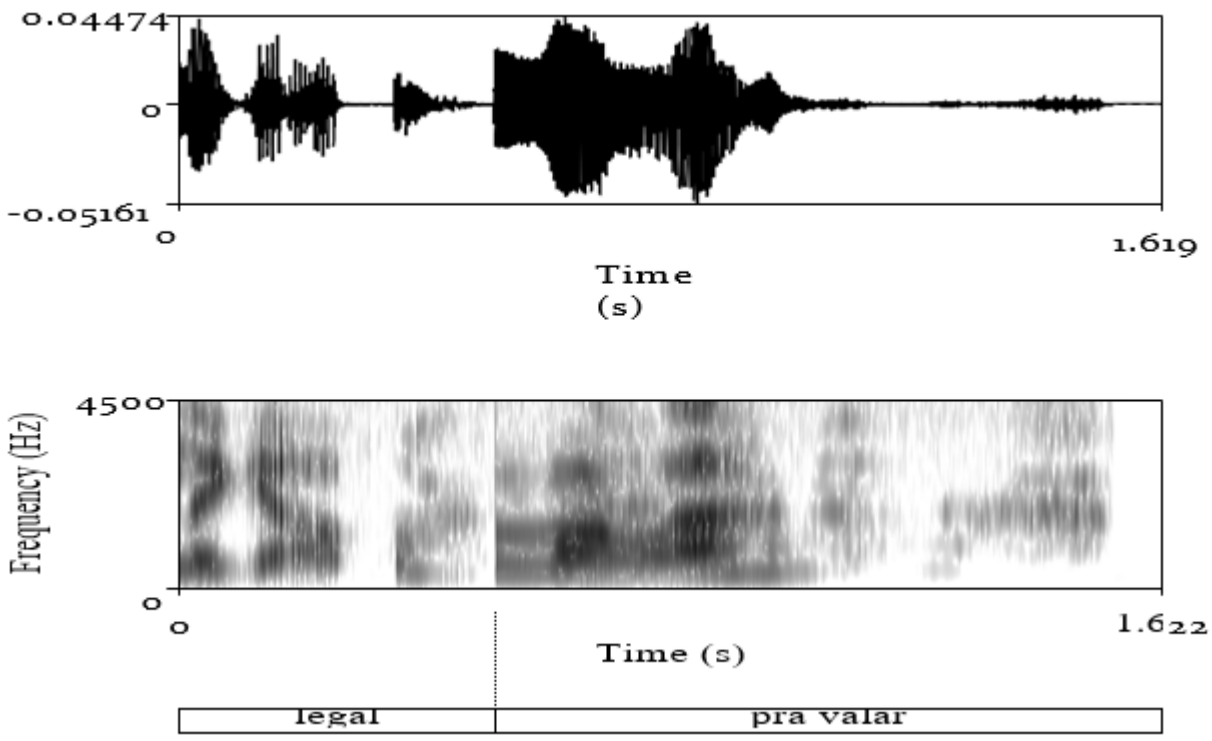

Figura 4.25 – Dois trechos de riso associados à fala em momentos distintos da locutora HD, um durante "legal" e o outro durante "pra falar".

Quanto à possibilidade de identificação de um indivíduo, RRA e correlatos acústicos como duração, intensidade e composição espectral podem ser discriminantes entre as pessoas (LINK, 2012; LAUF, 2001). Não é verdade que, ao ouvir a tosse de uma, entre outras pessoas co- nhecidas, sabemos de quem se trata?

Para ilustrar as diferenças duracionais e a frequência das diferentes VNV, tomamos, de um corpus com entrevistas informais, três locutores masculinos e três femininos entrevistados por seus amigos próximos. A razão de um corpus dessa natureza é o fato de assegurar um diálogo mais longo e a possibilidade de aparecerem eventos de riso, bem como VNV distintas dos RRA, pelo grau de familiaridade entre os interlocutores.

Da tabela 4.6 podemos ver que, nos homens, a duração de uma VNV foi cerca de 14% da duração total do diálogo, enquanto nas mulheres variou entre cerca de 10 a 21% com média semelhante à dos homens. A frequência de VNV tende a ser superior nas mulheres, va-

riando de cerca de 9 a 12 por minuto contra 5 a 10 por minuto nos homens. Os RRA são majoritariamente mais frequentes entre as VNV, com exceção de HD, que exibe frequência relativa de RRA semelhante à do riso. Em geral, o riso dura em média de duas (homens) a três vezes (mulheres, com exceção de HD que, por outro lado, ri muito mais frequentemente que todos os demais) mais do que duram os RRA. Nem todo riso é igual (TROUVAIN, 2014): pode ocorrer durante um trecho de fala, sob a forma de ruído respiratório mais forte antes ou, frequentemente, depois de um trecho de riso associado à fala, pode ser feito com uma sílaba curta repetida algumas vezes (o famoso "hahaha"). A locutora HD ilustra bem essa variação, como se vê nas ilustrações que seguem. Na Figura 4.25 vêem-se a forma de onda e o espectrograma de banda larga de trechos de riso ao longo das expressões "legal" e "pra falar". O trecho de fala é seguido de uma forte expiração nos dois casos, que completa a sensação de riso.

