Plínio A. Barbosa

Manual de Prosódia Experimental

EDITORA DA **ABRALIN**

Palavras dos Editores

Esta publicação, digital e gratuita, compõe o catálogo de livros digitais da Editora da ABRALIN, uma editora *open access*, criada em 2020, que busca oferecer mecanismos efetivos de publicação e circulação de obras de Linguística no país. A ideia que norteia seu funcionamento encontra melhor expressão nas palavras de seu idealizador, Prof. Dr. Miguel Oliveira Jr., então presidente da ABRALIN: "acreditamos que dar acesso livre à produção intelectual de excelência, que é fruto – na maioria das vezes – de investimento público, é o caminho mais democrático no contexto socioeconômico em que vivemos". Sem dúvida, essas palavras foram definitivas para o nosso engajamento na criação da Editora da ABRALIN. Queremos contribuir para fazer da Editora da ABRALIN um canal permanente de apoio à divulgação da sólida pesquisa feita nas muitas áreas da Linguística no Brasil.

Como todos sabemos, a ABRALIN desempenha papel fundamental na consolidação dos estudos linguísticos no Brasil, contribuindo de maneira crucial para a criação e a preservação de espaços de acolhimento da diversidade de ideias linguísticas, algo que tem urgência ética e é – no nosso entendimento – atitude necessária para manter o indispensável diálogo entre a sociedade e a comunidade científica. A Editora da ABRALIN nasce dentro desse contexto e com esse desígnio maior.

A excelência do trabalho da Editora e das obras por ela publicadas será garantida – disso temos certeza – pela esperada contribuição dos associados da ABRALIN. Tal contribuição constantemente vem em atendimento aos editais e aos critérios tornados públicos periodicamente, na forma de propostas de publicação, na colaboração junto ao Conselho Editorial e com as demais atividades envolvidas no funcionamento da Editora.

Nossa expectativa é que a Editora da ABRALIN possa fornecer obras de qualidade, acessíveis gratuitamente ao público-leitor interessado, fomentando, assim, a pesquisa em Linguística, contribuindo com o diálogo constante entre pesquisadores e sociedade.

Valdir do Nascimento Flores
Gabriel de Ávila Othero
Editores

Manual de Prosódia Experimental

Plínio A. Barbosa

2022

EDITORA DA **ABRALIN**

SUMÁRIO

A minha amada Rosinha, companheira doce de todas as horas.

O argumento mais forte não prova nada, se a conclusões não são verificadas pela experiência.

Roger Bacon

APRESENTAÇÃO

Um livro múltiplo, que introduz conceitos complexos de uma maneira simples, precisa, abrangente e atrativa. Não se limita a explicações e fornecimento de referências que respaldem a compreensão de fundamentos teóricos e a implementação de procedimentos metodológicos de análise, mas demonstra, passo a passo como fazer um estudo experimental com foco na prosódia da fala. Nesse sentido, constitui uma sequência primorosa sobre a exploração da temática da prosódia, introduzida em Barbosa (2019).

O carater múltiplo do livro advém de vários fatores. É múltiplo o público que pode se instruir com sua leitura, desde os que queiram adentrar na temática da prosódia experimental até os pesquisadores, professores, especialistas, peritos e estudantes da área de estudos da linguagem e afins.

É múltiplo o olhar do autor sobre os tópicos explorados: explicita os conceitos; fundamenta-os teoricamente e experimentalmente; demonstra como analisá-los por meio de técnicas instrumentais de análise acústica e pletismográfica, bem como por meio de testes perceptivos; explica como efetuar medidas acústicas relevantes ao estudo dos diversos elementos prosódicos e como considerá-las estatisticamente; fornece explanações sobre como interpretar os resultados de um experimento sobre a prosódia da fala; indica ferramentas e scripts de análise; e propõe exercícios para estimular a prática, proporcionando assim atividades que contribuem para a apropriação de conceitos e de procedimentos que levem à dirimição de dúvidas.

Também múltipla é a literatura de base referenciada, na qual o autor se faz presente com obras influentes, quer sob o prisma do modelamento da prosódia, quer sob o prisma descritivo.

O livro compreende 7 capítulos, constituindo-se o sétimo de exercícios para subsidiar a prática e o primeiro em uma introdução que

considera o fazer do trabalho experimental, explorando a relação ente teoria, procedimentos metodológicos, análise e interpretação de seus resultados à luz dos princípios teóricos de base e das propriedades que devem ser observadas na investigação de um problema de pesquisa e na testagem das hipóteses da pesquisa: a falsificabilidade, a preditividade e a explicitabilidade.

O segundo capítulo se debruça sobre a interação entre prosódia e segmentos, considerando modelos diversos de entoação e de ritmo da fala de inspiração fonológica e fonética.

O conteúdo do capítulo 3 é extremamente relevante, absolutamente precioso para aqueles que desejam realizar experimentos de prosódia, por incluir informações não encontradas em publicações da área. Aborda os procedimentos de metodologia experimental aplicados ao desenvolvimento de pesquisa em prosódia. Abrange a formulação e verificação das hipóteses de pesquisa e protocolos de investigação experimental. O capítulo 4 aborda questões rítmicas e explicita como fazer medições de duração de unidades linguísticas de extensão variada: a sílaba; a unidade VV; as pausas silenciosas; as pausas preenchidas; as pausas respiratórias; os grupos acentuais; os eventos de natureza dialógica; os eventos não linguísticos; e os grupos respiratórios. Considera os procedimentos metodológicos de medição, normalização e suavização de medidas, complementando a exposição textual com ilustrações, tabelas, fórmulas e referências a trabalhos que as utilizam.

O capítulo 5 aborda questões melódicas e de qualidade de voz, compreendendo exposição sobre sistemas de notação entoacional, procedimentos de medição acústica para análise de padrões entoacionais e de descritores acústicos de qualidade de voz.

O capítulo 6 é dedicado à exploração de procedimentos de estatística inferencial que permitem investigar a reprodutibilidade da amostragem dos dados da pesquisa experimental, ou seja, permitem, considerar os dados analisados pela estatística descritiva em relação a população de dados concernentes e, dessa maneira, contribuir para

subsidiar os achados de pesquisa.

Todos os capítulos contêm uma apresentação inicial do conteúdo a ser explorado e as exposições muito bem ilustradas com gráficos, tabelas, quadros, fórmulas e exemplos e acompanhadas de discussões de achados relatados em trabalhos pertinentes aos tópicos abordados.

Apesar de o livro conter 7 capítulos, a obra não se limita às fronteiras do livro, extrapola-as ao se fazer acompanhar por um conjunto de materiais suplementares de livre acesso, organizados em torno de três pastas nomeadas “Estatística”, “Exercícios” e “Metadados”.

O Manual de Prosódia Experimental, de autoria de Plinio Barbosa, vem, não apenas preencher uma lacuna no âmbito dos estudos da linguagem, mas também impactar positivamente a área, provocando desdobramentos pelo potencial que tem de impulsionar, com subsídios qualificados, os estudos sobre a Prosódia Experimental para todos os interessados que têm acesso à escrita em língua portuguesa.

São Paulo, 10 de maio de 2022
Sandra Madureira
Professora titular da PUC-SP.

PREFÁCIO

Este livro é fruto de mais de 25 anos de ensino e pesquisa experimental na área de prosódia, tendo por principal motivação de sua redação uma preocupação crescente em melhorar o aprendizado de meus alunos e das pessoas que me buscaram ao longo desses anos. Meu desejo é que seja um manual útil para quem quer fazer pesquisa em prosódia.

Esse desejo de ser mais didático sempre existe no professor mas, em pelo menos três ocasiões com que me deparei com alunos do Ensino Médio e Fundamental para falar de Fonética e Prosódia, marcou-me mais profundamente a necessidade de ser claro e compreensível, mesmo para leigos no assunto.

Neste manual, meu intento foi o de apresentar procedimentos de experimentação e teorias mais gerais de produção e percepção da prosódia para que o leitor forme uma compreensão que norteie sua interação com teorias mais específicas e possa adequadamente formular hipóteses científicas na área. O livro "Prosódia", que escrevi para a Graduação e Pós-Graduação a convite de meu amigo e colega Tommaso Raso, pode ser uma leitura que auxilie nesse crescimento na área, embora sua leitura não seja imprescindível para acompanhar este manual.

Todo o material suplementar do livro se encontra disponível no repositório do GitHub, neste endereço: https://github.com/pabarbosa/prosodia-experimental. Os scripts *SGDetector* e *ProsodyDescriptor* permitiram a geração dos dados acústicos usados como exemplo neste livro e estão disponíveis neste endereço: https://github.com/pabarbosa/prosody-scripts. Esse aspecto do livro faz parte de um modo de pensar condensado na expressão *Open Science*, Ciência Aberta, que deseja que todas as etapas do fazer

ciência estejam disponíveis para quem quiser refazer, verificar a correção do que foi feito, aprender com material disponível livre e gratuitamente.

Há muitas instituições e pessoas a agradecer, meu trabalho seria impossível sem umas e outras. Quero particularmente deixar meu débito a minha instituição de trabalho, a Universidade de Campinas, e às agências CNPq, CAPES e FAPESP, bem como a fontes externas quando de colaborações internacionais especialmente na Dinamarca e Suécia. Há muitos colegas que me ensinaram sobre prosódia e fazer experimentação a partir de nossa interação e aqui cito aqueles com quem tenho colaborações mais frequentes: Sandra Madureira, Oliver Niebuhr, Philippe Boula de Mareüil e Anders Eriksson. A Sandra agradeço ainda os cursos ministrados juntos, aqui na Unicamp e em congressos, bem como o curso ministrado em 2020 com o Leônidas Silva Jr que, juntamente com um futuro colega, Philipp Meer, me despertaram o interesse pela Fonética de segunda língua e língua estrangeira. Meus alunos me ensinam muito, pois são suas questões e dúvidas que nos fazem torna mais didáticos e a eles agradeço muito, não só os que orientei na Iniciação Científica, no Mestrado e no Doutorado, mas também todos os alunos da Graduação e Pós-Graduação da Unicamp e de outras instituições, especialmente aqueles que participaram da disciplina epônima e que revisaram partes deste livro. Foram eles Gustavo Silveira, Aline Benevides, Valdete da Macena, Carla Minello, Rafael Marques e Beatriz Freire. Ao Gustavo devo ainda me ter instruído e incentivado no uso do GitHub, ao qual sou especialmente grato. Há amigos queridos também fora do país e aos quais deixo minha gratidão pelo apoio constante, especialmente Oliver e Philippe.

Minha esposa é sempre a primeira incentivadora de meu trabalho, também tornado possível por sua dedicação a nossas tarefas quotidianas em casa, especialmente nos tempos de isolamento social em que a maior parte deste livro foi escrita. Agradeço ainda a meus queridos pais Cynthia e Lidio.

Toda obra é fruto do sustento de um outro. Eu seria um grande usurpador se não referir a confecção deste livro Àquele que me mantém em vida e me dá Suas luzes: meu querido Deus, para O qual se deve toda Glória. Sem os pulsos de Teu Coração, nada seria possível, *innefabile dolcezza*.

Material Suplementar

Este livro é acompanhado de um conjunto de arquivos suplementa- res de livre acesso, disponíveis no endereço https://github.com/ pabarbosa/prosodia-experimental, organizados em pastas.

A pasta **Estatistica** contém os roteiros de testes estatísticos e os dados utilizados no capítulo 6. Esse material está organizado segundo as seções do capítulo com arquivos TXT com dados e roteiros com as funções utilizadas no R para fazer os testes. Sendo assim, pressupõe-se que o leitor tenha o software R instalado para tirar o máximo proveito. Esse software está disponível gratuitamente em htttps://www.r-project.org. Os arquivos com os roteiros para os testes em cada pasta começam com a palavra "Roteiro" no nome do arquivo TXT. Os dados têm seus nomes referidos nas suas respectivas seções do capítulo.

A pasta **Exercicios** contém o material necessário para fazer os exercícios propostos no capítulo 7 organizados por pastas que se referem ao número do exercício mencionado no capítulo.

A pasta **Metadados** contém os áudios, arquivos de anotação TextGrid para o software Praat e eventuais textos referidos nos capítulos do livro. Esse material está organizados em pastas que se referem aos capítulos 3 a 6, onde esse tipo de material é referido. Algumas pastas contêm áudios e outros metadados do corpus referido no respectivo capítulo, para que o leitor consulte e faça um melhor juízo daquilo que é ilustrado no respectivo capítulo.

O software Praat, necessário para rodar os scripts referidos no livro e disponibilizados no GitHub, se encontra disponível gratuitamente em http://www.praat.org.

Capítulo 1

Introdução

Ao aliarmos no título deste livro os termos “prosódia” e “experimental”, está pressuposto que é possível utilizar o método experimental para realizar estudos prosódicos. Antes de situarmos os estudos de prosódia dentro do contexto mais geral dos estudos de fonética, vamos recordar alguns fundamentos da experimentação.

1.1 Ciclo Experimental

Toda investigação experimental pressupõe um percurso que se inicia a partir de observações guiadas por uma teoria científica concebida por um pesquisador inserido numa comunidade científica. Ao confrontar as observações com as hipóteses provindas da teoria científica elegida, inicia-se o ciclo experimental que vai interrogar hipóteses e modelo teórico.

Conforme esquematiza o diagrama da Figura 1.1, uma teoria científica pressupõe uma tríade (observação, modelo, hipóteses), sendo composta de um modelo associado a um conjunto de enunciados construídos a partir de hipóteses que são testadas no quadro desse modelo. Novas hipóteses podem ser enunciadas sempre que as atuais são refutadas ou refinadas a partir dos resultados de um ciclo experimental que começa a partir de observações novas; em nosso caso, observações de natureza prosódica.

Não há observação ingênua, pois toda observação é guiada pelo modo de proceder de um estilo de pensamento (FLECK, 1992) ou estilo de raciocínio (HACKING, 1992) e precisa ser guiada pelo corpo de pesquisadores de uma comunidade científica. Por exemplo,

um contorno de frequência fundamental (Fo) tem partes que não são consideradas picos ou vales com função linguística, quando são associados a efeito micromelódico, isto é, como reflexo fisiológico[1] (BARBOSA, 2019, p. 67-68). A própria noção de contorno é um construto teórico, uma vez que os valores de Fo são calculados a partir do valor de um ciclo glotal e, portanto, não formam um contínuo ao longo do tempo. Outro exemplo é a leitura adequada e, portanto, a mera possibilidade de observar o que informa um espectrograma. Antes de qualquer instrução, qualquer mancha escura ou cinza seria passível de "mostrar" algo, mas todo aquele que passou por essa instrução sabe que somente determinados contrastes de cinza ou padrões ao longo do tempo são relevantes.

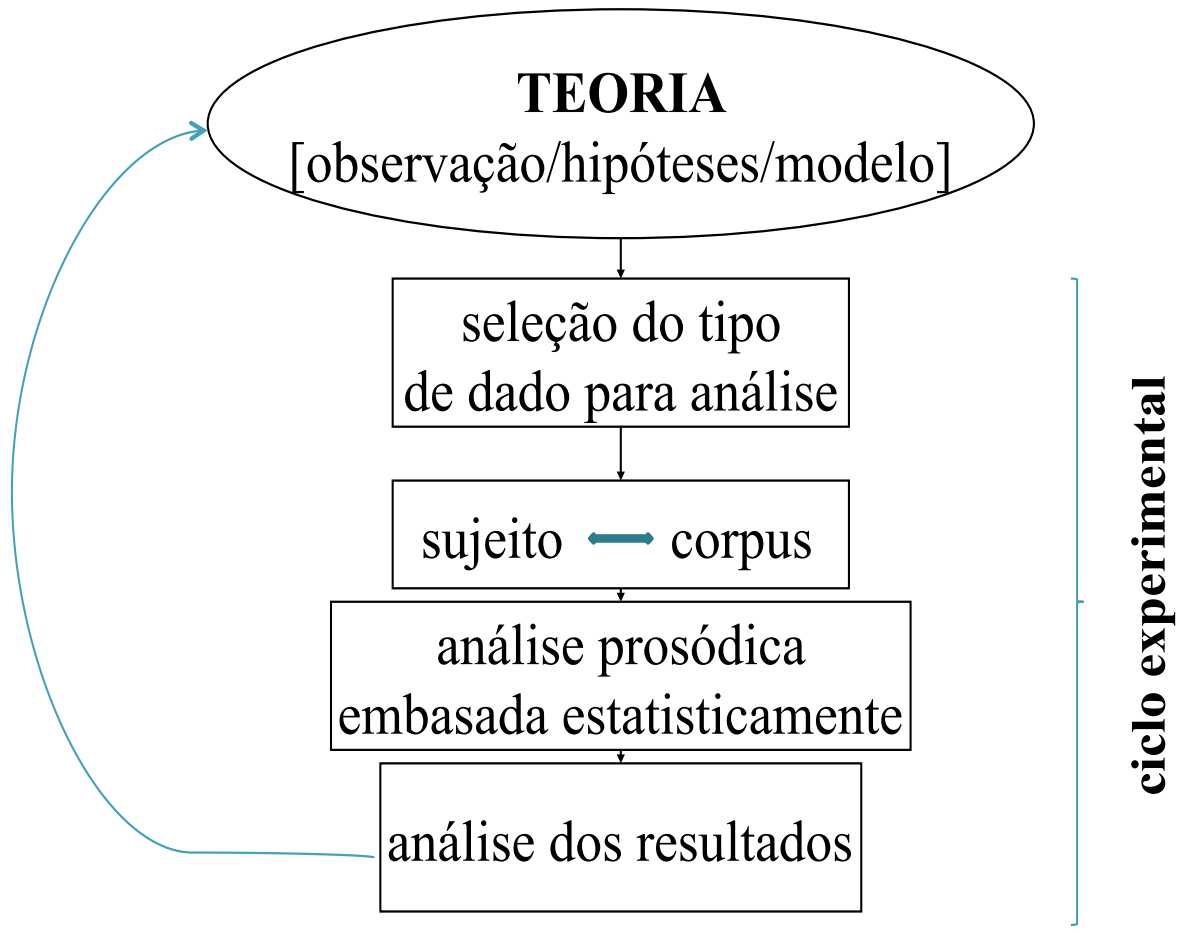

Figura 1.1 – Esquema do ciclo experimental aplicado a estudos prosódicos.

1 Vide seção 3.2.3 a esse respeito.

Essa observação instruída possibilita a investigação de algo novo do ponto de vista prosódico e a testagem de hipóteses fundamentadas em alguma teoria de produção ou percepção da prosódia que deve satisfazer as propriedades que seguem (XU, 2010).

1. *Falsificabilidade*. As hipóteses devem conter mecanismos para testar sua veracidade ou falsidade.
2. *Preditividade*. Toda teoria precisa prever observações em condições experimentais diversas a partir de um modelo.
3. *Explicitabilidade*. A forma de predição das observações deve ser explícita, reprodutível a partir de um modelo sob a forma de regras ou equações.

Uma teoria é tanto melhor quanto maior seu caráter explicativo, que toca a propriedade da preditividade. No caso da prosódia, a implicação é a possibilidade de prever as maneiras como ritmo e entoação da fala são realizados em situações comunicativas diversas, o que está atrelado às hipóteses que são feitas.

As hipóteses e o modelo determinam não apenas os tipos de corpora que deverão ser obtidos (e.g., enunciados de fala, de canto ou resultados de um teste de percepção), como também o tipo de análise estatística inferencial que deverá ser conduzida para examinar a significância de diferenças entre variáveis dependentes, como veremos no capítulo 6.

Segundo uma linha filosófica derivada do princípio de incerteza de Heisenberg[2] , o corpus é afetado pela intervenção do experimentador, pelo comportamento do sujeito que está sendo avaliado, bem

2 Este princípio da mecânica quântica enuncia que duas propriedades de uma mesma partícula não podem ser conhecidas com a mesma precisão pois são complementares, isto é, ao se esmerar em tirar a incerteza de uma medida, a incerteza sobre a outra medida correlata aumenta.

como pela situação e ambiente de obtenção dos dados de produção ou de percepção (MILROY, 1987, cap. 2).

No que diz respeito aos corpora de fala, conforme propomos anteriormente (BARBOSA, 2012) entende-se sua maior ou menor naturalidade a partir de dois eixos: gênero do material gravado e grau de intervenção do experimentador, como se vê na Figura 1.2.

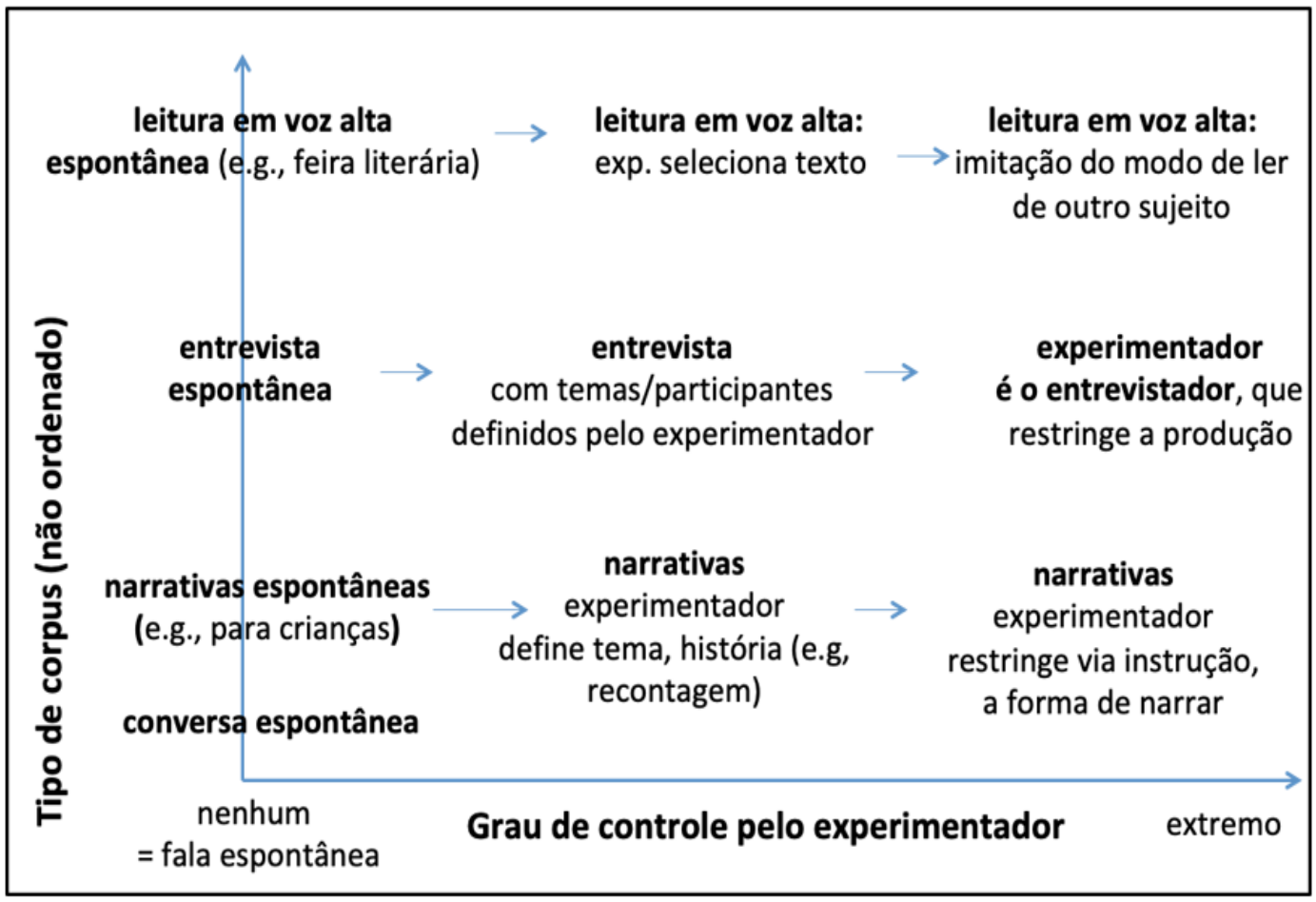

Figura 1.2 – Espaço de naturalidade dos corpora de fala.

Habitualmente chama-se de fala espontânea tudo que se refere à conversa espontânea. No entanto, não pode se restringir a espontaneidade a essa instância comunicativa se entendemos espontaneidade como evento comunicativo natural. Pois bem, todos os eventos que estão no grau zero de intervenção do experimentador são, a nosso ver, fala espontânea, pois são todos situações comunicativas de nossa cultura: o que muda é apenas o gênero. No momento em que o

experimentador faz alguma intervenção, por menor que seja, passa a ser uma fala que se configura como fala de laboratório, que pode chegar até um grau máximo de intervenção, como no caso das imitações da fala no extremo direito da figura. Entre esses exemplos de intervenção máxima e a fala espontânea se encontram os casos do que se chama fala semiespontânea, em que a fala não é completamente dirigida, mas se fornecem elementos para a obtenção do corpus, como nos casos de *map task* e *silent video task* que abordaremos no capítulo 3.

A consideração da relação entre sujeito e corpus é essencial para minimizar o efeito do protocolo experimental sobre o comportamento do indivíduo. Como veremos no capítulo 3, cada indivíduo pode se comportar de forma não esperada por conta de alguma falha do desenho experimental ou alguma não adequação do desenho a um indivíduo em particular. Vamos dar dois exemplos. Algumas pessoas são mais afetadas em sua fala durante a digestão do que outras. Assim, toda gravação após a ingestão de quantidade razoável de alimento pode resultar numa fala mal articulada ou numa atenção menor num teste de percepção em alguns indivíduos. Por isso, obter dados nesse momento do dia não é aconselhável. Como segundo exemplo consideremos o controle da taxa de elocução. Um metrônomo luminoso poderia ser, em princípio, uma ideia relevante para assegurar a mesma cadência de fala entre os diferentes sujeitos. No entanto, essa técnica só funciona bem em pessoas com experiência musical. Pode-se no entanto utilizar um trecho da mesma leitura de uma pessoa como modelo para se efetuar esse controle, como usamos há alguns anos (BARBOSA, 1994).

Após a obtenção dos dados de fala, é necessário anotá-los para se obter um corpus de fala. Essa anotação também depende de pressupostos teóricos, uma vez que podemos anotar desde sílabas a enunciados e parágrafos. Embora pareçam tarefas simples à primeira vista, não são. A noção de sílaba requer que se leve em conta duas possibilidades na cadeira da fala: a sílaba fonética ou a sílaba fonológica realizada. A de enunciado, por outro lado, não é simples nos casos de narrativas e

entrevistas, por exemplo e requer considerações de natureza pragmática, como reconhecer no enunciado um ato de fala completo (CRESTI, 2000). Também unidades intermediárias a essas duas requerem cuidado, como o caso do grupo acentual, que só se pode definir se se reconhece uma proeminência local, ponto de culminância de um mecanismo acentual. Esse reconhecimento pode requerer um conjunto de juízes ou um procedimento automatizado que espelhe a percepção humana.

Completada a tarefa de anotação pode-se passar à análise prosódica que pode se dar ao nível duracional, melódico ou intensivo, combinando ou não os três níveis de análise. Essa análise pode ser paradigmática ou sintagmática, a depender do que se deseja mostrar, como apontamos na seção seguinte.

A análise prosódica deve ser acompanhada do teste estatístico inferencial apropriado às hipóteses a serem testadas para se avaliar a reprodutibilidade dos achados. Somente dessa forma será possível avaliar as hipóteses levantadas antes do ciclo experimental desenhado para isso. Conhecer o teste estatístico adequado é fundamental para a boa condução do experimento, o que requer aprendizagem específica, agora disponível em alguns centros de excelência nas universidades brasileiras.

1.2 A Prosódia na Experimentação

Os estudos prosódicos formam parte da disciplina da Fonética que descreve e infere as características de nosso modo de falar. Enquanto a fonética segmental se ocupa do conteúdo, pois investiga as características de vogais e consoantes, a fonética prosódica investiga o ritmo e a entoação da fala, isto é, como algo foi dito e não o que foi dito.

Assim, dentro do campo de estudos prosódicos cabem aqueles dos

estilos de elocução, isto é, o estudo do que cada pessoa muda na fala ao adaptar sua forma de falar ao conversar com um amigo ou desconhecido; ao ler ou ao narrar; ao dar uma aula ou ao dar uma palestra; ao ler uma história para uma criança ou ao ler para um adulto; ao fazer um discurso em gêneros diversos como político, religioso, de formatura, persuasivo e tantos outros estilos do falar.

Em cada um desses estilos cabe a investigação de como funções tais como a marcação de unidades menores na fala (segmentação) e o modo de chamar a atenção para um trecho de fala em relação ao contexto (proeminência) são realizadas a partir do controle da articulação com imediatas consequências acústicas (línguas orais) ou gestuais (línguas de sinais).

Essas funções básicas se superpõem a elementos de nossa expressividade e afeto que são passíveis de estudo experimental, como as atitudes proposicionais, a confiança em e a dúvida do que se diz, as atitudes sociais como a hostilidade e a gentileza e ainda as diferentes emoções tais como tristeza, alegria, raiva ou medo. Esses elementos expressivos de nossa fala afetam o modo como falamos e devem ser de alguma forma controlados tanto para serem estudados por si quanto para não variarem quando se deseja avaliar um estilo ou uma função prosódica. Afinal, uma palestra ministrada com tristeza não pode ser diretamente com- parada a uma leitura com alegria, pois há dois elementos de natureza diversa variando simultaneamente.

É esse tipo de questão que devemos ter em mente para garantir a obtenção de um bom desenho experimental, como veremos no capítulo 3. Em experimentação essa condição em que só se modifica um dos elementos de um contraste é chamada de condição *ceteris paribus*. Por exemplo, se se quer estudar quais são as modificações prosódicas quando da realização do foco contrastivo em uma palavra (e.g., “Eu vi uma moto VERDE.” e não vermelha), embora o que acontece no eixo sintagmático seja primordial para a veiculação da função, a compara- ção paradigmática com uma frase dita de forma neutra

(e.g., a mera asserção “Eu vi uma moto verde.”) é importante para se entender os ajustes prosódicos que foram feitos para se realizar o foco contrastivo: aceleração da fala e tom baixo em “moto”, lentificação da fala e subida seguida da descida de contorno melódico em “verde”.

Uma das condições fundamentais da experimentação tem a ver com o chamado paradoxo do observador emprestado da Física Quântica, que diz respeito ao não uso de especialistas em linguagem em experimentos que envolvam a fala ou a própria linguagem. No que tange a fenômenos linguísticos ou paralinguísticos[3] específicos, não se pode pedir a linguistas e pesquisadores da fala para produzi-los nem para percebê-los. Também não cabe ao especialista qualquer avaliação chamada de “intuição do linguista”. São práticas que devem ser completamente banidas da experimentação, pois todo fenômeno de linguagem ou fala deve ocorrer em situações habituais de comunicação, isto é, percebidas e realizadas pelo sujeito comum. De que vale para a comunicação um fenômeno aparentemente percebido por um linguista? Por exemplo, no chamado deslocamento acentual, em que um acento lexical é produzido em expressões congeladas como “JEsus CRISto”, não se pode perguntar a linguistas se eles escutam um deslocamento em expressões não congeladas como “café quente”. Há que se realizar experimento controlado para inferir características de acento lexical na primeira sílaba de “café” e criar um protocolo de percepção para avaliar se leigos percebem algum indício de acento inicial. Há alguns protocolos disponíveis para isso.

Mesmo para coisas aparentemente simples como indicar se houve uma fronteira não terminal (um tipo de pausa subjetiva) num trecho de fala, não há concordância total nem mesmo entre especialistas. Por isso, o exame da percepção de qualquer pausa, terminal ou não termi-

3 São fenômenos que concernem a comunicação, mas não são explicitamente aspectos linguísticos. Um exemplo é uma atitude proposicional como a confiança ou a dúvida quanto à veracidade de uma asserção. Uma modificação global da prosódiase dá nesse contraste, como mostramos em Barbosa (2019).

nal, por um número razoável de ouvintes leigos tem sido proposto na literatura (BARBOSA, 2010; COLE; MO; BAEK, 2010), com o apoio da análise estatística para determinar a probabilidade da presença da fronteira.

Outro aspecto importante na montagem de corpora de fala ou em testes de percepção é a homogeneidade dos sujeitos envolvidos. Se determinamos características prosódicas de uma língua como o português brasileiro, deve-se ter em mente que há variação prosódica entre os diversos dialetos de nosso território. Assim, se não é possível ter um bom número de sujeitos de cada dialeto, pode-se ao menos descrever um dos dialetos, uma região mantendo a homogeneidade de características sociolinguísticas. Não se pode descrever aspectos sonoros do português brasileiro tendo um único representante de um dialeto e vários de outro, por exemplo. Nem mesmo vários sujeitos jovens e um único de faixa etária maior, porque "só se encontrou" aquele sujeito. Deve-se ter um controle adequado do grupo de sujeitos que se descreve para que se tenha uma descrição apropriada das características prosódicas que se investiga.

Todos esses aspectos serão examinados com vagar no capítulo 3. Ficam esses comentários gerais para a reflexão do leitor, assim estará mais amadurecido para compreender melhor as questões de desenho experimental. Antes, porém, convém apresentar as principais teorias de produção e percepção da prosódia e seu uso subsidiário em outras teorias linguísticas.

Capítulo 2

Teorias e modelos prosódicos

Na próxima seção apresentamos duas principais teorias de produção da prosódia que pressupõem uma separação entre a produção segmental e aquela acima do segmento. Nas duas seções seguintes, apresentamos respectivamente dois modelos de geração de contornos melódicos e duracionais que apresentam algumas vantagens didáticas para explicar sua relação com experimentos envolvendo melodia e ritmo da fala. Quando da exemplificação com desenhos experimentais no capítulo seguinte, lançaremos mão de teorias específicas de natureza fonética ou fonológica para deixar clara a relação entre teoria, hipóteses e experimentação embasada estatisticamente. No entanto, essas teorias têm alguma relação com as teorias apresentadas no capítulo em que nos encontramos.

2.1 Quanto à Separação entre Segmentos e Prosódia

A teoria *Frame*/*Content* de MacNeilage (1998) é uma teoria do desenvolvimento da fala que se fundamenta na separação essencial entre uma máscara silábica e os segmentos que a constituem. Ela se originou da análise de erros de fala em adultos, como nas trocas de sons em "sons of toil" para "tons of soil"[4] em que apenas as consoantes trocam de lugar. Já no exemplo "odd hack" no lugar de "ad hoc", as vogais é que trocam de lugar e as consoantes permanecem. Esse exemplo, acrescido de uma série de evidências apresentadas pelo autor, revelam que as posições de segmentos de natureza distinta, como as vogais e as

4 Em português podemos citar o exemplo autêntico de "mé e pão" no lugar de "pé e mão", numa conversa sobre pedicure e manicure, que me fora dado pela colega Ana Luísa Navas.

consoantes, são fixadas para esses segmentos de forma específica: uma posição para consoante não pode ser preenchida por vogal e uma posição definida para uma vogal não pode ser preenchida por consoante. Além disso, a tonicidade também impõe uma restrição de preenchimento: segmentos tônicos tendem a trocar de lugar entre si, da mesma forma que os átonos.

Essa natureza distinta está atrelada a mecanismos distintos de produção para vogais e consoantes: afastando-se da parte superior da boca na vogal e aproximando-se da mesma parte superior na consoante, criando, assim, um padrão típico de oscilação mandibular. Esse padrão oscilatório, segundo a teoria, já está presente em ciclos de mastigação e sucção e teria sido aproveitado filogeneticamente para as primeiras produções verbais nos hominídeos superiores em que a sílaba CV teria logo assumido seu papel canônico.

A sílaba canônica CV, que no balbucio começa por uma repetição de sequências idênticas (e.g., bababa, mamama), começa a variegar durante o período das primeiras palavras, se "colorindo" de diferentes segmentos. A sílaba funciona, assim, como uma máscara (*frame*) que condiciona o preenchimento de segmentos diversos (*content*) do balbucio até o fim da aquisição do sistema fonológico.

Uma teoria semelhante quanto ao papel de sílabas e segmentos e que também encontrou evidência para seu modelo a partir da análise de erros de fala foi proposta por Shattuck-Hufnagel e Klatt (1979). A teoria de *Slots*/*Fillers* proposta inicialmente pela primeira autora (SHATTUCK-HUFNAGEL, 1979) procura explicar a natureza dos erros de fala (*lapsus linguae*) por falhas de processamento que estariam relacionadas a uma entre três possibilidades: (1) aos próprios segmentos (*fillers*), (2) a posições (*slots*) a serem preenchidas por esses segmentos ou (3) ao modo de preencher os *slots*. O modelo proposto é serial e envolve três componentes assim ordenados: (a) a seleção dos segmentos ou fonemas a partir dos itens lexicais recuperados no léxico mental; (b) a sequência ordenada de posições (*slots*) estruturalmen-

te definidas do enunciado, processada independentemente dos segmentos e (c) um mecanismo para integrar as duas partes (segmentos e posições) que incluiria: uma ferramenta para "encaixar" os segmentos nas posições especificadas na etapa anterior, uma etapa de monitoramento que checa ou apaga segmentos e uma etapa final que monitora erros eventuais.

Erros como "mé e pão" (vs. "pé e mão") podem ser explicados na etapa de preenchimento da primeira posição por um segmento que viria depois (/m/ em "pé e mão"), mas que já estava disponível na memória de trabalho[5] quando da primeira etapa de seleção de segmentos. O segmento que deveria ter sido preenchido, o /p/, fica ainda disponível e acaba preenchendo a segunda posição.

Observe que, embora não façam referência direta a uma teoria de desenvolvimento da fala, tanto a teoria de *Slots*/*Fillers* quanto a de *Frame*/*Content* pressupõem a separação da sucessão silábica com relação aos segmentos que a constituem. Para explicar os erros encontrados, conta mais a sequência de sílabas em si do que sua estrutura, pois a maioria dos erros de fala encontrados no inglês ocorrem na parte CV inicial da estrutura silábica (tanto em sílabas CV quanto CVC, por exemplo), dando evidência de que a sequência canônica de transições CV é como que a coluna vertebral para a organização do que é dito.

Por serem considerados componentes independentes, a sucessão silábica e os segmentos podem ser mudados independentemente sem que um componente afete o outro. Dois exemplos ajudam a entender isto. A mesma curva melódica assertiva pode apresentar diferentes segmentos em contraste como "Pedro canta nesta noite" vs. "Paulo corre nesta pista". Caso essas frases sejam enunciadas com o único intuito de informar algo a respeito de Pedro e de Paulo, haverá similaridade entre as suas curvas melódicas, pois estas curvas expressam a mesma função

5 Trata-se do mecanismo cognitivo para reter informações enquanto fazemos uma tarefa. Ver COWAN (1997) para detalhes.

semântico-pragmática, apesar de possuírem conteúdos segmentais diferentes. Em contraponto, se a primeira sentença for pronunciada com ênfase no pronome demonstrativo “nesta”, a curva melódica nessa frase se distinguirá da curva melódica da forma neutra para veicular o fato de que Pedro cantará naquela noite específica e não em outra. Nesse caso os segmentos não mudam, mas sim a prosódia, pois tanto a organização temporal das sílabas quanto a melodia do enunciado com ênfase em “nesta” é modificada.

Examinemos agora modelos específicos que tratam de entender o que está em jogo para se produzir uma curva melódica ou um padrão duracional da sequência silábica.

2.2 Quanto à Melodia

As duas principais classes de modelos melódicos são de natureza ou fonológica ou fonética. Fundamentada nas fonologias métrica e autos-segmental, os modelos fonológicos da entoação mais usados na literatura derivam de ou têm semelhança com a proposta seminal de Pierrehumbert (1980). Esses modelos assumiram um papel prático por terem gerado sistemas de notação melódica que examinaremos em linhas gerais aqui, antes de fazer uso amplo no capítulo 5.

2.2.1 O modelo de Pierrehumbert

Pierrehumbert propôs que a entoação fosse especificada por uma sequência linear de tons simples (H, L) ou tons combinados (e.g., H+L, L+H), que implementariam os acentos de *pitch* e tons de fronteira, que representam as fronteiras prosódicas do enunciado. Há seis possi- bilidades para marcar os acentos de *pitch* representadas pelos símbolos H*, L*, H+L*, H*+L, L+H*, L*+H, em que o asterisco (*) indica a associação do tom com a sílaba tônica da palavra. Essa repre-

sentação é estática no sentido de que o que conta são os tons em si, mesmo nos casos bitonais. Assim, em L+H*, embora a sequência seja realizada por uma subida da curva melódica, apenas conta o fato de que há um tom baixo antes e se chega a um tom alto depois, nesse caso alinhado com a sílaba tônica.

Por exemplo, numa asserção neutra, a sentença "Marianna made the marmalade." pode ser enunciada com os seguintes tons e acentos, omitindo o nível do *phrase accent* [6]: "Maria$_{H^*}$*nna made the mar$_{H^*}$malade$_{L\%}$.", em que apenas se informa quem fez a geleia. Em contraste, no enunciado com ênfase no sujeito, "Maria$_{L+H^*}$nna made the marmalade$_{L\%}$, a mudança de notação representa a curva melódica correspondente que veicula o fato de que foi realmente Marianna quem fez a geleia. A diferença do movimento melódico dos dois enunciados é mostrada na Figura 2.1 onde se vê um nível melódico alto durante "Marianna" na asserção neutra e uma subida de um tom L para o tom H que se alinha na tônica de "Marianna" na asserção com ênfase no sujeito. Quanto à fronteira, ambas as asserções terminam em tom de fronteira baixo (L%).

6 Na proposta original, o *phrase accent* é um tom que explicaria uma forma melódica entre o acento de *pitch* e o tom de fronteira. No entanto, esse tipo de componente tem sido muitas vezes contestado na literatura. Para uma discussão,ver Grice, Ladd e Arvaniti (2000).

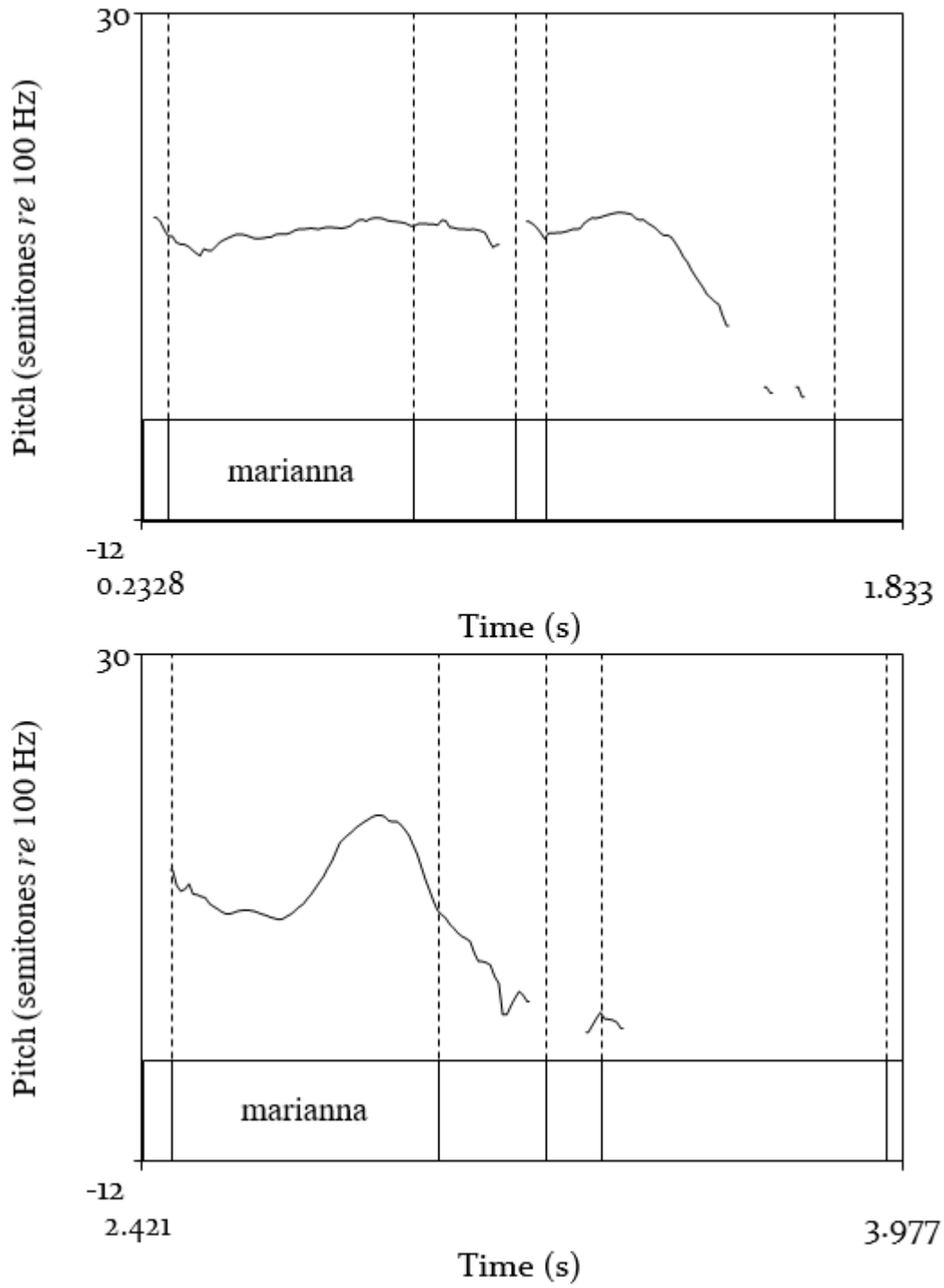

Figura 2.1 – Contraste entre dois enunciados correspondentes à sentença “Marianna made the marmalade” da oficina de aprendizado do ToBI, apenas com a marcação da palavra “Marianna”. A de cima é o enunciado neutro (tom H* na tônica de “Marianna”), a de baixo o enunciado com foco em “Marianna” (tom L+H*).