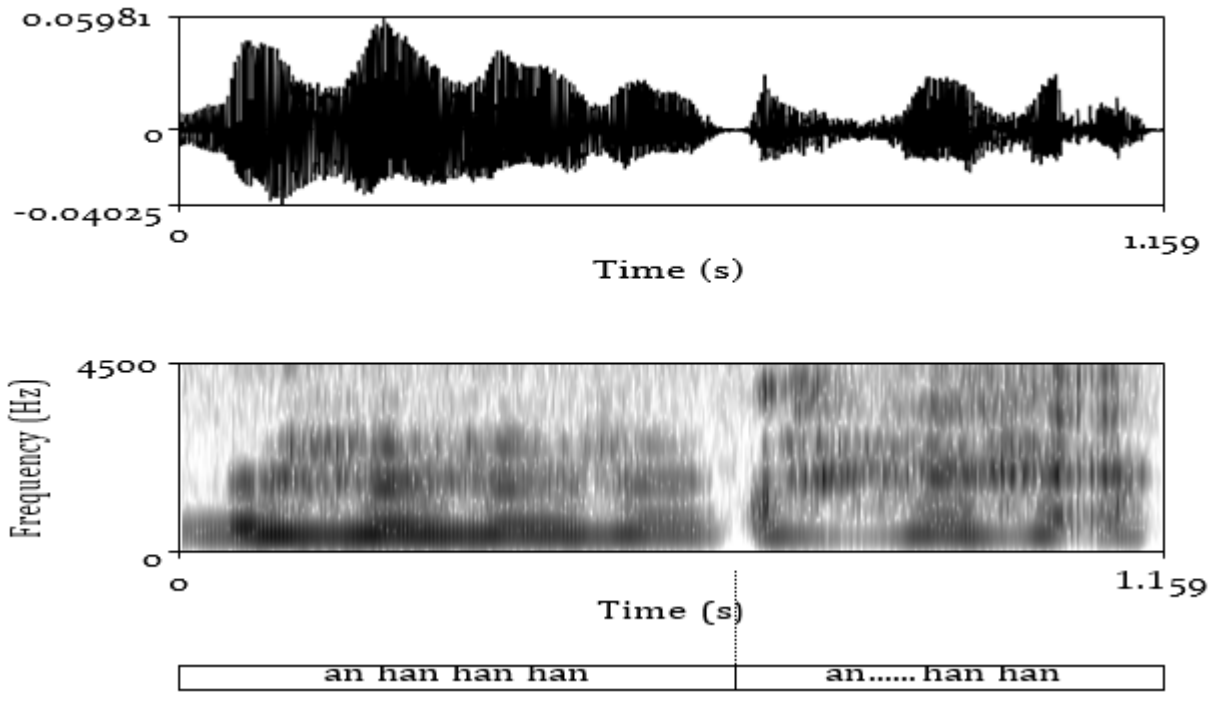

Figura 4.26 – Dois trechos de riso por repetição de sílaba da locutora HD.

Já na Figura 4.26, a mesma locutora repete uma sílaba semelhante a [hã] nas duas sequências extraídas de momentos diferentes com diferentes inícios. Além de riso, houve um episódio de gargalhada, com a locutora MM, que durou 887 ms.

Não houve episódio de riso no locutor FD, mas, por outro lado, bocejou, suspirou, puxou forte o nariz, soprou e fez um estalido com os lábios, embora um a dois eventos de cada. Vemos assim que a natureza das VNV pode ser bem distinta, bem como a frequência relativa de algumas delas, como os risos. A variação duracional do riso é, como se vê na tabela 4.6, maior do que dos RRA, com coeficiente de variação de cerca de 50% contra cerca de 30% no RRA.