Assim, nessa teoria, tudo não passa de uma simples sequência de eventos tonais, o que é atestado pela própria maneira de gerar automaticamente a entoação da fala em trabalho da mesma autora (PIERREHUMBERT, 1981). Funções matemáticas são utilizadas para

gerar as transições entre os eventos tonais. O que se passa entre esses eventos não seria linguisticamente informacional para a autora e os adeptos desse modo de representação: são propriedades atribuídas a restrições de natureza articulatória. Essa mesma forma de conceber a representação da entoação da fala é proposta por Ladd (1983b, 1996), que introduziu na proposta de Pierrehumbert uma forma de conectar a sequência de acentos de *pitch* e tons de fronteira por algoritmos de estilização da curva de Fo fundamentados na teoria de percepção da entoação do grupo de pesquisa holandês IPO (HART; COLLIER; COHEN, 1990).

Do modelo de Pierrehumbert surgiu em 1992 um sistema de notação entoacional para o inglês americano (SILVERMAN et al., 1992) chamado de *Tone and Break Indices* (ToBI), que usamos aqui para ilustrar as diferenças entre os dois enunciados acima. É uma notação prática, no entanto, tem baixo índice de acordo entre anotadores quando se trata de escolher um símbolo para um dos seis acentos de *pitch* possíveis, como confirma o índice inferior a 50% encontrado numa revisão de seu uso depois de 10 anos (WIGHTMAN, 2002). A razão desse baixo índice é que se exige do anotador que "escute" o evento tonal, isto é, que distinga de oitiva se é por exemplo um L*+H ou L+H*. De qualquer forma, por sua praticidade, usamos neste livro uma representação que, na superfície, é semelhante a essa para assinalar tanto proeminência quanto fronteira, mas numa concepção fonética da notação entoacional. Essa representação faz parte do sistema DaTo.

2.2.2 O Sistema DaTo de Notação Entoacional

Desenvolvido como parte do trabalho de doutorado de Lucente (2012), o sistema DaTo assume uma relação estreita entre os mecanismos laríngeos para a produção de frequência fundamental (Fo) e o

material linguístico, especialmente a sequência silábica. Radicalmente diferente dos modelos fonológicos apresentados na seção anterior, para esse sistema, as propriedades dinâmicas da curva melódica com suas restrições são fundamentais para a realização das diferentes funções comunicativas[7]. Essas propriedades dinâmicas dizem respeito aos limites de vibração das pregas vocais para a realização de acentos de *pitch* e tons de fronteira num determinado espaço de tempo que normalmente é o intervalo correspondente à silaba acentuada.

O sistema prescinde da necessidade de marcação de que parte da curva melódica está alinhada com a sílaba proeminente porque considera sempre um alinhamento à direita. Concebe também tons dinâmicos e estáticos. Os primeiros são movimentos melódicos como subidas e descidas com característico alinhamento da taxa máxima de subida/descida com a sílaba tônica. Já os tons estáticos são níveis baixo ou alto alinhados com a sílaba proeminente ou marcando fronteira prosódica. Diferentemente do sistema ToBI, o DaTo requer apenas que o anotador reconheça primeiro se a palavra é proeminente ou não (ou se há fronteira ou não), para somente depois observar no traçado da curva melódica visível, através de um programa de análise da fala que extraia essa curva, o tipo de tom, a partir de sua forma e seu alinhamento com a vogal.

As frases contrastadas acima pelo sistema DaTo são transcritas das seguintes maneiras: “Maria$_{H}$ nna made the mar$_{H}$malade$_{L\%}$.” e “Maria$_{>LH}$ nna made the marmalade$_{L\%}$. Não se trata apenas da re- tirada do sinal de alinhamento (*), mas também de uma concepção dinâmica do tom, em que a descida que precede o tom ascendente LH é parte constitutiva de sua implementação e o sinal > indica que o tom está atrasado em relação ao início da vogal, um atraso que está associ- ado a uma grande variedade de funções quando associado a diferentes parâmetros prosódicos (WARD, 2019, p. 91-92), como veremos no ca- pí-

7 Veja também os mesmos pressupostos no modelo entoacional de Kiel (KOH LER, 1991).

tulo 5.

Os sistemas ToBI e DaTo são sistemas de notação que representam a curva melódica. Mas há na literatura modelos de geração da curva melódica fundamentados numa análise fonética dos enunciados. Esses modelos são importantes porque permitem formular hipóteses que con- sideram as restrições do sistema laríngeo. Os modelos mais conhecidos são o desenvolvido há muitos anos por Hiroja Fujisaki e o modelo mais recente de Yi Xu. Esses dois modelos são bem distintos do modelo de Pierrehumbert, pois são ditos superposicionais. Eles propõem a curva melódica como resultado da composição de componentes distintos, enquanto no modelo da autora americana a sequência de tons é linear, um se segue ao outro sem influência de alguma unidade em outro nível.

2.2.3 O Modelo de Fujisaki

O modelo de Fujisaki(HIROSE; FUJISAKI, 1982) estabelece que a curva melódica (curva de Fo) é composta aditivamente de três componentes na escala logarítmica. Como se vê na Figura 2.2, esses componentes são a frequência de base ou valor mínimo F_b; o componente relativo ao sintagma entoacional (*phrase component*) e o componente relativo ao acento de *pitch* (*accent component*). Por essa forma de gerar a curva melódica supor a superposição de três componentes, esse modelo fonético faz parte da classe de modelos superposicionais.

O resultado da adição desses três componentes pode ser visto à direita da figura: a linha tracejada horizontal é o valor mínimo, a linha tracejada superior define os limites dos sintagmas entoacionais a partir dos comandos de sintagma (*phrase commands*) e a linha cheia é obtida com a soma dos três componentes com a aplicação final dos comandos de acento (*accent commands*), que são elementos do modelo que permitem a geração da curva a partir de equações matemáticas

que usam valores associados a suas magnitudes e extensão temporal (no caso do comando de acento) para gerar a curva melódica.

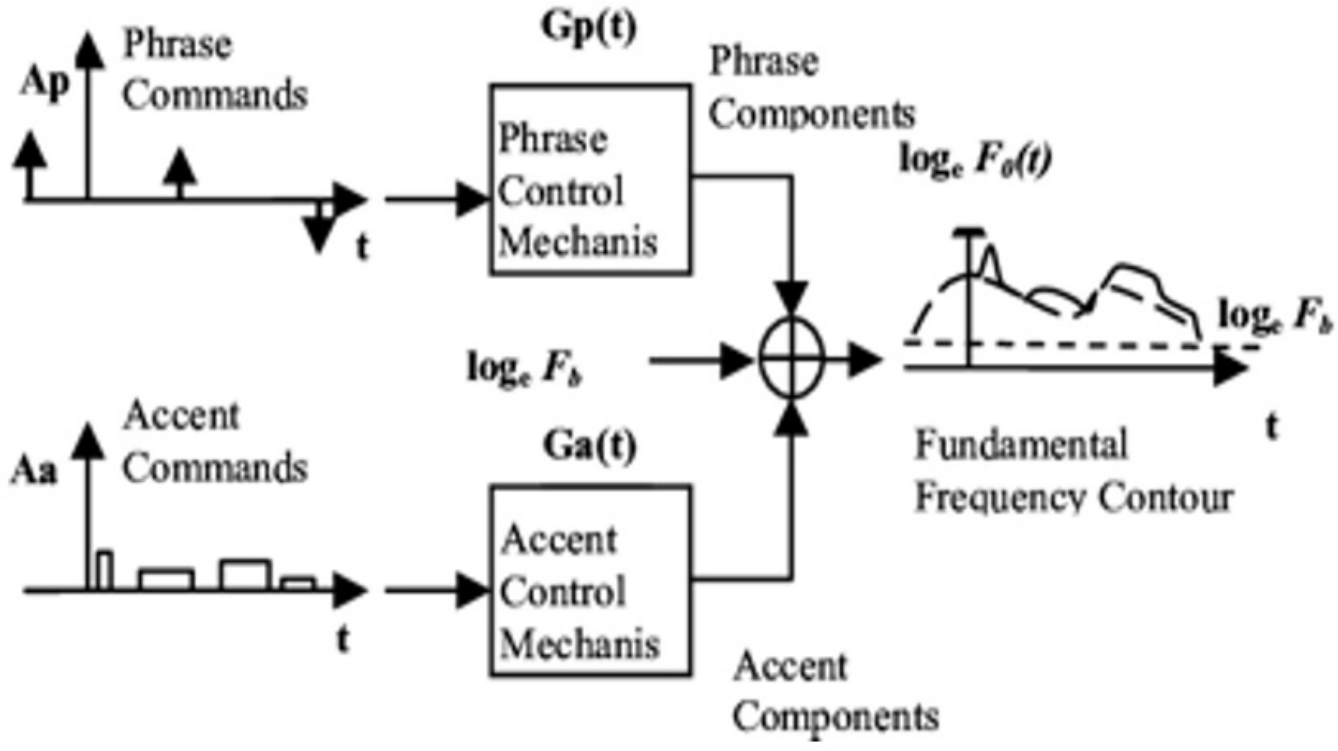

Figura 2.2 – Componentes do modelo de Fujisaki explicado no texto, reproduzida com autorização do autor, Keikichi Hirose. Fonte: Hirose e Fujisaki (1982).

Um exemplo de geração da curva melódica com esse modelo pode ser visto para o enunciado lido “Quando ouvia os sinos a chamá-los, enroscava-se debaixo da manta com os joelhos quase chegando à testa e pensava: ‘talvez se esqueçam de mim’ ” por locutora paulista nas Figuras 2.3 e 2.4. Este exemplo é parte do trabalho experimental desenvolvido por Barbosa, Mixdorff e Madureira (2011) para comparar as diferenças melódicas entre as falas lida e narrada em português brasileiro e alemão padrão.

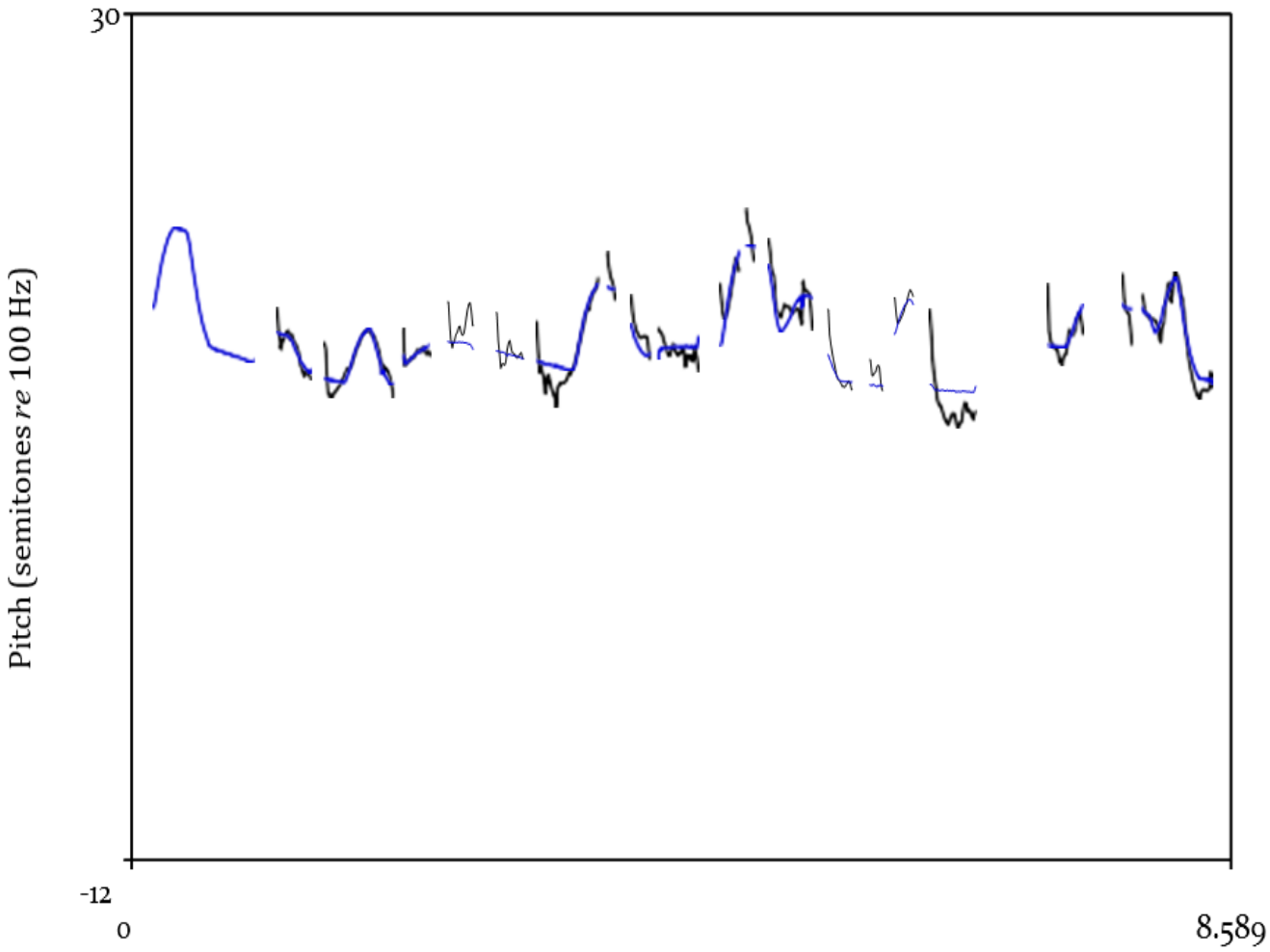

Figura 2.3 – Curvas melódicas (Fo) do enunciado “Quando ouvia os sinos a chamá-los, enroscava-se debaixo da manta com os joelhos quase chegando à testa e pensava: ‘talvez se esqueçam de mim’ ” lido por locutora paulista. Em preto a curva original e, mais clara, a curva gerada pelo modelo de Fujisaki.

Como se pode ver nas figuras, há muito detalhe no traçado da curva melódica original e o modelo de Fujisaki a simplifica sem perda na percepção da entoação. A curva é gerada a partir dos três componentes mencionados com a especificação das posições temporais e valores dos comandos que são obtidos a partir de uma fase de minimização da diferença entre a curva do modelo e a curva original. É assim um mecanismo de aprendizado automático por minimização de erro. O resultado desse procedimento de minimização é visto no trecho do enunciado mostrado na Figura 2.4 onde se vêem dois comandos de sintagma nos instantes de tempo 2,03 e 3,84 s com suas respectivas amplitudes, 0,16 e 0,19, e dois comandos de acento que são intervalos

que duram 0,18 e 0,29 s. Os comandos de sintagma geram a forma geral da curva logo após sua posição, enquanto os comandos de acento geram os acentos de *pitch* dentro de seus intervalos correspondendo aos trechos “chamá-los” e “manta”.

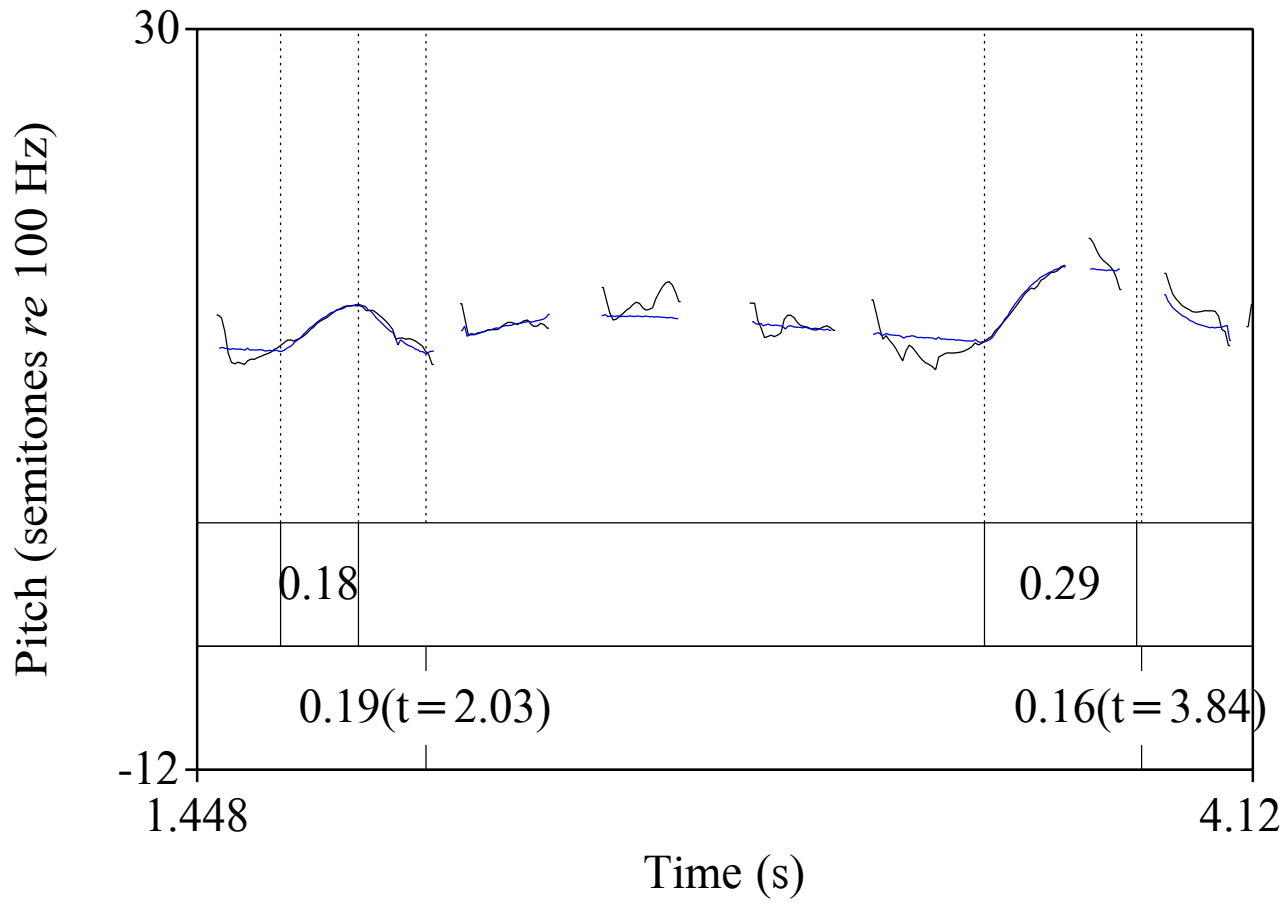

Figura 2.4 – Trecho **a chamá-los, enroscava-se debaixo da manta**do enunciado da Figura 2.3 comparando a curva melódica original (preto) com a gerada pelo modelo de Fujisaki (mais clara) com os comandos de sintagma (camada de baixo) e de acento (camada de cima) assinalados.

A vantagem desse tipo de modelamento é que a curva melódica para cada enunciado pode ser especificada apenas pelos valores dos comandos, possibilitando, por meio desses valores, comparar os estilos lido e narrado nas duas línguas. De fato, em estudo anterior, Mixdorff e Barbosa (2012) mostraram que o modelo de Fujisaki dá melhor conta das proeminências em alemão do que em PB, uma vez que nesta segunda língua a duração é mais frequentemente usada para assinalar essa função prosódica. Pela análise dos comandos de acento, mostramos que

as narrativas nas duas línguas envolvem uma taxa de subidas de Fo mais elevada e valores bem mais variados para esses comandos, assinalando maior variação melódica nos trechos de narrativa em comparação com os de leitura. No estudo de 2011 (BARBOSA; MIXDORFF; MADUREIRA, 2011), comparamos o modelo de Fujisaki com o modelo PENTA, apontando algumas vantagens do segundo em sua relação direta com unidades linguísticas como sílabas e palavras fonológicas.

2.2.4 O Modelo PENTA

O modelo PENTA desenvolvido por Yi Xu (XU; WANG, 2001; XU, 2005) é um modelo superposicional de geração da curva de Fo também na escala logarítmica. Nesse modelo, as funções comunicativas afetam de forma paralela e independente a forma geral da curva final levando em conta alvos estáticos ou níveis e alvos dinâmicos ou inclinações. Um exemplo do tipo de curva de Fo gerada pelo modelo pode ser visto na Figura 2.5.

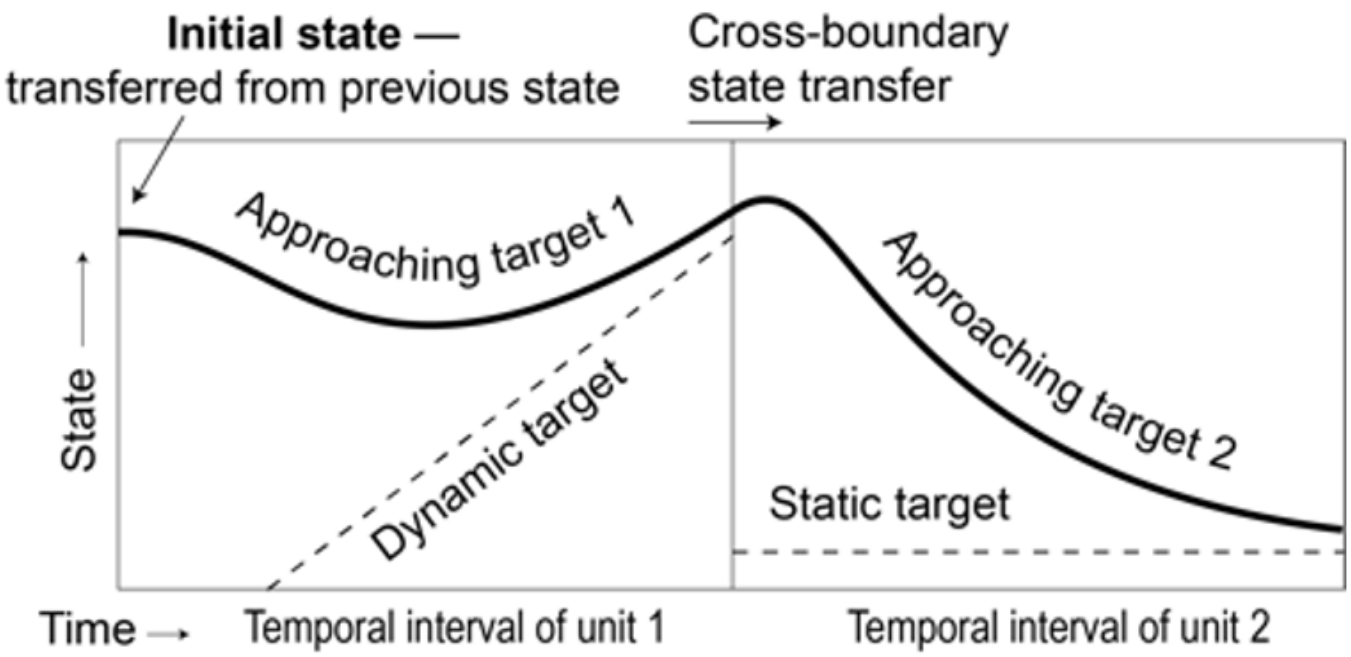

Figura 2.5 – Curva melódica básica gerada pelo modelo PENTA, conforme explicado no texto, reproduzida com autorização do autor, Yi Xu. Fonte: Xu e Wang (2001).

Observe que a curva melódica no esquema da figura se aproxima do alvo dinâmico (*dynamic target*) ao longo do intervalo da unidade linguística 1 e depois se aproxima do alvo estático (*static target*) ao longo do intervalo da unidade linguística 2. Alvos dinâmicos são aqueles que assinalam subida ou descida de Fo, enquanto alvos estáticos assinalam um valor fixo a ser atingido. As unidades linguísticas são os domínios para a realização de uma determinada função prosódica da língua. No caso da implementação do acento de *pitch*, podem ser sílabas, palavras fonológicas, grupos acentuais ou outros domínios relevantes para a função na língua. Além do acento de *pitch*, o modelo realiza o tom de fronteira, aplicando paralelamente uma modificação na curva melódica para a realização de um tom final alto ou baixo.

Por conta de as funções comunicativas nesse modelo serem implementadas de forma paralela, isto é, a realização de uma é independente da realização da outra e o efeito de uma se superpõe ao das outras, o modelo se classifica como superposicional. Por serem pautados em princípios articulatórios de produção da curva de Fo, tanto o modelo de Fujisaki quanto o de Xu acabam sendo modelos fisiologicamente plausíveis que se coadunam com a percepção da melodia. Ambos se servem de uma representação logarítmica da curva de Fo que aponta para a percepção da sensação de *pitch*, uma vez que a percepção do som tem características não lineares próximas da forma logarítmica. Embora o modelo de Fujisaki não imponha de antemão os limites das unidades prosódicas, seus comandos de acento e de sintagma podem ser definidos de forma alinhada com essas mesmas unidades, alcançando plausibilidade linguística.

Os modelos acima pressupõem que a duração silábica seja especificada previamente, daí a necessidade de modelos que tratam da organização temporal.

2.3 Quanto à Organização Temporal

Os modelos de geração da curva melódica que vimos na seção anterior tratam a duração de maneira secundária, atrelada às próprias unidades prosódicas, o que não nos permite entender como a duração silábica, tão fundamental nos modelos apresentados na primeira seção, é gerada. Os modelos de duração seguintes, por estarem focados numa questão tecnológica, a da geração da duração para sistemas de síntese da fala, não levam em conta o papel da sílaba na fala.

2.3.1 Modelos Segmentais

A duração das unidades sonoras começou a ser tratada de forma bastante prática, por conta da necessidade da geração automática de enunciados para a síntese da fala. Os modelos iniciais geravam a duração de unidades isomórficas ao fonema e, por isso, são chamados de modelos segmentais da duração. O modelo mais referenciado da literatura, usado como ponto de partida para os modelos que se seguiram é o modelo de Klatt (KLATT, 1979, 1987).

Em seu modelo, a duração de cada fone do inglês americano é obtida pela equação 2.1.

$$Dur = MinDur + \frac{(InhDur - MinDur) \times PRNCT}{100} \qquad (2.1)$$

Em que *Dur* é a duração gerada; *InhDur* é a duração intrínseca do fone, obtida de uma tabela; *MinDur* é a duração mínima calculada a partir da duração intrínseca[8] e *PRNCT* é a porcentagem de modificação determinada de forma cíclica pela aplicação de um conjunto de

8 Em geral *MinDur* = 0, 45 *InhDur* para todos os fones não acentuados, sendo que esse valor mínimo é dobrado nos fones de sílabas acentuadas.

onze regras.

Dois exemplos de regras, sem detalhamento do fator *PRNCT*, ilustram o tipo de contexto examinado para modificar a duração dos fones: (a) a regra 7, de encurtamento de segmentos átonos, foi proposta a partir de trabalhos experimentais como os de Fry (1958) e é resumida pelo autor assim: segmentos não acentuados são mais curtos que os acentuados, e (b) a regra 8 foi obtida a partir dos trabalhos de Bolinger (1972) e Umeda (1975) sobre a ênfase, resumida assim: uma vogal sob ênfase deve ser bastante alongada. As regras são especificadas de forma matematicamente explícita a partir de valores distintos de *PRNCT*, começando pela aplicação da primeira: atribuir pausas silenciosas de 200 ms antes de cada sintagma no interior da sentença e cada vez que na ortografia tiver uma vírgula. Qualquer influência do ritmo da fala na duração dos fones é totalmente relegada a um ajuste ulterior por proposta do próprio autor (KLATT, 1975). Além do caráter fixo e ad hoc da regra de inserção de pausa silenciosa, o ritmo tem papel secundário no modelo.

Os modelos de O'Shaughnessy (1981, 1984) e Bartkova e Sorin (1987) para o francês e de Santen (1994) para o inglês também são segmentais e, mesmo que as regras sejam obtidas por procedimentos distintos, aplicam fatores de correção contextual para obter a duração final do fone, como no modelo de Klatt em quem se inspiraram. Esses fatores podem ser exemplificados pela lista dada pelo último autor, citando os trabalhos de Klatt: acento lexical, ênfase, fonemas precedentes e seguintes, posição no sintagma e na palavra e natureza do fonema.

Talvez por estarem envolvidos com sistemas de síntese da fala, os autores acima não se preocuparam com os níveis acima do segmento como nos modelos seguintes.

2.3.2 Modelos Acima do Segmento

Buscando integrar como parte constitutiva da duração silábica o aporte do ritmo da fala, os modelos seguintes partem da especificação da duração de unidades superiores ao segmento.

Em seu modelo para gerar a duração do inglês, Witten (1977) procura integrar aspectos prosódicos como a taxa de elocução, a pausa, o ritmo e a curva melódica, tirando o máximo proveito do pé como ponto de partida de geração.

Calculado a partir do início da vogal, como vários foneticistas faziam (cf. autores como André Classe nos anos 1940 e Ilse Lehiste nos anos 1960), Witten parte da duração de pé básica de 480 ms, procurando encaixar as sílabas que constituem cada pé nesse intervalo. Se essa operação produz uma sílaba de tamanho menor que um valor mínimo, o pé é então alongado. A taxa de elocução é modificada a partir de restrições tanto de alongamento máximo possível para os fones, para as taxas lentas, quanto de limite de compressão silábica, para a taxa mais rápida possível (para o autor, cerca de 7 sílabas por segundo). O autor considera, no entanto, um modelo simplificado que admite três tipos de pé: iambos (curta/longa), troqueus (longa/curta) e espondeus (longa/longa).

O modelo de Kohler (1986) para a geração da duração silábica do alemão parte da duração do pé e de restrições quanto à duração dos fones, sendo modelos distintos para as sílabas acentuadas e átonas. Seu modo de conceber a relação entre a pesquisa básica e a geração da duração é apresentado em trabalho ulterior que se resume nestes três pontos (KOHLER, 1991, p. 122): (1) fundamentação da pesquisa aplicada na pesquisa sobre a fala natural aos níveis da produção e da percepção da fala; (2) modelamento da fala com base em pressupostos teóricos motivados e dados empíricos (e não soluções ad hoc) e, no caso dos sistemas de síntese e reconhecimento da fala, (3) agrupar o conhecimento espalhado em várias áreas e fomentar seu aperfeiçoa-

mento antes de criar regras para os sistemas de tecnologia de fala.

Embora utilize um modelo de geração da duração implementado a partir do aprendizado por redes neurais, Campbell (1992, 1993) considera a sílaba como a unidade básica do ritmo da fala. Seu modelo gera assim a duração da sílaba para o inglês britânico para depois distribuir essa duração entre os segmentos que a constituem assumindo uma distribuição uniforme[9].

A busca por construir modelos de duração ecologicamente relevantes, isto é, que espelhem nossos mecanismos de produção e percepção da fala, é a agenda dos modelos dinâmicos do ritmo da fala que apresentamos a seguir, retendo aqui os que permitem uma melhor didática para o entendimento do protocolo de pesquisa experimental, como veremos ao final deste livro.

2.3.3 Modelos Dinâmicos do Ritmo da Fala

Os efeitos de alongamento segmental provocados pela proximidade a uma fronteira prosódica são explicados no modelo de Byrd e Saltzman (2003) assumindo a hipótese de um relógio abstrato externo às pautas gestuais da fonologia articulatória de Browman e Goldstein (1990). Nessa fonologia, os sons da fala são produto de gestos articulatórios dispostos teoricamente numa pauta dita gestual em que cada linha representa o intervalo em que uma determinada ação no trato se dá, como fechar os lábios para um som labial. A teoria explica muitos processos fônicos ao nível lexical (BROWMAN; GOLDSTEIN, 1992), mas carecia uma relação transparente e seguindo pricípios dinâmicos para a prosódia, o que procuraram fazer Dani Byrd e Elliot Saltzman em seus trabalhos. Sendo assim, essa fonologia puramente lexical carecia de um planejamento do tempo a longo termo, muito embora os trabalhos em produção de fala que fundamenta-

9 Essa assunção é contestada no trabalho de Barbosa (1994) que mostra empiricamente que essa uniformidade ocorre na unidade que vai do início de uma vogal ao início da próxima.

ram defendessem a primazia de uma unidade articulatória de vogal a vogal (KELSO; SALTZMAN; TULLER, 1986; LÖFQVIST, 1986).

Para dar conta dos efeitos duracionais na proximidade de fronteira prosódica, Byrd e Saltzman identificaram quatro níveis de organização temporal (*Ibidem*, p. 156), sendo o nível transgestual aquele que se refere às propriedades temporais locais em porção específica do enunciado que os autores exploram para explicar o efeito prosódico. Esse efeito é disparado pelo chamado *π-gesture* ou *gesto prosódico*, conforme proposta oriunda de estudos anteriores (BYRD, 2000; BYRD et al., 2000), que desacelera os gestos de constrição do seu domínio com um grau definido por um valor real positivo denominado de nível de ativação. Nesse domínio, tanto vogais quanto consoantes têm sua duração alterada, mais especificamente a vogal pré-fronteiriça e a consoante imediatamente seguinte, limites que definem uma unidade que vai de uma vogal a outra (unidade VV).

Outro aspecto importante do modelo é que o nível de ativação do gesto prosódico é proporcional à força da fronteira prosódica, o que faz com que fronteiras mais fortes provoquem efeitos de alongamento maior nos gestos segmentais sob o domínio do gesto prosódico. Com esse modelo, os autores simularam quais seriam as consequências para a duração segmental da variação de fatores como (1) a presença ou não do gesto prosódico em seu domínio, causando alongamento dos segmentos ou não; (2) o alinhamento do gesto prosódico com relação aos gestos segmentais, causando alongamento apenas onde se encontra o domínio desse gesto; (3) a força da fronteira prosódica, causando maior alongamento quanto maior o seu valor; e (4) a forma do gesto prosódico, que produz efeitos variados sobre os gestos segmentais. Nenhum dado natural foi, no entanto, apresentado para comparar com as simulações.

Os efeitos duracionais são concebidos de forma distinta no modelo de osciladores acoplados de Barbosa (2006). Esse modelo é uma formulação matemático-computacional que integra uma teoria

dinâmica da produção do ritmo da fala, pressupondo três níveis de acoplamento em três escalas temporais distintas[10] com as seguintes propriedades:

1. O ritmo da fala advém do acoplamento (influência mútua) entre um componente estruturante implementado por um oscilador acentual com parâmetros modificáveis por informação sintática local e componentes regularizadores implementados pela oscilação inicialmente periódica tanto do oscilador silábico quanto do oscilador acentual;
2. A estruturação e a regularidade rítmicas, implementadas pelo acoplamento dos dois osciladores do modelo, operam em escalas temporais distintas, a primeira, da ordem da magnitude do grupo acentual, a segunda, da magnitude da sílaba;
3. O oscilador silábico tem seus ciclos ancorados na sequência de inícios das vogais;
4. O oscilador silábico induzido pelo acentual gera padrões temporais complexos que reproduzem aqueles encontrados em enuncia- dos naturais do português brasileiro;
5. A taxa de elocução, especificada pelo período do oscilador silábico na condição em que não está acoplado com o oscilador acentual, taxa que é uma propriedade dinâmica básica do modelo;
6. Plausibilidades linguística e biológica que possibilitam integrar outros componentes, como o sistema entoacional e os mecanismos de percepção do ritmo e da entoação.

10 O acoplamento entre o oscilador silábico e o acentual, o acoplamento entre os níveis linguísticos acima do oscilador acentual e esse oscilador e o acoplamento entre o oscilador silábico e os gestos da pauta gestual. Suas escalas temporais são respectivamente a da duração silábica, a da duração do grupo acentual e a da duração do fone isomórfico ao fonema.

A figura 2.6 ilustra os componentes do modelo dinâmico do ritmo da fala, em relação ao qual se foca aqui apenas a parte que apresentaos osciladores acentual e silábico e sua interação por meio da força de acoplameto ω_0. O papel dos níveis linguísticos mais elevados, do léxico e da pauta gestual é discutido amplamente em Barbosa (2006).

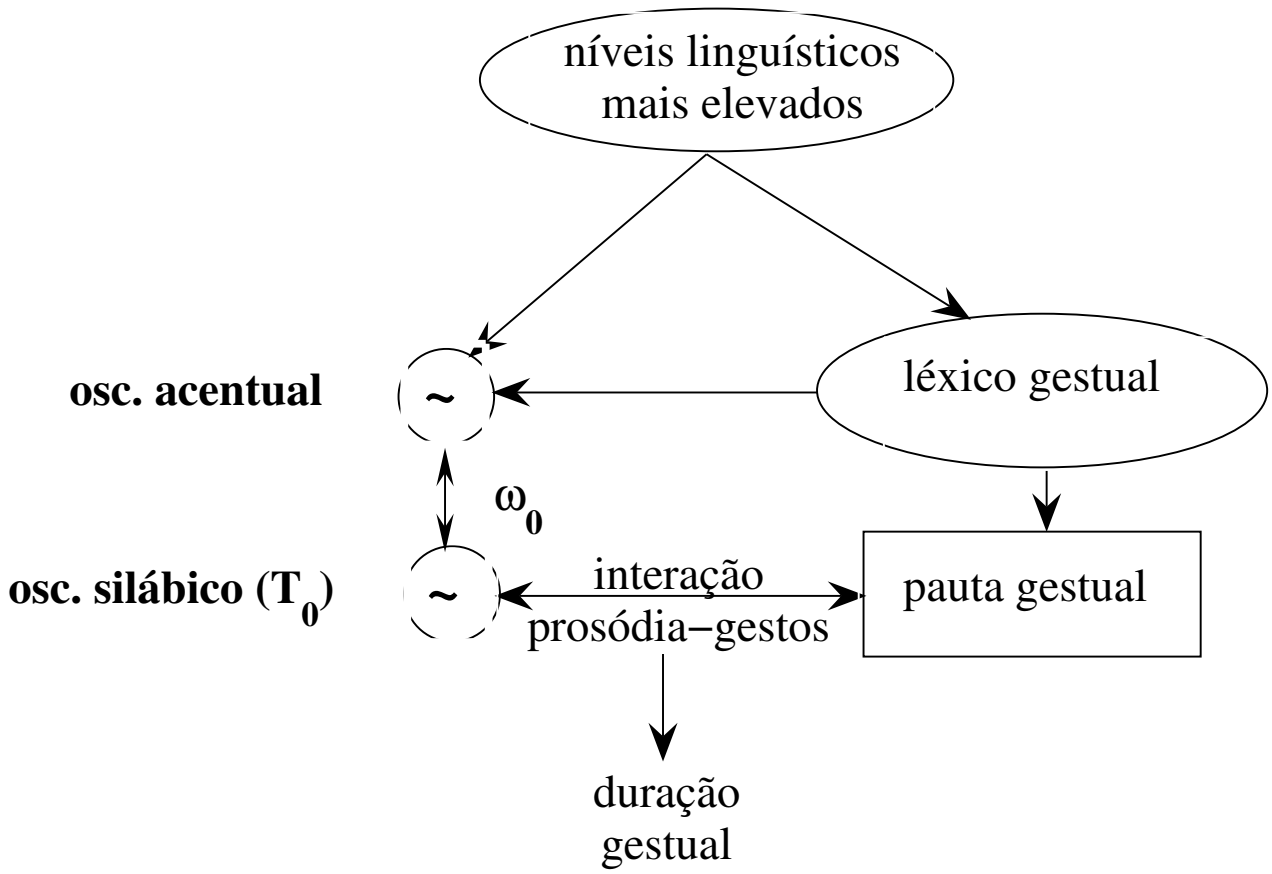

Figura 2.6 – Diagrama do modelo dinâmico do ritmo da fala de Barbosa.

A estimação dos parâmetros desse modelo tomou como referência um *corpus* de frases isoladas, lidas de forma neutra e em três taxas de elocução por um locutor masculino do Recife de cerca de 35 anos na época da gravação. A equação de acoplamento de período do oscilador silábico contém uma função exponencial de sincronismo entre os dois osciladores, cuja forma foi determinada empiricamente a partir

do mesmo *corpus*. Essa caráter exponencial pode ser visto na parte superior da Figura 2.7.

No modelo, as durações das sílabas fonéticas que são as unidades VV são modificadas ao longo do grupo acentual por equações que im- plementam um crescimento duracional até a realização do acento frasal parametrizado por uma força de acoplamento. O crescimento é tanto maior quanto maior a força desse acento, que depende da forma como o falante divide seu enunciado em constituintes e de como faz as proemi- nências prosódicas. A maneira como o oscilador silábico se deixa afetar pelos acentos frasais especificados pelo oscilador acentual que se vê na figura é controlada pelo valor da força de acoplamento ω_0. O modelo é capaz de gerar a duração da ordem da sílaba de maneira próxima à natural a partir de simulações experimentais (BARBOSA, 2007), como ilustra a Figura 2.7. O acento frasal são as posições de proeminência das unidades do tamanho da sílaba ao longo do enunciado.

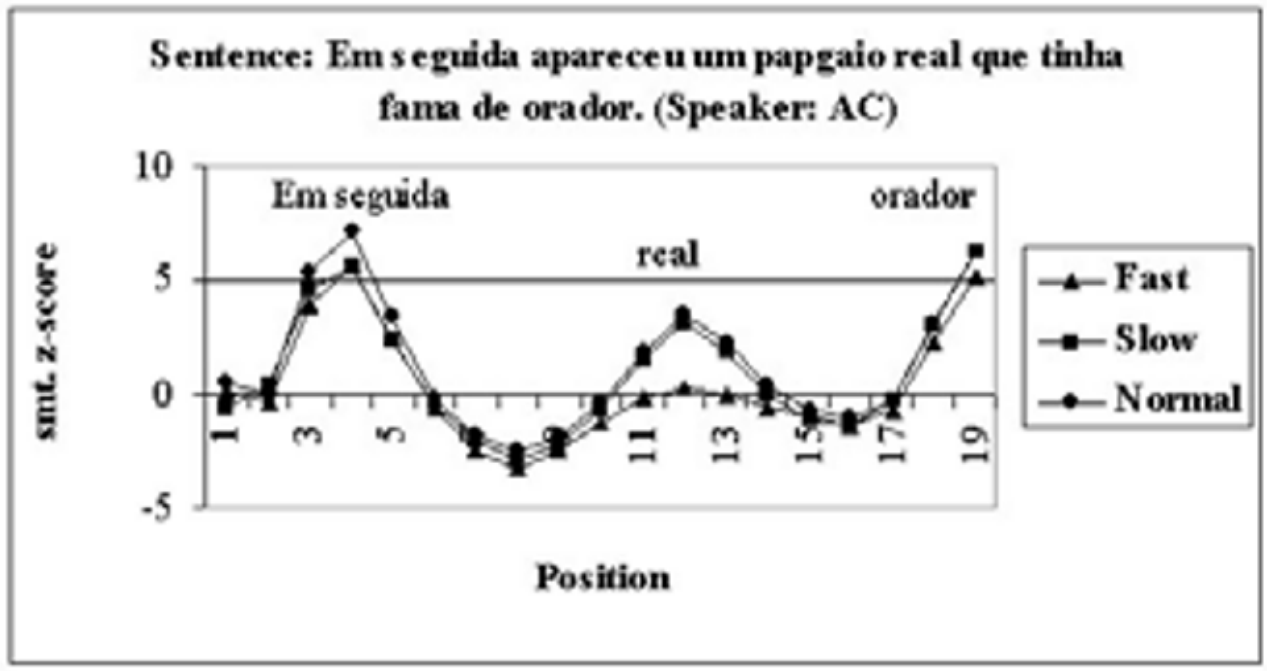

Figura 2.7 – Comparação de duração natural normalizada (acima) e duração gerada pelo modelo dinâmico (abaixo) para as unidades VV da sentença "Em seguida apareceu uma papagaio real que tinha fama de orador." em três taxas de elocução.

Observe na parte superior da figura a duração das unidades VV normalizada pela técnica de *z-score* de três enunciados da mesma sentença em taxas de elocução distintas produzidas por locutor paulista. O padrão obtido com o modelo no painel abaixo é bastante similar. É notório como a duração natural tem um padrão que é de

subida em cada grupo acentual, delimitados pelos três picos que se vêem no painel acima e claramente reproduzidos no painel abaixo.

De interesse experimental é o contraponto que se pode fazer entre os modelos apresentados nesta e nas duas últimas seções (que levam em conta o componente prosódico da fala para explicar a duração silábica) e os modelos segmentais. Esses últimos foram implementados tendo em vista a geração da duração dos segmentos para sistemas de síntese da fala e, por isso, tiverem pouco interesse em entender os mecanismos de produção e percepção da fala que, como vimos na seção 2.1, propõem a prosódia como princípio de organização temporal.

O uso de procedimentos matemáticos e estatísticos para obter a melhor aproximação para a duração dos fones nos modelos segmentais é uma via que não permite uma modificação flexível dessa duração em contextos mais gerais dos que os que são normalmente considerados.

Os modelos que se fundamentam em unidades superiores ao segmento consideram, ainda que localmente, no caso do modelo de Byrd e Saltzman, o aporte das unidades prosódicas, especialmente a sílaba para explicar os padrões duracionais.

Capítulo 3

Metodologia Experimental

Nas próximas seções apresentamos os elementos principais da metodologia experimental aplicada à área de prosódia. Alguns experimentos serão descritos em detalhe, precedidos de uma apresentação sucinta das teorias e observações que os motivaram para que o leitor possa acompanhar todas as fases do ciclo experimental e possa formar um senso crítico. As seções começam por uma discussão geral para entrar em detalhes a partir de um estudo experimental. O título da seção evoca o tema principal, aquele que tomará mais tempo da discussão e especulação de alternativas, mas toda seção conterá todos os aspectos do ciclo experimental, incluindo uma rápida apresentação das observações que motivaram cada estudo. Outras teorias serão elencadas resumidamente aqui, além daquelas mais gerais apresentadas no capítulo anterior.

Para toda análise acústica que requeira um grau de a profundamento em fonética acústica experimental, o leitor encontra informação detalhada no livro de Barbosa e Madureira (2015). Todos os dados usados neste e nos demais capítulos foram obtidos rodando os scripts *SGDetector* e *ProsodyDescriptor*, disponíveis no repositório neste endereço: https://github.com/pabarbosa/prosody--scripts.

3.1 Hipóteses Científicas em Prosódia Experimental

Toda hipótese de pesquisa decorre da teoria científica que procura explicar os fatos observáveis e isso não é diferente na área da fala e da linguagem. As hipóteses formam uma ponte entre a teoria

e a metodologia que será empregada para confirmá-las, refutá-las ou refiná-las, por isso devem ser formuladas de tal forma que possam ser testadas por técnicas de medida e de inferência estatística.

Uma hipótese adequadamente formulada expressa, sob a forma de uma asserção, o que deve ser testado. Por exemplo, se admitirmos por uma teoria geral de percepção da fala, que o acento numa língua é percebido por se destacar do contexto imediato e que esse destaque é realizado por meio de parâmetros como a duração silábica, podemos emitir a hipótese de que a duração da sílaba tônica é maior do que a da sílaba átona. Autores como Massini (1991) e Barbosa (1996) mostraram que essa hipótese se confirma, desde que se assegurem condições de igualdade de contexto, uma vez que diversos fatores afetam a duração silábica. Por exemplo, em enunciados como "Não quero que ela apareça", a sílaba final é átona e dura mais do que a sílaba tônica anterior, por motivos específicos. Nesse caso, a átona dura mais por um efeito chamado de "alongamento final" (GAITENBY, 1965; OLLER, 1973; KLATT, 1975) que estende a duração de segmentos que precedem uma pausa. Além disso, a duração também depende da natureza dos segmentos e, no caso da sílaba átona desse exemplo, a duração também é maior por conta do segmento [s], que é dos mais longos do português brasileiro.

Veremos na seção 3.2 que, para testar uma hipótese, devemos ter condições experimentais em que o contexto imediato seja o mesmo, como no constraste entre os enunciados "Parece que casou sábado" vs. "Parece que caso sábado", em que, de fato, desde que pronunciadas com mesma entoação e com o mesmo ritmo, a sílaba "-sou" do primeiro enunciado (pronunciada como [zo]) é mais longa do que a sílaba -"so" (pronunciada como [zU]) do segundo.

Observe que, admitindo a mesma estrutura prosódica nos dois enunciados num determinado locutor, a única coisa que difere entre eles é a troca entre as palavras "casou" e "caso". As primeiras sílabas das duas palavras podem também ser comparadas para avaliar diferenças

duracionais entre tônica e pré-tônica pois, na palavra “caso”, a primeira sílaba é tônica e, na palavra “casou”, a primeira sílaba é átona. Em estudo comparando apenas as vogais das sílabas em entrevistas e trechos lidos, mostramos que as duas categorias de átonas se comportam da mesma forma (BARBOSA; ERIKSSON; ÅKESSON, 2013), contrariando resultados com frases isoladas em que a pós-tônica é mais curta do que a pré-tônica.

Passamos a detalhar dois exemplos de experimentos para indicar a forma como as hipóteses científicas são construídas. No primeiro exemplo, utilizaremos duas teorias concorrentes para investigar a existência do desfazimento do chamado encontro acentual, enquanto, no segundo exemplo, utilizaremos uma teoria do papel crucial da transição C-V para o processamento da sílaba, no intuito de mostrar que nossos sistemas de produção e percepção sonora estão vinculados e ancorados temporalmente nesse evento silábico.

3.1.1 Hipóteses em Pesquisa sobre Encontro Acentual

Na Fonologia Métrica de Liberman e Prince (1977), o aspecto relacional do acento pode ser indicado por uma estrutura chamada de “grade métrica”. Esse tipo de representação em grade, que representa em coluna o “grau” de saliência silábica, fornece uma ferramenta para explicar a necessidade de preservar a alternância de proeminências acentuais, alternância que caracteriza o chamado ritmo linguístico e que é garantida por uma “regra de ritmo”. A regra de ritmo se impõe na teoria quando a relação fraco-forte que existe em iambos como *thìrtéen* da grade 3.1 se inverte, soando como um troqueu (padrão forte-fraco) quando inserida em sequências como *thírtèen men*, em que a palavra *men* porta o acento frasal. O papel da regra do ritmo é o de desfazer o encontro acentual (*stress clash*) entre elementos adjacentes na grade, como se vê pelas duas colunas mais altas na grade

3.1, em que os marcadores de posição ‘x’ ocupam lugar sobre a sílaba correspondente, indicando um grau de acento que é proporcional à altura da coluna de ‘x’.

Tabela 3.1 – Grade 1, com choque acentual.

		x
	x	x
x	x	x
thir	teen	men

Pode-se ver que, na linha média da grade 3.1, os dois *x* consecutivos não têm nenhum *x* numa coluna que os separasse e que permitisse um “relaxamento” da “tensão” (termos dos autores) criada pela adjacência das duas colunas de *x* mais à direita que, assim, marcam uma contiguidade de dois níveis de saliência acentual, configurando o que os autores chamam de choque acentual[1]. Esse choque é desfeito, segundo eles, pela regra do ritmo, que age para criar a relação da grade 3.2, que soaria como se o acento estivesse na primeira sílaba da primeira palavra (observe agora que os dois *x* na linha média são intercalados pelo *x* de uma sílaba na linha inferior).

Tabela 3.2 – Grade 2, com choque acentual desfeito.

		x
x		x
x	x	x
thir	teen	men

Se adotarmos exemplo similar em português, a mesma regra do ritmo atuaria para modificar a relação métrica numa sequência como

1 Preferimos o termo “encontro acentual” por não entendermos que essa contigui dade seria sempre desfeita.

"café quente", produzindo uma palavra "café" que soaria como paroxítona, algo bastante improvável. No entanto, um desfazimento de cho- que acentual parece se dar na pronúncia fossilizada da expressão "Jesus Cristo", tão facilmente ouvida na canção de Roberto Carlos. Voltaremos a essa questão depois. Por ora, passemos a examinar a previsão da teoria dinâmica do ritmo apresentada na seção 2.3.3.

Nessa seção, vimos que o modelo dinâmico gera durações que aumentam até a realização do acento frasal, que ocorre em torno de unidade VV lexicalmente acentuada em palavra que o locutor enunciou como proeminente. Assim, se a sentença "Tomam|os um café quent|e" for produzida com dois acentos frasais, um na primeira palavra e outro na última, teremos um grupo acentual final, o segundo do enunciado, que começa depois da realização do primeiro acento frasal em "-ma-" e vai até "quen-", que é caracterizado por um movimento ascendente de duração. Por conta desse movimento de crescimento de duração, haverá um reforço de duração na segunda sílaba de "café" e não um reforço da duração de 'ca-". Observe que a previsão teórica do modelo dinâmico é oposta àquela prevista pela Fonologia Métrica em caso de desfazimento do encontro acentual. Isso ocorre porque, no modelo de osciladores acoplados, as posições de acento lexical são apenas pontos de ancoragem eventual de acento frasal.

Em estudo anterior (BARBOSA, 2002), mostramos que quatro locutores do português paulista realizam situações de encontro de acentos lexicais da forma hipotetizada pelo modelo de osciladores acoplados, isto é, com aumento de duração na sílaba mais à direita, contígua ao segundo acento lexical da sequência de duas palavras em análise, delimitadas abaixo por colchetes. Usamos pares de frases em que a primeira é uma frase-controle, sem encontro acentual, e a segunda é a frase experimental, para a qual ocorre encontro de acentos lexicais entre as palavras-chave "comi" e "bolo", entre "bordeaux" e "xucro", entre "falou" e "baixo" e entre "bebê" e "calvo", conforme abaixo, onde se indica em negrito a unidade onde incidiu o acento frasal.

- Eu [comi bol**or**] sexta-feira à n**oit**e. vs. Eu [comi b**ol**o] sexta-feira à n**oit**e;
- O [bordeaux chin**ês**] derramou-se pela m**es**a. vs. O [bordeaux x**ucr**o] derramou-se pela m**es**a;
- Parece que [falou 'baix**ou**'], e não 'ca**iu**'. vs. Parece que [falou 'b**aix**o'], e não '**alt**o';
- Um l**ind**o [bebê carm**im**]. vs. Um l**ind**o [bebê c**alv**o].

Para o locutor paulista analisado e utilizando-se um teste de ANOVA[2] com nível de significância de 5%, não foram encontradas diferenças significativas na duração média ao comparar tanto as primei- ras sílabas da palavra-chave quanto ao comparar as segundas sílabas (observe que são sílabas idênticas do ponto de vista fonológico). Na comparação entre sílabas, apenas o segundo par de frases teve diferença significativa na segunda sílaba ([do]) com valor de duração média de 151 ms na frase-controle e de 167 ms na frase experimental ($p < 0.02$). Na comparação com as unidades VV, em todos os pares a segunda VV é sempre mais longa que a primeira, na palavra-chave.

Com base nesse e em outros experimentos conduzidos (BARBOSA, 2002; BARBOSA; ARANTES, 2003; BARBOSA; ARANTES; SILVEIRA, 2004; MADUREIRA et al., 2004), a hipótese de desfazimento acentual da Fonologia Métrica foi refutada e a do modelo de osciladores acoplados confirmada. Observe como as hipóteses nas duas teorias conduziram a uma metodologia experimental que foi capaz de decidir entre uma e outra, uma vez que faziam previsões exatamente opostas. Além do mais, no caso do inglês americano, a cuidadosa investigação de Grabe e Warren (1995) revelou que existe um padrão de alternância de sílabas fortes e fracas que independe de qualquer noção

2 O teste ANOVA avalia a significância da diferença de média entre dois ou mais grupos de valores, desde que se obedeçam determinadas condições para a sua realização. Detalhes sobre esse tipo de teste no capítulo 6, seção 6.1.

de choque acentual, resultado confirmado em estudo ulterior de Kimball e Cole (2014).

Passemos agora a exemplificar um segundo experimento, sobre sincronização fala-metrônomo.

3.1.2 Hipóteses em pesquisa sobre o *p-center*

Estudos de autores como Fraisse (1982, p. 153) mostraram que a solicitação de produção espontânea de batidas repetidas do dedo indicador sobre a mesa em experimentos realizados desde a década de 1930 revela períodos com valores em torno de 600 ms que são representativos desse tipo de controle por nosso sistema motor.

Uma vez que a atividade motora da fala e a da batida do dedo indicador seriam controladas pelo mesmo mecanismo temporal gerado no córtex cerebelar (LEINER; LEINER; DOW, 1991), é de se esperar que a oscilação silábica produza períodos dessa ordem de grandeza quando somos solicitados a produzir sílabas repetidamente, como, de fato, ocorre (BARBOSA et al., 2005). Isso nos leva a pensar que podemos produzir essa repetição silábica em sincronismo com um metrônomo ou sequência de tons puros[3], ficando a questão de que lugar da sílaba ocorreria esse sincronismo, questão de pesquisa que norteou estudos na década de 1970. Esses estudos chamaram esse lugar de *perceptual-center* ou simplesmente *p-center* (MORTON; MARCUS; FRANKISH, 1976; MARCUS, 1976; POMPINO-MARSCHALL, 1989, 1991), definido como o momento no sinal acústico em que o ou- vinte se ancora para perceber uma sequência sonora como ocorrendo a intervalos regulares no tempo.

Vimos na seção 2.1 que a transição C-V é uma candidata para esse ponto de ancoragem temporal na fala, o que nos faz hipotetizar que a sílaba se sincronizaria com a batida de um tom puro exata-

3 O tom puro corresponde a um som periódico simples, formado por uma única frequência.

mente na transição C-V, isto é, no início da vogal, ponto em que há mudança brusca de energia. Consequentemente, não haveria distância entre o instante de tempo da transição C-V e esse tom.

Observe como essas asserções determinam o modo de conduzir a metodologia que deve conceber: (1) um modo de realizar essa sincronização com locutores do PB; (2) um modo de medir a distância entre a batida do metrônomo sonoro e o início da vogal; (3) um teste estatístico para avaliar se, ao menos em média, essa distância é nula. Visto que o foco de nosso sistema cognitivo na transição C-V está relacionado a transições bruscas de energia entre consoante e vogal, é importante testar o grau de sincronismo ao variar a discrepância dessas energias variando modos de articulação da consoante e altura da vogal. Tudo isso fizemos num experimento sobre p-center em PB (BARBOSA et al., 2005).

Para testar a hipótese principal sobre o sincronismo fala-metrônomo em torno da transição C-V, concebemos uma tarefa de produção de uma sequência de sílabas que o participante, um estudante paulista de cerca de 20 anos, tinha que fazer em simultaneidade com uma sequência de tons puros tocada via fone de ouvido. A primeira etapa do experimento foi aferir a taxa de elocução confortável de produção de uma sequência silábica pelo participante. Isso foi feito pedindo apenas que ele produ- zisse uma sequência de sílabas [pa], como achasse melhor, o que ele fez com um intervalo médio entre inícios de vogal de 556 ms (108 bpm).

Para a tarefa propriamente dita, cada sílaba a ser produzida repetidamente foi apresentada visualmente num cartão, numa ordem aleatória ao início de cada produção. Foram 21 sílabas CV distintas produzidas pela combinação das consoantes /p f r s ʃ l m/ com as vogais /i ɛ a/. Em torno de dez sílabas idênticas foram usadas para as análi-ses, sendo descartadas as cinco primeiras, pois foram consideradas um tempo de adaptação à tarefa. Para a realização do experimento, uti- lizamos um metrônomo analógico, de marca Matrix, modelo MR-

500, com taxas-limite de 40 a 208 bpm e a fala do participante foi capturada com microfone unidirecional com amostragem a uma taxa de 22,05 kHz com sinal do metrônomo gravado simultaneamente.

A figura 3.1 mostra a sequência de sílabas [pɛ] produzida em sincronismo com o metrônomo à taxa de 108 bpm. Observe como os pulsos do metrônomo, visíveis na parte negativa do gráfico, ficam em torno da transição CV.

A distância entre cada pulso e a sílaba correspondente foi medida calculando a fase φ definida pela equação 3.1, em que $t_{trans.CV}$ é o instante de tempo em que se dá a transição CV, no início acústico da vogal, $t_{p-center}$ é o instante da batida do pulso mais próximo do metrônomo e M é o período do metrônomo.

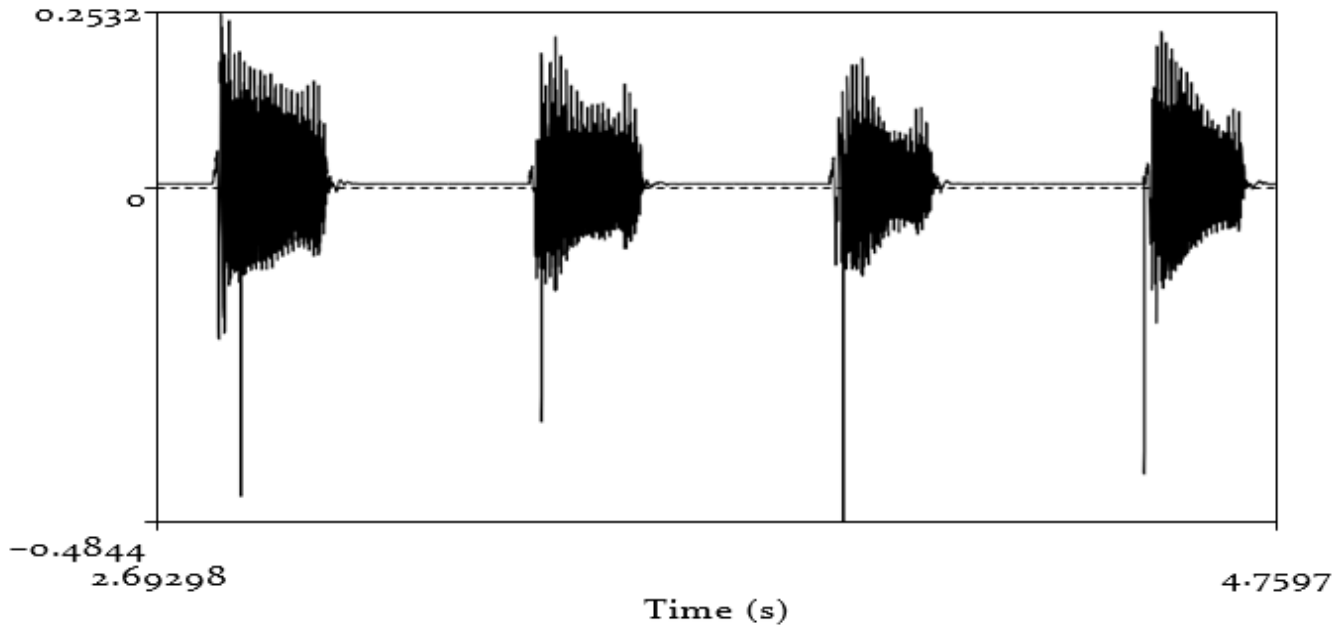

Figura 3.1 – Formas de onda de sequência de sílabas produzidas em sincronismo com o metrônomo a 108 bpm. Pulsos do metrônomo visíveis na região negativa do gráfico.

(3.1)

$$\varphi = \frac{(t_{trans.CV} - t_{p-center}) \cdot 360^{o}}{M}$$

Embora os resutados do experimento tenham revelado uma grande diferença na realização da tarefa, dependendo da sílaba considerada e da taxa do metrônomo, é importante ressaltar essa "atração" da produção do participante pela transição C-V, exceção feita quando de presença de consoante com muita energia como [ʃ]. A figura 3.2, reproduzida do capítulo 2 de Barbosa (2006), ilustra os valores médios (e desvios-padrão) da diferença de fase para todas as sílabas com o metrônomo a 108 bpm. Um teste t de amostra única[4] foi realizado para cada sílaba, adotando-se como hipótese nula que a média de $\varphi = 0$. Essa hipótese se manteve para as sílabas [pɛ], [pi], [fa], [sa], [la], [lɛ], [xa], [xɛ], [mi], isto é, todas elas tiveram seus inícios de vogal síncronos com as batidas do metrônomo.

Seis dos tipos silábicos que seguem a hipótese nula apresentam uma discrepância de energia total entre consoante e vogal seguinte das mais altas do rol de sílabas analisadas. Além disso, quando a taxa do metrônomo se lentifica, passando a 80 bpm, o sincronismo mantém um padrão semelhante à produção espontânea em relação a seu afastamento da transição C-V.

4 Este teste avalia a significância da diferença entre a média de um conjunto de dados e uma única média teórica. No caso aqui, 0. Uma apresentação formal desse teste será feita no capítulo 6, seção 6.1.

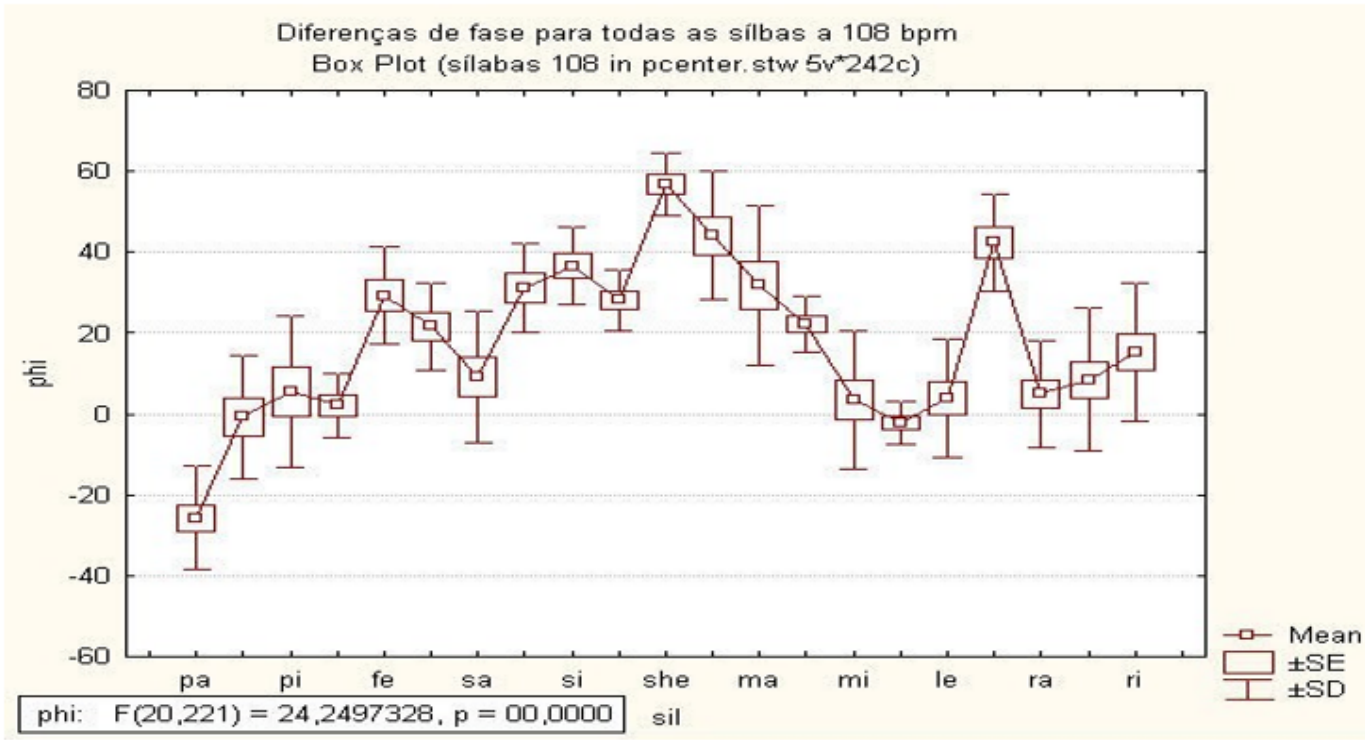

Figura 3.2 – Média e desvio-padrão (em graus) da posição do p-center em relação à transição C-V para as sílabas do experimento de sincronização fala-metrônomo com metrônomo a 108 bpm. O valor o para φ corresponde à posição da transição C-V.

Sobre os três pontos gerados pela hipótese principal, mostramos com esse exemplo (1) um modo de realizar a sincronização com locutores do PB que foi estendida a outros locutores no trabalho de Melo (2016), (2) um modo de medir a distância entre a batida do metrônomo e o início da vogal com a equação de fase acima e (3) usamos o teste t de variável única para avaliar se a hipótese de distância nula pode ser mantida.

Quanto ao último ponto, o experimento revelou um comportamento mais complexo com dependência do tipo de consoante e vogal e da taxa de metrônomo. Mas, em regra geral, os experimentos sugerem que há uma regularidade silábica que é organizada por uma sequência de transições CV. Novos experimentos podem ser feitos para confirmar esses resultados com participantes com e sem experiência musical, para ver como essa experiência afetaria o desempenho na tarefa de sincronização. Indivíduos de outras faixas etárias e outros dialetos e línguas testariam a universalidade do fenômeno. A

produção de sílabas distintas na sequência testaria como o sujeito se adaptaria à mudança de padrão de discrepância de energia em cada sílaba. Outras estruturas silábicas avaliariam se, de fato, a consoante de coda em sílabas CVC não afetaria o fenômeno de sincronização mantendo a atração pela transição C-V. O leitor pode ver que a teoria do *p-center* gera uma série de questionamentos concretos, uma característica de todo estudo experimental bem conduzido.

Vimos nesta seção e na anterior que as hipóteses foram formuladas de tal forma que conduziram à montagem de experimentos em que as observações vinculadas às teorias puderam ser verificadas a partir de medidas acústicas e de testes estatísticos inferenciais. Enquanto no experimento sobre encontro acentual apenas uma das teorias concorrentes explica o que acontece em PB em caso de encontro acentual, o experimento sobre *p-center* confirma parcialmente o sincronismo fala-metrônomo na transição C-V, pois é afetado por condições específicas. Prosseguindo com a metodologia, abordaremos os protocolos para a realização de experimentos na área.

3.2 Protocolos de Investigação em Prosódia Experimental

Os equipamentos para gravação e reprodução para estudos prosódicos são os mesmos dos usados para qualquer estudo fonético. Por isso, recomendamos a seção "Instrumentos de gravação e reprodução da fala" do livro de Barbosa e Madureira (2015) que traz a recomendação de que o microfone a ser usado seja unidirecional com uma resposta em frequência relativamente uniforme na faixa entre 30 e 16000 Hz. Para a reprodução sonora em testes de percepção da prosódia, o uso de fones de ouvido de alta qualidade é recomendado e existe uma gama grande de produtos no mercado.

O objetivo desta seção é orientar o leitor quanto a protocolos para

a realização de experimentos que dizem respeito especialmente a (1) escolha do participante[5]; (2) escolha do material a ser gravado e seleção de material para um teste de percepção; (3) protocolos experimentais para gravação de corpora e preparação de instruções para testes de percepção; (4) técnicas para obter material comparável em estilos de elocução distintos; (5) técnicas úteis para testes de percepção, como a deslexicalização.

3.2.1 Escolha do Participante

A escolha do participante depende da pesquisa que se faz e das hipóteses vinculadas à teoria adotada, mas, considerando a disponibilidade de cada um e a manutenção do bem-estar de cada pessoa, deve-se levar em conta os seguintes aspectos gerais.

Caso a pesquisa não diga respeito ao estudo de alguma patologia de fala afetando a prosódia, os participantes da pesquisa não devem ter problemas fonoarticulatórios ou auditivos. A depender do grau de importância para a pesquisa, essa constatação pode ser feita por autodeclaração ou com o auxílio de um fonoaudiólogo.

O participante deve ser capaz de realizar a tarefa que se pede e nem sempre isso é óbvio. Assim, um pequeno teste antes da coleta de dados é importante. Por exemplo, numa determinada leitura, um par- ticipante pode ter problema de fluência e, dependendo do caso, deve ser dispensado do experimento. Claro que, se as consequências da fluência em leitura para a prosódia da fala for o tema do trabalho, a escolha do participante se guiará justamente pelos níveis de fluência. Nesse caso em particular, uma ferramenta ou protocolo que avalie essa fluência é necessário para a classificação de cada um num determinado nível. A variação de fluência é inevitável em estudos de prosódia

5 Usamos o termo “escolha” de forma intercambiável com “seleção”, uma vez que,embora “seleção” assinale a obediência a um conjunto de critérios, uma parte de aleatório deve sempre ser considerada para um experimento que comporta um teste estatístico inferencial. Assim, entre dois participantes que obedecem a determinados critérios de inclusão, escolhe-se um deles para o experimento.

de língua se- gunda (L2) ou estrangeira (LE), exigindo, nesse caso, a avaliação dessa fluência ou da proficiência para que o aporte desse fator nos resultados possa ser avaliado adequadamente. Em tarefas com narrativas ou com jogos, é necessário pré-avaliar a habilidade do participante com essas tarefas, incluindo o uso dos equipamentos e das ferramentas do proto- colo experimental. Também se deve levar em conta que há pessoas que não têm muita habilidade em manusear o mouse; outras, a depender da faixa etária, não têm a capacidade de fazer determinadas tarefas por falta de treino ou maturidade motora ou cognitiva. Damos alguns exemplos.

Num experimento que fizemos, era necessário que um texto fosse lido de forma persuasiva, mas nem todos os participantes contactados foram capazes de fazer isso de forma adequada. Noutro experimento ainda, foi preciso simular uma atitude sarcástica a partir de um cenário imaginado. Mais uma vez, alguns participantes não foram capazes de fazer isso satisfatoriamente e foram descartados. A avaliação da adequação dos enunciados produzidos pela realização desses tipo de tarefa pode ser feita num teste de percepção em que se pergunta se determi- nado enunciado veicula persuasão, sarcasmo ou outra atitude ou afeto.

Mesmo quando o participante tem condições físicas e tem habilidade para fazer as tarefas, é preciso verificar se os parâmetros prosódico-acústicos são adequadamente mensuráveis em sua fala. Em caso de vozes soprosas, roucas ou com muita laringalização, por exemplo, haverá muitas falhas na medida de Fo, impossibilitando trabalhar com esse parâmetro. Para tanto, é essencial fazer um teste de gravação com verificação subsequente em programas de análise acústica para ver a continuidade dos traçados de Fo e se as fronteiras dos segmentos no espectrograma de banda larga são claramente delimitáveis na pessoa gravada. Se houver muitas falhas na obtenção desse traçado e na de- limitação de fronteiras na fala, o melhor é escolher outro participante. Nem sempre uma fala que soa bonita, agradável, é boa para

análise acústica.

Numa tarefa de percepção, por outro lado, uma tarefa de familiarização é indispensável para avaliar se as instruções são bem executadas e se a tarefa testará realmente o que se deseja. A seção 3.2.9, devotada a experimentos de percepção da prosódia, examinará esses cuidados com mais detalhe.

3.2.2 Distratores, Aleatorização e Deslexicalização

Em protocolos experimentais que envolvam o uso de frases ou palavras isoladas, é imprescindível garantir duas coisas. A primeira delas é que se intercalem frases (nos experimentos com frases) ou palavras distratoras (nos experimentos com palavras), para que o participante não infira os objetivos do experimento, pois isso afeta a forma de pronunciar ou o desempenho num teste de percepção. O número de distratores deve ser maior do que o das frases experimentais, para que seu efeito seja efetivo. Por exemplo, se um modo de realizar a entoação de questões for o tema do experimento, o número de frases interrogativas pode suscitar um comportamento desviante para evitar a monotonia ou por incomodar o participante, no caso de ele ter dificuldades com interrogativas. Para remediar isso, frases de outras modalidades como assertivas e imperativas devem ser inseridas em quantidade apropriada entre as interrogativas do experimento. Essas frases distratoras serão descartadas depois, não tomando tempo algum da análise. O mesmo se dá em testes de percepção, pois o ouvinte deve realizar uma tarefa de tal forma que o que faz num momento não influencie o que vai fazer depois.

O segundo aspecto imprescindível em leitura é que nas repetições dos trechos pelo mesmo participante e por participantes diferentes, a sequência tenha ordens distintas. Isso evita que o comportamento não desejado gerado numa leitura por conta do que se leu antes seja

reproduzido em todas as repetições pela mesma pessoa ou em todas as pessoas. Por exemplo, se numa frase um participante usa de uma ênfase numa palavra e encontra numa frase seguinte a mesma estrutura sintática, pode tender a fazer uma ênfase semelhante, enquanto, se ti- vesse lido essa frase muito mais adiante, não teria feito assim. Com a mudança de ordem das frases garante-se que o efeito indesejado não se repetirá. Esse efeito em que uma tarefa determina o comportamento na seguinte se chama efeito de "prompt". Por isso, o procedimento de aleatorização de frases ou palavras a serem lidas deve ser sempre adotado. A cada repetição, mesmo num mesmo participante, uma ordem aleatoriamente distinta deve ser usada. Se o material estiver escrito em cartões ou preparado em slides, a aleatorização simples pode ser feita, respectivamente, como se embaralhassem cartas ou com uma função de aleatorização do programa de apresentação de slides. Em ambos os casos, cada frase e palavra deve estar num cartão ou slide distintos.

No caso de testes de percepção, a aleatorização dos estímulos do teste é feita por instrução do programa que se usa para fazer o teste. O Praat, por exemplo, tem quatro métodos de aleatorização, a depender do que se deseja obter: (1) simples com ou (2) sem reposição, (3) por blocos evitando ou (4) não evitando o mesmo estímulo ouvido imediatamente antes. Convido o leitor a examinar esses procedimentos no *Help* do programa com a chave de busca *Randomization strategies*.

Ainda no caso de testes de percepção de prosódia da fala, um outro tipo de distrator pode ser necessário: fazer com que o ouvinte se concen- tre nos aspectos prosódicos e não nos segmentais ou ainda, evitar que reconheça uma língua ou um locutor conhecido. O reconhecimento de uma língua pode prejudicar o experimento, pois pode induzir um com- portamento específico no participante. Reconhecer um locutor pode facilitar de forma não desejada um teste de percepção. Por exemplo, num teste de reconhecimento de estilos de elocução entre jornalístico e político, se o participante ouve o enunciado sem nenhuma modificação e reconhece que quem fala é o Bóris Casoy, res-

ponderá que o estilo é jornalístico. Para fazer com que o participante se concentre na forma como o locutor fala, em sua prosódia, existem técnicas de deslexicali- zação.

Um dos métodos usados mais remotamente é o da inversão do sinal de áudio que o leitor pode ouvir como exemplo no arquivo **VentoSu- lInvertido**. A vantagem é que não se entende o que foi dito, mas, fora isso, apresenta duas desvantagens principais. O locutor pode ainda ser identificado pelo tom da voz e a curva melódica fica invertida, impos- sibilitando o reconhecimento da entoação do enunciado.

Outro método usado até hoje é o da filtragem passa-baixas, que usa um filtro digital para manter apenas as frequências do sinal de fala abaixo de determinada frequência de corte. O leitor pode ouvir dois exemplos com cortes distintos nos arquivos **VentoSulFiltrado200** e **VentoSulFiltrado400**. Note que é muito difícil reconhecer o que foi dito, embora, talvez, na filtragem a 400 Hz, se soubéssemos o assunto antes, pudéssemos inferir algo. Para que esse método funcione, deve-se preservar a curva de Fo e, para tanto, saber a frequência máxima de Fo no trecho que a pessoa fala, sob o risco de alterar a curva de Fo e tornar equívoca a percepção da entoação. Neste locutor a frequência máxima é de 210 Hz, sendo o corte a 200 Hz algo que deve ser evitado. Aqui parece não ter prejudicado tanto, pois a entoação e ritmo da fala parecem semelhantes nos dois áudios. A filtragem passa-baixa pode ser feita no Praat selecionando o objeto de áudio (Sound no Praat), escolhendo no menu *Filter* a opção *Filter (pass hann band)* para então escrever os limites de frequência a ser preservada entre 0 e a frequência de corte.

Ainda outro método de deslexicalização é o PURR (*Prosody Unveiling through Restricted Representation*), método que propõe preservar a prosódia do enunciado substituindo o sinal original por uma versão que usa uma onda senoidal. Baseando-se nessa filosofia, Petra Wagner implementou um script para o Praat que usa a função $sinc(x) = sen(x)/x$ que é alterada para se iniciar a cada pulso glotal

do sinal de fala com valor de frequência previamente extraído por um algoritmo específico do Praat. A intensidade é também preservada, no entanto, nenhuma transição formântica entre consoantes e vogais é mantida nesse método. O mesmo exemplo de áudio modificado por esse método pode ser ouvido no áudio **VentoSulPurr**. O script pode ser obtido neste lugar: <PURR-2004>. Na seção 3.2.9 daremos exemplos de experimentos que usaram esse método.

A figura 3.3 mostra as curvas de Fo do áudio original, que pode ser ouvido no site do livro com o nome **VentoSulOriginal**, e dos áudios deslexicalizados pelos métodos de inversão do sinal, de filtragem passa- baixas nas duas frequências e do algoritmo PURR.

3.2.3 Escolha e Cuidados com o Material para Gravar

Tendo em vista que a maior parte da pesquisa em prosódia envolve algum aspecto de controle experimental, é necessário obter gravações dos participantes ou fazer com que se submetam a um teste de percepção. Tanto uma tarefa quanto outra não deve passar de 30 minutos seguidos. Havendo necessidade de mais material gravado ou de um teste de percepção mais longo, é preciso organizar esses períodos de coleta em sessões em que esse tempo-limite seja respeitado. Evita-se num caso o cansaço vocal e, no outro, a sobrecarga cognitiva.

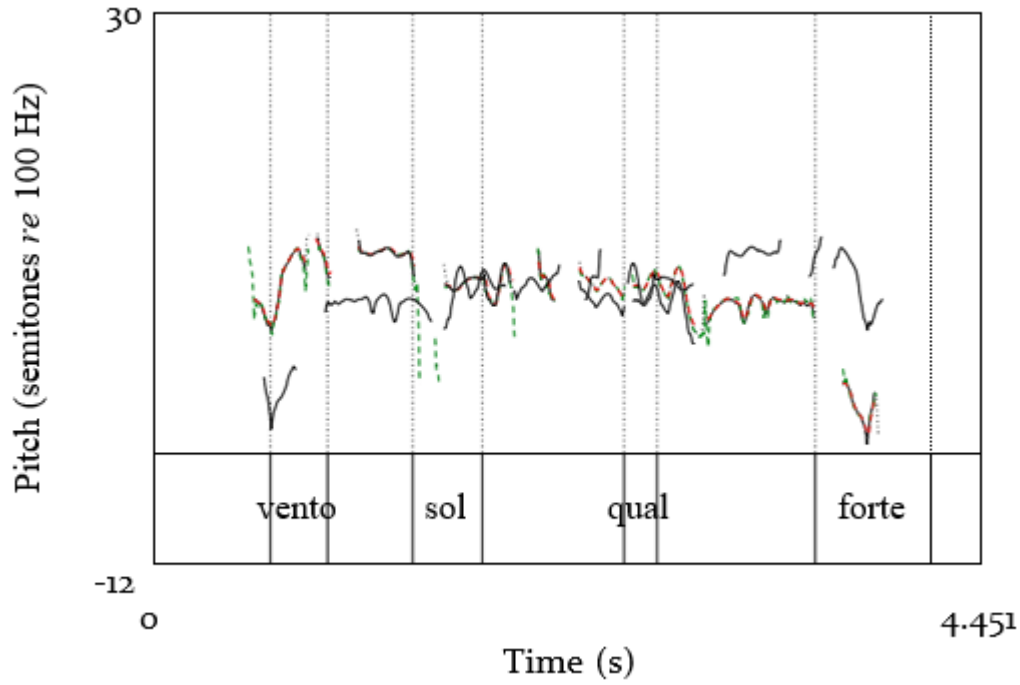

Figura 3.3 – Curvas de Fo do enunciado "O vento sul e o sol discutiam qual dos dois era o mais forte." após três procedimentos de deslexicalização. A curva original e aquelas pelo método de filtragem passa-baixas (curva pontilhada vermelha - 400 Hz - e curva tracejada verde - 200 Hz) e o método PURR coincidem, enquanto a por inversão tem a curva invertida também. Apenas algumas palavras são mostradas para facilitar a visualização de trechos da curva melódica.

Como discorremos no primeiro capítulo, a fala de laboratório abrange muito mais do que a leitura de textos: é toda fala obtida com algum grau de intervenção do pesquisador (XU, 2010). Em qualquer experimento em que se deseja compreender a forma e a função prosódicas em situações espontâneas, é inevitável que um procedimento de controle experimental não entre em jogo. A leitura de frases isoladas foi, durante muito tempo, o tipo de material gravado mais usado na investigação fonética tanto segmental quanto prosódica. Mas, se de um lado permite investigar com eficácia o que se altera na forma em contrastes de determinada função comunicativa nos eixos paradigmático ou sintagmático, por outro lado dificilmente o contraste se dá espontaneamente. Vejamos um exemplo.

Num experimento sobre os correlatos acústicos do acento lexical em PB, contrastamos dois estilos de elocução, fala lida e fala de entre-

vista, para ver se os correlatos se mantinham os mesmos para assinalar o grau de tonicidade das vogais de oxítonas, paroxítonas e proparoxítonas com número de sílabas variando de 2 a 6 (BARBOSA; ERIKSSON; ÅKESSON, 2013). Neste estudo usamos do seguinte procedimento: gravamos as entrevistas entre amigos próximos para assegurar mais material. Em seguida, transcrevemos as entrevistas ortograficamente e escolhemos trechos das próprias entrevistas com frases mais apropriadas para leitura a serem lidas pelas mesmas pessoas duas semanas depois. Examinamos três parâmetros prosódicos em vogais em posição tônica, pré-tônica e pós-tônica nos três padrões acentuais lexicais do PB: duração, desvio-padrão da Fo e intensidade relativa (ênfase espectral). Seguindo esse procedimento, foi possível comparar diretamente os parâmetros em dois estilos de elocução, fala lida e fala de entrevista. Os resultados mostraram que os parâmetros mantiveram a mesma hierarquia de importância em revelar o acento lexical, sendo a duração bem superior aos demais parâmetros prosódicos.

Por conta de muitas vezes haver necessidade desse contraste entre fala não planejada de antemão (como na entrevista, em narrativas) e fala lida, a leitura é ainda muito usada nos estudos prosódicos, desde a frase isolada até textos de diversas naturezas.

Tanto para estudos em uma língua pouco investigada, quanto para estudos de estilos de elocução em que o papel de determinadas funções prosódicas bem como as formas prosódicas a elas associadas são pouco conhecidas, é importante utilizar material a ser lido nos experimentos. Num trabalho sobre o estilo jornalístico em PB e em francês da França (MAREÜIL; BARBOSA, 2018), utilizamos um texto curto, de cerca de 100 palavras nas duas línguas que foi lido em estilo habitual e jornalístico por quatro profissionais do jornalismo em Campinas e em Paris. O texto lido foi exatamente o mesmo, somente os enunciados foram produzidos de forma a reproduzir uma leitura habitual e uma leitura jornalística por pessoas acostumadas a fazer isso em sua profissão. Pelo estudo dos grupos acentuais realizados pe-

los participantes, chegamos à conclusão de que a proporção de uso de proeminências iniciais aumenta no estilo jornalístico, sobretudo em francês, que a taxa de elocução diminui de 10 a 43% nesse estilo e que a frequência fundamental mediana é superior no estilo jornalístico, mais nos homens em francês e nos dois sexos em PB.

A frase isolada, por sua vez, pode ser usada para aumentar a compreensão de aspectos básicos da prosódia, como efeitos segmentais, realização de diferentes tons de fronteira, realização de diferentes tipos de foco, diferenças na realização de modalidades frásticas como assertiva, interrogativa, imperativa e mesmo estudos dos efeitos de diferentes atitudes e emoções.

Os efeitos segmentais se referem tanto à interferência da produção da prosódia nos segmentos fônicos (consoantes e vogais) quanto a dos segmentos fônicos na prosódia. É, portanto, uma interferência de mão dupla. Desse modo, verificam-se modificações tanto nos parâmetros prosódico-acústicos (Fo, duração e intensidade) diante da ocorrência de segmentos fônicos com determinadas características, quanto nas propriedades acústicas dos segmentos fônicos por modificações na estrutura prosódica de um enunciado. Um dos mais conhecidos fenômenos desse tipo de interferência mútua é a chamada micromelodia, uma modificação local na curva de Fo que se verifica sob o efeito da ocorrência de um fone vozeado ou não vozeado. Assim, no contraste “Fizeram a sobremesa usando a nata disponível.” vs. “Fizeram a sobremesa usando nada de caro.”, a realização de [t] em “nata” aumenta localmente a vibração das pregas vocais na vogal seguinte, fazendo com que a curva de Fo no início dessa vogal comece com valor mais alto. Exatamente o movimento contrário, ocorre depois do [d] em “nada”, ou seja, no início da vogal que segue essa consoante a curva de Fo se inicia com valor mais baixo.

Um exemplo do segundo tipo de efeito segmental que pode ser estudado pelo contraste de frases isoladas é a mudança de intensidade, de duração e de primeiro formante do [s] na palavra “saco” em condi-

ções distintas de foco em: “Ele comprou um SACO de estopa.” (após alguém ter falado que era uma caixa) vs. “Ele comprou um saco de estopa.” (após uma pergunta sobre o que a pessoa fez naquele dia). Observe que o uso de frases isoladas permite assegurar o mesmo contexto fonético para aquilo que se quer verificar, o efeito da prosódia sobre o segmento [s].

É também fundamental, para se construir um conhecimento de como se realiza acusticamente uma fronteira prosódica, a comparação entre unidades linguísticas de sentenças distintas. Veja o seguinte contraste entre diferentes tipos de fronteira prosódica ao fim da palavra “cedo”, sendo as duas primeiras terminais e as duas seguintes não terminais. O locutor é de identidade paulista com cerca de 50 anos na época da gravação.

- Logo cedo? Paulo foi pra São Paulo.
- Logo cedo. Paulo foi pra São Paulo.
- Logo cedo Paulo foi pra São Paulo.
- Logo cedo, Paulo foi pra São Paulo.

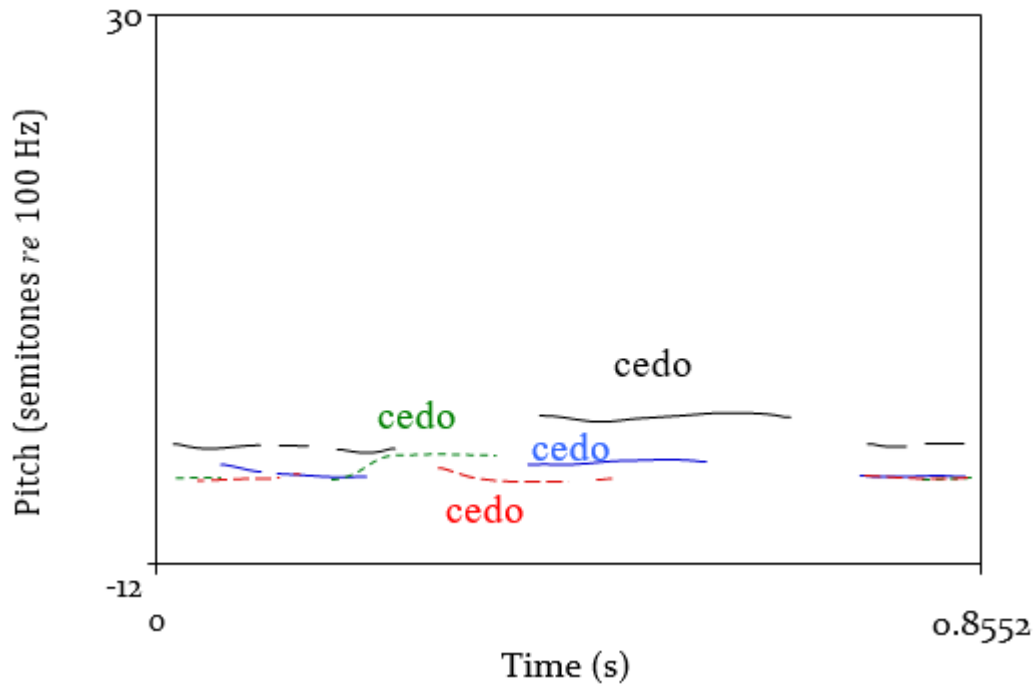

Figura 3.4 – Curvas de Fo do trecho “Logo cedo Paulo” de quatro enunciados de locutor paulista extraídos de “Logso cedo. Paulo foi pra São Paulo.” (vermelho, tracejado); “Logo cedo? Paulo foi pra São Paulo.” (verde, pontilhado); “Logo cedo Paulo foi pra São Paulo.” (cheia, mais alta); “Logo cedo, Paulo foi pra São Paulo.” (cheia, mais baixa).