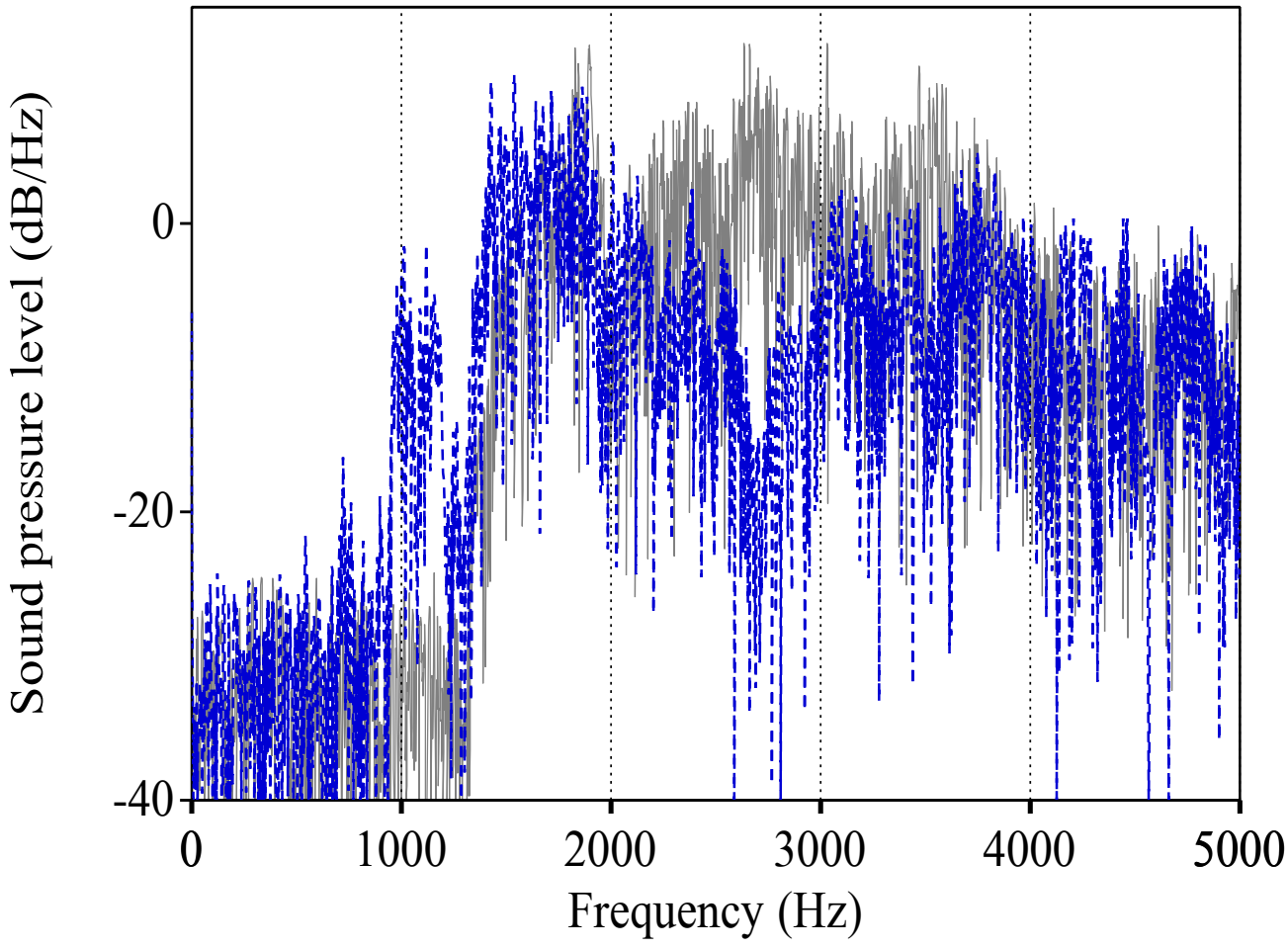

Figura 4.27 – Espectros de Fourier de um evento de RRA dos locutores AX (claro) e FD (escuro), onde se vê claramente que as frequências mais intensas estão acima dos 1000 Hz, com regiões de ressonâncias mais afastadas em FD, com dois grandes lobos (entre 1500 e 2000 Hz e depois de 3000 a 4000 Hz) do que em AX, com uma concentração maior entre 2000 e 4000 Hz.

Além da importância para a descrição do ritmo, as VNV podem ser usadas como pistas para diferenciação entre locutores. Por exemplo, a Figura 4.27 mostra os espectros de Fourier de um evento único de RRA em dois locutores masculinos distintos onde se vêem nítidas diferenças espectrais. Para além de revelar aspectos individuais pelo ruído que ressoa de forma audível no trato vocal, a atividade respiratória em si pode ser observada por dispositivo específico, permitindo entender como se dá a coordenação entre respiração para a fala e a própria fala.

4.10 Medidas de grupos respiratórios

A Figura 4.28 ilustra os sinais respiratórios de uma locutora alemã fluente em inglês, lendo de forma persuasiva um texto nessa língua para vender um produto. O sinal de cima é da variação de expansão do tórax ao longo do tempo e, o de baixo, da variação de expansão do abdômen, ambos em medidas arbitrárias. Concentrando-nos nos movimentos expiratórios e, portanto, de diminuição dos valores do sinal ilustrado na figura, vê-se que há movimentos simultâneos de expansão do abdômen, revelando um não sincronismo entre as duas cavidades durante a fala. Para os grupos respiratórios 1, 2 e 4, há no entanto uma forte proximidade entre os valores máximos e mínimos dos movimentos de ambas as cavidades.