Na Figura 3.4, ilustramos o contraste das curvas de Fo dos quatro exemplos para o trecho “Logo cedo Paulo”. Observe nos exemplos de fronteiras não terminais indicadas pelos contornos de linhas cheias que a curva sobe e fica nivelada em “cedo” caindo depois, no início de “Paulo”. Já nos exemplos de fronteiras terminais, indicadas pelos contornos de linhas tracejada e pontilhada, se pode ver em “logo cedo” assertivo que a curva desce durante “cedo” enquanto em “logo cedo” interrogativo a subida da curva está contida na palavra “cedo”. Nesses dois últimos exemplos há uma pausa silenciosa entre “cedo” e “Paulo”.

Muito frequentemente, o conhecimento de como são os perfis de F0 numa situação controlada como essa, com pelo menos três funções distintas, permite identificar esses mesmos perfis ou componentes deles na fala espontânea, como se vê no exemplo da Figura 3.5. Nesse exemplo chama-se a atenção para os perfis de F0 das palavras que precedem fronteiras não terminais, “opção” e “público”. Pode-se

ver que, tanto na leitura quanto na entrevista, são perfis majoritariamente ascendentes. O perfil é de uma subida mais acentuada em "opção", na leitura em contraste com a entrevista, e mais semelhante na segunda palavra, embora tenha uma descida curta ao final por conta de uma laringalização durante as pós-tônicas de "público", na entrevista.

A entrevista foi sobre tópico relacionado aos estudos e à vida profissional, entre amigos próximos. E a leitura, de trechos selecionados da entrevista transcritos ortograficamente e lidos pela mesma pessoa duas semanas depois, três vezes em ordem aleatória entre os trechos selecionados.

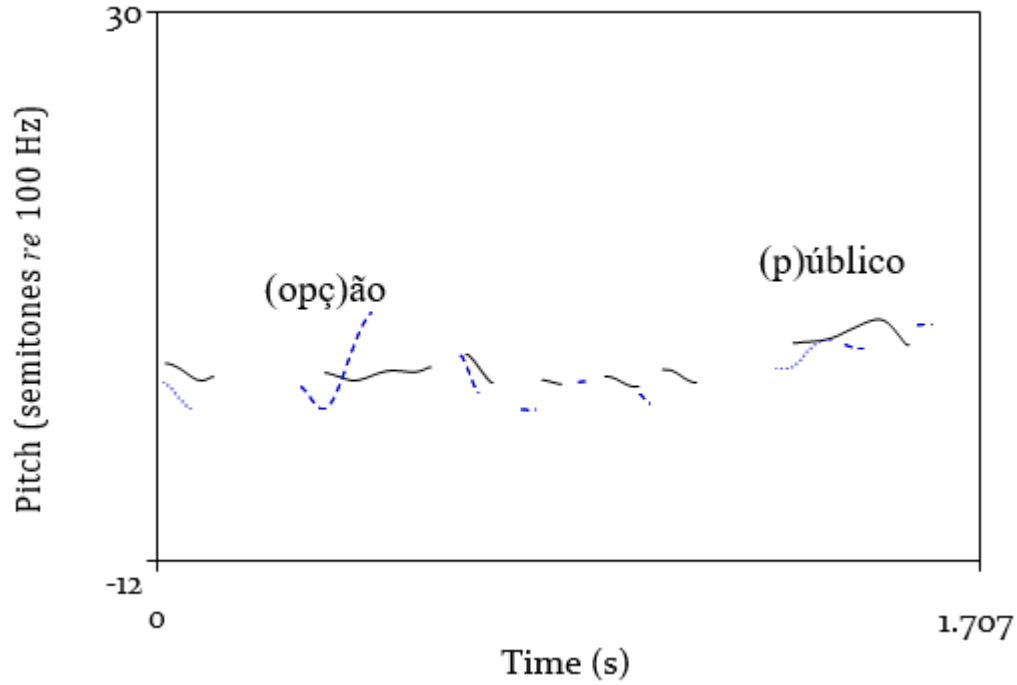

Figura 3.5 – Curvas de F0 do trecho "de opção, prestei concurso público [. . .]." de enunciado mais amplo tirado de uma entrevista entre amigos (linha cheia) e de leitura duas semanas depois do mesmo trecho pelo mesmo locutor paulista de cerca de 25 anos.

O mesmo se dá em diferentes tipos de foco estreito. Reconhecer na fala espontânea um foco contrastivo, por exemplo, é uma tarefa facilitada se se conhece como se realiza em laboratório. No exemplo acima, em que usamos a palavra "saco" para exemplificar o efeito da

prosódia sobre o segmento [s], pode-se montar um protocolo experimental para o estudo dos diferentes perfis de F0 em enunciados contrastando a mesma palavra com diferentes tipos de foco: foco informacional, foco contrastivo, ausência de qualquer foco ou em posição pós-focal. Para tanto, bastaria instruir o participante a primeiramente ler a frase que será a controle, sem qualquer tipo de foco. Para as demais, instrui-se da seguinte forma: na com foco informacional se apresenta para leitura a frase que contém a palavra "saco" grafada em maiúsculas e se pede para que leia o que está em maiúsculas com ênfase. Para o foco contrastivo, por sua vez, diz-se ao participante que a palavra em maiúsculas sublinhada, por exemplo, deve ser lida para deixar claro ao interlocutor que não é um "monte" e sim um "saco" de estopa. E para obter a frase com a palavra "saco" de forma desfocada, colocam-se as maiúsculas na palavra seguinte, "estopa", instruindo o participante para dar ênfase nessa outra palavra. As frases devem ser repetidas algumas vezes (pelo menos cinco) e ser intercaladas com frases distratoras.

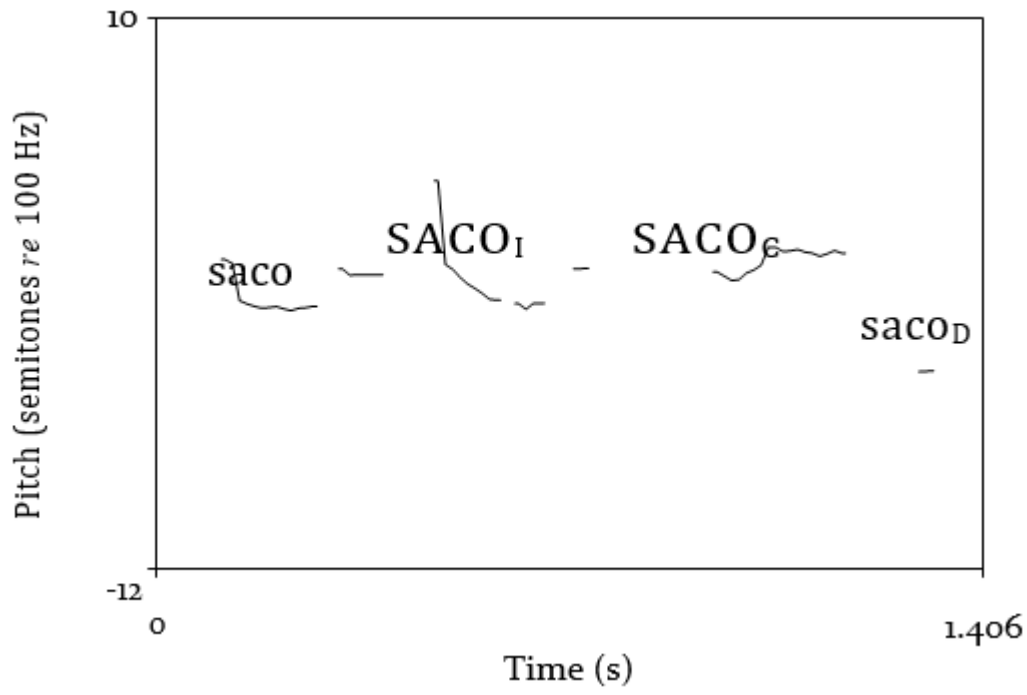

Figura 3.6 – Curvas de F0 extraídas da sentença "Ele comprou um saco de estopa." com quatro condições de foco, da esquerda para a direita: sem foco, com foco informacional, com foco contrastivo e desfocado.

Na Figura 3.6 pode-se ver que, no primeiro perfil de F0, sem foco algum, após o efeito elevador de F0 do [s], o contorno fica nivelado na tônica [a] e sobe na pós-tônica. Na mesma tônica, o perfil é descendente na palavra com foco informacional, enquanto é ascendente no foco contrastivo. A palavra desfocada é muito curta e foi dita com alguma soprosidade, tendo apenas alguns valores válidos e baixos de F0. Esses padrões poderão ser verificados, em seguida, na fala espontânea, a partir do conhecimento adquirido pelo pesquisador na situação controlada apresentada aqui. Vale a pena, sobretudo, observar os movimentos ascendente e descendente nos dois tipos de foco, pelo fato de a mudança na direção do movimento determinar diferenças importantes na percepção de proeminência melódica.

O efeito de atitudes e emoções na fala é o de afetar normalmente a prosódia como um todo nos enunciados, não tendo efeitos majoritariamente de domínio restrito como no caso de foco estreito mesmo que, como vimos, o foco estreito em "estopa" afete as condições de realização da palavra "saco". O problema do estudo de atitudes e emoções é a possibilidade de obter resultados satisfatórios em fala não espontânea, pois é algo que requer capacidade de interpretação. Vejamos o caso do estudo de Silva (2019) na seção seguinte.

3.2.4 Protocolo Experimental: Atuação de Atitudes

O corpus de fala montado por Silva (2019) em sua tese de doutorado foi formado por gravações de dez sentenças, produzidas por onze participantes brasileiros, sendo seis mulheres e cinco homens. Essas sentenças foram produzidas em atitudes de ironia sarcástica, sarcasmo puro e fala neutra, com enunciados correspon-dentes validados por teste de percepção. Por esse procedimento, 236 enunciados foram considerados válidos e retidos para as análises prosódico-

-acústicas. A validação é necessária justamente para reter apenas os enunciados que realmente passam as atitudes requeridas no estudo, tendo em vista que nem sempre é simples fazer essa interpretação. Silva utilizou para isso de um texto escrito apresentando um cenário em que a sentença experimental era usada para marcar a atitude naquela situação específica num contexto dialógico. Vejamos o exemplo para produzir sarcasmo puro e ironia sarcástica. Dois participantes, A e B, liam e procuravam "sentir" as situações seguintes para produzir de forma apropriada as duas atitudes nas sentenças em negrito.

No caso de sarcasmo puro foi esta a situação usada para a sentença "Você deveria passar protetor antes de sair de casa":

> A: Eu já estou vermelha no corpo inteiro novamente. E isso porque eu apliquei uma espessa camada de protetor solar antes de sair para a praia hoje.
>
> B: Sério? Qual você usa?
>
> A: Aquele que eu comprei na cidade recentemente. Você estava lá, lembra?
>
> B: Mas aquele lá só tem fator de proteção 5! Isso é muito pouco! **Você deveria passar protetor antes de sair de casa.** (E quero dizer um protetor de verdade.)

enquanto, para a ironia sarcástica, a situação foi a seguinte:

> A: Finalmente o inverno acabou! Vou à praia esta tarde.
>
> B: Está apenas 5º C! Não acho que esteja bom lá ainda!
>
> A: Mas o tempo está tão lindo! Só por acaso vou pegar o guarda-sol e meu calção de banho.
>
> B: Claro... Porque está super quente. **Você deveria passar protetor antes de sair de casa**.

No seu primeiro estudo de produção, Silva investigou se a expressão das atitudes modifica um ou mais entre 17 parâmetros prosódico-acústicos calculados globalmente nos enunciados. Os parâmetros acústicos foram descritores estatísticos da frequência fundamental, intensidade global e relativa, duração e qualidade de voz, extraídos dos enunciados validados pelo uso de um script para o Praat. Esses parâmetros serão trabalhados nos capítulos 4 e 5.

Quatro dos cinco parâmetros relacionados à frequência fundamental foram significativamente distintos para a ironia sarcástica quando comparada à fala neutra, sendo que a mediana, o valor máximo e o valor mínimo de Fo[6] tiveram médias menores em relação à atitude neutra, resultado condizente com o obtido para línguas germânicas como o alemão e o inglês e o oposto dos encontrados para línguas românicas como o italiano e o francês, mostrando que aspectos atitudinais dependem mais de convenções sócio-culturais, conforme explicado em detalhe na tese de Silva (2019).

Para a qualidade de voz, o resultado mais interessante é que, em relação à atitude neutra, tanto a ironia sarcástica quanto o sarcasmo puro tiveram redução da relação harmônico-ruído, o que aponta para um aumento do ruído espectral nesses enunciados, isto é, mais soprosidade na fala. Observe que as modificações estudadas afetam os enunciados como um todo. Esse aspecto foi mais importante do que modificações locais (em palavra específicas, por exemplo). Observe ainda que, embora sejam contrastadas nesse estudo frases, elas só se encontram isoladas, fora de contexto, na condição neutra. Esse aspecto pode explicar uma leitura encontrada mais rápida nas frases experimentais (as com sarcasmo puro e ironia sarcástica) obtidas a partir da situação colocada para os participantes. O uso de um mesmo texto em diferentes estilos pode ser uma alternativa para explicar essa diferen-

6 Seis descritores apresentados no capítulo 5.

ças nos contrastes examinados aqui.

Em dois outros estudos mostrados na próxima seção apresentamos protocolos em que se faz um paralelo entre diferentes estilos de fala.

3.2.5 Protocolo Experimental: Entrevista Informal pareada a leitura

No estudo de Barbosa, Eriksson e Åkesson (2013), pareamos fala lida e entrevista informal de forma a obter material comparável para o estudo dos correlatos acústicos do acento lexical nas vogais do PB. Embora a duração silábica (e consequentemente seu principal componente, a vogal) já tenha sido apontada como parâmetro mais importante em apontar o acento lexical em PB, algumas questões haviam ficado em aberto do ponto de vista experimental, uma vez que os estudos se restringiram à fala lida (BARBOSA, 1996; MASSINI, 1991; MORAES, 1987; FERNANDES, 1976). As principais questões dizem respeito (1) ao grau de importância dos diferentes parâmetros prosódico-acústicos para assinalar o acento lexical, (2) se a hierarquia entre esses graus se mantém em outros estilos de fala e (3) a eventuais diferenças entre homens e mulheres no uso dos parâmetros relativos ao acento lexical.

Assim, as hipóteses científicas foram assim enunciadas: (1) a duração é o principal parâmetro que assinala o acento lexical, independentemente do estilo; (2) a intensidade global é o parâmetro na sequência de importância, tendo em vista os estudos de Massini (1991), Moraes (1987), Fernandes (1976); (3) a F0 teria uma importância menor, tendo sobretudo uma função entoacional; (4) não há diferença no uso e hierarquia dos parâmetros do acento entre homens e mulheres.

Para responder a essas hipóteses, montamos um corpus com as seguintes características e participantes: 5 homens (de 21 a 30 anos) e 5 mulheres (de 18 a 26 anos) do Estado de São Paulo, todos univer-

si- tários com Graduação completa que deram uma entrevista informal a algum amigo muito próximo, que fazia a gravação. Cada participante teve, assim, seu próprio entrevistador, com todas as gravações feitas na sala do Grupo de Estudos de Prosódia da Fala, e usando um microfone Shure SM58 conectado à placa de som ProTools com amostragem de 22050 Hz. Em seguida, fizemos a transcrição completa das entrevistas. Do material transcrito dos 10 entrevistados, selecionamos 15 trechos com uma sintaxe compatível com a de texto escrito e montamos 15 cartões escritos sem eventuais hesitações. Esses trechos foram se-lecio- nados a partir de palavras neles contidas que assegurassem, no total, uma proporção de oxítonas, paroxítonas e proparoxítonas semelhante à encontrada no PB (CINTRA, 1997). Os trechos foram lidos três ve- zes pelas mesmas pessoas, duas semanas depois, com leitura em ordem aleatória. De cada trecho foi escolhida uma palavra para ser analisada, que também foi transcrita isoladamente para ser lida. Assim, no total, para cada participante, foram analisadas 15 palavras distintas nos três estilos de elocução: Entrevista Informal entre amigos próximos (EI), Leitura de trechos da Entrevista transcrita (LE) e Leitura de Palavras isoladas (LP).

Para o total de 150 palavras (10 participantes x 15 palavras por participante), a distribuição do padrão acentual lexical foi de: 70% de paroxítonas, 20% de oxítonas e 10% de proparoxítonas, compatível com a distribuição desse padrão em PB segundo Cintra (1997). Quanto à extensão das palavras, foi de 2 a 6 sílabas, sendo 84% de 3 e 4 sílabas. As palavras selecionadas se encontravam na proporção de 62% em posição medial na frase, sendo 82% dessas em situação de proeminência nos dois estilos. Para cada palavra produzida foram medidas nas vogais as variáveis dependentes que seguem.

- Duração em ms. O número e a variedade das vogais dispensam o procedimento de normalização duracional explicado no capítulo

4: 1610 vogais para os homens e 1728 vogais para as mulheres.

- Mediana e desvio-padrão da Fo em Hz e em semitons, uma medida logarítmica que simula a sensação de *pitch* (mais no capítulo 4);
- Ênfase espectral em dB. A ênfase espectral (EE) foi definida por Traunmüller e Eriksson (2000) pela equação EE = L - L_o, em que L é a energia de todo o espectro e L_o é a energia da banda baixa do espectro da frequência 0 até o valor limite de 1,43 x média da Fo na vogal. Essa medida é correlato do esforço vocal.

Para mostrar as diferenças entre os valores médios dos parâmetros acima entre os estilos e os níveis acentuais em cada gênero, utilizamos uma ANOVA de dois fatores e um teste *post hoc*.[7] Para avaliar o quanto um parâmetro explica as diferenças de médias entre os níveis de acento e entre os estilos, usamos um teste estatístico chamado de tamanho do efeito (*effect size*).

O cálculo do tamanho do efeito mostrou que a duração é o correlato principal do acento lexical independentemente de estilo e que, diferentemente dos trabalhos anteriores, vogais pré-tônicas e pós-tônicas não diferem em duração a não ser na situação de palavra isolada (ainda apenas para os homens), via teste post hoc. Mostramos ainda que o acento lexical explica uma maior percentagem da variância dos parâmetros “duração” e “desvio-padrão” de F0 do que o estilo de elocução. As vogais pós-tônicas têm menos ênfase espectral nos três estilos com relação à tônica. Assim, uma queda em ênfase espectral é sinal de que a vogal precedente é acentuada lexicalmente, ainda que a F0 varie mais em tônicas e pós-tônicas, especialmente nas palavras isoladas. Em mulheres, a variação de F0 é mais importante para explicar acento lexical do que a ênfase espectral.

7 O primeiro teste é mais geral e avalia a significância da diferença de média entre os conjuntos de valores para as duas variáveis independentes, grau de acento e estilo. Quanto ao segundo teste, ele avalia entre quais conjuntos de dados existe a diferença. Detalhes sobre esses tipo de testes no capítulo 6, seção 6.1.

3.2.6 Protocolo Experimental: Estilos de Elocução

Dois estudos experimentais que fizemos permitem levantar uma discussão sobre outros aspectos da pesquisa experimental em prosódia. O primeiro deles abordou a questão de imitação de fala (BARBOSA; MAREÜIL, 2018). É sabido que a imitação da fala retém apenas os aspectos mais salientes do locutor ou a representação fonológica do enunciado (COLE; SHATTUCK-HUFNAGEL, 2011). Sendo assim, pensamos que a imitação do estilo telejornalístico seria marcado por uma tendência para a proeminência inicial, pois essa tem sido observada tanto no estilo telejornalístico francês quanto em PB em palavras com grande carga semântica (e.g., BIlhões de reais), mas não somente nessas, quando se pensa nas frases de abertura das notícias.

No que tange o estilo profissional de locução em geral, resultados de um estudo prévio sobre locução de rádio (CAMPOS, 2012) indicaram que as principais mudanças quando da imitação desse estilo por um profissional do rádio em comparação com sua própria entrevista informal foi um aumento de 12% da mediana de F0 e um aumento da taxa de acentos de *pitch*. Propusemo-nos então a comparar imitações do estilo jornalístico em duas línguas/culturas, o francês da França e o português brasileiro, nas variedades mais disseminadas: a locução em telejornais de Paris e aquela no eixo Rio-São Paulo, respectivamente, por representarem a norma de pronúncia dos respectivos países. Para tanto examinamos quais parâmetros prosódico-acústicos são mo- bilizados para assinalar dois tipos de imitação que lançam mão seja da memória de longo termo, seja daquela de curto termo.

Como hipóteses de trabalho, tendo em vista resultados de estudos prévios (CAMPOS, 2012; CASTRO, 2008), esperamos, na imitação do estilo de telejornal: (1) um aumento na mediana e no desvio-padrão de F0; (2) um aumento na proporção de proeminência inicial; (3) a con-

vergência de alguns parâmetros, especialmente taxa de elocução, com os parâmetros do modelo imitado consecutivamente; (4) uma igualdade de resultados para ambas as línguas.

Para essa pesquisa selecionamos quatro jornalistas em cada país, de Paris e de Campinas, dois homens e duas mulheres em cada caso. O número diz respeito, sobretudo em Paris, à dificuldade de conseguir profissionais disponíveis e com tempo para colaborar na pesquisa. Cada um deles leu um texto de três maneiras em sua língua. O texto foi *La Bise et le soleil* e sua correspondente tradução para o PB, "O vento sul e o sol". As três maneiras foram: (1) de forma neutra - NE; (2) no estilo telejornalístico de cada país, de acordo com a internalização de cada um do que é o estilo telejornalístico - JM; (3) no estilo de uma telejornalista de cada país, em que, após ouvi-la, fazia-se a locução sobre outro assunto tentando imitá-la (imitação consecutiva) - JC. As leituras foram uniformemente divididas em 36 grupos acentuais (AP) em PB e 39 em francês com pelo menos 3 sílabas no grupo acentual em cada língua. Textos e grupos acentuais analisados podem ser vistos abaixo, primeiro em PB e depois em francês. Somente os grupos acentuais entre colchetes foram analisados.

[O vento] sul e o sol [discutiam] qual dos dois era [o mais forte], quando passou [um viajante] [envolto] [num casaco]. [Ao vê-lo], [apostaram] que [aquele] que [primeiro] [conseguisse] [obrigar] [o viajante] [a tirar] [o casaco] [seria] [considerado] [o mais forte]. [O vento] sul [começou] [a soprar] [com muita força], mas quanto [mais soprava], [mais o viajante] [se embrulhava] [no seu casaco], [até que] [o vento] sul [desistiu]. O sol brilhou então [com toda intensidade], e [imediatamente] [o viajante] tirou [o casaco]. [O vento] sul teve assim [de reconhecer] [a superioridade] do sol.

La bise et [le soleil] [se disputaient], chacun [assurant] [qu'il était] [le plus fort], [quand ils ont vu] [un voyageur] [qui s'avançait], [enveloppé] [dans son manteau]. [Ils sont tombés] d'accord [que celui] [qui arriverait] [le premier] [à faire ôter] [son manteau] [au voyageur] serait [regardé] [comme le plus fort]. Alors, la bise s'est mise [à souffler] [de toute sa force] mais [plus elle soufflait], [plus le voyageur] serrait [son manteau] [autour de lui] et [à la fin,] la bise [a renoncé] [à le lui faire ôter]. Alors [le soleil] a commencé] [à briller] et [au bout d'un moment], [le voyageur], [réchauffé], [a ôté] [son manteau]. Ainsi, la bise [a dû reconnaître] [que le soleil] était [le plus fort] des deux.

Dentro de cada grupo acentual realizado pelos quatro jornalistas em cada língua, medimos as variáveis seguintes:

- Proporção de proeminências iniciais em cada leitura. Essa proeminência inicial foi toda palavra com saliência em borda esquerda que não fosse a posição de acento lexical. Por exemplo, em "discutiam", foi contado como proeminência inicial se havia essa saliência na sílaba "dis" em PB e, no caso, do francês nas sílabas *se* ou *dis* do grupo acentual *se disputaient* ;
- Descritores estatísticos da F0: mediana, máximo, amplitude de variação (máximo - mínimo), desvio-padrão bruto e normalizado

pelo valor da mediana, semi-amplitude entre quartis[8] bruta e normalizada pela mediana, com todos os valores em semitons;
- Tempo total de leitura e tempo total de pausa silenciosa (soma dos valores das durações das pausas silenciosas);
- Intensidade relativa calculada pela fórmula da ênfase espectral, conforme visto na seção anterior.

Para a análise estatística, usamos o teste não paramétrico de dois fatores de Scheirer Ray Hare (SHR), equivalente à ANOVA de dois fatores, com os fatores SUJEITO (4 níveis) e ESTILO (3 níveis) e, em seguida, utilizamos o teste *post hoc* não paramétrico de Wilcoxon[9]. Para todos os casos o nível de significância α foi fixado em 1%. A razão do uso do teste não paramétrico é que, em nenhum dos casos, os resíduos passaram no teste de normalidade e a razão de se usar um nível de significância mais baixo é diminuir a chance de erro do tipo I, por termos um número baixo de participantes.

Os resultados significativos para o PB revelaram uma proporção de proeminência inicial de cerca de 57% nos estilos neutro e imitação de cor contra 67% no estilo imitação consecutiva. Para F0, a mediana é maior em 2 semitons no estilo imitação de cor em relação aos demais, tendo uma amplitude de variação também maior de 2 semitons nos estilos de imitação em relação ao neutro, mas apenas nos grupos acentuais contendo proeminência inicial. Para o tempo de leitura, 10% a mais de duração na imitação de cor com relação ao neutro e até 31% a mais na imitação consecutiva.

Já para o francês, os resultados significativos revelaram uma proporção de proeminência inicial de cerca de 50% no estilo neutro contra 65% nos dois estilos de imitação. Quanto à F0, encontramos uma mediana maior de 3 semitons nos estilos de imitação nos homens com

8 Medida não paramétrica do desvio-padrão, definida como a metade da diferença entre os quartis 1 e 3.

9 Esses testes serão apresentados no capítulo 6, seção 6.1.

relação ao neutro e do mesmo montante no estilo de imitação consecutiva em relação ao neutro, mas apenas para uma das jornalistas. A amplitude média de variação de F0 é maior em 3 semitons nos estilos de imitação, mas apenas para a mesma jornalista que teve mediana de F0 maior na imitação consecutiva. Houve 32% a mais de duração em dois jornalistas no estilo de imitação consecutiva, lentificando assim a fala. Esse resultado é inesperado, uma vez que a jornalista cujo modo de falar lhes foi apresentada como modelo para imitar fala muito rapidamente. Nas duas línguas houve 2 a 3 dB a mais de ênfase espectral nas imitações em relação ao neutro, sinalizando maior esforço vocal na imitação independentemente de língua.

Concluímos com esse trabalho que a proeminência inicial é um traço importante do estilo imitado nas duas línguas, sendo sinalizado também por maior valor de mediana de F0 e amplitude de variação, isto é, mais agudo e mais variável. A maior diferença entre as línguas diz respeito ao estilo de imitação consecutiva por conta das particularidades de cada jornalista: a francesa que falou rapidamente e a brasileira que falou lentamente (essa escolha não foi, em princípio proposital, nem muito menos caracteriza a locução de todo jornalista nesses países). As hipóteses (1) e (2) foram confirmadas, mas não a hipótese (3), pois não observamos nenhuma convergência no estilo imitado consecutivamente em francês. Quanto à hipótese (4), as diferenças foram mais relacionadas aos participantes ou gêneros e não tanto às línguas.

3.2.7 Protocolo Experimental: Variação da Taxa de Elocução

Em um dos experimentos que compuseram nosso trabalho sobre o ritmo da fala (BARBOSA, 2006), utilizamos uma passagem do livro "A Menina do Nariz Arrebitado" de Monteiro Lobato, lida por quatro participantes masculinos paulistas em três taxas de elocução para es-

tudar como se modificam as durações de sílabas fonéticas e grupos acentuais no PB sob a demanda de falar mais lentamente ou mais rapidamente.

Para esse corpus, as taxas de elocução foram eliciadas por instruções dadas pelo experimentador. Esse solicitou a cada participante que começasse a ler a passagem com uma taxa de conversação confortável (taxa normal). Em seguida, que lesse o mais lentamente possível (taxa lenta), mas preservando o sentido dos enunciados e assim evitando o estilo de ditado e, finalmente, que lesse o mais rapidamente possível sem cometer lapsos (taxa rápida). É claro que, apesar das instruções, não há impedimento de que taxas estatisticamente iguais fossem produzidas, o que, de fato, ocorreu depois da constatação de diferença com um teste de ANOVA. Para evitar isso, pode-se dar como modelo para escuta prévia um áudio de fala natural ou sintetizada em que haja taxas de elocução estatisticamente distintas para serem reproduzidas. Utilizamos esse procedimento no trabalho de doutorado (BARBOSA, 1994) obtendo cinco taxas de elocução distintas por participante. Não usamos esse procedimento aqui para garantir um certo conforto nas produções. O excerto de Monteiro Lobato segue abaixo.

Em seguida apareceu um papagaio real que tinha fama de orador. Subiu a tribuna de um poleiro de ouro e fez um belo discurso a respeito da arte de falar. Nesse discurso provou que os homens tinham aprendido a falar com os papagaios, e não os papagaios com os homens, como diz a ciência destes. Uma chuva de palmas acolheu suas palavras.

O mesmo não aconteceu, porém, com a poetisa Lagartixa, que principiou a recitar uma longa poesia e engasgou no meio, acabando o recitativo em choro e faniquito. Para destruir essa má impressão vieram três vagalumes mágicos que fizeram várias sortes, sendo muito apreciada a sorte de comer fogo.

Foram segmentadas todas as sílabas fonéticas (unidades VV) por detecção automática de início de vogal com o script *Beat Extrac- tor* (BARBOSA, 2006) seguido de correção manual e, em seguida, com o script *SG Detector*, foram obtidos os picos de duração normalizada considerando todo pico local como fronteira à direita de grupo acentual (para aprender como obter duração normalizada veja o capítulo seguinte). Com isso pudemos obter, por participante e por taxa, os valores das durações de unidades VV e dos grupos acentuais com os quais calculamos média e desvio-padrão, mostrados nas Tabelas 3.3 e 3.4.

A Tabela 3.3 revela uma extensão de médias de duração das uni- dades VV de 152 ms a 283 ms (correspondentes respectivamente a 6,6 e 3,5 unidades VV/s de taxa de elocução). O intervalo entre os per- centis 5% a 95% é de 95 ms a 570 ms para AP, o locutor mais lento, e de 87 ms a 292 ms para FA, o locutor mais rápido, revelando valores mínimos semelhantes e valores máximos de praticamente o dobro para AP em relação a FA. Isso representa uma grande variabilidade para o alongamento silábico entre diferentes locutores do PB e diz respeito a variações individuais.

Tabela 3.3 – Valores médios (e desvios-padrão) em milissegundos da duração das unidades VV no *corpus Lobato* em quatro participantes paulistas masculinos, para três taxas de elocução.

	Participante			
Taxa de Elocução	AP	AC	DP	FA
Lenta	283 (185)	243 (203)	190 (154)	189 (111)
Normal	235 (169)	223 (166)	188 (165)	165 (88)
Rápida	201 (144)	194 (138)	165 (119)	152 (74)

Tabela 3.4 – Valores médios (e desvios-padrão) em milissegundos da duração dos grupos acentuais no *corpus Lobato* em quatro participantes paulistas masculinos, para três taxas de elocução.

	Participante			
Taxa de Elocução	AP	AC	DP	FA
Lenta	1504 (694)	1518 (562)	1107 (527)	1154 (389)
Normal	1370 (496)	1353 (521)	1233 (588)	931 (363)
Rápida	1180 (543)	1348 (525)	1077 (546)	889 (395)

A Tabela 3.4 revela uma extensão de médias de duração de grupo acentual de cerca de 1 s a 1,5 s para diferentes taxas, revelando uma certa tendência em realizar um acento frasal a uma cadência semelhante. Grosso modo, essa ordem de grandeza é da leitura de um hemistíquio de um verso alexandrino, verso de 12 sílabas. Pode-se calcular das duas tabelas, de fato, que o valor médio do número de unidades VV em cada grupo acentual para os quatro locutores é de 6,5 unidades VV, muito próximo ao número de sílabas do hemistíquio, que é de 6. Esses aspectos revelam o caráter universal da sucessão de proeminências na fala.

3.2.8 Protocolo Experimental: Leitura e Narrativa

Há alguns anos gravamos o corpus Belém com o fim de cotejar a prosódia da leitura e da narrativa consecutiva de um texto sobre a origem dos pastéis de nata de Belém[10]. O trabalho que detalhamos aqui, ressaltando o seu protocolo experimental, é o de Barbosa e Silva (2012), trabalho que procurou responder duas perguntas: o que faz com que dois enunciados sejam percebidos como prosodicamente distintos e que parâmetros contribuem para que dois enunciados difiram no

10 O texto foi proposto pelo INESC de Lisboa quando de uma colaboração empreendida em 2009. O texto original, em português europeu, narra a história dos pastéis de Belém. A mesma equipe solicitou na época a um brasileiro que adaptasse o texto para o português brasileiro e foi a versão adaptada que usamos nos experimentos com o PB.

modo de falar?

Considerando que as duas funções básicas da prosódia são a marca- ção de fronteiras durante a fala, bem como o assinalamento de unidades proeminentes, nos propusemos a examinar as diferenças produzidas e percebidas relacionadas a essas duas funções. No trabalho de Barbosa e Silva (2012), examinamos unicamente parâmetros temporais, muito embora um deles diga respeito à entoação stricto sensu, pois medimos a taxa de acentos de *pitch* por segundo.

Para avaliar diferenças rítmicas, comparamos a fala lida e a narrada, por suas características distintas. A escolha da fala narrada se deve ao fato de ela apresentar elementos comuns com a conversa espontânea, um estilo de elocução muito frequente nas instâncias comunicativas. Alguns desses elementos são as hesitações causadas pelo macro e microplanejamento do discurso, que dizem respeito respectivamente ao conteúdo e organização sintática do que se vai dizer (LEVELT, 1989). Embora haja hesitações na fala lida, essas são bem menos frequentes do que na narracão de uma história lida, devido à maior demanda para a memória de trabalho[11].

O corpus usado foi formado pela leitura e pela narração consecutiva do texto dos pastéis de Belém adaptado para o PB. O texto tem cerca de 1600 palavras e foi, na época, lido por duas mulheres e um homem, sendo os três estudantes do curso de Linguística com cerca de 30 a 45 anos. Imediatamente após sua leitura, a história foi narrada pelos três participantes com suas próprias palavras.

O corpus foi primeiramente segmentado em excertos de 9 a 18 se- gundos por conta dos testes de percepção que foram feitos e para a avaliação do vínculo entre produção e percepção da prosódia. A esco- lha dessa extensão é fundamentada em testes anteriores com trechos mais curtos e na literatura, pois é necessário um trecho mais longo para que o ouvinte infira o modo de falar de uma pessoa. Cada

11 Trata-se do mecanismo cognitivo para reter informações enquanto fazemos uma tarefa. Ver (COWAN, 1997) para detalhes.

excerto foi segmentado em sílabas fonéticas cujas fronteiras foram definidas por inícios de vogais, como repetidamente explicado na literatura sobre prosódia (LEHISTE, 1970; CLASSE, 1939; BARBOSA, 2006, 2019). Essa segmentação foi feita semi-automaticamente em camada de anotação do software Praat em duas etapas. Para a primeira etapa usamos o script Beat Extractor para a detecção automática dos inícios de vo- gal e, na segunda etapa, corrigimos os erros de detecção manualmente, introduzindo manualmente a transcrição fonética para possibilitar a normalização da duração silábica.

A partir da segmentação realizada, um script feito para esse trabalho calculou 10 parâmetros prosódico-acústicos, entre eles: a taxa de elocução em unidades VV por segundo, os três primeiros descritores estatísticos (média, desvio-padrão e assimetria da distribuição) e a taxa por segundo dos picos de duração normalizada ao longo dos excertos, a taxa de picos da curva de Fo suavizada por segundo[12], os coeficientes de variação (desvio-padrão dividido pela média) da duração do grupo acentual, do número de unidade VV por grupo acentual e da duração da unidade VV e, por fim, a taxa de unidades VV não salientes (isto é, não são picos locais de duração normalizada).

Os excertos analisados foram associados em pares de áudio separados pelo áudio de um tom puro (de frequência de 1000 Hz) para a confecção de um teste de discriminação no Praat do qual participaram 10 estudantes de Linguística. A finalidade do tom puro é apenas assinalar a fronteira entre os dois áudios a serem avaliados quanto à diferença rítmica. Em seu conjunto, os excertos foram formados pela narrativa de uma das participantes e pela leitura dos outros. Foram utilizados 44 pares de excertos combinados aleatoriamente para o teste, que durou até cerca de 25 minutos para ser completado.

Cada excerto foi também deslexicalizado (vide seção 3.2.2 sobre

12 Essa taxa foi obtida dividindo o número de picos de F0 suavizado num trecho por sua duração. Esses picos são máximos locais da curva suavizada de F0 com a frequência de corte de 5 Hz. Em seguida interpola-se a curva e conta-se automaticamente o número de picos no excerto.

deslexicalização) usando o algoritmo desenvolvido por Vainio et al. (2009). Assim, cada par de excertos foi combinado na ordem AB e BA tanto na versão original quanto deslexicalizada (cada participante ouviu os áudios nas duas ordens, o que foi necessário para avaliar a consistência das respostas). A inversão da ordem permite avaliar o grau de consistência das respostas dos ouvintes, pois, sendo o mesmo par avaliado, espera-se que a resposta quanto ao grau de discriminação seja a mesma. Cada ouvinte avaliou primeiramente a versão deslexicalizada e depois, em ordem aleatória, a versão original a partir da seguinte instrução: "Avalie quão diferente é o modo de falar dos trechos de áudio no par separados por um tom numa escala de 1 a 5, sendo 1 'mesmo modo de falar' e 5 'modos de falar completamente diferentes', usando qualquer nível entre os dois a partir da sua percepção." O tom usado foi um sinal senoidal com amplitude descendente de 1000 Hz de cerca de 30 ms. As respostas de 1 a 5 foram posteriormente recodificadas linearmente de -1 a 1, sendo 0 considerado uma resposta neutra.

Quanto ao desempenho dos ouvintes, duas hipóteses foram consideradas: (1) que a consistência nas respostas seria maior na fala deslexicalizada e que haveria melhor desempenho para diferenças entre estilos diferentes, tendo em vista maior atenção na prosódia por conta da deslexicalização; (2) que a diferença de valores médios de pelo menos um parâmetro acústico seria capaz de prever as respostas dos ouvintes, por conta do esperado vínculo entre produção e percepção da prosódia.

Quanto à consistência das respostas, a média das respostas no mesmo par de excertos em diferentes ordens foi menor e menos va- riável para a versão original (diferença entre médias com $t_{gl=398} = 4,2$, $p < 10^{-4}$), contradizendo a primeira hipótese. Isto é, a informação lexi-cal ajudou na manutenção da mesma resposta para o par apresentado em ordem distinta. Quanto à resposta ao grau de diferença no modo de falar, não há diferença alguma quer se use a versão deslexicalizada quer se use a versão original.

Para avaliar a segunda hipótese, tomamos apenas pares de excertos com respostas com consistência inferior a 0,5 (a consistência foi definida como a diferença das respostas nas duas ordens de apresentação que teoricamente deveria ser 0) e com desvio-padrão entre respostas de diferentes ouvintes também menor do que 0,5, com o fim de utilizar apenas os 15 pares com respostas homogêneas e consistentes que permitissem uma melhor avaliação de sua relação com os parâmetros prosódico-acústicos. Usamos testes de regressão linear múltipla[13] para prever a resposta média dos ouvintes para cada par a partir da diferença dos valores médios dos 10 parâmetros acústicos extraídos dos excertos em cada par.

O melhor modelo explicou 71% da variância das respostas dos ouvintes (resp)[14]: $resp = -1,5 + 10,4.pr + 2,65.sr - 10,75.pr \times sr$ em que *sr* é a taxa de elocução e *pr* é a taxa de picos de duração normalizada, isto é, respectivamente um parâmetro relacionado à sucessão de sílabas fonéticas e outro relacionado à sucessão de sílabas fonéticas proeminentes. Esse resultado confirma a segunda hipótese e aponta para o papel crucial da taxa de elocução e da taxa de produção de sílabas proeminentes para a percepção do modo de falar.

3.2.9 Testes de Percepção da Prosódia

Conforme acabamos de ver, os testes de percepção são muito úteis para avaliar a relação entre produção e percepção da prosódia. Além de testes de discriminação que discutiremos mais detalhadamente aqui, há também os testes de classificação. Se o teste de discriminação requer, cognitivamente, uma avaliação de elementos presentes na memória de trabalho, o teste de classificação faz uso da memória de longo

13 Este teste avalia a correlação entre um conjunto de variáveis preditoras e uma variável predita.

14 Valor de p de pelo menos 0, 009 para todos os coeficientes da regressão ($F_{3,11} = 12,4$, $p < 0,0008$).

termo, pois requer a associação de um estímulo a uma classe que nos é conhecida em menor ou maior grau, em função de nossa experiência comunicativa.

Perceber elementos na fala envolve a capacidade de avaliar semelhanças e diferenças entre estímulos, evocando eventualmente duas categorias de memória. A percepção pode assim se dar de duas maneiras: (1) pela comparação da informação acústica armazenada temporariamente na memória de trabalho para os dois estímulos[15] ou (2) pela comparação da informação acústica do estímulo que se ouve com elementos de alguma categoria construída e armazenada na memória de longo termo para o estímulo que se ouviu anteriormente. Por conta disso, é necessário falar de categorias também na investigação em prosódia experimental.

Pelo relato testemunhal de Repp (1984), a pesquisa sobre percepção categórica na fala começou nos Haskins Laboratories depois da construção do primeiro sintetizador de fala, o *Pattern Playback* com o trabalho de Liberman et al. (1957), que criaram um contínuo acústico de sílabas sintéticas do tipo /Ce/ em que C = /b d g/. O desempenho de teste de discriminação do tipo ABX (ouvem-se os três estímulos e é preciso responder se X é A ou B) revelou que os ouvintes tinham mais facili- dade em responder quando os estímulos proviam claramente de duas categorias distintas do que quando provinham da mesma categoria. Por exemplo, se dois estímulos distintos eram exemplos de /be/, foi mais difícil discriminá-los do que se um era exemplo de /be/ e outro de /de/. Essas categorias de um dos três fonemas foram avaliadas antes com um teste de classificação. Sendo assim, propunha-se que o desempenho dos ouvintes no teste de classificação poderia prever seu desempenho no teste de discriminação. Essa relação estreita entre os desempenhos nas duas tarefas sempre foi considerado

15 A memória de curto termo retém informação acústica por cerca de no máximo 500 ms, mas essa informação pode ser categorizada de alguma forma e permane- cer na memória de trabalho até seu limite temporal, que é de cerca de 20 segundos (COWAN, 1997).

como necessária para se verificar se houve ou não percepção categórica. No entanto, essa necessidade foi contestada mais de uma vez em trabalhos como os de Pollack e Pisoni (1971), Schouten, Gerrits e Hessen (2003), Gerrits e Schouten (2004), que mostraram que além de não ser estreita a relação dos desempenhos nos dois testes, o desempenho do participante depende do tipo de teste, sendo ABX apenas um deles. Para um estudo de outros tipos de teste, que não consideraremos neste livro, ver também a tese de Gerrits (2001) e as excelentes recomendações e apanhado geral dos testes mais úteis no relatório de McGuire (2010).

3.2.9.1 Testes de Discriminação

O teste de discriminação pode ser feito em diversos paradigmas (ABX, AX, AXB, 2IFC, 4IAX and 4I-oddity), mas todos envolvem responder a uma pergunta sobre a similaridade ou dissimilaridade entre estímulos. O mais usado na área de prosódia é o teste do tipo AX, em que se pergunta se o segundo estímulo (X) é igual ou distinto do primeiro (A). Vários pares de estímulos como esses são habitualmente apresentados para os ouvintes e, como vimos, é preciso montar o experimento numa plataforma que possibilite a aleatorização dos estímulos, o controle do tempo entre a resposta dada e o próximo estímulo, o tempo para dar a resposta desde a apresentação do estímulo (tempo de reação), a inclusão de estímulos distratores que dificultem ao ouvinte inferir os objetivos do experimento e ter um comportamento enviesado, entre eles número de falsos alarmes[16] em demasia, entre outras coisas. Várias plataformas estão disponíveis gratuitamente, entre elas o Praat.

Para testes de discriminação de trechos de fala para estudos

16 Um falso alarme é identificar um fenômeno ou estímulo que não pertence a uma categoria como dessa categoria. Assim, por exemplo, se é solicitado a identificar fronteiras prosódicas num enunciado, o acerto é quando uma fronteira de fato é identificada como tal e um falso alarme uma posição em que não tem fronteira identificada como sendo de fronteira.

prosódicos recomenda-se que a extensão desses esteja entre 10 e 20 segundos, pois valores inferiores a 10 segundos não permitem adequada avaliação de um modo de falar e valores superiores a 20 segundos estão em geral além do limiar temporal da memória de trabalho para dados acústicos.

Além do exemplo da seção 3.2.8, comparamos trechos de fala de estilos jornalístico e político em quatro línguas (BARBOSA; MADUREIRA; MAREÜIL, 2017): português brasileiro e europeu, francês da França e alemão da Alemanha, além de leitura e narrativas de não profissionais. Como os ouvintes foram brasileiros com português nativo, foi necessário deslexicalizar os trechos de fala política e jornalística, para que a discriminação não se desse pelo reconhecimento de quem falava. Esse trabalho revelou uma discriminação entre os estilos profissionais, sendo o discurso político o mais facilmente reconhecido nas quatro lín- guas. Além do teste de discriminação, um dos experimentos realizados envolveu um teste de classificação.

3.2.9.2 Testes de Classificação

Os testes de classificação ou identificação possibilitam saber se o ouvinte é capaz de classificar um estímulo dentro de um conjunto fechado ou aberto de possibilidades. Se o conjunto for fechado, o teste se denomina teste de classificação de escolha forçada. Por exemplo, no caso de identificação de estilos de elocução entre três possibilidades: sermão religioso, discurso político ou fala telejornalística, o ouvinte é convidado a ouvir um estímulo e escolher imediatamente depois entre uma dessas três possibilidades.

Se uma das opções de resposta permite que o ouvinte diga que não é nenhum dos estilos propostos, tem-se um teste de classificação de escolha não forçada que permitiria avaliar a ambiguidade de determinados estímulos. A esse respeito, vale a pena incluir no teste de

classificação uma avaliação do quão típico representante da categoria escolhida é aquele estímulo, uma avaliação denominada em inglês *goodness of fit* (qualidade de ajuste). O resultado do teste permite ao experimentador uma reavaliação dos estímulos considerados maus representantes de sua classe.

Algo semelhante ao julgamento de qualidade de ajuste é a avaliação de alguma grandeza perceptiva numa determinada escala que avalie aspectos como "quão enfático soa a palavra x" numa escala de 1 a 5, "quão agudo soa a palavra x" numa escala gradativa de nada agudo a extremamente agudo, "quão agradável soa este enunciado" numa escala de 5 pontos como "nada agradável", "pouco desagradável", "nem agradável nem desagradável", "agradável" e "muito agradável". Há vários modos de pedir a um ouvinte para gradar uma qualidade para fins de investigação prosódica com algumas propostas que fazem um apanhado de escalas perceptivas no trabalho de Rietveld e Chen (2006, p. 286- 302). A importância de avaliar os resultados desse tipo de avaliação é a determinação de quais parâmetros acústicos expressariam uma determinada qualidade perceptiva. Se qualidades como "agudo" trazem imediatamente à lembrança o valor médio da Fo; outras, como "enfático", sugerem maior intensidade, Fo e duração, outras ainda, como "agradável", não são clara e decisivamente associadas a um parâmetro prosódico-acústico.

Outro aspecto fundamental para tirar o máximo proveito dos resultados de um teste de percepção é a análise da resposta dos participantes, avaliando sua sensibilidade ao teste, bem como a coerência entre suas respostas. A sensibilidade a um teste leva em conta não apenas a taxa de respostas de um determinado tipo, como também o viés de resposta de um participante. Suponhamos que se queira verificar se um trecho de fala telejornalística é, de fato, percebido como tal e se faça um teste de classificação. Suponha-se ainda que, entre os estímulos, haja cerca de 60% de estímulos de fala de um estilo de elocução bem distinto como distratores. Se um dos participantes tiver

respondido a todos os estímulos, nos dois estilos, como sendo todos de fala telejornalística, esse participante teria 100% de "acerto" para o estilo telejornalístico, embora com 60% de falsos alarmes, pois teria dito que todos os distratores são do estilo telejornalístico. Esse participante teria introduzido um viés, não tendo sensibilidade ao teste. Para corrigir esse tipo de resultado, é necessário considerar acertos e falsos alarmes, o que é proposto na Teoria da Detecção (GREEN; SWETS, 1966; MACMILLAN; CREELMAN, 2004) com o conceito de d' (d linha). Essa grandeza é definida pela equação 3.2 e mede uma resposta "líquida", isto é, que considera a diferença em unidades de *z-score*[17] entre proporções de acerto e de falso alarme.

$$d' = z(p_{acertos}) - z(p_{\text{falsos alarmes}}) \tag{3.2}$$

Sendo assim, um participante que respondesse a um estímulo com proporções idênticas de acerto e falso alarme teria um d' nulo, enquanto nosso participante hipotético, que tem 60% de falsos alarmes e 40% de acertos teria um d' negativo, mais precisamente de -0, 5. Valores de d' que considerem uma sensibilidade razoável a um estímulo devem ser pelo menos maiores ou iguais a 1. Participantes com valores de d' nulos e negativos devem, em princípio, terem seus resultados desconsiderados num determinado experimento, pois não foram sensíveis ao teste.

Outro aspecto a se considerar quanto aos resultados de um teste de percepção é a coerência de respostas dos participantes, se aplicável. É evidente que, num teste de identificação de fronteiras prosódicas, por exemplo, determinadas fronteiras fracas não são percebidas por todos os participantes e isso é um fato da percepção, pois a não coincidência das respostas não é um problema, é uma informação importan-

17 Medida normalizada de um valor na distribuição gaussiana, definido com arazão entre a diferença de um valor e a média da distribuição pelo desvio-padrão.

te sobre a saliência dessa fronteira.

Por outro lado, nos casos em que se quer determinar as características prosódico-acústicas de um estilo de elocução, de uma atitude ou de uma emoção, por exemplo, é necessário validar perceptivamente esses estímulos. A validação consiste em saber se veiculam, de fato, para os ouvintes aquele estilo, aquela atitude ou aquela emoção. Para tanto, testes de classificação devem ser feitos e, em seguida, deve-se medir a consistência dos participantes em suas respostas, a fim de saber se respondem não aleatoriamente e da mesma forma a um mesmo estímulo.

O teste estatístico que mede essa coerência é o teste de coerência entre juízes, proposto por Cohen para dois juízes e por Fleiss para várias juízes, como extensão do teste de Cohen. Recomendamos ao leitor que se inteire em livros como o de Rietveld e Hout (1993) sobre esse tipo de teste.

3.3 Prelúdio para o Próximo Capítulo

Tendo visto os aspectos metodológicos mais importantes da área de pesquisa em prosódia experimental, vamos apresentar as técnicas de medida de parâmetros prosódico-acústicos que permitirão ao leitor fazer seus próprios experimentos. Os aspectos vistos aqui serão retomados adiante, com estudos de caso e apresentação sucinta de técnicas para análise estatística inferencial.

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.

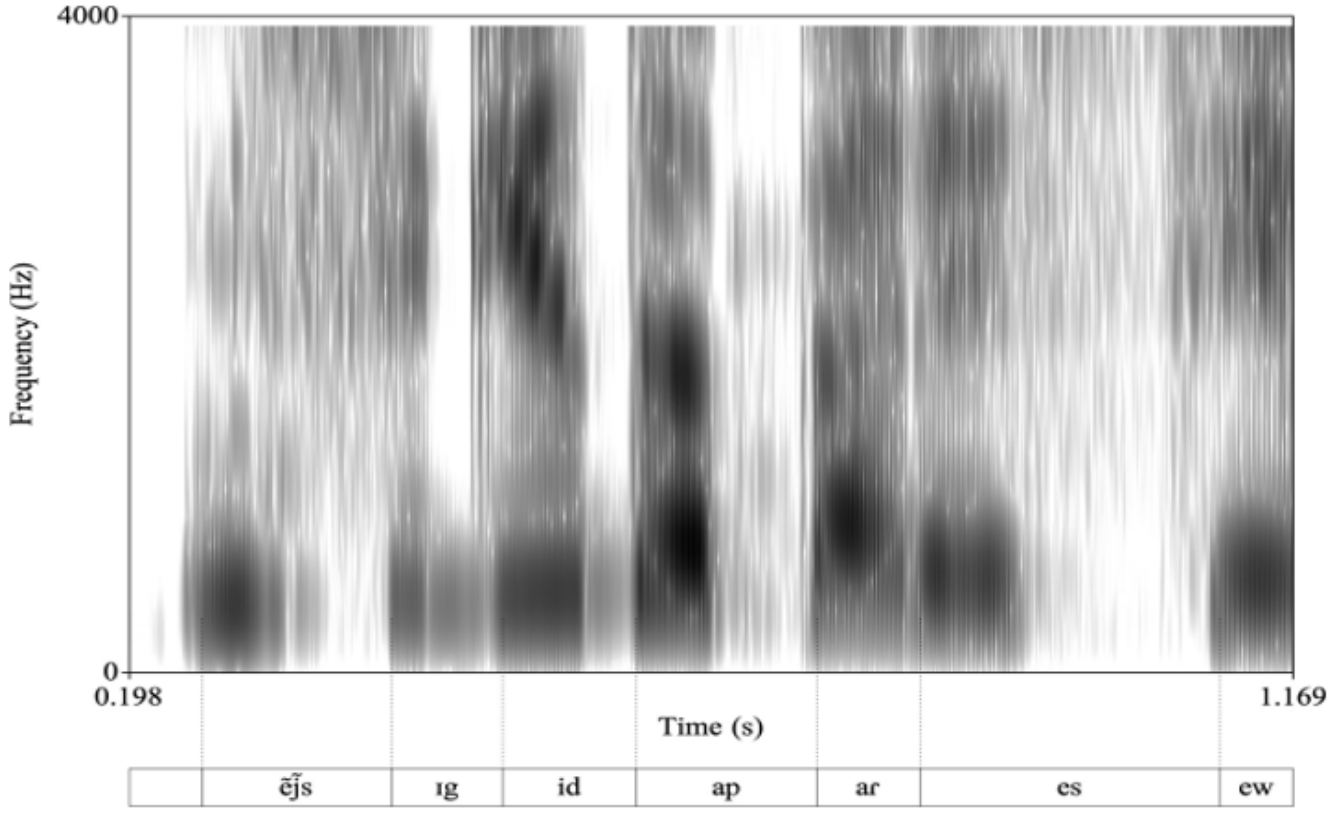

Figura 4.1 – Espectrograma de banda larga e camada de anotação para o trecho “Em seguida apareceu” do enunciado “Em seguida apareceu um papagaio real que tinha fama de orador”. Ver texto para saber como reproduzir a segmentação e etiquetagem.

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.

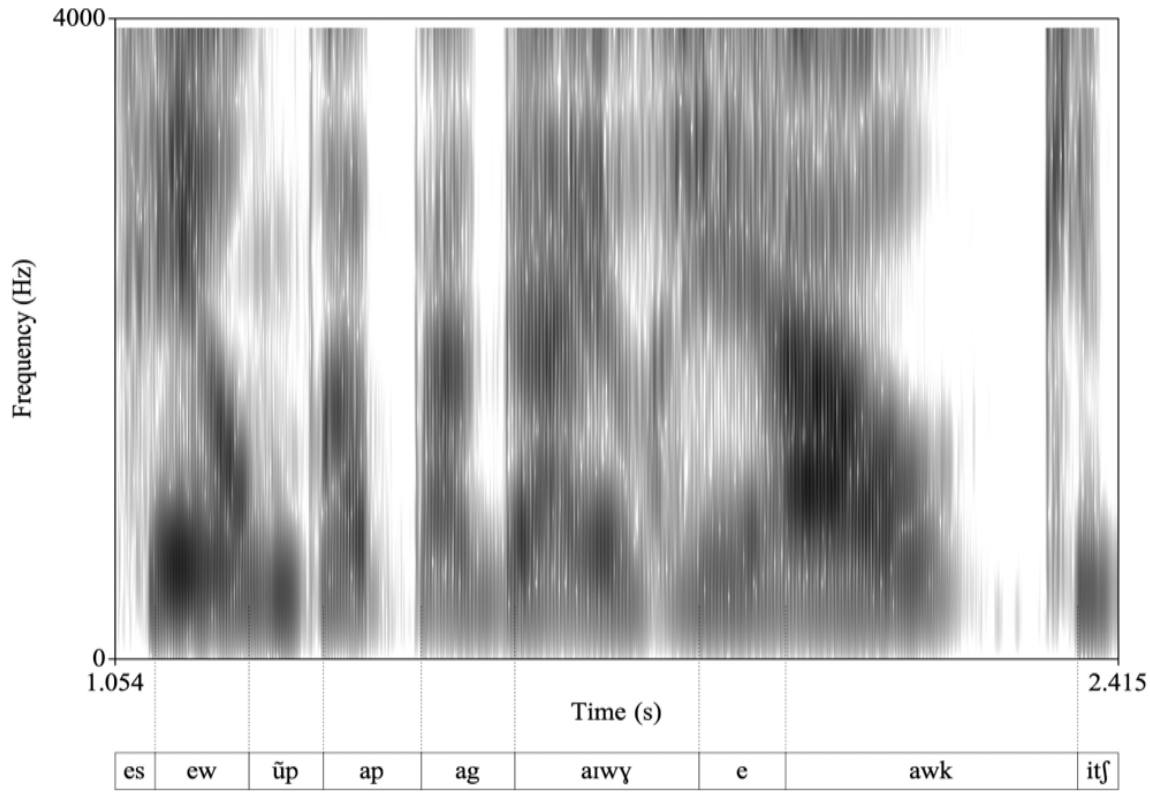

Figura 4.2 – Espectrograma de banda larga e camada de anotação para o trecho "-ceu um papagaio real" do enunciado "Em seguida apareceu um papagaio real que tinha fama de orador". Ver texto para saber como reproduzir a segmentação e etiquetagem.

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.

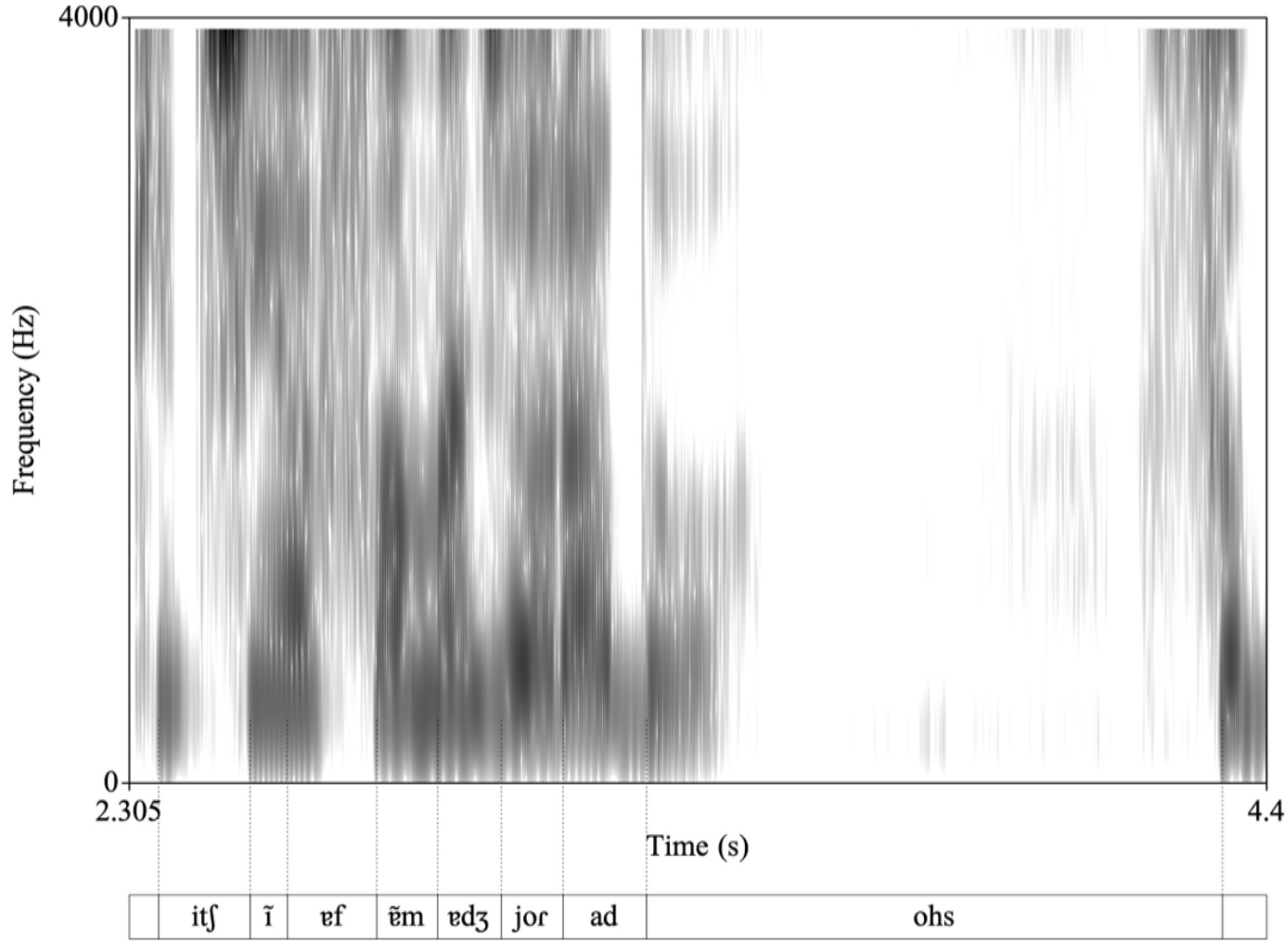

Figura 4.3 – Espectrograma de banda larga e camada de anotação para o trecho “que tinha fama de orador” do enunciado “Em seguida apareceu um papagaio real que tinha fama de orador”. Ver texto para saber como reproduzir a segmentação e etiquetagem.

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.

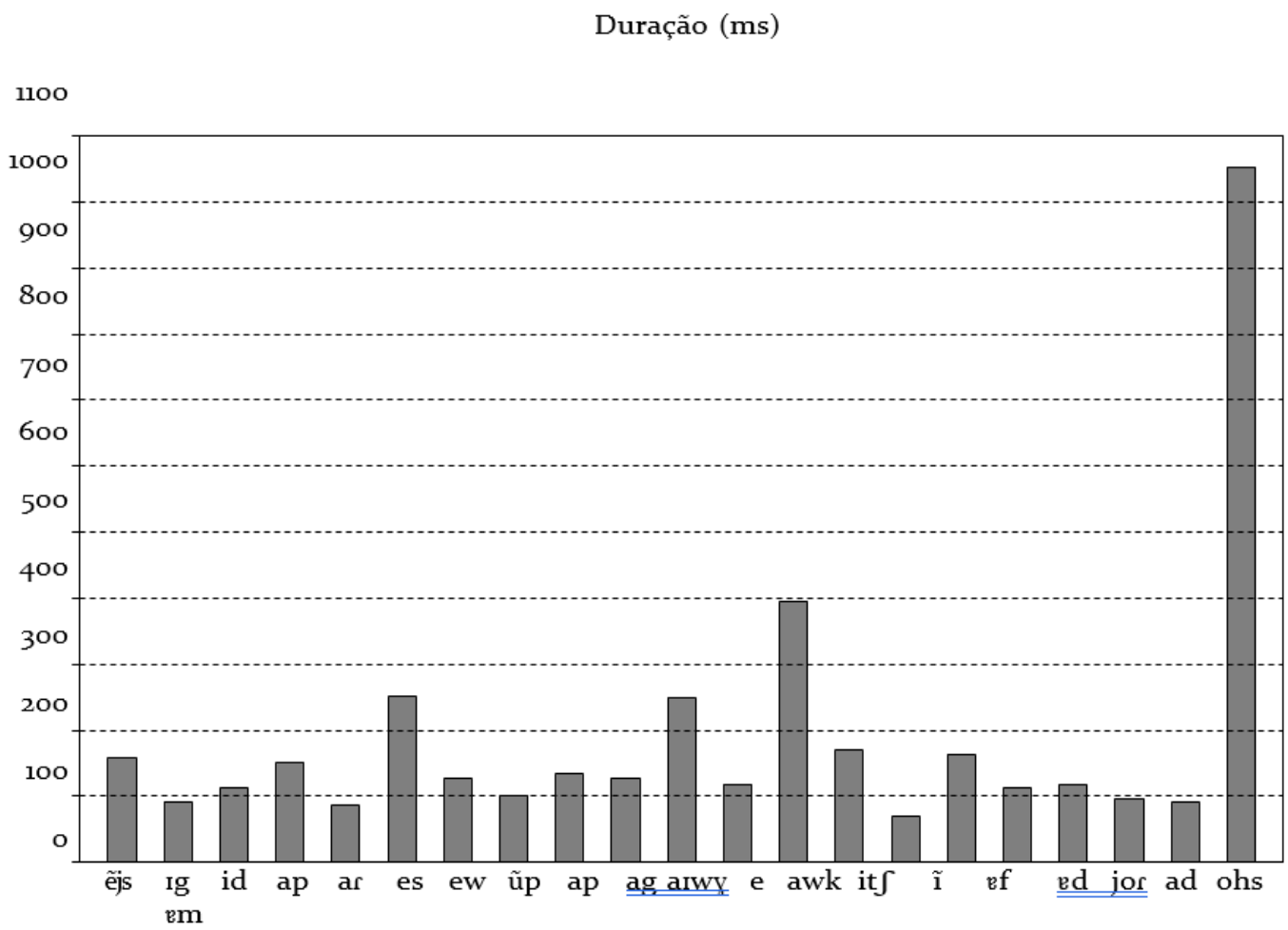

Figura 4.4 – Durações brutas, em milissegundos, do enunciado “Em seguida apareceu um papagaio real que tinha fama de orador” de locutora carioca.

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.

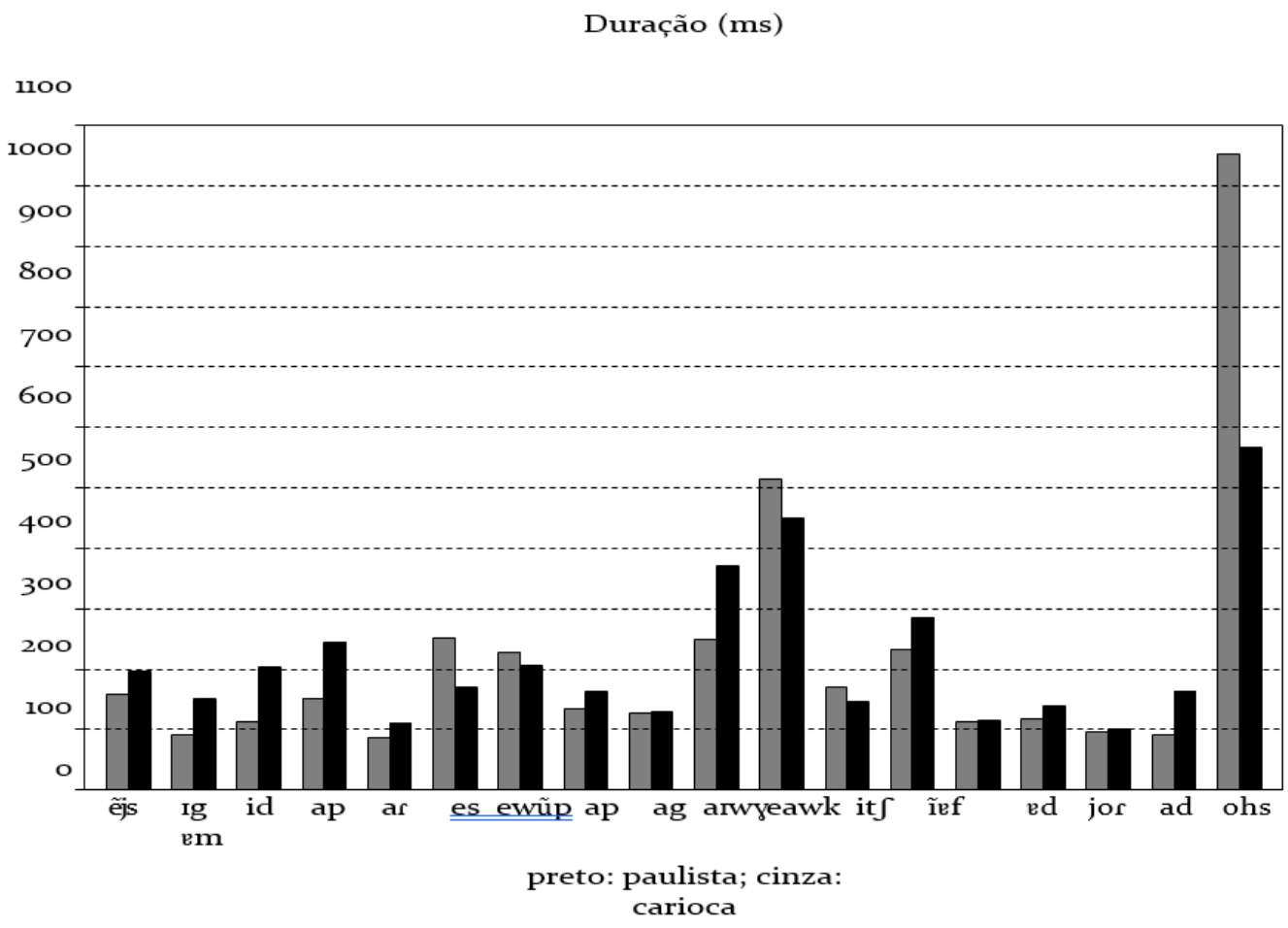

Figura 4.5 – Durações brutas, em milissegundos, do enunciado "Em seguida apareceu um papagaio real que tinha fama de orador" por locutoras carioca (cinza) e paulista (preta).

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.

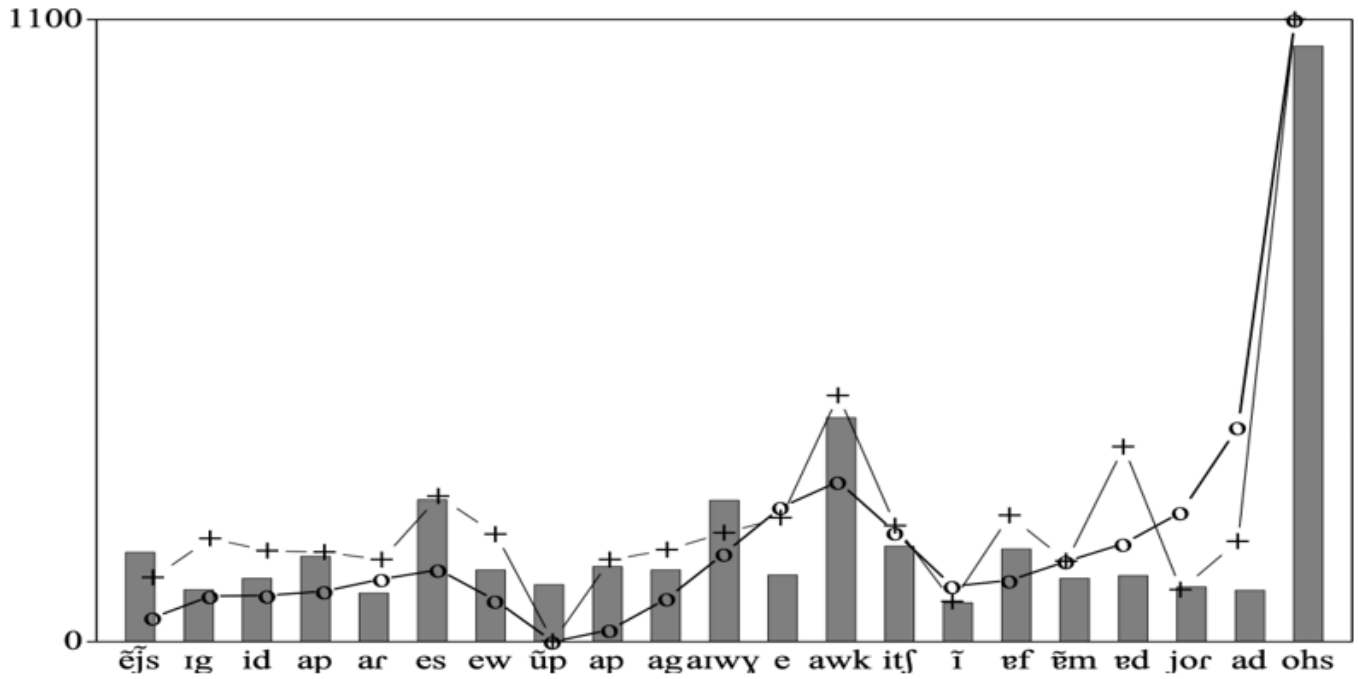

Figura 4.6 – Durações brutas em milissegundos (barra cinza), por *z- score* (pontos + conectados) e por *z- score* suavizado (pontos *o* conectados) do enunciado “Em seguida apareceu um papagaio real que tinha fama de orador” por locutora carioca.

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.

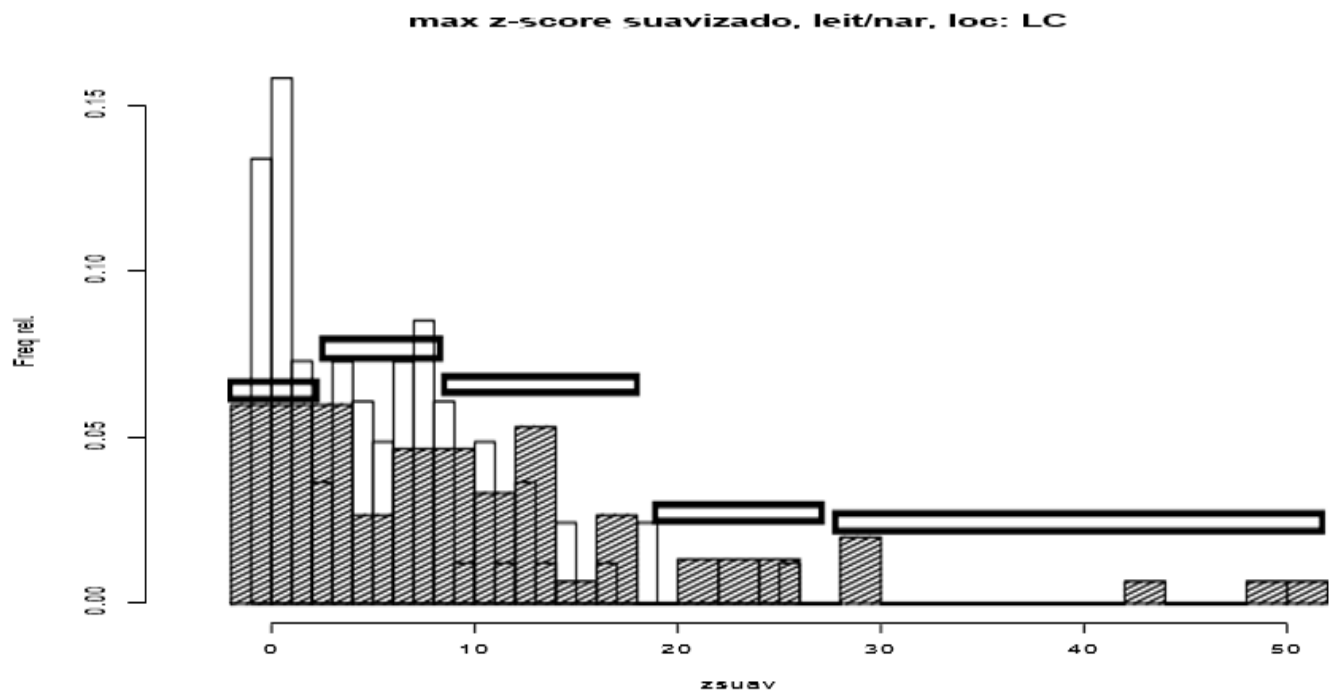

Figura 4.7 – Histogramas superpostos dos picos de z-score suavizados de leitura (barras claras) e narração (barras hachuradas) da locutora paulista LC.

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.

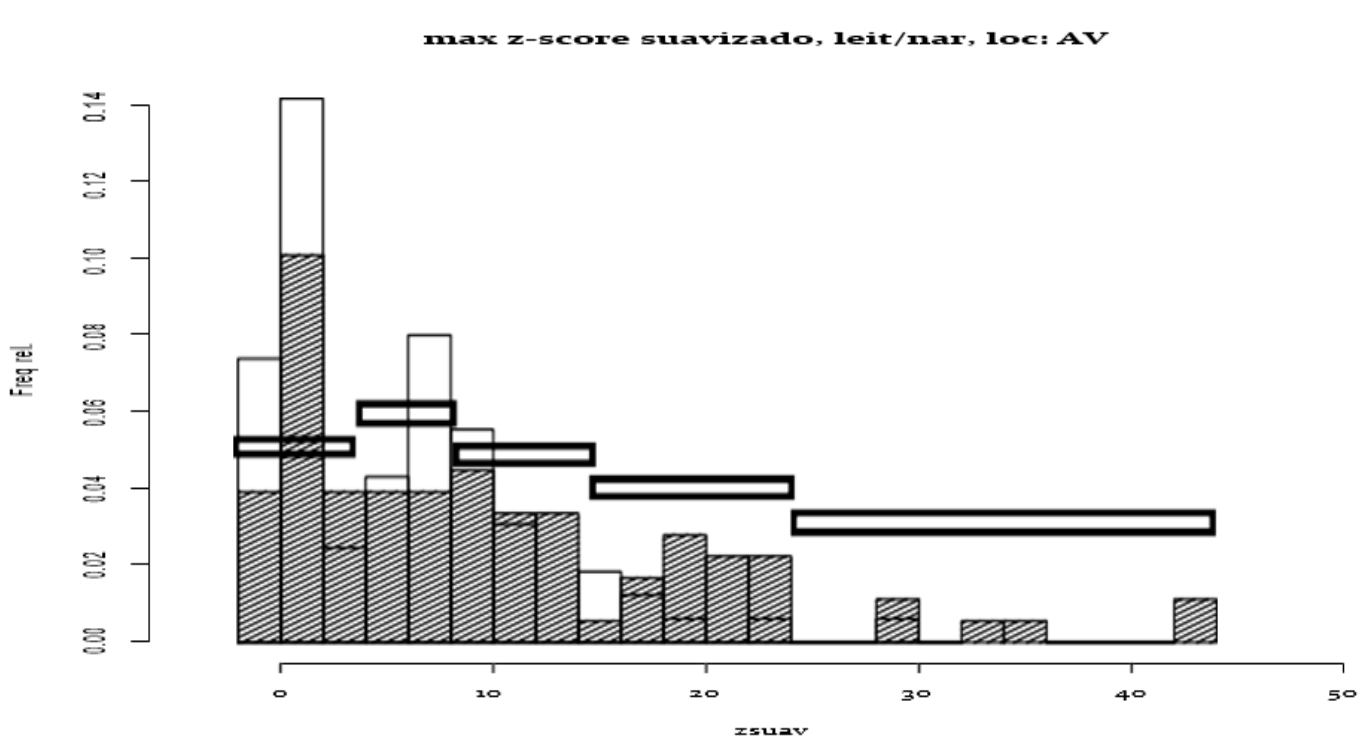

Figura 4.8 – Histogramas superpostos dos picos de *z-score* suavizados de leitura (barras claras) e narração (barras hachuradas) da locutora paulista AV.

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.

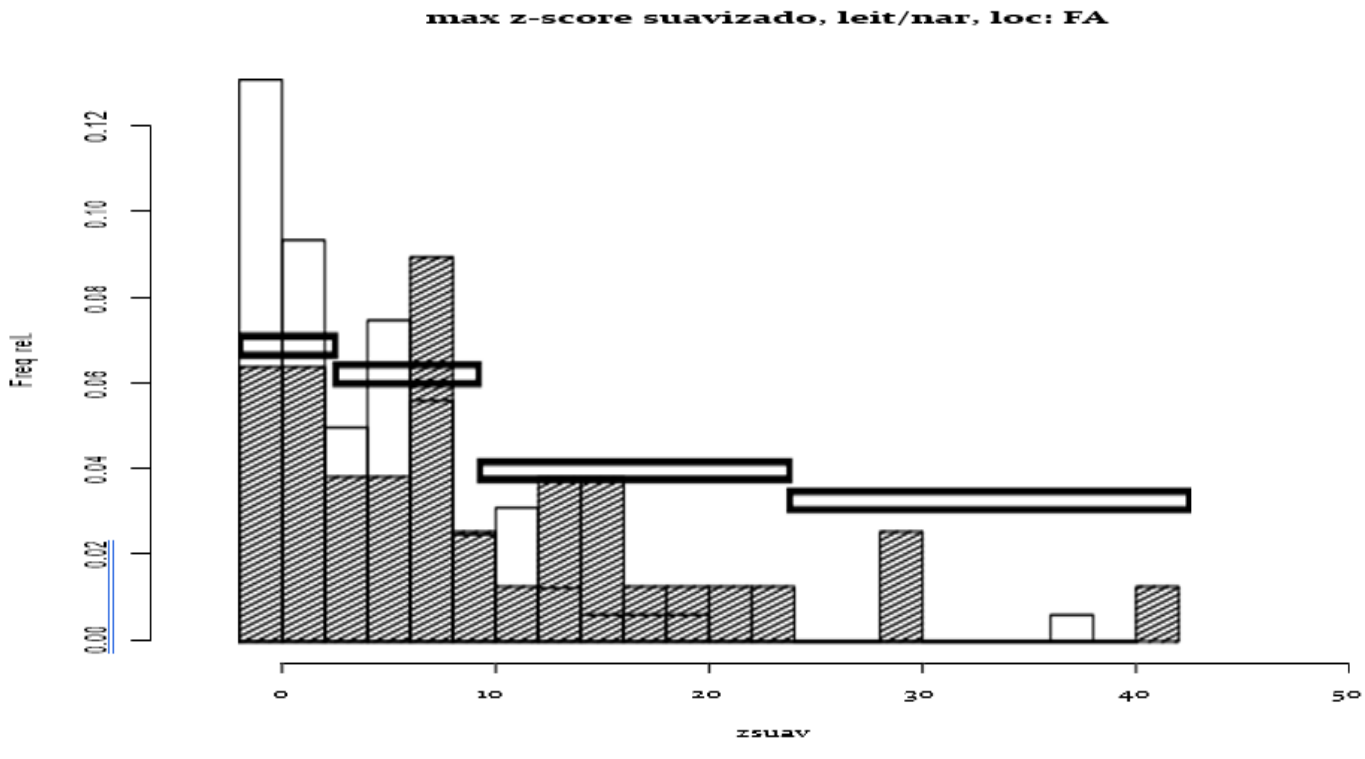

Figura 4.9 – Histogramas superpostos dos picos de *z-score* suavizados de leitura (barras claras) e narração (barras hachuradas) do locutor paulista FA.

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.

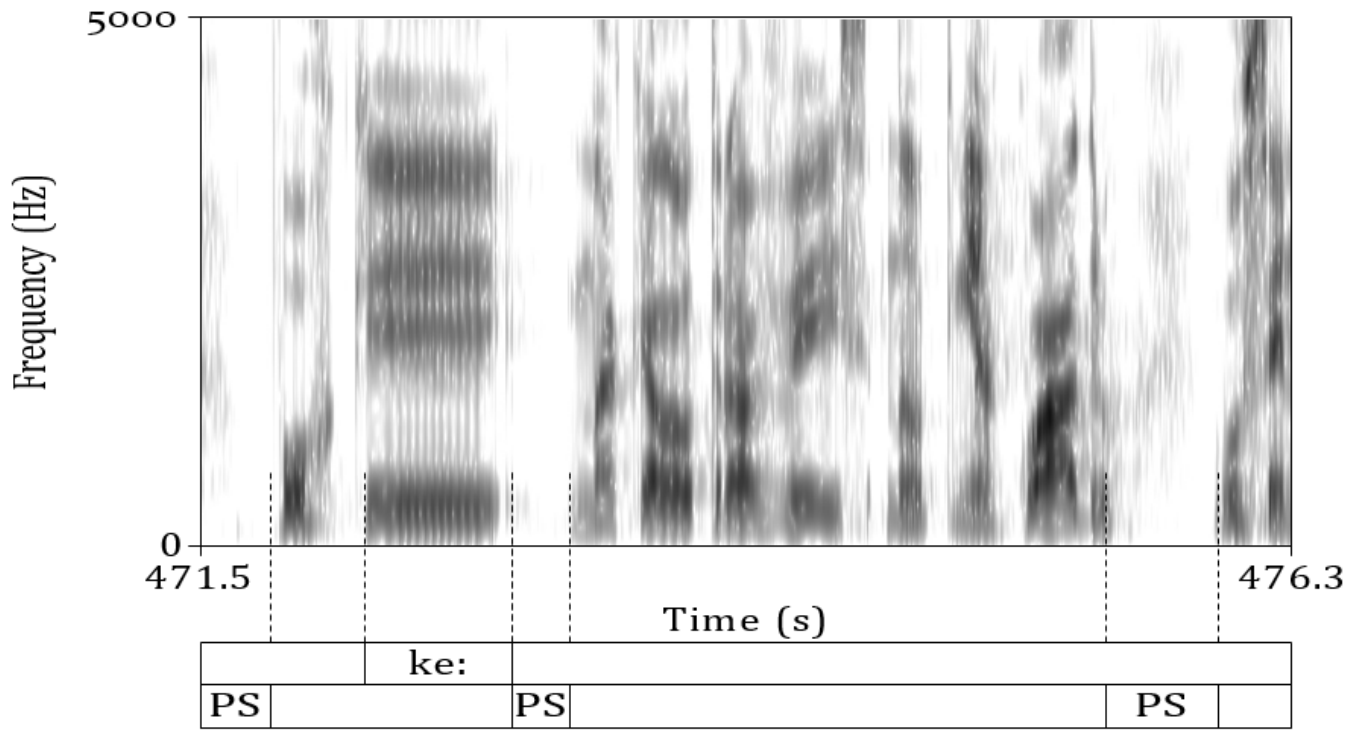

Figura 4.10 – Espectrograma de banda larga e segmentação de pausas preenchida (acima, indicando o segmento produzido) e silenciosa (abaixo, identificada por PS).

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.

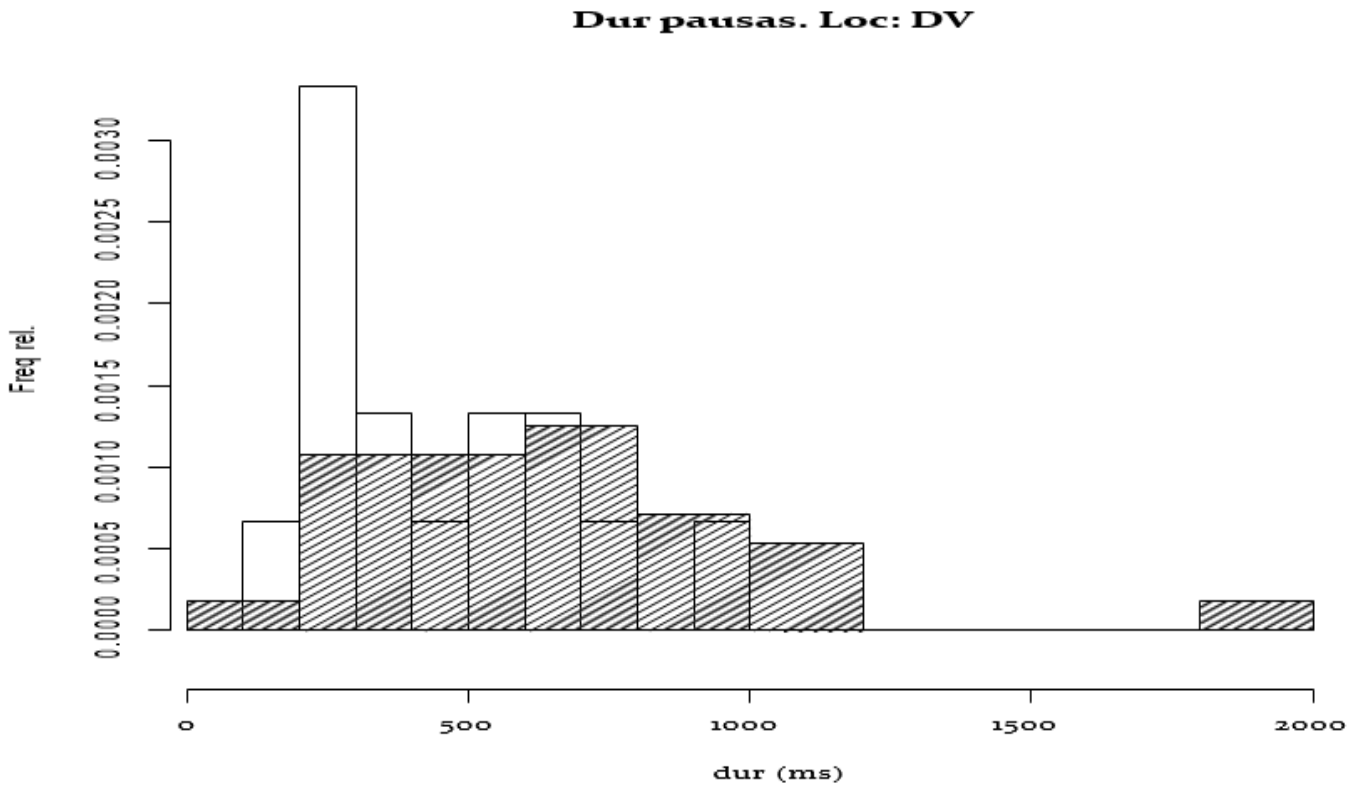

Figura 4.11 – Histogramas superpostos das durações das pausas preenchidas (retângulos claros) e silenciosas na fala do locutor DV em milissegundos.