Tabela 4.6 – Descritores duracionais de VNV em diálogos de seis locutores paulistas com seus amigos respectivos. Apenas as VNV mais frequentes, RRA e risos, são mostradas, mas todas foram medidas. As medidas são: duração total de VNV em segundos (durT), porcentagem em relação à duração do diálogo (%Dial), o número de VNV por minuto (#/min). Para cada tipo de VNV, são informadas a média, o desvio-padrão (entre parênteses) e o intervalo de confiança a 95% da duração em ms. Para todos os locutores, com exceção da locutora HD, a frequência relativa de RRA é superior a 95%. Para HD, os RRA são 50% de todos as VNV, com 41% de risos e o restante dividido entre quatro suspiros e um pigarro. Os três primeiros locutores são masculinos e os três seguintes, femininos.

fal.	durT	%Dial	#/min	RRA	risos
AX	51,2	14,0	8,3	419 (154)	
				223 a 840	950 (301) 762 a 1324
FD	23,8	14,2	4,7	512 (177)	
				275 a 1050	-
MD	40,6	14,3	9,5	307 (111)	
				165 a 548	592 (302) 388 a 794
HD	19,0	10,4	11,9	294 (129)	
				124 a 570	333 (149) 186 a 669
MI	70,5	20,7	8,9	358 (118)	
				181 a 694	1237 (881) 320 a 2596
MM	50,4	14,5	10,4	308 (130)	
				134 a 618	948 (405) 433 a 1500

Eles foram obtidos em gravação simultânea da fala com microfone unidirecional no contexto de pesquisa sobre a coordenação entre fala e respiração quando da persuasão (BARBOSA; NIEBUHR, 2020). Para a gravação dos movimentos de expansão do tórax e do abdômen, foi usado o dispositivo Resp Track, projetado e construído na Universidade de Estocolmo por Johan Stark. Os sinais aqui mostrados foram obtidos com a locutora de pé, com o texto apresentado à sua frente, na altura dos olhos. O dispositivo é fundamentado no princípio do *Respiratory Inductance Plethysmography* (RIP), pletismógrafo respiratório de indutância, que mede mudanças na área da seção transversal tanto da caixa torácica quanto do abdômen por meio de duas cintas, uma na altura das axilas e outra na altura do umbigo.

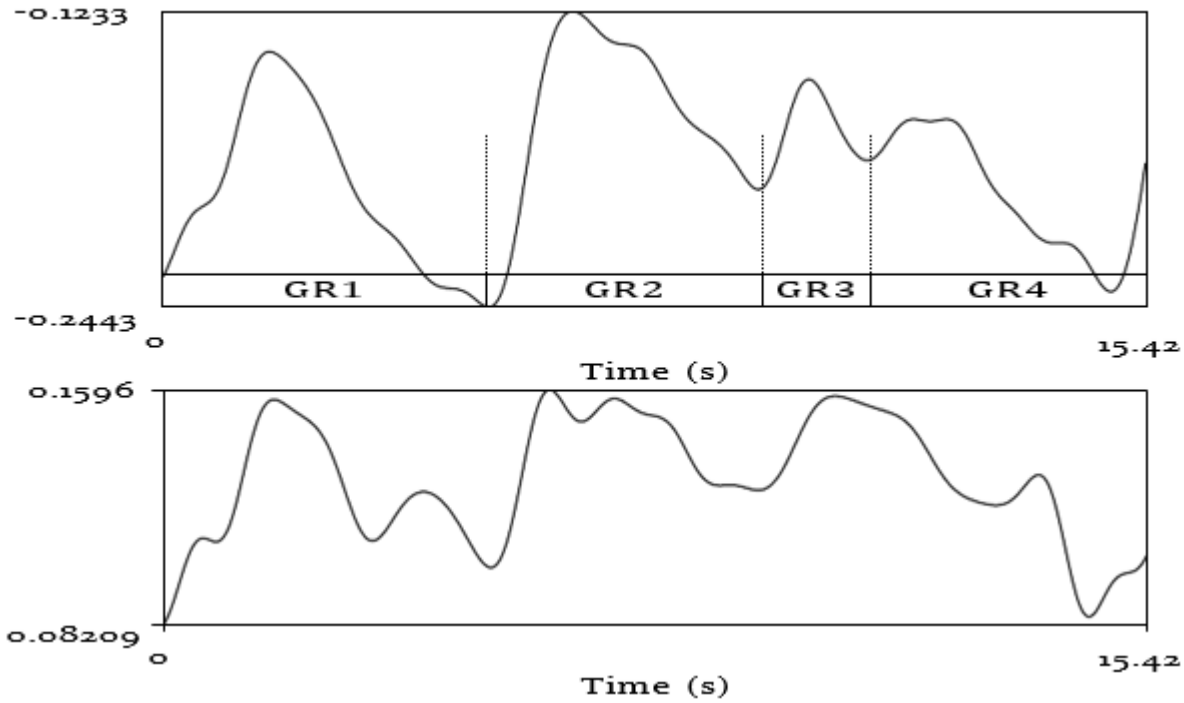

Figura 4.28 – Sinais respiratórios do tórax (acima) e do abdômen (abaixo) de locutora alemã lendo um trecho de texto em inglês de forma persuasiva.