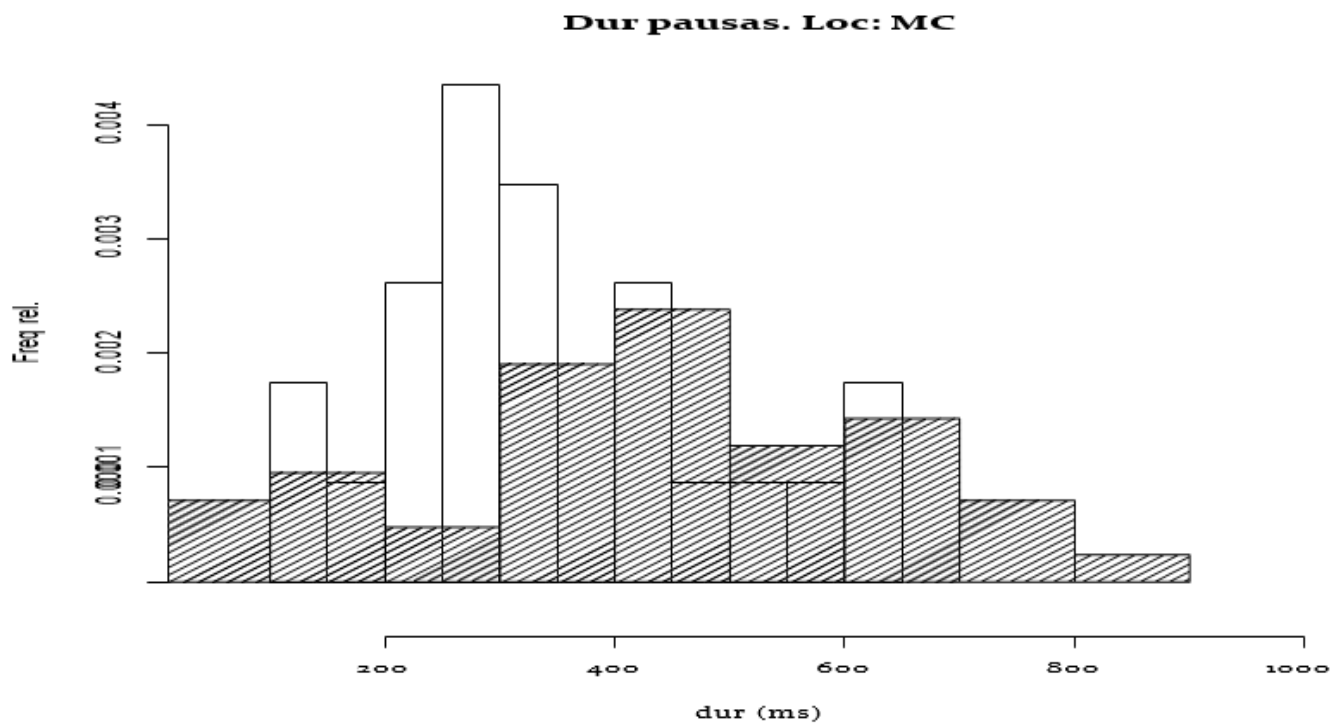

Figura 4.12 – Histogramas superpostos das durações das pausas preenchidas (retângulos claros) e silenciosas na fala do locutor MC em milissegundos.

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.

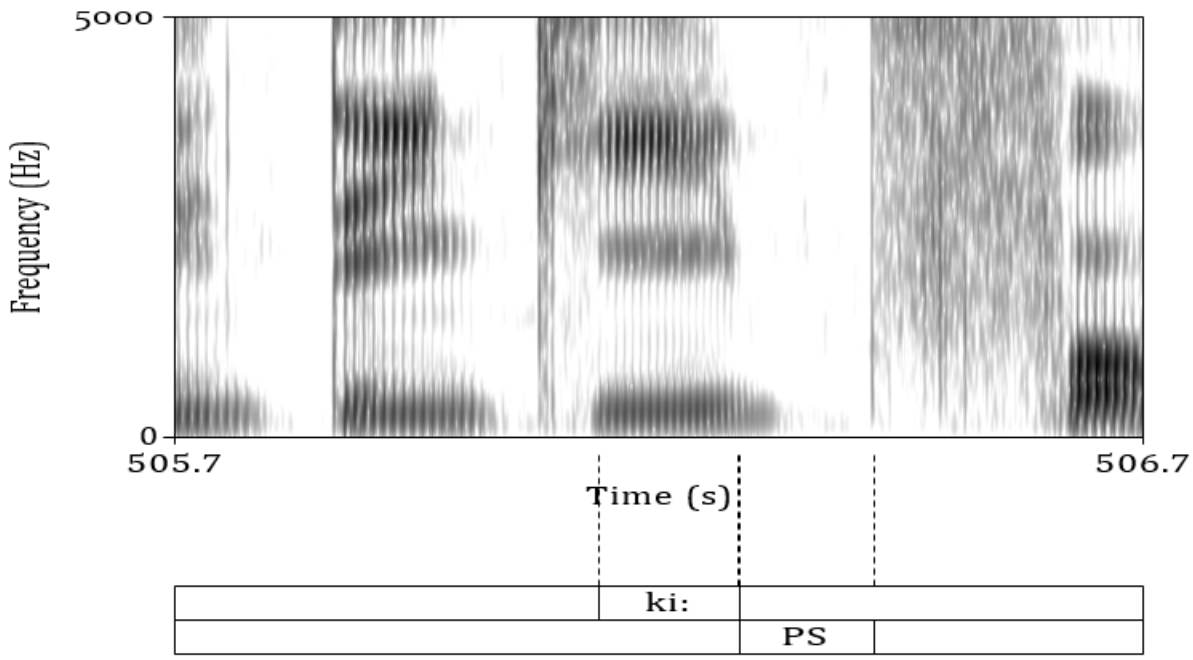

Figura 4.13 – Espectrograma de banda larga e segmentação de pausa silenciosa de duração 136 ms na fala do locutor MC.

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.

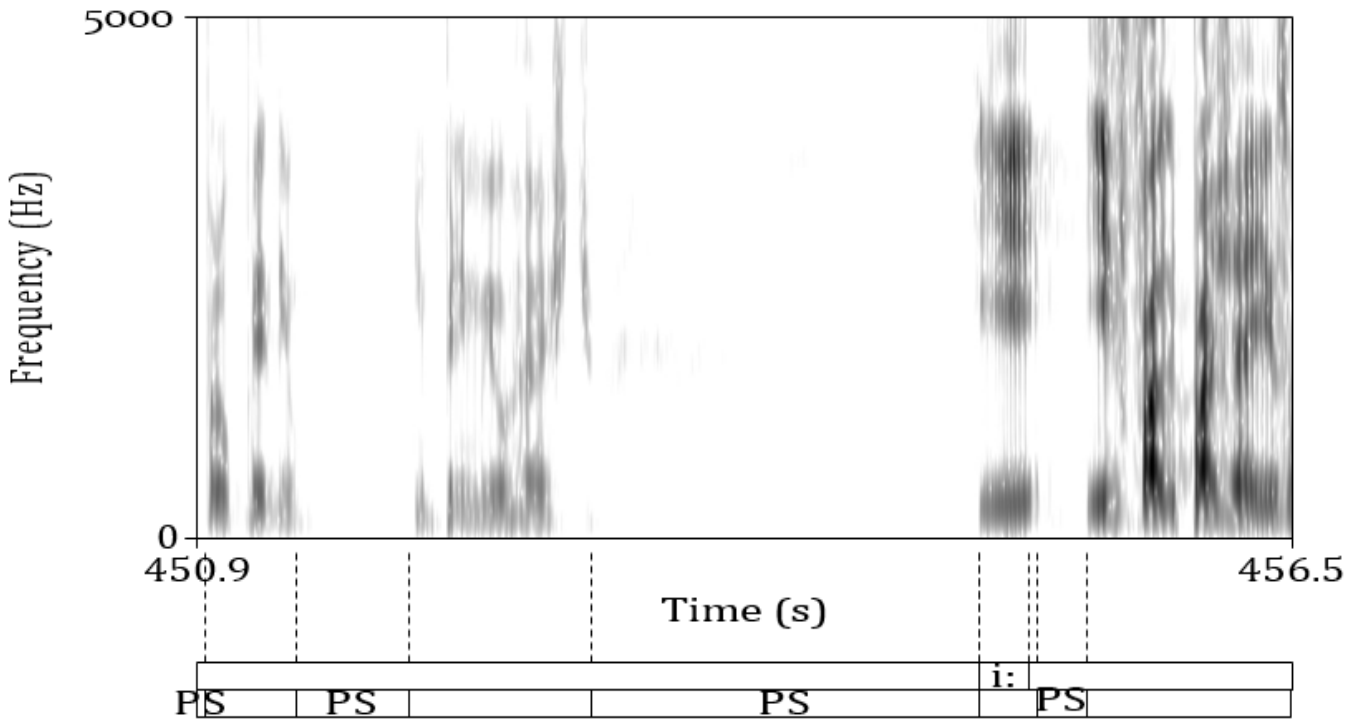

Figura 4.14 – Espectrograma de banda larga e segmentação de pausa silenciosa de duração 2000 ms na fala do locutor DV no trecho central da figura.

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.

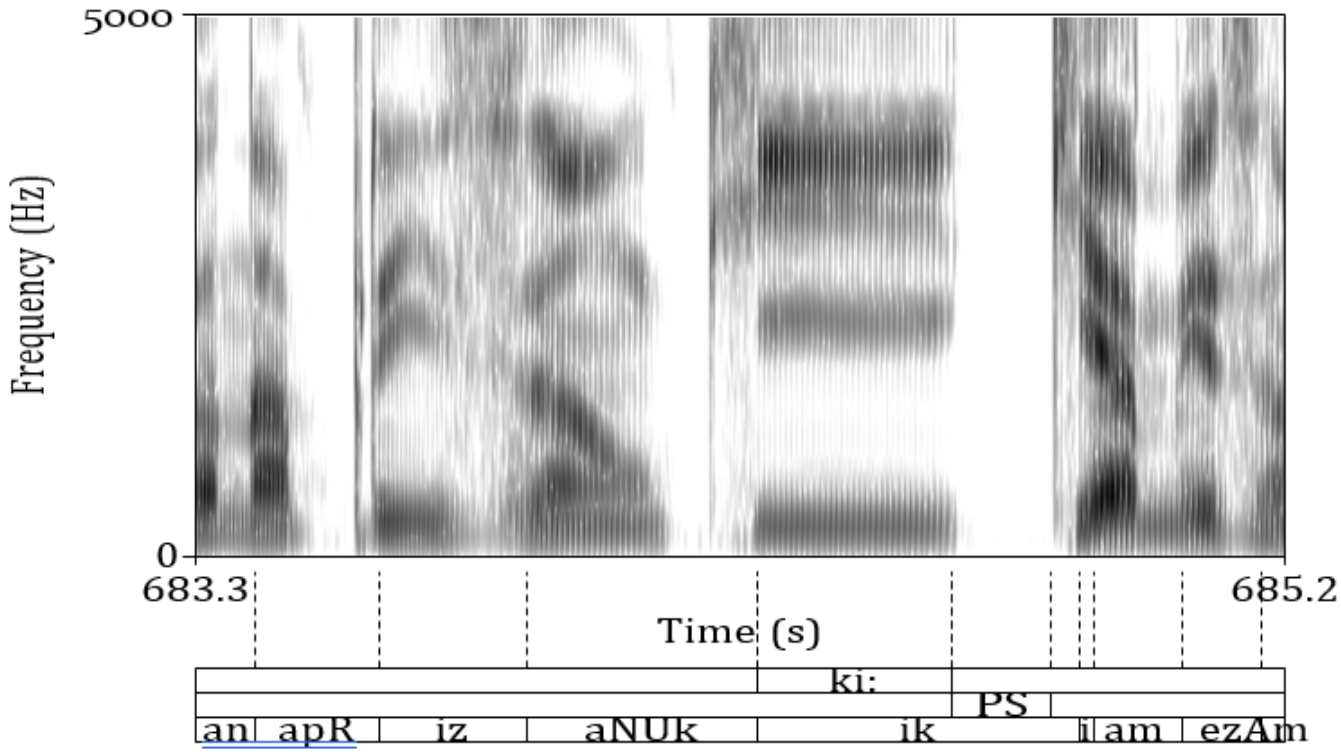

Figura 4.15 – Espectrograma de banda larga e segmentação de pausa preenchida (acima), pausa silenciosa (meio) e unidades VV (abaixo). Trecho da fala de MC, “na prisão que a mesa”.

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.

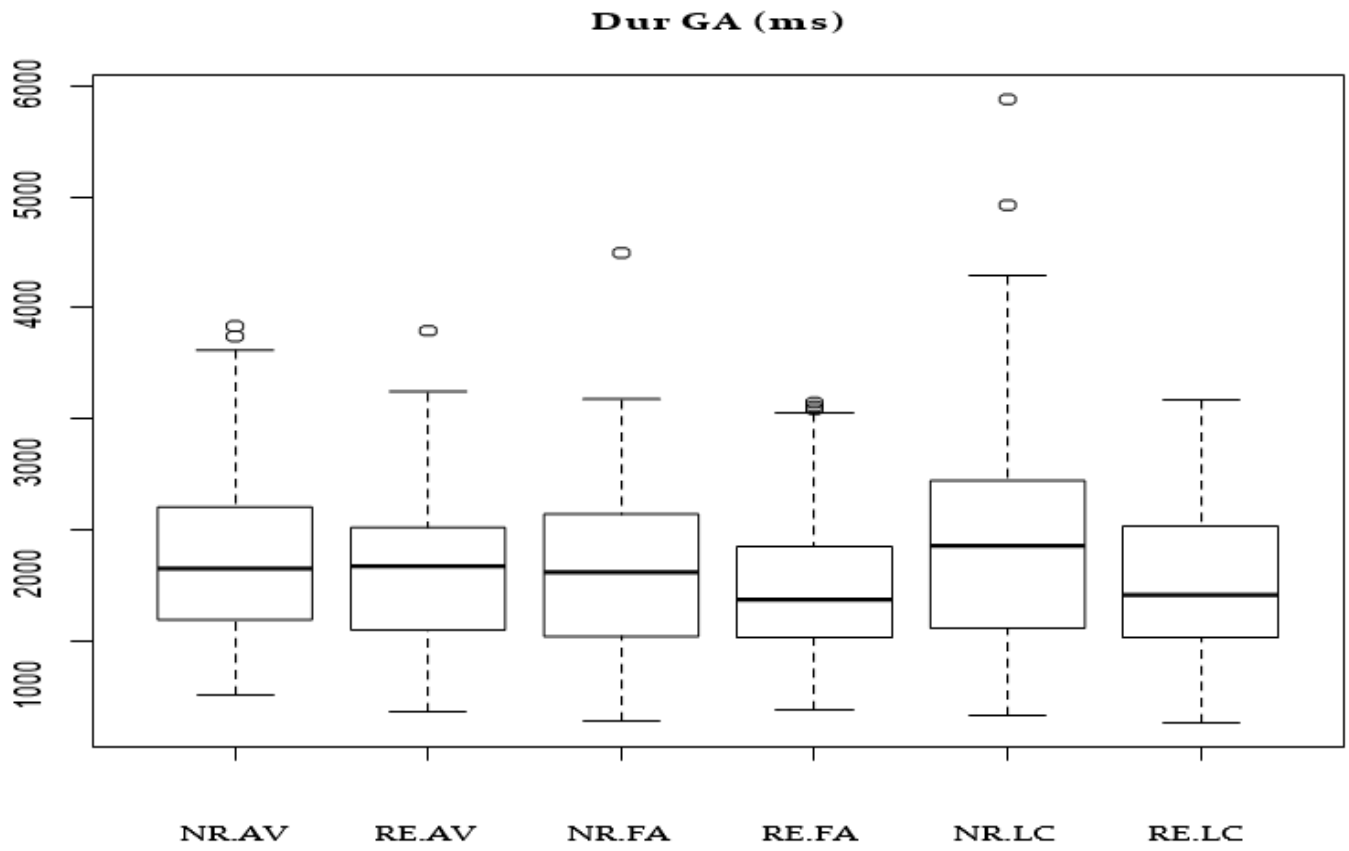

Figura 4.16 – Diagramas de blocos da duração dos grupos acentuais em milissegundos de três locutores paulistas (AV, LC, FA) nos estilos lido (RE) e narrado (NR).

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.

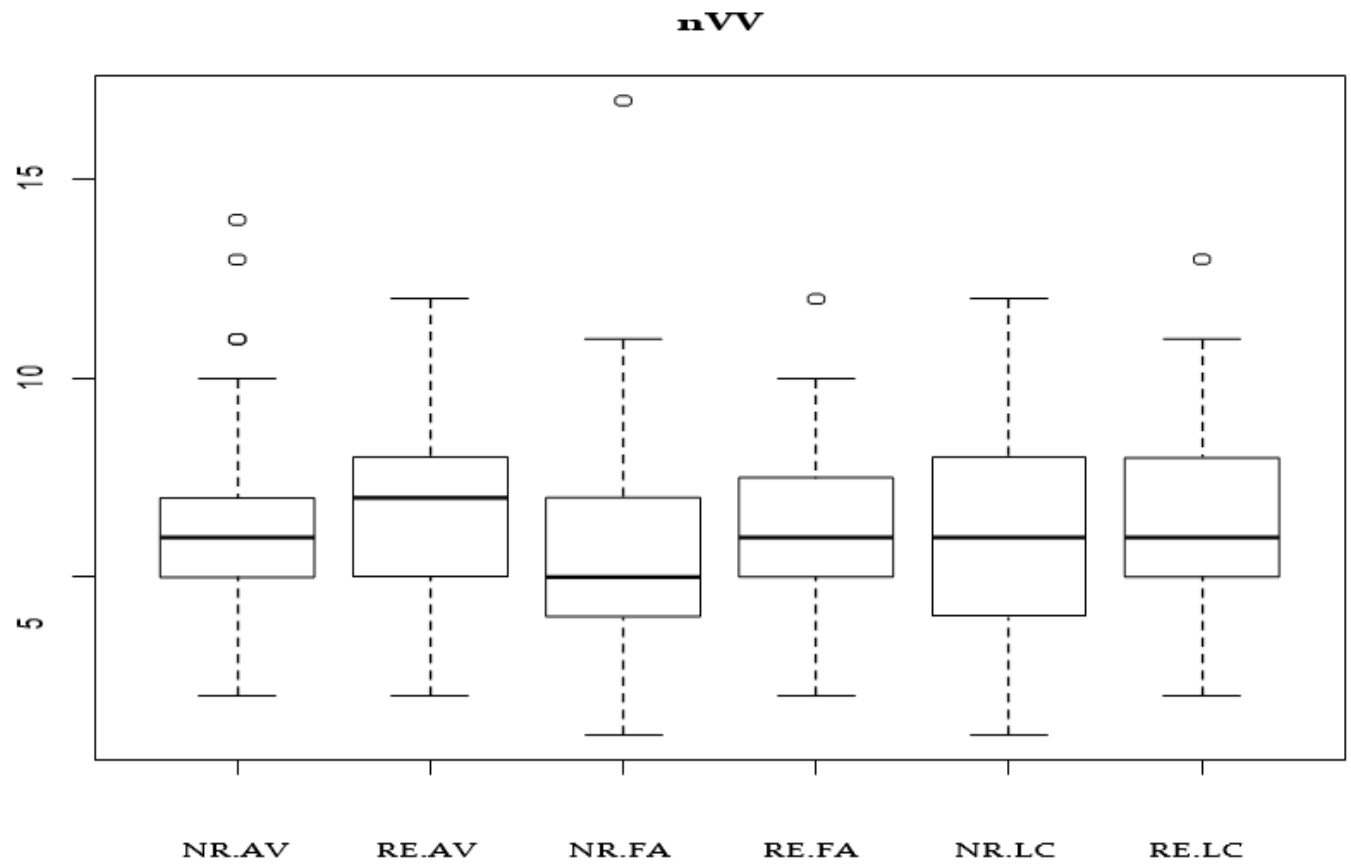

Figura 4.17 – Diagramas de blocos do número de unidades VV nos grupos acentuais de três locutores paulistas (AV, LC, FA) nos estilos lido (RE) e narrado (NR).

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.

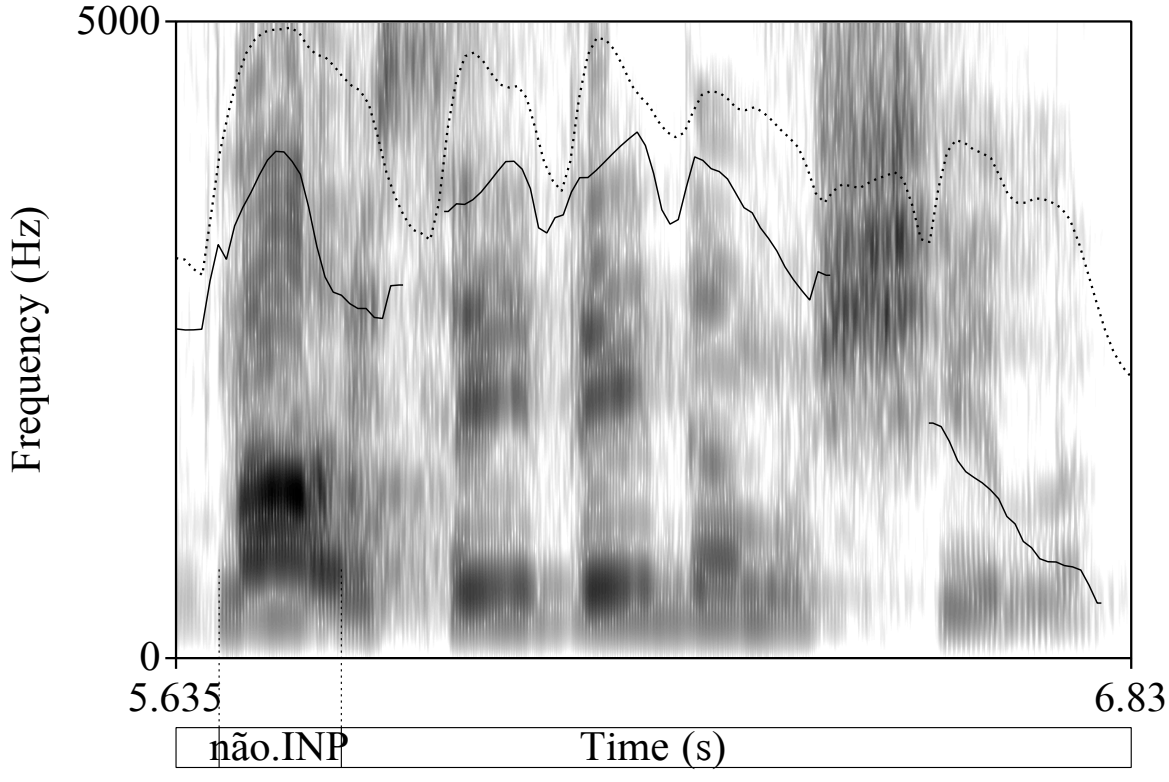

Figura 4.18 – Espectrograma de banda larga e curvas de Fo (cheia) e intensidade (pontilhada) do trecho “não, isso aí veio da mochila” do locutor 1 tendo sido segmentado o incipitário “não”.

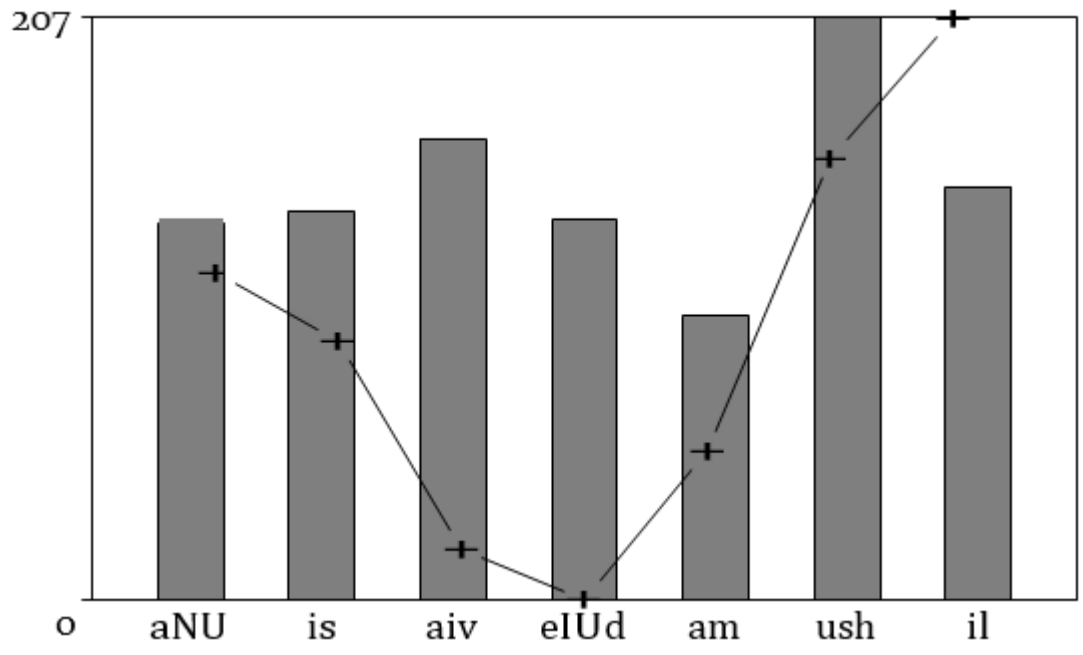

Figura 4.19 – Valores de duração bruta (ms) e *z-score* suavizado das unidades VV do trecho “não, isso aí veio da mochila” do locutor 1. A primeira unidade VV corresponde à rima de “não”.

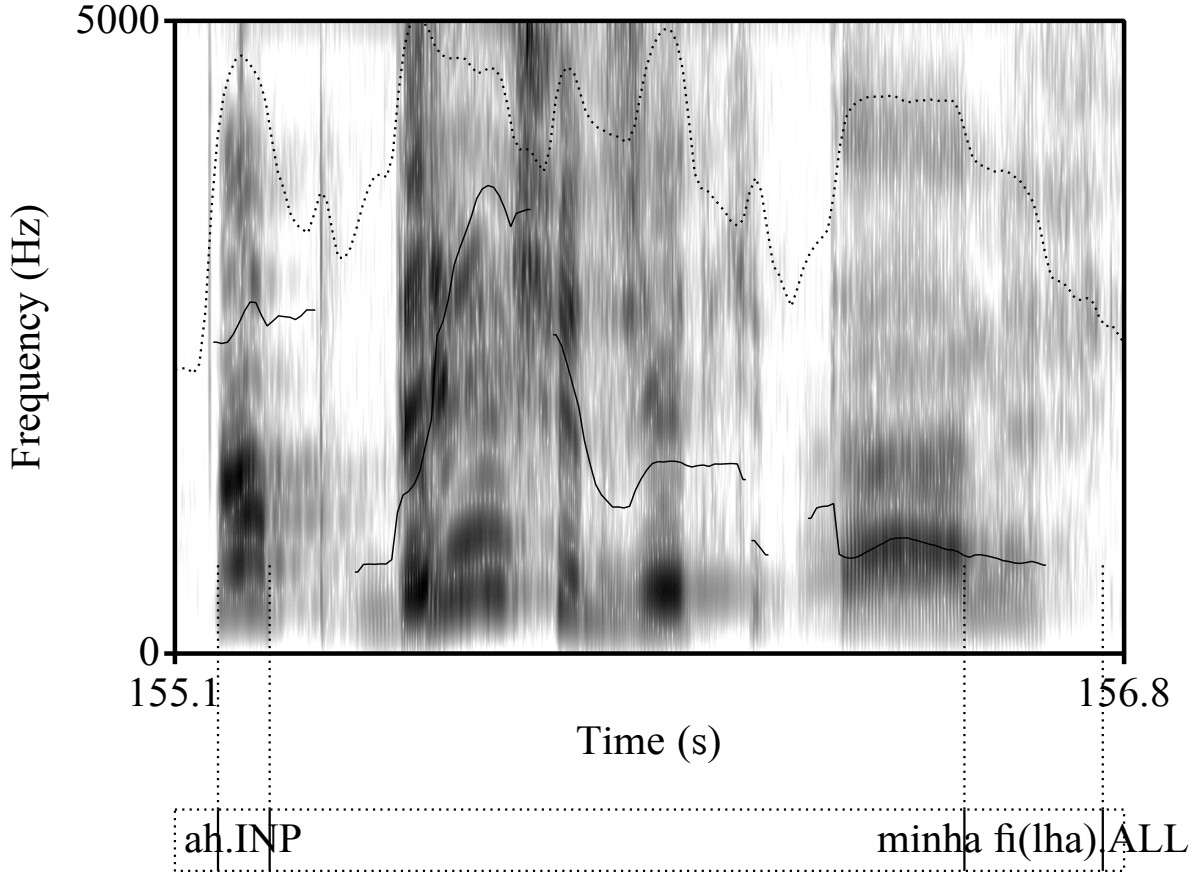

Figura 4.20 – Espectrograma de banda larga e curvas de Fo (cheia) e intensidade (pontilhada) do trecho “ah! deixa do jeito que tá, minha fi(lha).” do locutor 1 tendo sido segmentados o incipitário “ah” e o alocutivo “minha fi(lha)”.

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.

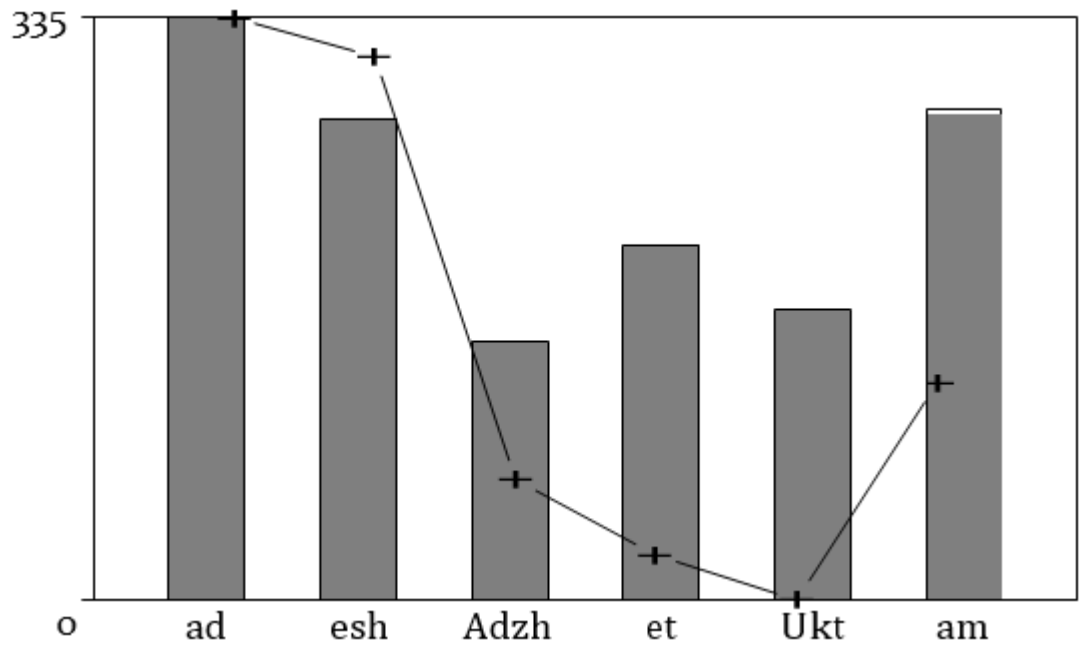

Figura 4.21 – Valores de duração bruta (ms) e *z-score* suavizado das unidades VV do trecho "ah! deixa do jeito que tá, minha fi(lha)." do locutor 1.

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**.

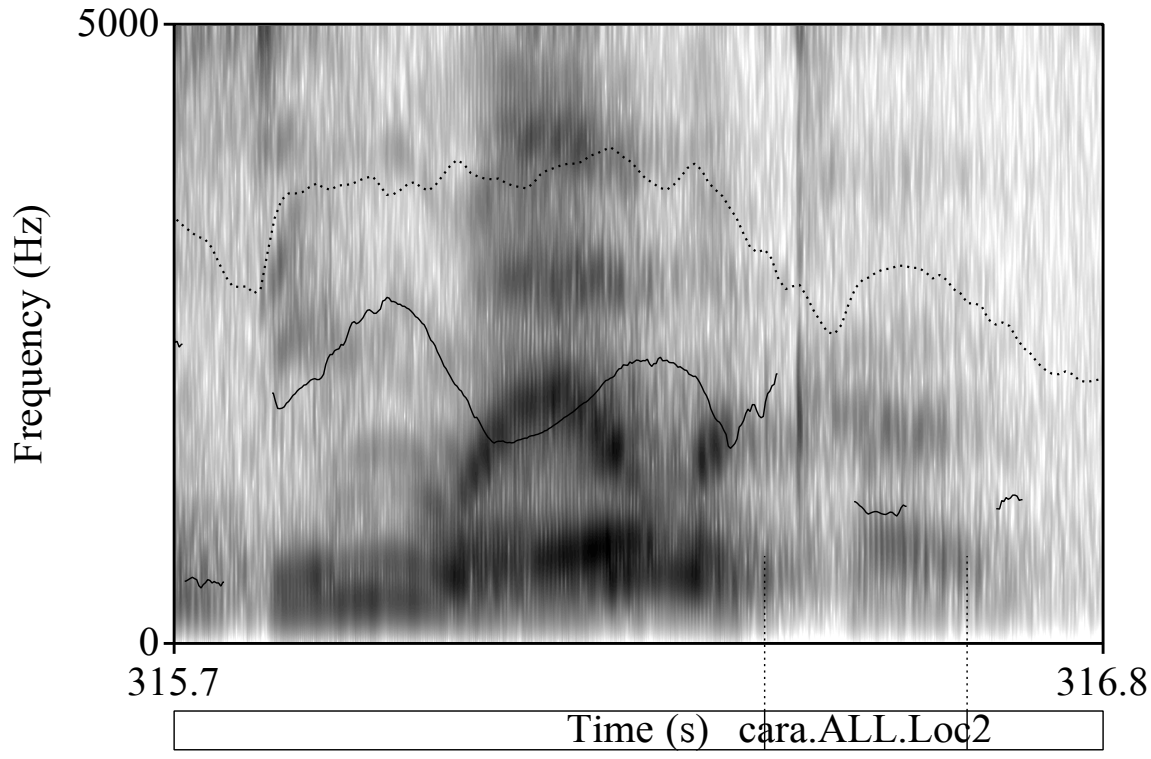

Figura 4.22 – Espectrograma de banda larga e curvas de Fo (cheia) e intensidade (pontilhada) do trecho “cê tá igualzinho ela, cara” do locutor 2 tendo sido segmentado o alocutivo “cara”.

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.

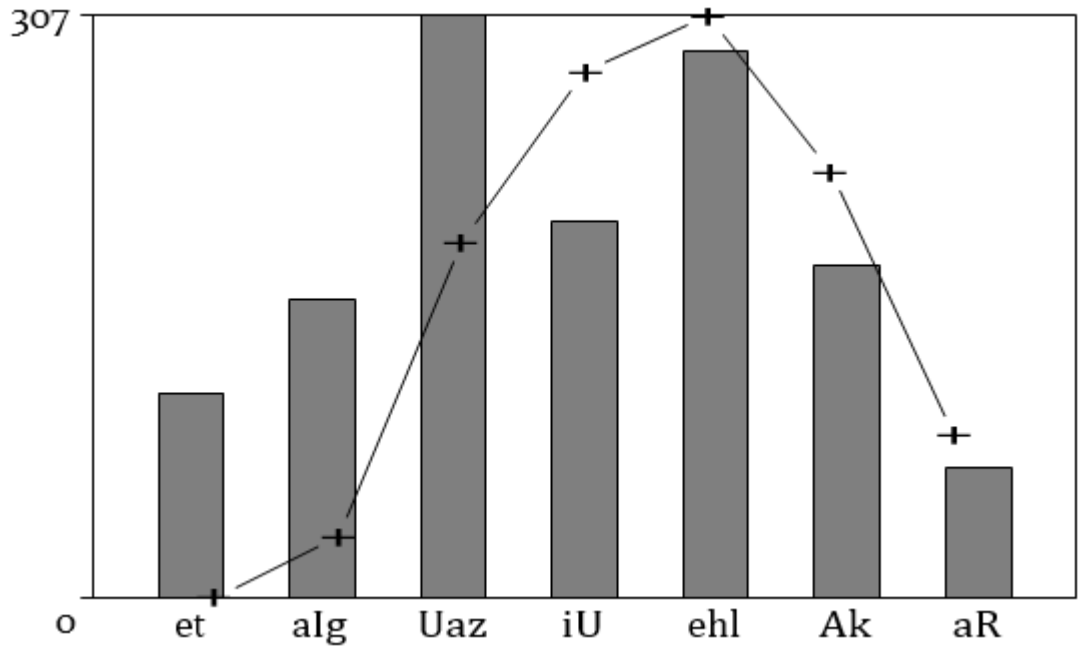

Figura 4.23 – Valores de duração bruta (ms) e *z-score* suavizado das unidades VV do trecho "cê tá igualzinho ela, cara" do locutor 2.

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

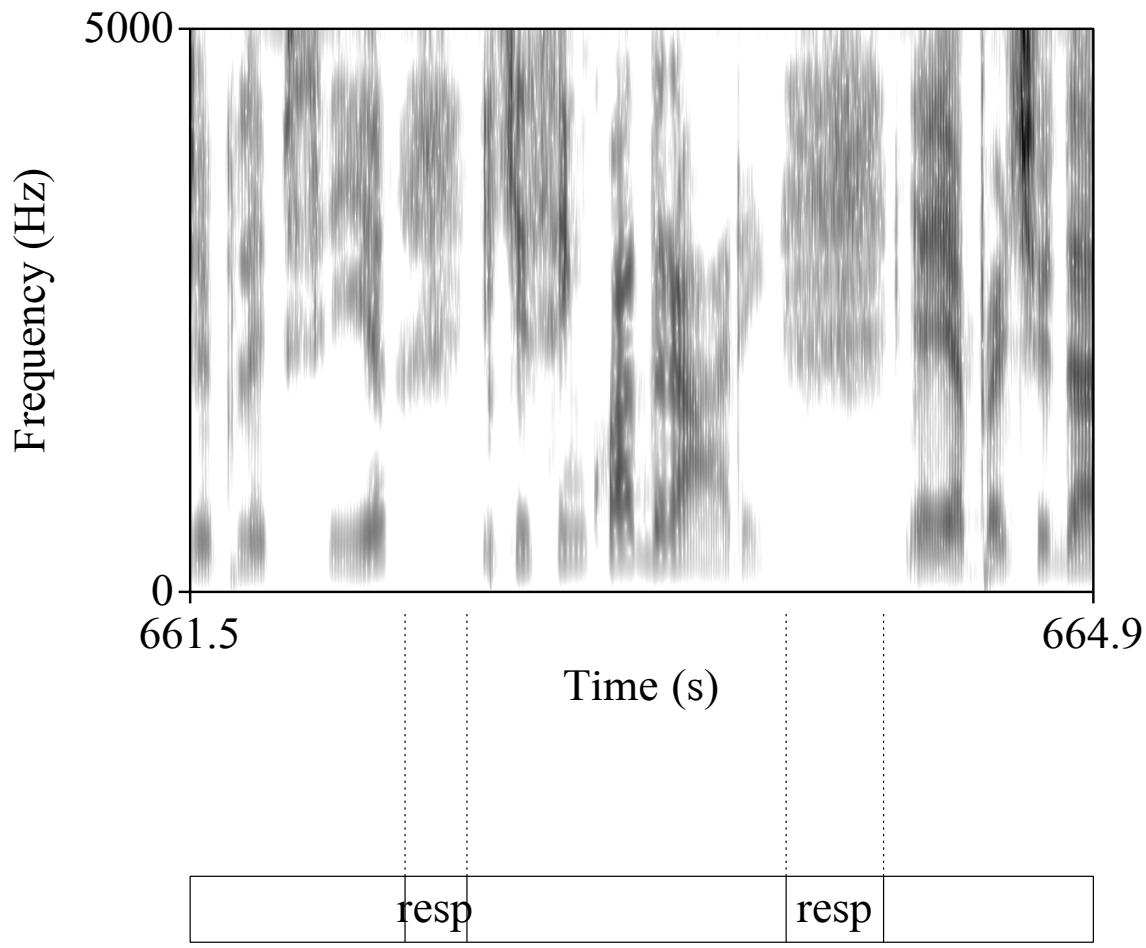

Figura 4.24 – Trechos com RRA em locutora feminina paulista de cercade 25 anos (MI da tabela 4.6). O primeiro ruído tem menor duração eé menos intenso porque marca uma fronteira prosódica mais fraca do que aquela em que a locutora produz o segundo ruído.

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.

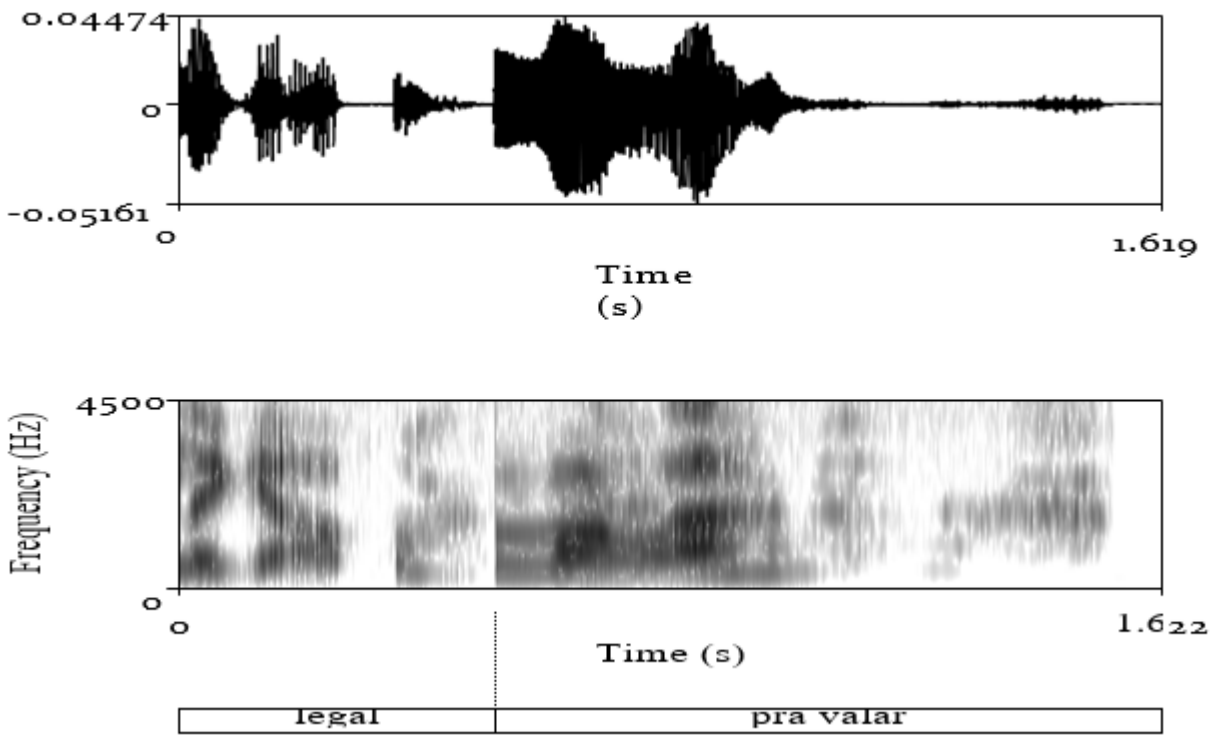

Figura 4.25 – Dois trechos de riso associados à fala em momentos distintos da locutora HD, um durante "legal" e o outro durante "pra falar".

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.

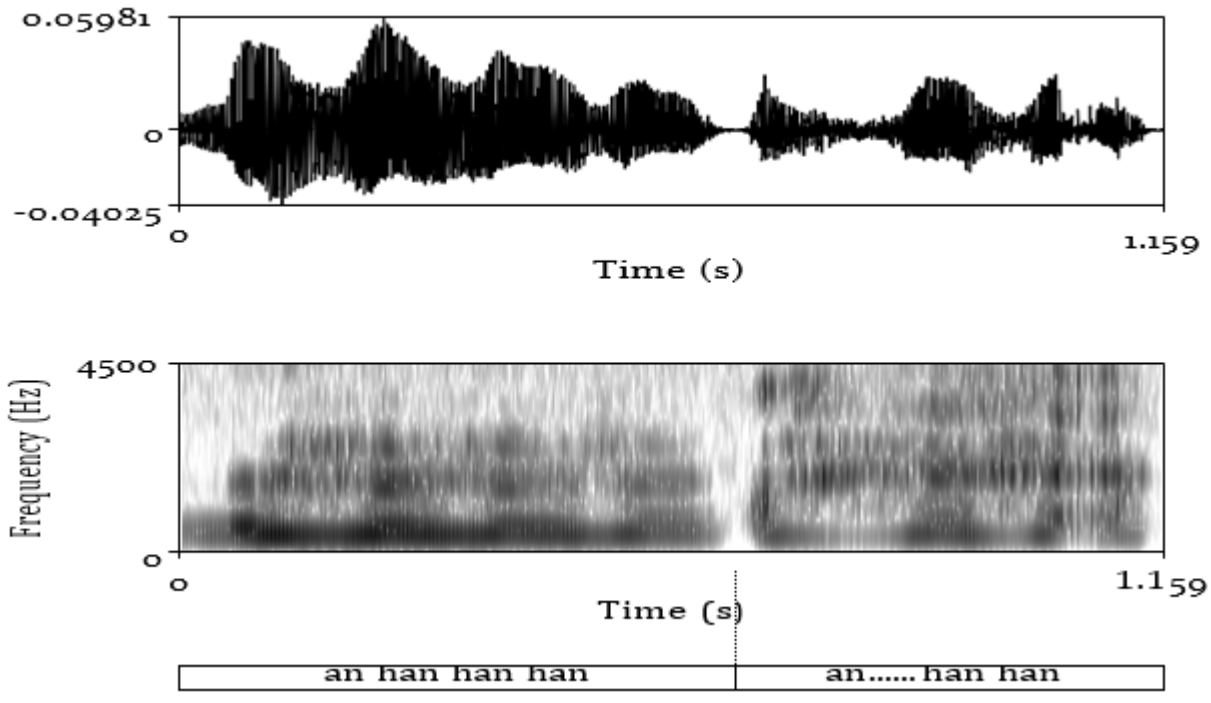

Figura 4.26 – Dois trechos de riso por repetição de sílaba da locutora HD.

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.

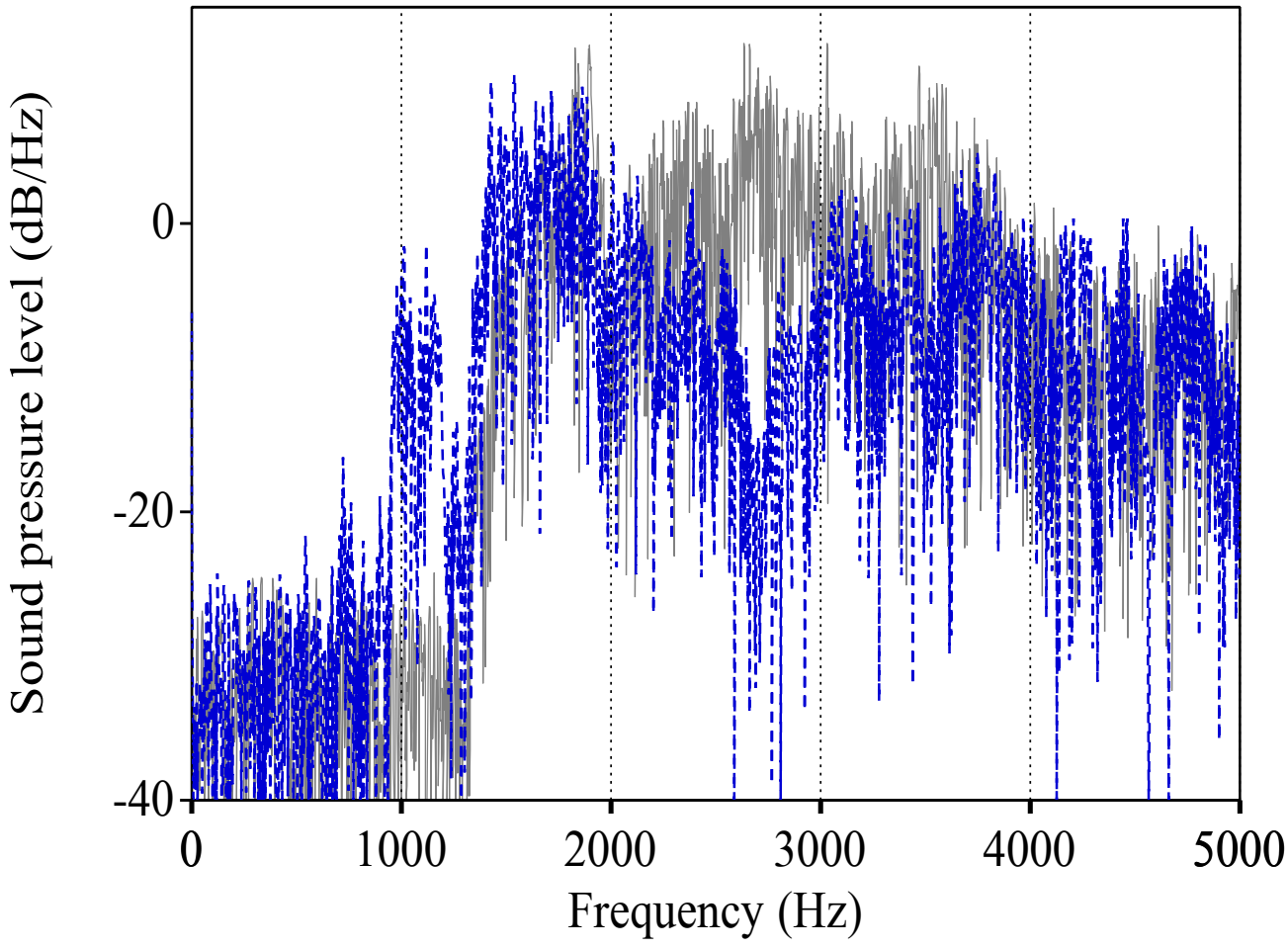

Figura 4.27 – Espectros de Fourier de um evento de RRA dos locutores AX (claro) e FD (escuro), onde se vê claramente que as frequências mais intensas estão acima dos 1000 Hz, com regiões de ressonâncias mais afastadas em FD, com dois grandes lobos (entre 1500 e 2000 Hz e depois de 3000 a 4000 Hz) do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

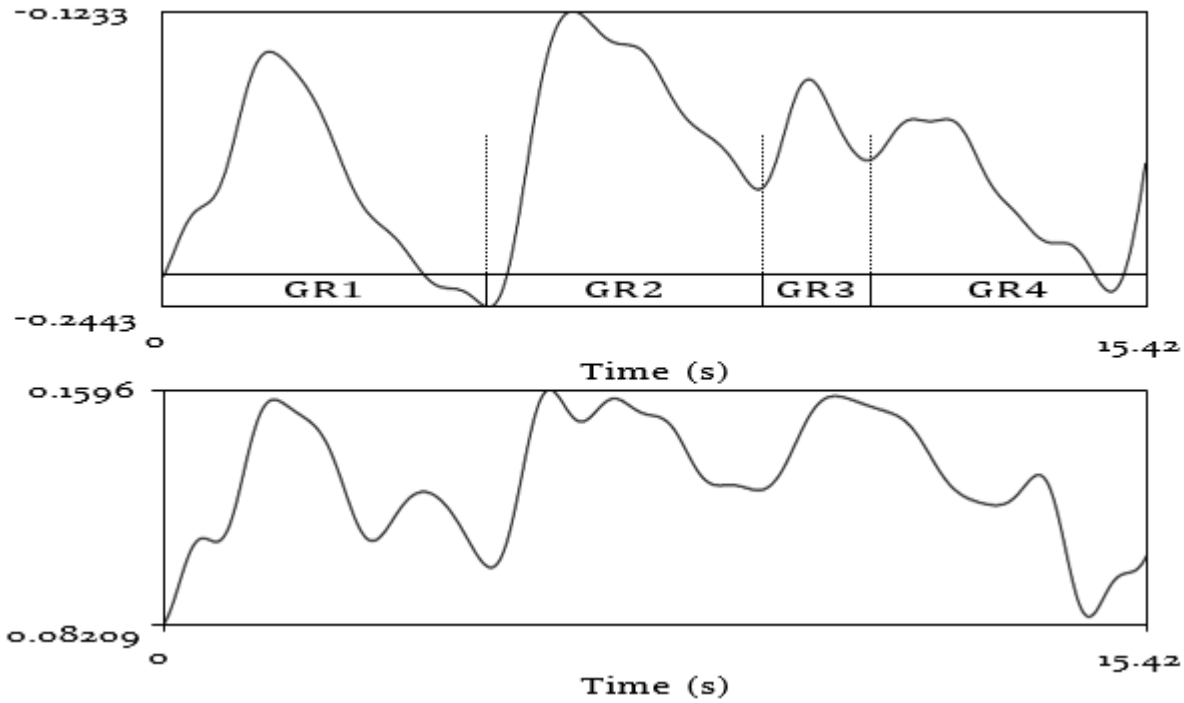

Figura 4.28 – Sinais respiratórios do tórax (acima) e do abdômen (abaixo) de locutora alemã lendo um trecho de texto em inglês de forma persuasiva.

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.

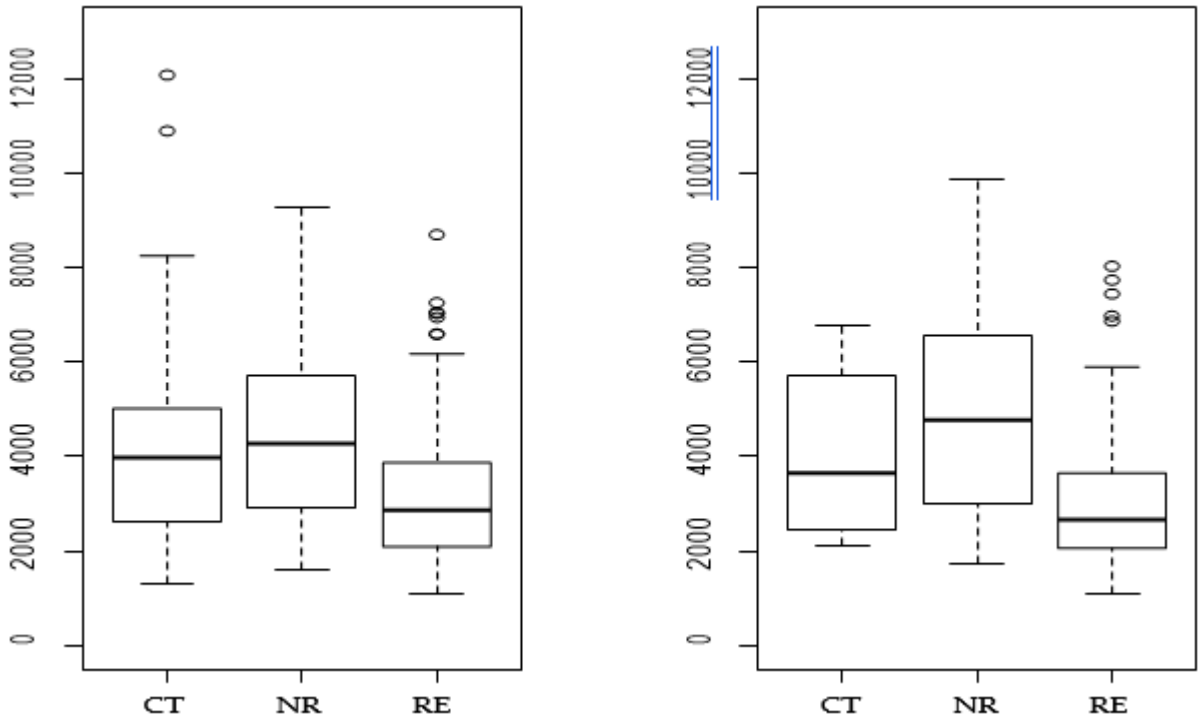

Figura 4.29 – Diagramas de bloco da duração do grupo respiratório para os estilos leitura (RE), narração (NR) e comentário (CT) por sexo, sendo os três blocos à esquerda dos homens e os da direita, das mulheres.

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.

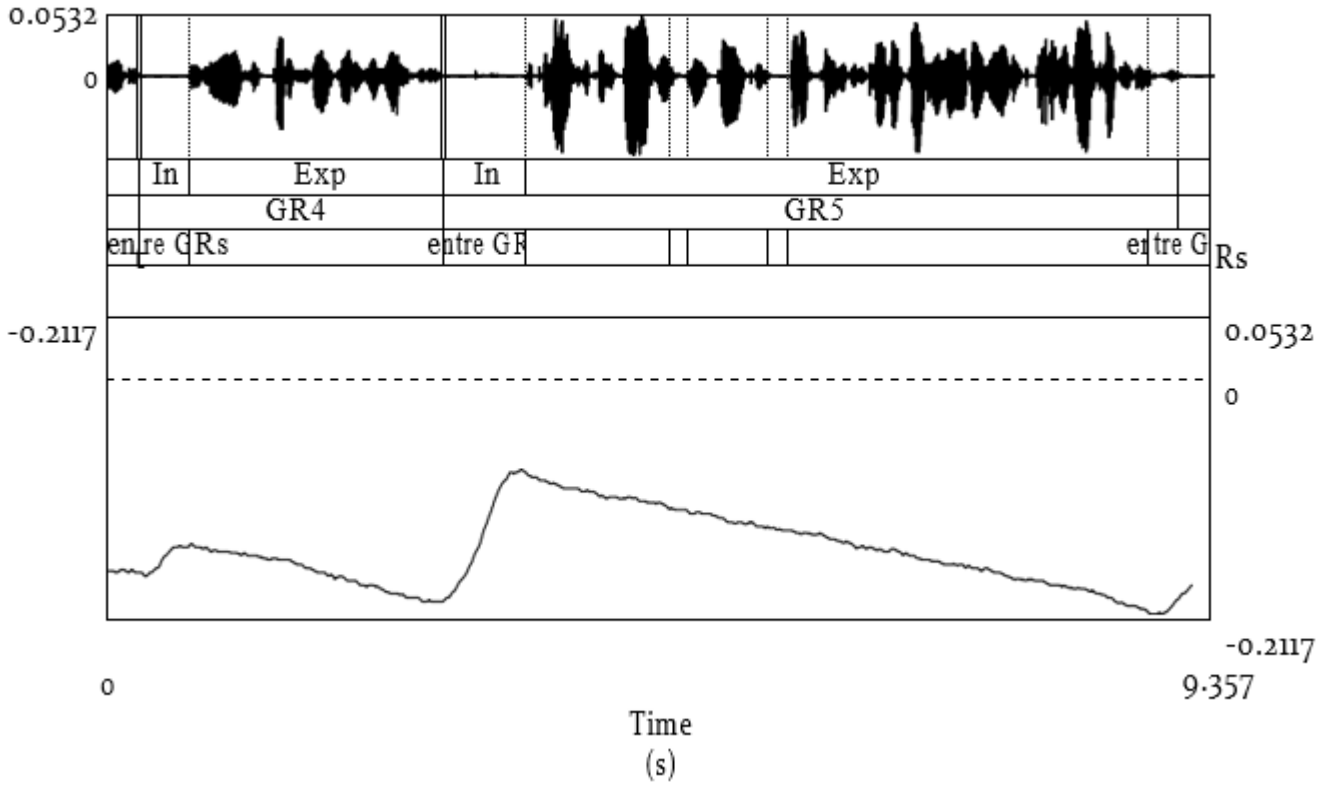

Figura 4.30 – Forma de onda, camadas de anotação e sinal de expansão do tórax de locutor masculino no estilo leitura. O trecho lido é: “os dias pareciam todos iguais. [entre GRs] O que mais custava no entanto era ter de se levantar no meio da noite para rezar as matinas.”. Na anotação, In é fase de inalação, 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.

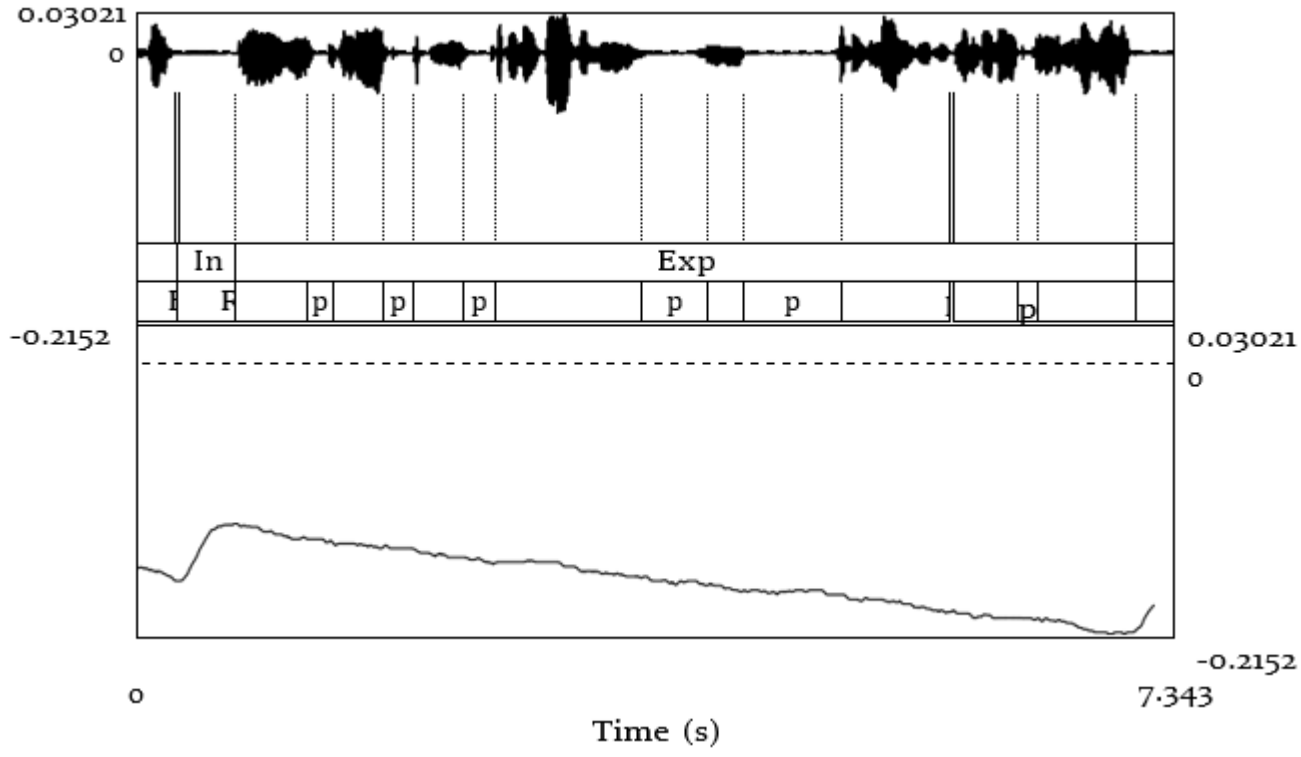

Figura 4.31 – Forma de onda, camadas de anotação e sinal de expansão do tórax de locutor masculino no estilo narração. O trecho lido é: “vida é... que ela... que... era levada pelos monges... o nome dele é Manuel.”. Na anotação, In é fase de inalação, Exp é a fase de expiração e o trecho etiquetado como “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.

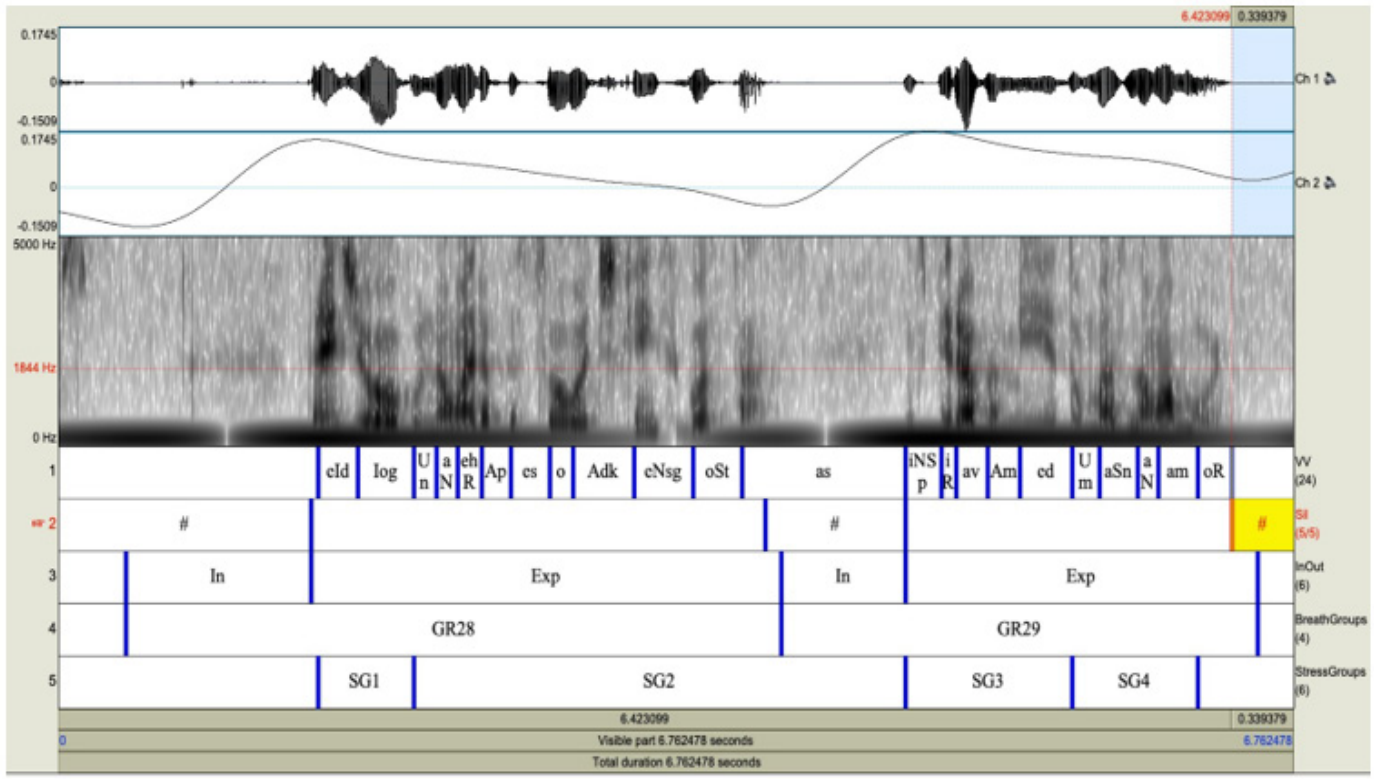

Figura 4.32 – Forma de onda, sinal de expansão do tórax de locutor feminino no estilo leitura e camadas de anotação. O trecho lido é: “Frei Diogo não era pessoa de quem se gostasse. Inspirava medo, mas não amor.” Vide texto para indicação das camadas.

Capítulo 5

Medidas melódicas e de qualidade de voz

Após uma apresentação, na primeira seção, de sistemas de notação para a entoação *stricto sensu*, delinearemos as formas de se obterem medidas fundamentadas tanto em aspectos qualitativos quanto em aspectos quantitativos do contorno melódico em torno das funções de acento de *pitch* e marcação de fronteira, bem como da realização de diferentes estilos de elocução. O capítulo também apontará o interesse dessas medidas para a pesquisa experimental.

5.1 Sistemas de notação melódica

Ao longo dos anos, muitas formas de notação da curva melódica, curva primariamente relacionada à veiculação da entoação da fala, foram propostas. Segundo Crystal (1997), as primeiras formas de notação foram propostas no séc. XVIII por Joshua Steele, a partir de um sistema semelhante a notas musicais, algo próximo ao que, já no séc. XX, foi usado por Fónagy (FÓNAGY; MAGDICS, 1963) para a descrição melódica da emoção na fala e da música. Nos anos 1940, Pike (1945) propôs um sistema que considerava quatro níveis de *pitch* para o inglês, enquanto as representações icônicas de Bolinger (1986, 1989) assinalavam, a partir da década de 1960, que o sistema pikeano tinha muitas limitações, por conta da riqueza melódica em diversos contextos comunicativos.

A chamada *British School* de Crystal (1969) trabalha com a noção de "configuração" que propõe, como parte obrigatória de um

sintagma entoacional, a configuração do núcleo[1], que pode ser um movimento de descida, descida-subida ou subida em nível baixo no quadro de uma gramática entoacional que pressupõe outras configurações opcionais que ocorrem na sequência de elementos: pré-cabeça, cabeça, núcleo e cauda (TAYLOR, 1992).

Tanto Bolinger (1951) quanto Ladd (1983a) apresentaram críticas aos sistemas propostos pelas escolas americana (Pike) e britânica. O sistema americano pikeano de níveis, por ser muito restritivo, não dá conta de curvas melódicas globais ou mesmo da declinação de Fo, muito frequente nos enunciados assertivos em inglês (e em português brasileiro, PB). Por outro lado, o sistema britânico não considera questões como a recorrência de alinhamento dos acentos de *pitch* com as sílabas acentuadas, por exemplo. O sistema de Bolinger, por ser uma espécie de cópia estilizada da curva de Fo, não tem as vantagens de uma concepção enxuta e analítica de eventos melódicos que pudessem servir de norte à construção de uma economia da entoação, embora seja muito interessante para a apreciação da expressividade da fala.

Os sistemas que examinaremos aqui procuram dar conta tanto do caráter combinatório dos acentos de *pitch* quanto da importância em se respeitar o alinhamento da curva melódica com pontos singulares como as sílabas acentuadas.

5.1.1 O sistema ToBI de notação

O sistema ToBI (de *Tone and Break Indices*) de notação entoacional deriva dos trabalhos de Pierrehumbert e colaboradores (PIERREHUMBERT, 1980; PIERREHUMBERT; HIRSCHBERG, 1990; SILVERMAN et al., 1992) e se propõe a capturar dois aspectos prosódicos: (1) o ritmo, pelo emprego de números assinalando quatro "forças"

1 O chamado acento nuclear, o último acento de *pitch* do sintagma entoacional.

de fronteira prosódica, daí a expressão *Break Indices* da sigla (BI) e (2) os eventos melódicos de tons de fronteira e acentos de *pitch* (BECKMAN; ELAM, 1993) que explicam o termo Tones da sigla (To). Os tons de fronteira são marcados pelos símbolos L- e H- para fronteiras de sintagmas intermediários (*intermediate phrase*) e pelos símbolos L% e H% para fronteiras de sintagma entoacional *(intonational phrase*). Quanto ao acento de *pitch*, há cinco tipos, conforme a descrição que segue. Essa descrição é reproduzida do apêndice A das instruções de transcrição do ToBI (BECKMAN; ELAM, 1993) para fins de comparação com o sistema DaTo que será apresentado na próxima seção.

- H* alvo tonal que está na parte superior ou média da gama de variação de Fo do locutor no respectivo sintagma;
- L* alvo tonal que está na parte inferior da gama de variação de Fo do locutor no respectivo sintagma;
- L*+H alvo tonal na parte inferior da gama de variação de Fo do locutor em sílaba acentuada, seguido de subida pronunciada para um pico na parte superior da mesma gama de variação de Fo;
- L+H* alvo tonal alto em sílaba acentuada imediatamente precedido de subida íngreme a partir de um vale de Fo na parte inferior da gama de variação do locutor;
- H+!H* descida de tom a partir de valor elevado em sílaba não acentuada precedente.

O número e tipos de acentos de *pitch*, reiteram em mais de uma publicação os proponentes do ToBI, foram concebidos para o inglês americano[2]. O exemplo da Figura 5.1, que pode ser ouvido em

2 Por isso os sistemas de notação baseados no ToBI para outras línguas tiveram que fazer adaptações, como nos casos do G-ToBI para o alemão e o Sp-ToBI para o espanhol europeu.

To-BImadeH, ilustra o uso do tom H* no enunciado *Marianna made the marmalade* nas palavras proeminentes “Marianna” e “marmalade”. Observe que ambos os tons estão altos e em nível semelhante de F0.

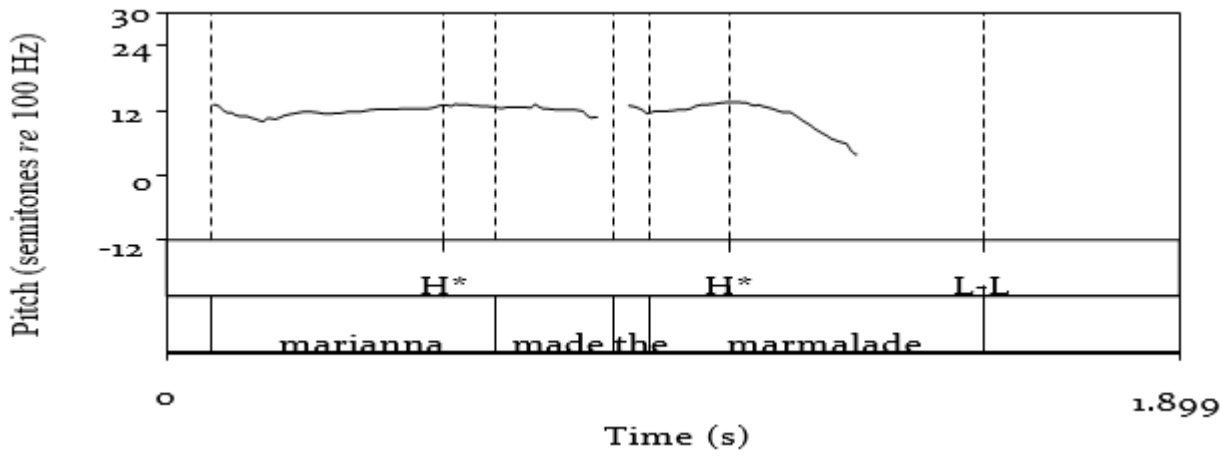

Figura 5.1 – Curva de Fo do enunciado *Marianna made the marmalade* com dois tons altos H, exemplo da oficina de aprendizado do sistema ToBI.

Ao compararmos com o evento bitonal L+H* da Figura 5.2, que pode ser ouvido em **ToBImadeLH**, no mesmo tipo de sentença, mas pronunciada de forma a assinalar foco contrativo em “Marianna”, a curva melódica da palavra em foco começa com uma subida a partir de nível baixo de F0 para atingir o pico que se vê na figura, como na instrução acima para esse tipo de acento de *pitch*.

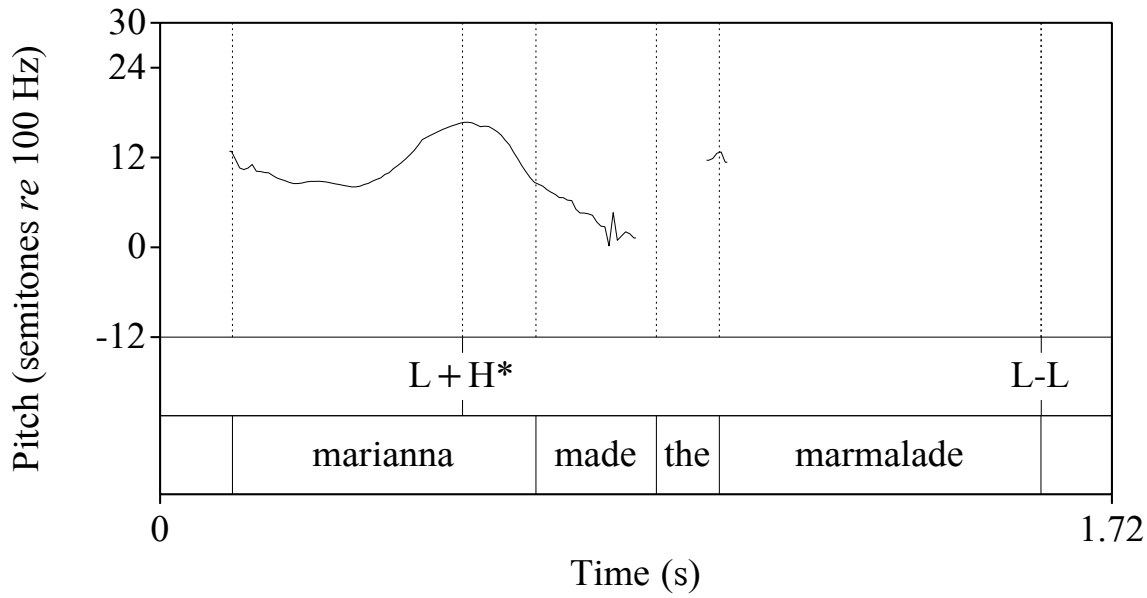

Figura 5.2 – Curva de F0 do enunciado *Marianna made the marmalade* com o evento bitonal L+H*, exemplo da oficina de aprendizado do sistema ToBI.

O último exemplo, na Figura 5.3, que pode ser ouvido em **ToBI-ma- deHH**, ilustra o uso do fenômeno de *downstep* no enunciado *Sublime mnemonic rhyme and free meter* nas palavras proeminentes “sublime”, “mnemonic”, “rhyme” and “meter”. Observe que houve uso do marcador ’!’ indicando queda do valor de Fo em relação ao nível precedente. A diferença entre os dois últimos acentos de *pitch* reside no fato de que em H+!H*, o tom alto que precede se encontra em sílaba não proeminente (“free”), o que não é o caso do tom !H* de “rhyme”, que é precedido da sílaba acentuada na palavra “mnemonic”.

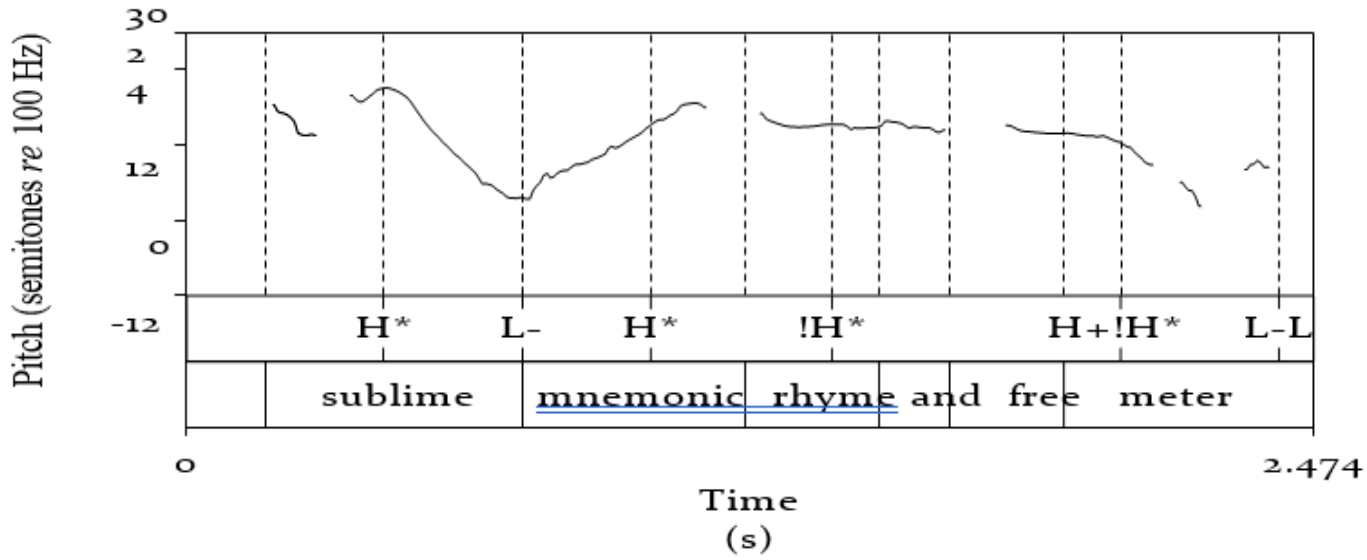

Figura 5.3 – Curva de Fo do enunciado *Ssublime mnemonic rhyme and free meter* para ilustrar o uso de *downstep* (!), exemplo da oficina de aprendizado do sistema ToBI.

Uma das principais críticas a sistemas como o ToBI veio do próprio grupo, dez anos depois de seu aparecimento (WIGHTMAN, 2002), por conta de um problema grave: o baixo acordo entre transcritores quanto ao tipo de tom a serem usados, por falta de instruções claras dos procedimentos de anotação. Esse resultado levou o autor a insistir para que se transcreva tão somente o que se "ouve" (sic), sem se guiar também pela informação dada pelos parâmetros acústicos, procurando se concentrar na função. Mas isso não vem sem trazer outros problemas, como colocou Hirst (2005) ao dizer que o que certamente essa nova proposta quer dizer se refere à interpretação do que se ouve, que diz respeito a sua função, algo que parecia não claramente separado da forma da curva melódica nos dez anos de aplicação do ToBI.

Essa mesma crítica é apresentada por Xu (2011), que assinala ainda que a forma de uma curva melódica na superfície pode estar associada a diferentes funções atuando em paralelo, como uma função de modalidade interrogativa na última palavra de um enunciado e, simultaneamente, uma função de foco na mesma palavra, que

afetam a forma do acento de *pitch* na superfície. No mesmo artigo e em outras publicações, Xu ainda aponta a necessidade de se conjugar um sistema notacional, se seu uso for realmente necessário, com abordagens de aprendizado como a dos modelos de geração da curva de Fo que vimos nas seções 2.2.3 e 2.2.4, que permitem inferir parâmetros que representam cada curva, sendo também passíveis de generalização. De fato, mostramos que é possível inferir características entoacionais a partir de abordagens fundamentadas no modelamento entoacional, como fizemos em PB (BARBOSA; MIXDORFF; MADUREIRA, 2011; BARBOSA,2016), com o modelo PENTA, e em alemão padrão (BARBOSA; MIXDORFF; MADUREIRA, 2011), com os modelos PENTA e de Fujisaki.

Antes de mostrar como combinar uma abordagem qualitativa fundamentada num sistema notacional com uma abordagem quantitativa a partir de descritores estatísticos da curva melódica, apresentamos o sistema DaTo, que se fundamenta numa abordagem dinamicista que leva em conta o ancoramento da curva melódica com pontos singulares da sílaba.

5.1.2 O sistema DaTo de notação melódica

O sistema DaTo de notação melódica, cuja sigla significa *Dynamic Tones* (LUCENTE; BARBOSA, 2009; LUCENTE, 2012), pressupõe a existência de um sincronismo entre a curva melódica e os movimentos articulatórios que geram os padrões espectrais, apesar de serem movimentos controlados por mecanismos distintos. Essa pressuposição o distancia do sistema ToBI, por conceber o trecho de curva melódica associado a uma determinada função como um perfil integral e não como uma composição de tons (e.g., como na notação L+H*).

O sistema DaTo não marca graus distintos de fronteiras prosódicas, deixando isso a cargo de um algoritmo semi-automático

de marcação de fronteira pela via da duração da unidade VV normalizada que foi apresentada na seção 4.2. Essa decisão também permite que o transcritor se concentre na função melódica e na descrição da forma da configuração melódica e seu alinhamento com o material linguístico. Do ponto de vista melódico, deve-se marcar apenas se a curva melódica precedendo imediatamente uma fronteira terminal ou não terminal é alta (H%) ou baixa (L%).

Antes da marcação de qualquer contorno melódico, o sistema requer que se reconheçam as palavras proeminentes e as fronteiras prosódicas do trecho sendo anotado, o que ressalta seu aspecto funcional. Tendo feito isso, então se passa a identificar o tipo de contorno ou tom. Para selecionar criteriosamente a palavra proeminente, recomenda-se que essa função de proeminência seja feita a partir de um conjunto de ouvintes leigos que assinalariam as palavras que se destacam do "fundo". As palavras assinaladas em destaque pela maioria dos ouvintes são então consideradas como proeminentes. O mesmo se faz com as fronteiras, solicitando aos ouvintes que indiquem como o locutor agrupou as palavras no trecho falado.

Quanto aos tipos de acentos de *pitch* no sistema DaTo, eles pertencem a duas classes, contornos dinâmicos, por se referirem a um movimento de subida (LH, >LH e HLH) ou de descida (HL, >HL e LHL), e tons estáticos alto (H) e baixo (L). Para todos esses contornos e tons, o aspecto crucial, que diz respeito ao sincronismo articulatório mencionado acima, é o alinhamento do pico de F0 (nos contornos ascendentes e no tom alto) ou do vale de F0 (nos contornos descendentes e no tom baixo) com a sílaba lexicalmente acentuada, bem como o alinhamento do movimento que prepara a subida ou a descida dos contornos dinâmicos com a sílaba átona precedente, como veremos adiante.

Estudos conduzidos em corpora de fala espontânea do PB (LUCENTE, 2012, 2017), sugerem que o contorno ascendente LH seja o contorno *default* do enunciado assertivo, marcando um foco

estreito. Em posição final de enunciados interrogativos esse contorno também pode aparecer, mesmo sendo mais comum a ocorrência de >LH (LUCENTE; BARBOSA, 2009).

O alinhamento do pico de F0 ao final da subida presente nos contornos LH e HLH se dá em algum ponto da vogal tônica da palavra proeminente, enquanto esse alinhamento se dá ao final ou mesmo depois da vogal tônica no contorno >LH, fazendo com que a subida desse contorno fique contida inteiramente no intervalo da vogal tônica. Essa diferença de alinhamento do pico de Fo em relação à tônica foi estudada por Kohler (2006a) em línguas como o alemão e o inglês, com o pico no início da vogal tônica sendo interpretado como marca de finalidade, no meio da tônica como abertura para novo argumento e, ao final da tônica, como no caso de >LH (*late peak* para Kohler), suscita uma interpretação de surpresa ou de algum tipo de expectativa. O número de possíveis significações desse atrasodo pico é bastante ampliado no trabalho de Ward (2019, p. 75-95), com aspectos como incredulidade, sugestão, pedido, oferta, convite, especulação, entre muitas outras. Esses estudos mostram como o alinhamento do pico ou vale (KOHLER, 2006b) com a vogal tônica é crucial para a interpretabilidade de um enunciado.

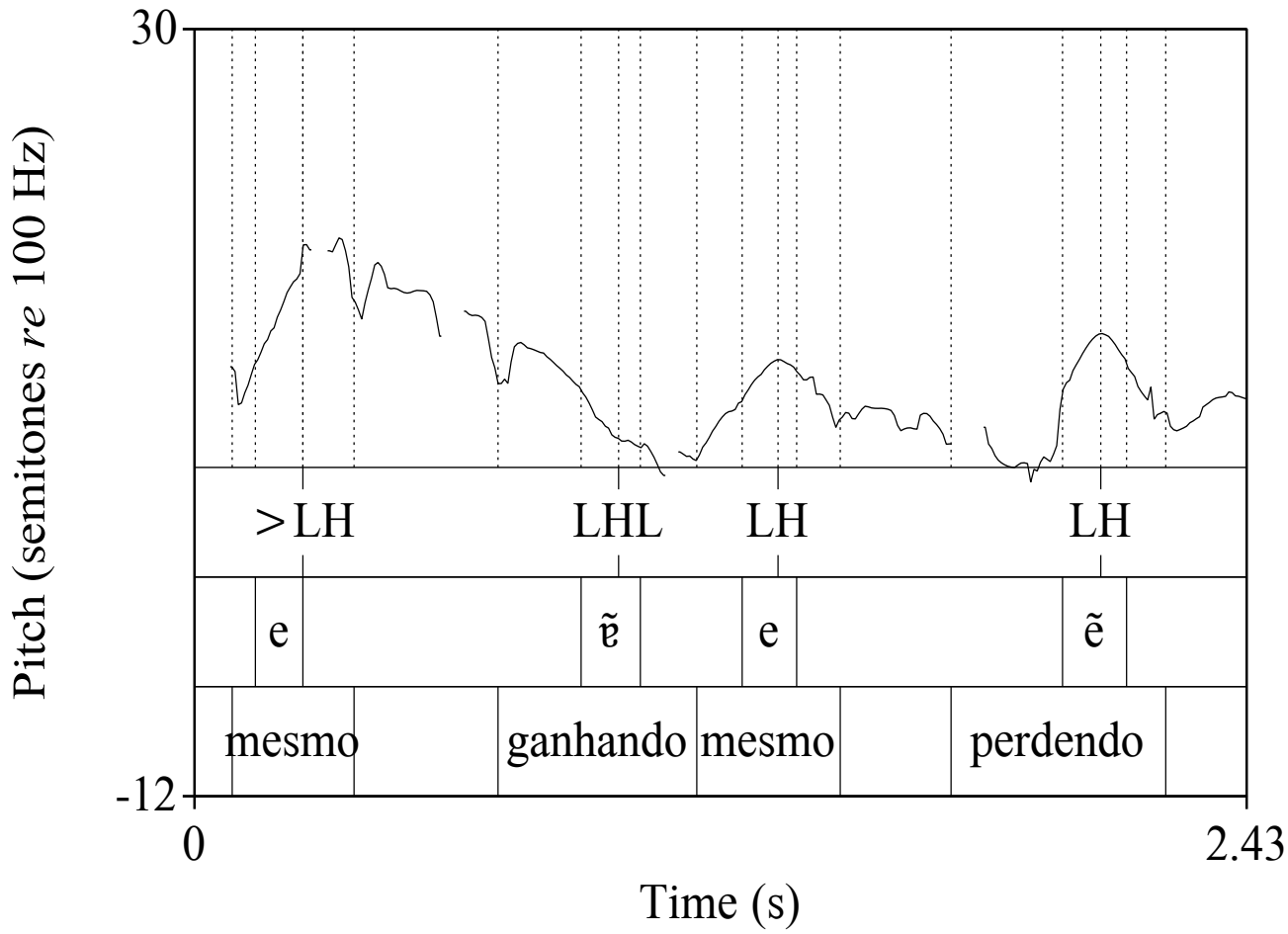

Figura 5.4 – Curva de Fo e camadas de anotação do trecho de enunciado"Mesmo o Brasil ganhando, mesmo o Brasil perdendo", ilustrando os contornos ascendentes LH e >LH. Somente as palavras proeminentes estão transcritas para facilitar a visualização. Indicam-se também os intervalos das vogais tônicas. Trata-se de uma locutora paulista durante um programa da rádio Você de Campinas.

Uma comparação entre os contornos LH e >LH pode ser vista na Figura 5.4, na qual se observa que o contorno >LH tem seu pico no extremo direito da vogal tônica de "mesmo", enquanto o contorno LH tem seu pico mais próximo ao meio da vogal tônica, especialmente a segunda ocorrência na palavra "perdendo". O contorno LHL é um contorno descendente, com descida lenta que se estabiliza na vogal tônica de "ganhando". Pode-se escutar o enunciado inteiro no arquivo **MesmoMesmo**.