O programa de software que o acompanha registra a mudança de área da cavidade a partir da mudança de corrente elétrica gerada pela mudança de extensão do indutor em forma de mola conectado à parte

interior das cintas. Observe na mesma figura quatro grupos respiratórios delimitados pelo movimento combinado de aumento e diminuição da área da seção transversal do tórax: a inalação é a porção em que os valores de área aumentam e a expiração, a porção em que os valores de área diminuem. A duração da fase de inalação varia com o estilo de elocução, sendo menor na fala persuasiva, pela necessidade de tomar mais ar em menos tempo para garantir fluxo expiratório para as ênfases próprias à persuasão.

O mesmo dispositivo foi usado num estudo sobre coordenação fala-respiração em três estilos de elocução no PB (BARBOSA; MADUREIRA, 2018). Quatro locutores, dois homens e duas mulheres, leram um trecho de cerca de 700 palavras sobre a origem dos pasteis de Belém, o corpus Belém já mencionado neste livro. Logo em seguida, narraram a história com suas palavras. Ao final, teceram comentários sobre os dois personagens principais, com temperamentos opostos. Os estilos são respectivamente leitura (LE), narração (NR) e comentário (CT).

Os números da Tabela 4.7 assinalam que a duração dos grupos respiratórios durante narração e comentário duram mais do que durante a leitura nos dois sexos, sendo a média dos dois primeiros estilos superior em 1, 2 segundos nos homens e em 1, 6 segundos nas mulheres, como indicado na Figura 4.29. Uma vez que a sucessão dos grupos respira- tórios corresponde à taxa de inalação ou tomada de ar, essa taxa seria menor nos estilos de narração e de comentário pelo fato de o locutor precisar planejar o que se vai dizer a intervalos mais afastados do que na leitura, estilo em que o que se vai dizer está à frente dos olhos de quem lê. Por outro lado, como a maior parte do ciclo respiratório para a fala é formado pela parte expiratória, essa é em média maior na narração e no comentário, como se espera intuitivamente.

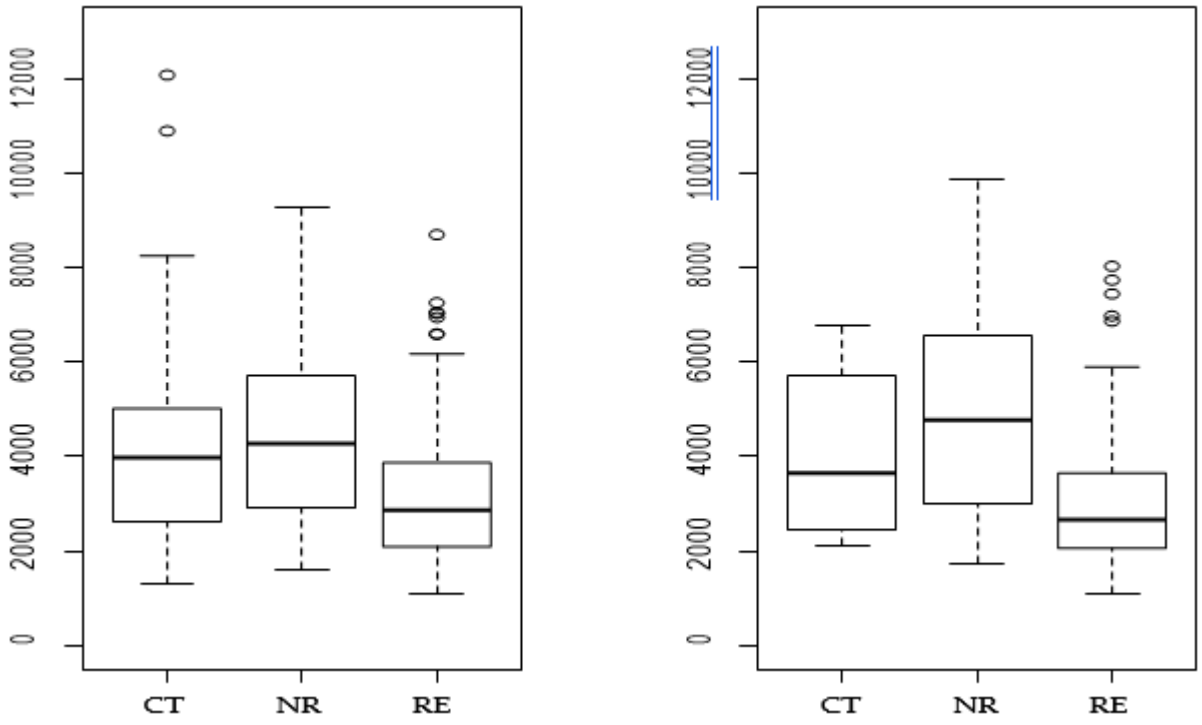

Figura 4.29 – Diagramas de bloco da duração do grupo respiratório para os estilos leitura (RE), narração (NR) e comentário (CT) por sexo, sendo os três blocos à esquerda dos homens e os da direita, das mulheres.

As Figuras 4.30 e 4.31 ilustram, respectivamente, o sinal de fala alinhado com os ciclos torácicos em trecho de leitura e de narração de um locutor masculino. Para além de indicarem os trechos de inalação e de expiração obtidos a partir do sinal do tórax, revelam ainda que há pausas silenciosas internas à fase expiratória nos dois estilos que não requerem inalação prévia. Sendo assim, nem toda pausa demarca um grupo respiratório, complementando o conhecimento adquirido na seção 4.5, em que medimos pausas silenciosas e preenchidas sem nos referir ao ciclo respiratório. Por outro lado, se é verdade, como vimos na seção 4.9, que ao menos parte considerável dos Ruídos Respiratórios Audíveis ocorre durante a fase de inalação, não podemos verificar, sem um dispositivo como o Resp Track, os momentos em que se dão inalações ou expirações inaudíveis.

Tabela 4.7 – Médias e desvios-padrão (entre parênteses) em milissegundos para a duração do grupo respiratório para quatro locutores do PB agrupados por sexo. A desigualdade ou igualdade ao final de cada bloco na coluna estilo indica se a diferença entre as médias é ou não é significativa e em qual direção.

sexo	estilo	média (desv-pad)
homens	LE	3169 (1415)
	NR	4478 (2044)
	CT	4266 (2341)
	LE< (NR=CT)	
mulheres	LE	2976 (1306)
	NR	5015 (2196)
	CT	4119 (1786)
	LE< (NR= CT)	

O estudo da duração dos ciclos respiratórios completa as medidas de duração das unidades da fala, pois avaliamos desde a unidade do tamanho da sílaba até o grupo respiratório. A Figura 4.32 resume alguns aspectos vistos aqui pela comparação da extensão das unidades, de cima para baixo nas camadas de anotação: a segmentação em unidades VV na primeira camada, a segmentação de pausas na segunda, a segmentação das fases de inalação e expiração na terceira, a partir do sinal do tórax na segunda posição no painel acima, a segmentação dos grupos respiratórios na quarta camada e, por fim, a camada final mostrando os grupos acentuais obtidos automaticamente a partir dos picos de duração normalizada dos intervalos da primeira camada (com o script *SGDetector*).

4.11 Prelúdio para o próximo capítulo

As medidas de duração revelam especialmente a organização rítmica da fala, em diversos domínios, da sílaba ao grupo respiratório. Uma compreensão da prosódia da fala não prescinde, no entanto, da

medida de seus aspectos estritamente melódicos e de qualidade de voz, que passamos a ver no capítulo seguinte.

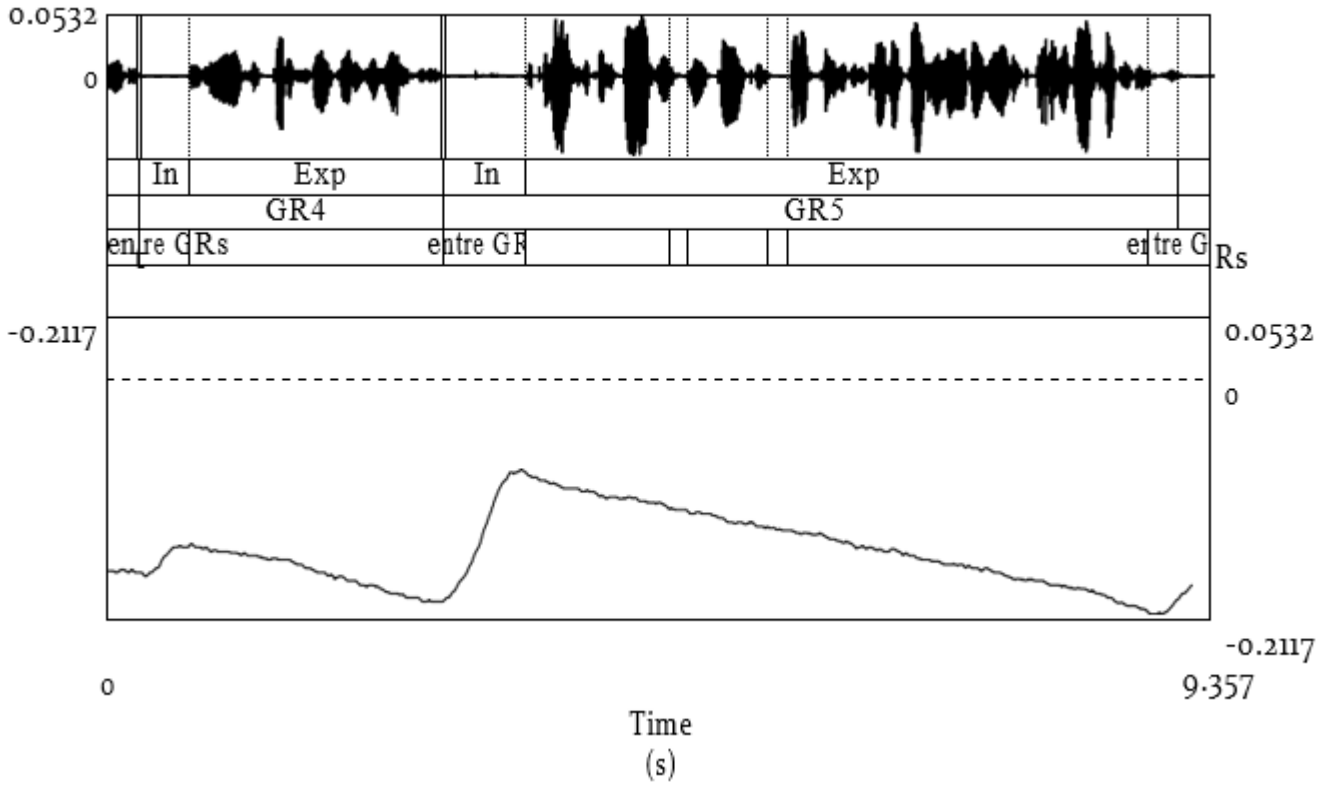

Figura 4.30 – Forma de onda, camadas de anotação e sinal de expansão do tórax de locutor masculino no estilo leitura. O trecho lido é: “os dias pareciam todos iguais. [entre GRs] O que mais custava no entanto era ter de se levantar no meio da noite para rezar as matinas.”. Na anotação, In é fase de inalação, Exp é a fase de expiração , GR4 e GR5 dois grupos respiratórios consecutivos de sua leitura e o trecho etiquetado como “entre GRs”, a pausa silenciosa entre o fim da expiração anterior e o final da inalação do grupo respiratório em questão.

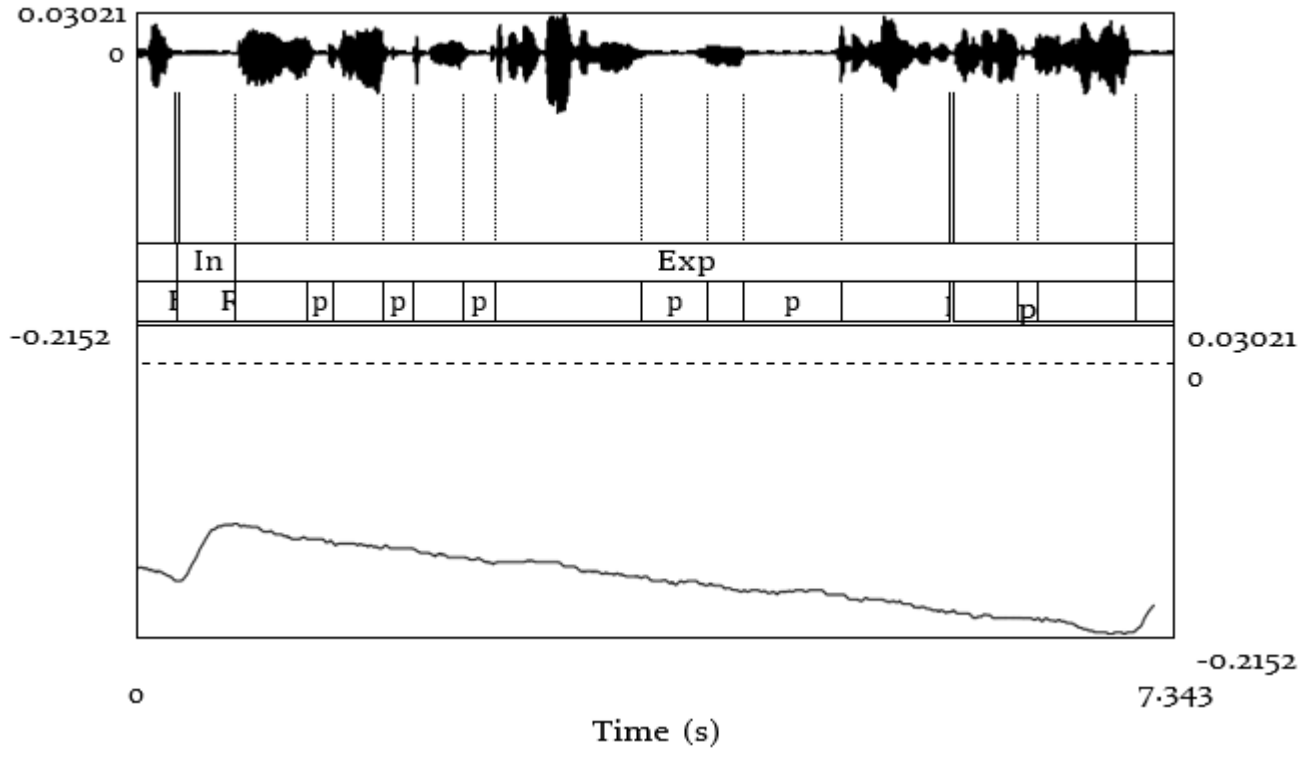

Figura 4.31 – Forma de onda, camadas de anotação e sinal de expansão do tórax de locutor masculino no estilo narração. O trecho lido é: “vida é... que ela... que... era levada pelos monges... o nome dele é Manuel.”. Na anotação, In é fase de inalação, Exp é a fase de expiração e o trecho etiquetado como “EGR”, a pausa silenciosa entre o fim da expiração anterior e o final da inalação do grupo respiratório em questão. Observe as várias pausas silenciosas (p) durante a expiração.

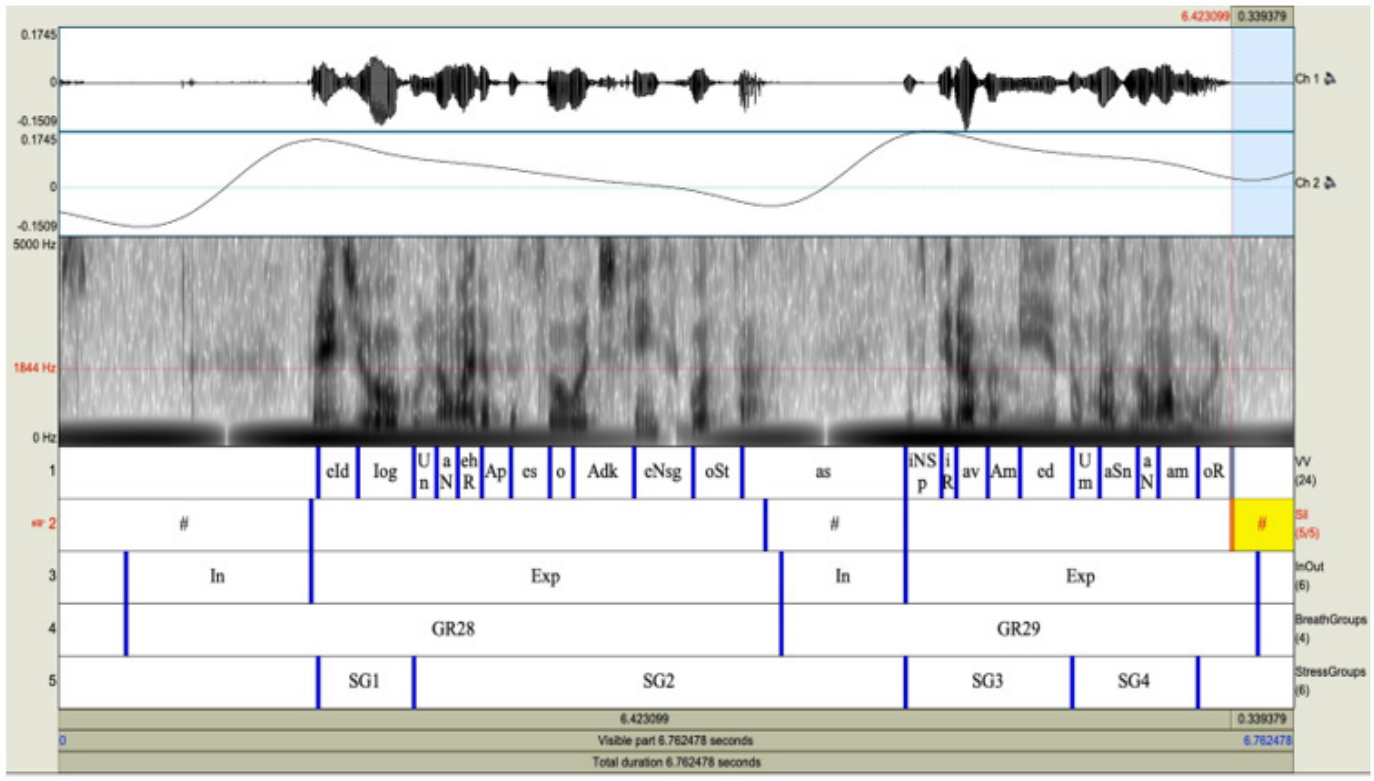

Figura 4.32 – Forma de onda, sinal de expansão do tórax de locutor feminino no estilo leitura e camadas de anotação. O trecho lido é: “Frei Diogo não era pessoa de quem se gostasse. Inspirava medo, mas não amor.” Vide texto para indicação das camadas.