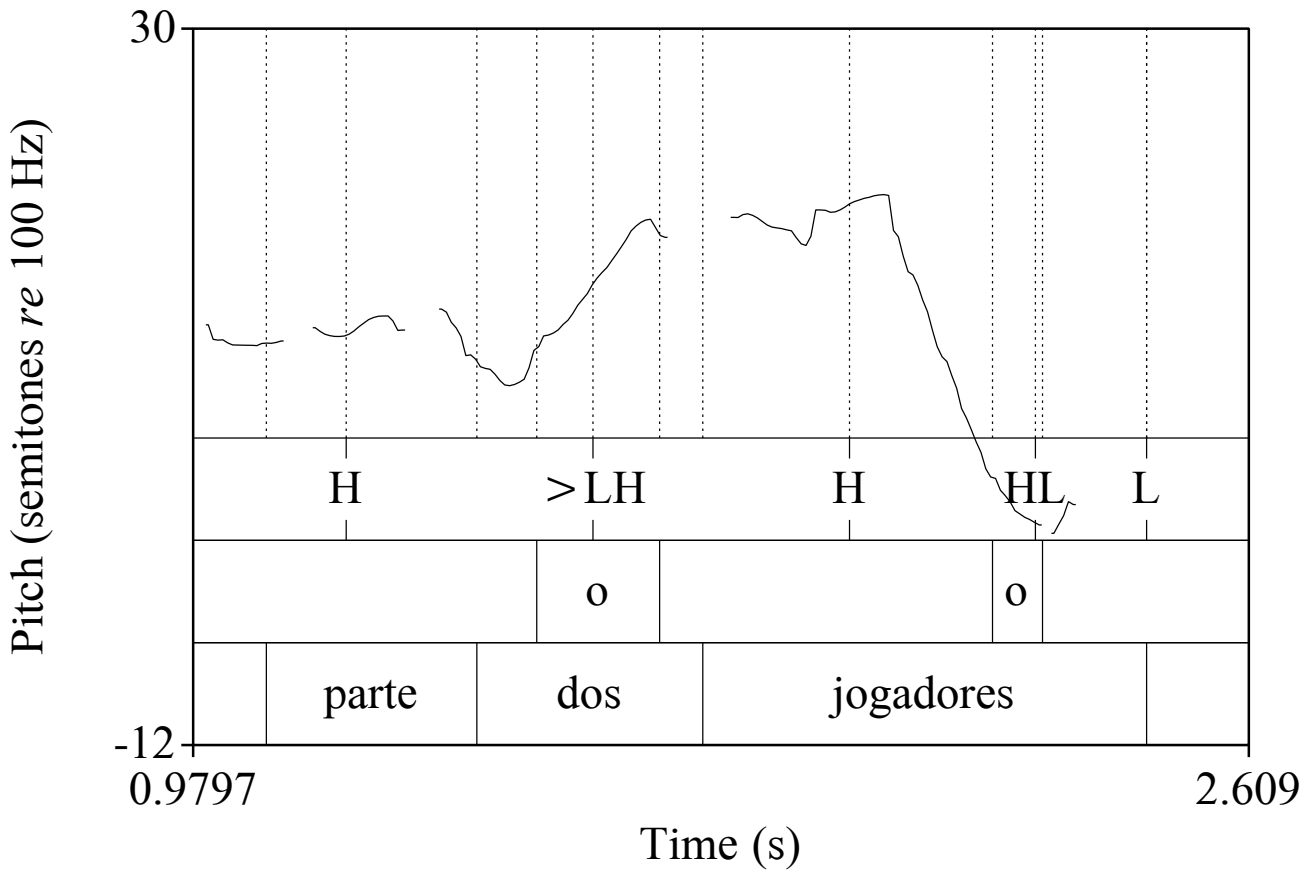

Figura 5.5 – Curva de F0 e camadas de anotação do trecho de enunciado “da parte dos jogadores”, ilustrando o contorno descendente HL, que pode ser comparado ao contorno também descendente LHL, porém com descida mais lenta da Figura 5.4. Indicam-se também os intervalos das vogais tônicas (ou proeminente, no caso de “dos”). Trata-se de uma locutora paulista durante um programa da rádio Você de Campinas.

O contorno HL, que habitualmente precede a fronteira de um enunciado assertivo, é composto por dois estágios: (1) subida da curva melódica em sílaba pré-tônica, podendo ser o clítico precedente, que culmina em pico de F0 alinhado normalmente à parte medial de vogal pré-tônica e (2) descida de F0 para alinhamento da curva em tom baixo durante a tônica. A subida precedendo a descida, de modo especular ao contorno LH, é necessária para a definição desse contorno e o diferencia do tom de nível L.

O contorno >HL tem a mesma forma que HL, mas se encontra atrasado, tendo seu vale mais à frente, fazendo com que seu pico se situe habitualmente no início da vogal tônica. O contorno LHL, por sua vez, assinala uma descida lenta de F0 própria de finais de enunciados assertivos e se alterna com HL, sendo o último mais enfático. Para

ser marcado como LHL, o contorno melódico também deve ter uma subida de F0 antecedendo sua descida.

O contraste entre as descidas dos contornos LHL e HL pode ser visto comparando-os nas Figuras 5.4 e 5.5, na qual se pode ver a rapidez com que a curva de F0 cai para atingir um mínimo ao final da vogal tônica de "jogadores". Observe no trecho que a contração "dos" foi realizada com ênfase e com contorno >LH, cujo pico está alinhado com o final da vogal.

Para apontar os aspectos abstratos da notação, a despeito de diferenças de implementação fonética, a Figura 5.6 mostra a mesma sequência de contornos empregados no mesmo enunciado da locutora paulista:

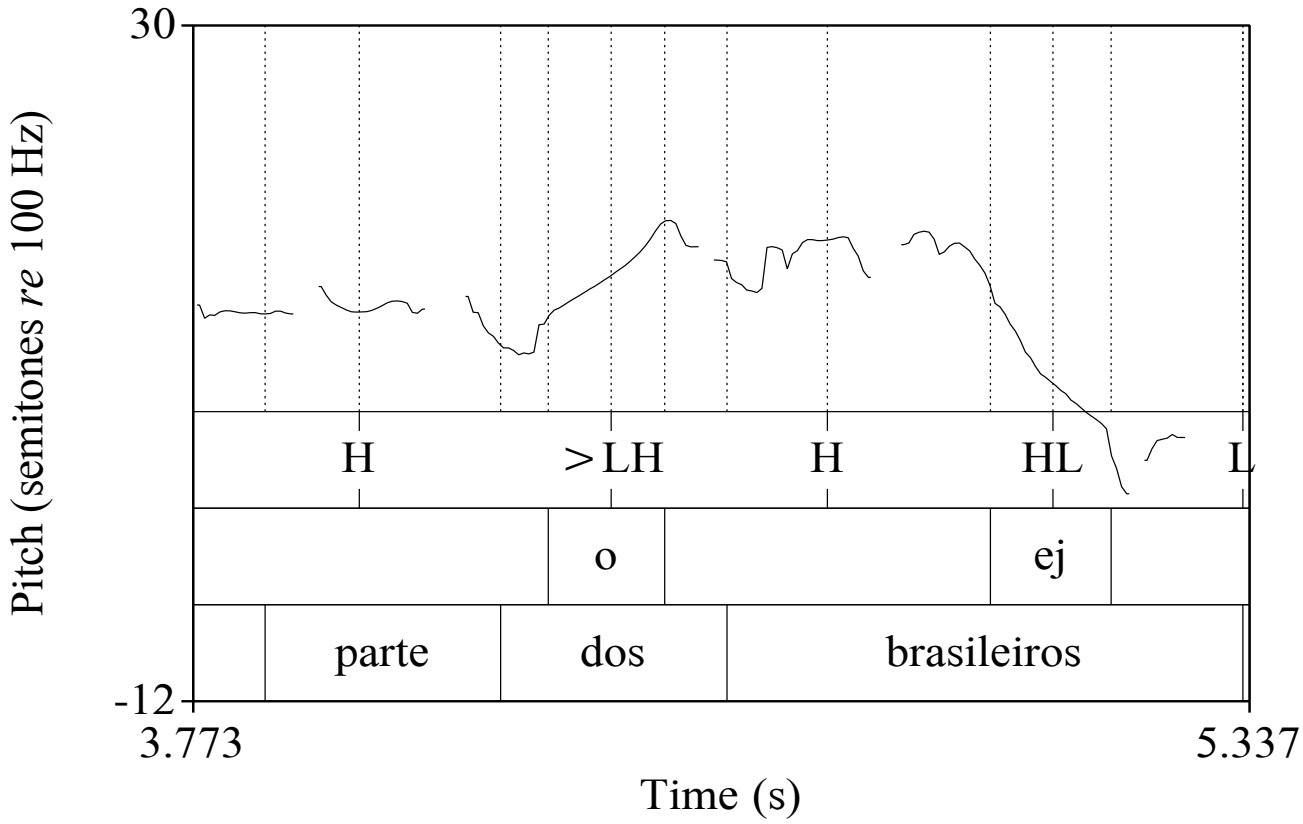

Figura 5.6 – Curva de F0 e camadas de anotação do trecho de enunciado "da parte dos brasileiros", ilustrando as regularidades do uso dos contornos, observando seu paralelismo com o uso na Figura 5.5. Indicam-se também os intervalos das vogais tônicas (ou proeminente, no caso de "dos"). Trata-se de uma locutora paulista durante um programa da rádio Você de Campinas.

"Não tenho percebido isto da parte DOS jogadores, eu te-

nho percebido isto da parte DOS brasileiros." (contração "dos" em maiúsculas por ter sido pronunciada enfaticamente nas duas ocorrências), como se pode escutar do arquivo **Jogadores**. Nas duas sequências o perfil melódico não é exatamente o mesmo, mas a função é a mesma.

O contorno ascendente HLH, que integra uma proeminência secun dária, é ilustrado na Figura 5.7, realizado por jornalista masculino da CBN de São Paulo na palavra "JULgar". E pode ser comparado na Figura 5.8 ao uso que ele faz desse mesmo contorno em "o governo", com pico inicial no artigo, bem como o uso de >LH para dar ampla ênfase na palavra "toda" (o enunciado pode ser escutado no arquivo **Dinheirama**).

Uma representação esquemática dos contornos dinâmicos e sua relação com a vogal tônica podem ser vistas nas Figuras 5.9 e 5.11 para as subidas e descidas, na Figura 5.10 para o contorno HLH e na Figura 5.12 para o contorno LHL. Essas representações foram extraídas do trabalho de Lucente (2017)[3].

3 Os contornos comprimidos vHL e vLH são propostas ulteriores de Lucente (2017) para representar uma compressão da curva melódica durante a vogal (com tanto o pico quanto a descida de F0 na vogal, em vHL, e adiantamento da subida de F0 numa pré-tônica em vLH), nos casos em que, logo após a realização de uma proeminência, segue uma outra com apenas uma sílaba de intervalo. A compressão está ligada ao sincronismo entre produção segmental e realização da curva de F0 mencionado acima.

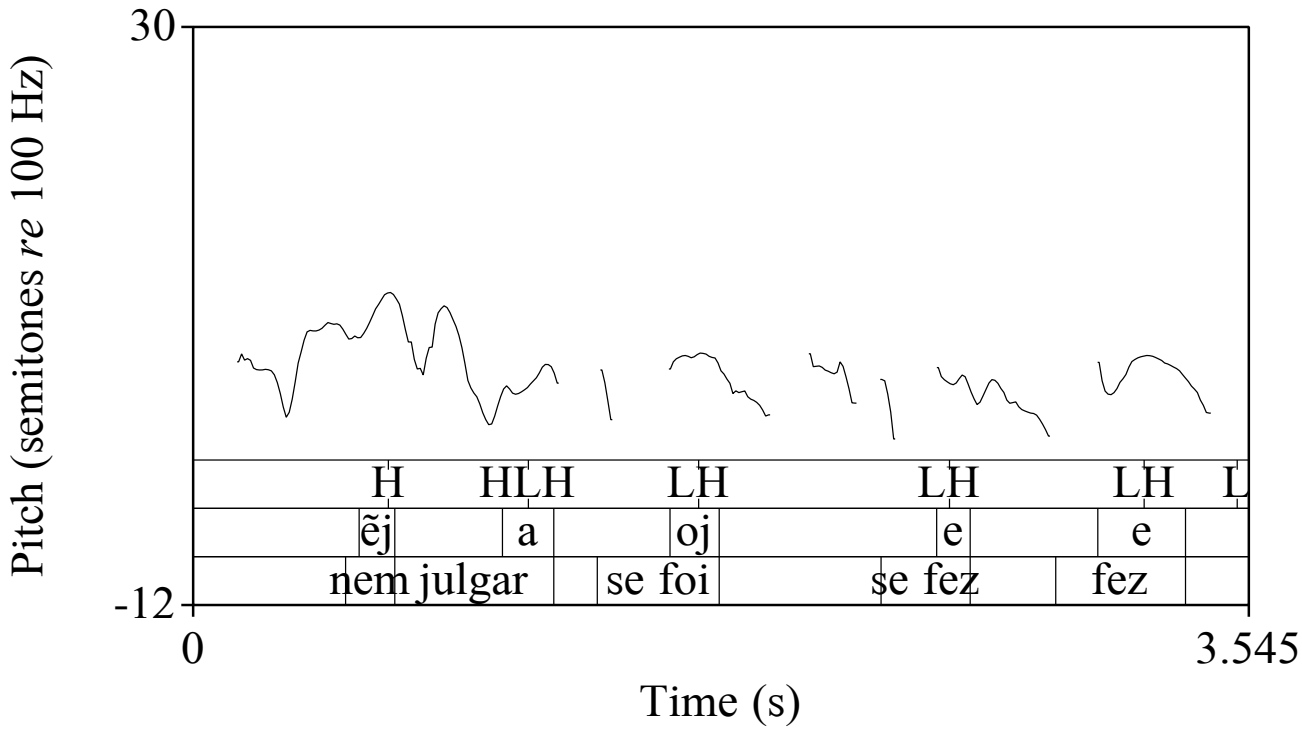

Figura 5.7 – Curva de Fo e camadas de anotação do trecho de enunciado "Não vamos nem JULgar se foi ou não foi, se fez ou não fez." para ilustrar o contorno HLH e o comparar com três instâncias do contorno LH, com a característica descida que precede a subida na partícula "se". Indicam-se os intervalos das vogais tônicas. Trata-se de um locutor paulista que comanda um programa na rádio CBN de São Paulo.

Os contornos de nível, H e L, representam alvos estáticos. Esses contornos são associados a uma proeminência que não tenha a subida obrigatória precedente dos contornos descendentes ou a descida obrigatória precedente dos contornos ascendentes. Podem aparecer acompanhados dos diacríticos ‘ !’ e ‘¡’, indicando *downstep* e *upstep*, respectivamente.

Uma forma de avaliar diferenças no emprego dos contornos anotados com o sistema DaTo é calcular a frequência relativa de cada um deles em trechos de fala. Em seu trabalho de mestrado, Freire (2020) anotou com o sistema DaTo a fala de imigrantes holandeses moradores da Holambra (SP), brasileiros e holandeses moradores da Holanda lendo uma história infantil curta. Os brasileiros não imigrantes leram a tradução da mesma para o PB, enquanto imigrantes e holandeses leram a história em holandês. É preciso ter em mente que, se os brasileiros não tiveram contato com o holandês, nem os holandeses

com o por- tuguês, os imigrantes de Holambra se consideravam holandeses e eram bilíngues ou trilíngues, falando, mesmo que de forma não simétrica, o PB, o holandês e seu dialeto original da Holanda.

O trabalho apontou uma diferença significativa entre as proporções dos contornos >LH e do tom L entre as mulheres imigrantes e as holandesas (nenhuma diferença entre homens holandeses nativos e imigrantes para essa comparação). Quanto à comparação entre imigrantes e brasileiros, o tom H mostrou diferença significativa para os homens. Os resultados desse trabalho apontaram que as mulheres imigrantes estão se aproximando da entoação das brasileiras, pelo uso mais frequente do contorno >LH, característico de marcação de proeminências no PB. Os homens, por sua vez, estão em situação intermediária entre a entoação dos holandeses e a dos brasileiros, embora tendam a se aproximar dos holandeses (FREIRE, 2020).

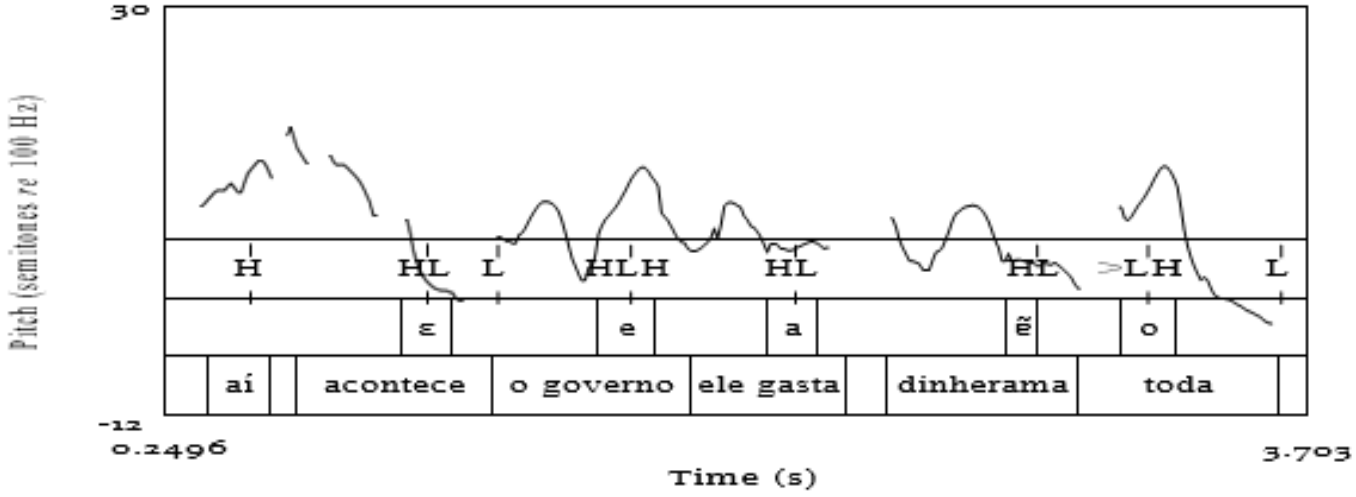

Figura 5.8 – Curva de Fo e camadas de anotação do trecho de enunciado “E aí o que acontece.... O governo, ele gasta aquela dinheirama toda...” para ilustrar a consequência da ênfase em “toda” para o perfil melódico, bem como o emprego do tom de fronteira baixo e a proeminência secundária no sintagma “o governo”. Indicam-se os intervalos das vogais tônicas. Trata-se de um locutor paulista que comanda um programa na rádio CBN de São Paulo.

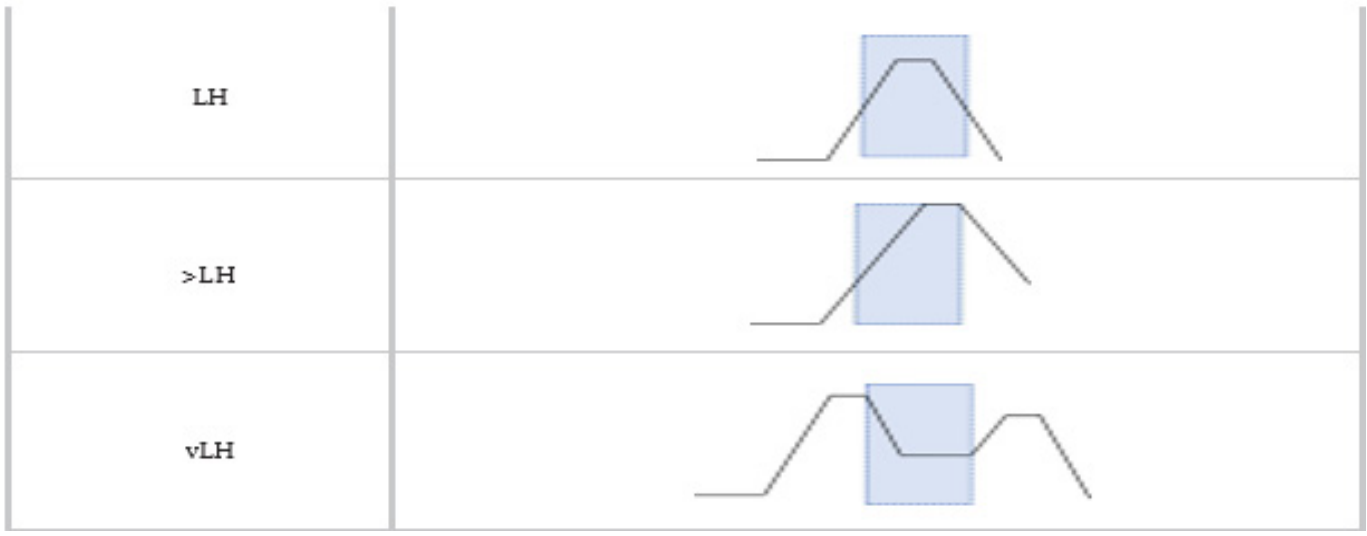

Figura 5.9 – Representação esquemática de contornos ascendentes mostrando forma e alinhamento com a vogal tônica (retângulo sombreado).

A combinação de descrições qualitativas, como essas reveladas pelos sistema de notação melódica, com medidas quantitativas relacionada à F0 fornece elementos muito ricos para a compreensão das diversas funções e usos comunicativos da entoação da fala. Passamos, assim, a apresentar descritores estatísticos de medidas melódicas para apontar diferenças entre situações comunicativas distintas.

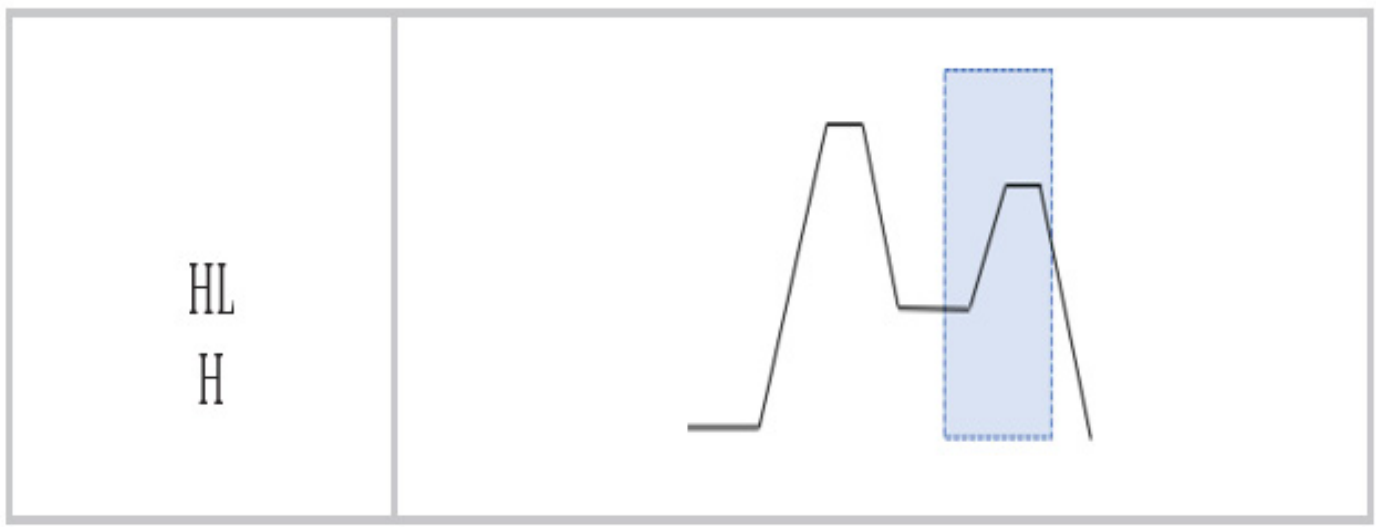

Figura 5.10 – Representação esquemática do contorno HLH mostrando forma e alinhamento com a vogal tônica (retângulo sombreado).

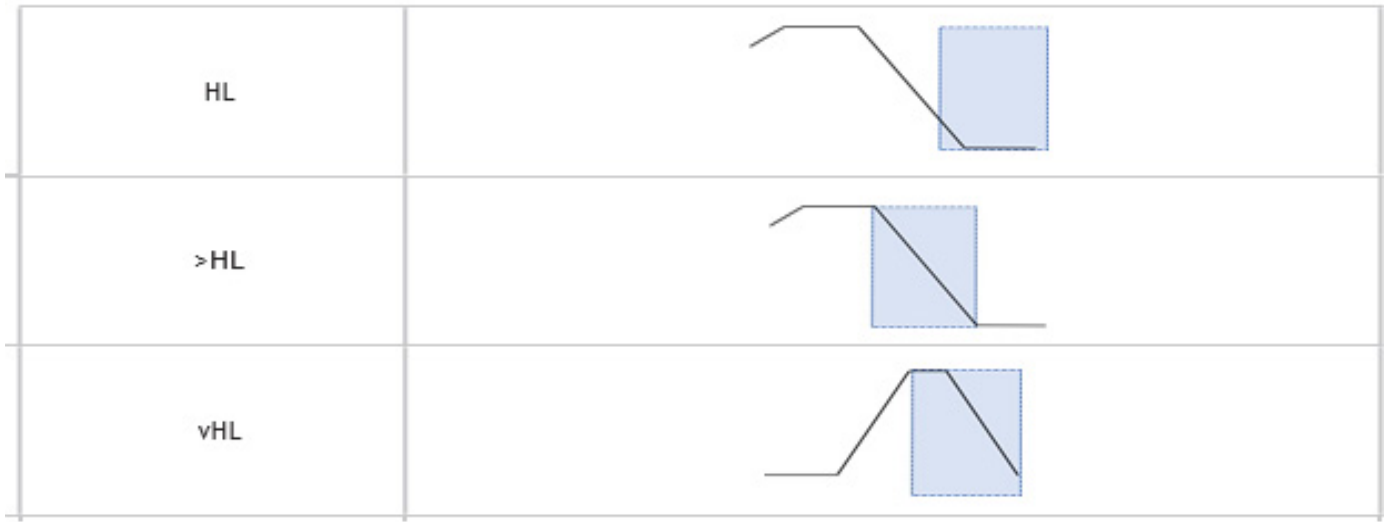

Figura 5.11 – Representação esquemática de contornos descendentes mostrando forma e alinhamento com a vogal tônica (retângulo sombreado).

5.2 Descritores melódicos

Os descritores estatísticos de uma variável assinalam diversos aspectos das amostras de valores da variável. Com o intuito de revelar semelhanças e diferenças entoacionais entre locutores e estilos de elocução, elencamos e explicamos abaixo o interesse de alguns descritores da F0, da primeira derivada da F0 (taxa de mudança da F0), além de outros que podem se revelar interessantes para a pesquisa experimental. Todos esses descritores permitem a realização de testes de estatística inferencial (vide seção 6.1).

5.2.1 Descritores de centralidade

Os descritores de centralidade de uma amostra de valores são medidas que refletem a região dos dados mais frequentes e que estão no centro de uma distribuição assumida como normal (gaussiana), por isso a referência à centralidade. O mais conhecido de todos é a média, mas há ainda a mediana. A mediana é o valor ou ponto central que divide a quantidade de dados em 50% à esquerda e à direita desse

descritor. Embora ambas revelem algo sobre a maior frequência dos valores, a mediana é mais robusta do que a média em pelo menos dois sentidos. Quando há erros de medida, a mediana continua refletindo os valores mais frequentes, enquanto a média é afetada pelo erro de medida, que pode ocorrer para determinada curva de F0. Além disso, se a amostra tem, por exemplo, uma cauda à direita, isto é, alguns poucos valores válidos de F0 bem mais altos do que os demais, a média refletirá esses valores, enquanto a mediana não, desde que os valores mais altos de F0 sejam em pequeno número.

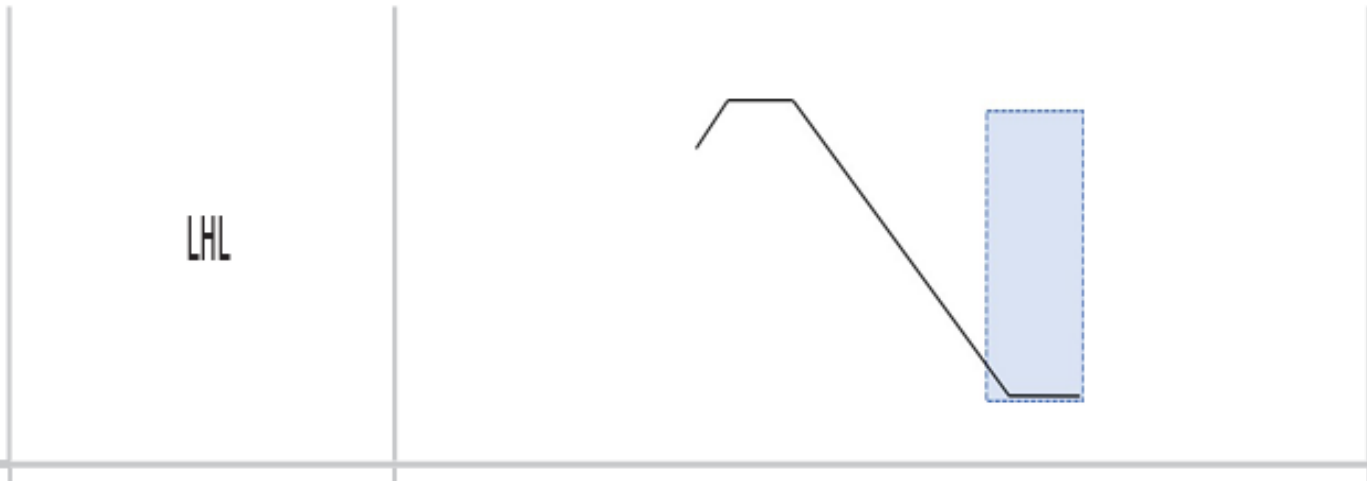

Figura 5.12 – Representação esquemática do contorno LHL mostrando forma e alinhamento com a vogal tônica (retângulo sombreado).

Para exemplificar, suponhamos que o algoritmo que calcula os valores de Fo tivesse dado a seguinte sequência de valores em Hertz (120, 127, 132, 136, 138, 140, 280), com claro erro de salto de oitava (a frequência dobra) no valor de 280. Para essa sequência, a média é 153 Hz e a mediana é de 136 Hz, pois divide exatamente à metade o número de valores à sua esquerda e à sua direita. Observe que 136 Hz é mais semelhante aos demais valores do que 153 Hz, por conta do efeito do erro de salto de oitava que entrou no cálculo da média. É sempre mais seguro, em dados sujeitos a erro, usar a mediana para estimar a centralidade da amostra.

Para ilustrar a utilidade das medidas de centralidade, observe a Figura 5.13, que mostra a curva de F0 de um homem e uma mulher lendo o trecho “Subiu a tribuna”. Os valores são expressos em Hertz (acima) e em semitons (abaixo), uma medida logarítmica de herança musical mais próxima da percepção da frequência (BARBOSA, 2019).

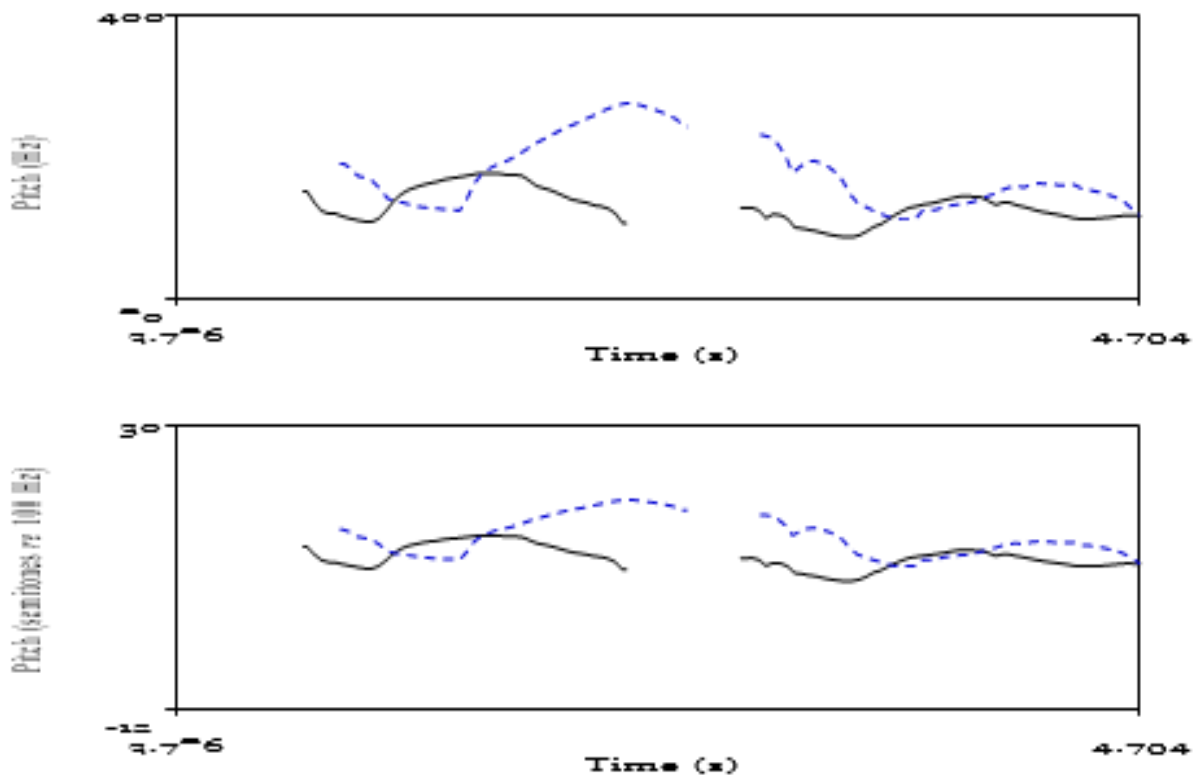

Figura 5.13 – Curva de F0 de um homem (linha cheia) e uma mulher (linha pontilhada) lendo o trecho ‘Subiu a tribuna”. Acima, em Hertz, e, abaixo, em semitons relativos a 100 Hz. Observe os valores da mulher mais altos nos dois casos.

Para o homem, a média de F0 no trecho é de 185 Hz (ou 11 semitons rel. a 100 Hz) e, na mulher, de 222 Hz (ou 14 semitons rel. a 100 Hz). Já o valor da mediana é, para o homem, de 183 Hz (ou 10 semitons relativos a 100 Hz) e para a mulher, de 210 Hz (ou 13 semitons relativos a 100 Hz). Observe que ambas as medidas de centralidade, em ambas as unidades físicas, expressam o mesmo: que o valor médio feminino é maior do que o masculino, refletindo a sensação de *pitch* mais agudo na mulher.

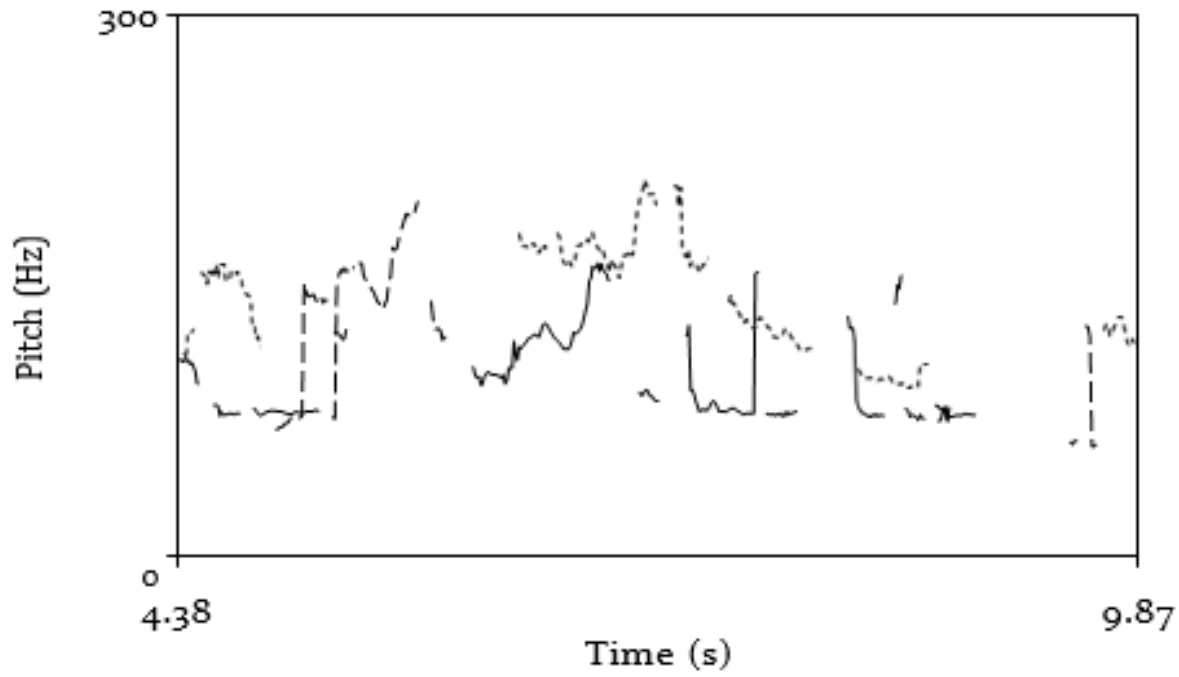

Figura 5.14 – Curva de F0 de um homem de cerca de 50 anos lendo o trecho do Primo Basílio 'Era a primeira vez que lhe escreviam aquelas sentimentalidades" numa leitura para informar (linha cheia) e outra interpretada de forma bem enfática (linha tracejada).

As medidas de centralidade podem ainda ser bem úteis para revelar a mudança global da F0 com a mudança de estilo de elocução, como se vê na Figura 5.14, considerando duas leituras, uma para informar e outra, interpretando o trecho de modo bem enfático. Essa mudança de interpretação faz com que a média passe de 97 Hz, no primeiro caso, para 137 Hz, no segundo. Há também alterações na variabilidade melódica, mas antes vejamos o exame da taxa de variação da F0, calculada sua derivada primeira. A derivada da F0 é relevante para o estudo me- lódico porque essa taxa de variação revela muito sobre a forma como o locutor realiza um acento de *pitch* ou marca uma fronteira em diferentes situações. Essa taxa tem uma relação direta com a articulação dos sons, pois eventos como acentos de *pitch* devem ter seus picos ou vales realizados na proximidade da sílaba tônica, para se fazerem mais audíveis. E para que isso se dê num determinado intervalo silábico, é preciso mudar a taxa de subida ou de descida da F0.

A Figura 5.15 retoma o traçado da F0 do homem que leu o trecho “subiu a tribuna”, acrescentando-se duas curvas: (1) uma curva de F0 suavizada com frequência de corte de 5 Hz para eliminar da primeira derivada valores bruscos sem significado fonético-linguístico, e (2) a primeira derivada de F0, obtida pelo script *fo_extrema* (ARANTES, 2008). Ao longo dessa curva, os valores acima de zero correspondem aos trechos de subida da curva de F0 e, os valores abaixo de zero, aos trechos em que a curva de F0 desce. Além disso, as taxas máximas das subidas e das descidas da curva de F0 são dadas respectivamente pelo picos e vales da derivada.

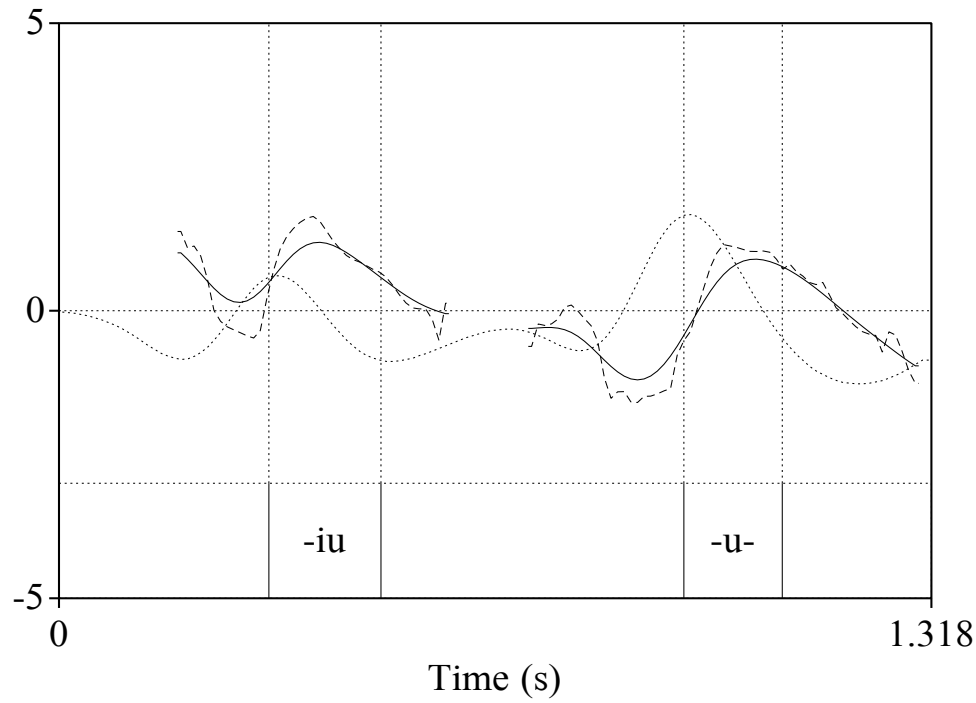

Figura 5.15 – Curva de F0 suavizada com frequência de corte de 5 Hz (linha cheia) e primeira derivada da mesma curva (linha pontilhada) de um homem que leu o trecho ‘Subiu a tribuna”. A curva tracejada é a curva original de Fo, para comparar com a suavizada.

Observe algo muito interessante: os picos das taxas de subida da curva de F0 coincidem com o início das tônicas de “subiu” e “tribuna”, cujos intervalos estão assinalados na figura. O valor máximo da taxa de subida se dá na segunda palavra lexical, “tribuna”, muito embora o pico da curva de F0 seja mais elevado na primeira palavra. A

subida rápida contribui para uma percepção de maior ênfase e para a percepção de um ritmo mais rápido, por contribuir com uma sucessão mais alta de acentos de *pitch*. Por isso, para bem estudar diferenças entre estilos de elocução e entre locutores, é importante o cálculo da média das taxas de subida, bem como das taxas de descida da curva de F0 nos trechos de fala. O script *ProsodyDescriptor*, que desenvolvemos para cálculo de parâmetros prosódicos em trechos previamente segmentados pelo pesquisador, faz esses cálculos automaticamente. Seu funcionamento é descrito no repositório em https://github.com/pabarbosa/prosody-scripts. A variabilidade tanto dessas taxas quanto dos valores da curva de F0 e os valores de seus pontos extremos, que dão a tessitura do locutor, são medidas úteis para o estudo prosódico, e são também calculadas pelo script, juntamente com os demais descritores desse capítulo, e explicados na próxima seção.

5.2.2 Descritores de dispersão e valores extremos

Os descritores de dispersão e os valores-limite de uma amostra de um conjunto de medidas refletem a variabilidade da medida. No caso da F0, quanto mais dispersa, menos monótono soa o trecho de fala. A mais conhecida dessas medidas é o desvio-padrão, que tem uma definição precisa, sendo a média quadrática das distâncias dos valores em relação à média.

Da mesma forma que a média, o desvio-padrão é sensível aos erros de medida e, como alternativa, se pode calcular a semi-amplitude entre quartis (SAQ), definida pela equação 5.1, em que Q1 é o primeiro quartil, o valor que divide o número de dados em 25% à esquerda e 75% à direita, e Q3 é o terceiro quartil, o valor que divide o número de dados em 75% à esquerda e 25% à direita.

$$SAQ_{F0} = \frac{Q3_{F0} - Q1_{F0}}{2} \quad (5.1)$$

Os valores mínimo e máximo de F0 de um trecho de fala definem a amplitude de variação da F0 nesse trecho, enquanto se calculada para um enunciado inteiro, esses limites definem a tessitura do locutor, seus limites superior e inferior da F0. Embora o máximo valor seja condicionado a estilo de elocução e emoção na fala, o mínimo varia bem menos com esses fatores.

Retomando o exemplo da Figura 5.13, o desvio-padrão da F0 no trecho é, para o homem, de 19 Hz (ou 2 semitons relativos a 100 Hz) e, na mulher, de 39 Hz (ou 3 semitons relativos a 100 Hz). Já o valor da SAQ, é, para o homem, de 12 Hz (ou 1,5 semitom relativos a 100 Hz) e, na mulher, de 31 Hz (ou 2,5 semitons relativos a 100 Hz). Quanto aos valores mínimo e máximo, entre 150 e 222 Hz no homem (72 Hz de amplitude de variação no trecho) e entre 171 e 301 Hz na mulher (130 Hz de amplitude de variação no trecho). Em semitons, os extremos estão entre 7 e 14 semitons no homem e entre 9 e 19 semitons na mulher.

O desvio-padrão é um descritor que pode ser usado para calcular a variabilidade das taxas de subida e de descida da F0 num trecho de fala, pois permite revelar aspectos relevantes do modo de falar de alguém numa certa circunstância de comunicação. Além dessa descrição de variabilidade da fala, outros descritores melódicos, como os que seguem, permitem revelar aspectos da vivacidade da fala.

5.2.3 Outros descritores melódicos

A taxa de picos (máximos locais) da F0 por segundo, desde que a curva melódica seja suavizada de forma a ressaltar os picos salientes para a percepção, está ligada ao ritmo da fala também, uma vez que

assinala a maior ou menor produção de acentos de *pitch* por unidade de tempo.

Tanto os valores desses picos locais da Fo quanto os momentos no tempo em que ocorrem podem variar, assinalando uma maior vivacidade ou criatividade do modo de falar, quanto maior for essa variabilidade. Assim, o cálculo dos desvios-padrão dos valores e dos intervalos de tempo de ocorrência de picos locais da Fo podem revelar semelhanças e diferenças melódicas.

Retomando o exemplo da Figura 5.13, há maior variabilidade na fala feminina, pois o desvio-padrão entre os dois picos da Fo é de 64 Hz na mulher e de 18 Hz no homem, significando que ela fez os dois acentos de *pitch* com valores máximos bem mais distintos que o homem.

Um outro exemplo permite observar os valores de taxas em diferentes trechos de fala. Trata-se de trecho inicial da leitura de uma fábula de Esopo em PB, "O vento sul e o sol discutiam qual dos dois era o mais forte", e em francês como língua estrangeira (FLE), *La bise et le soleil se disputaient, chacun assurant qu'il était le plus fort* na Figura 5.16.

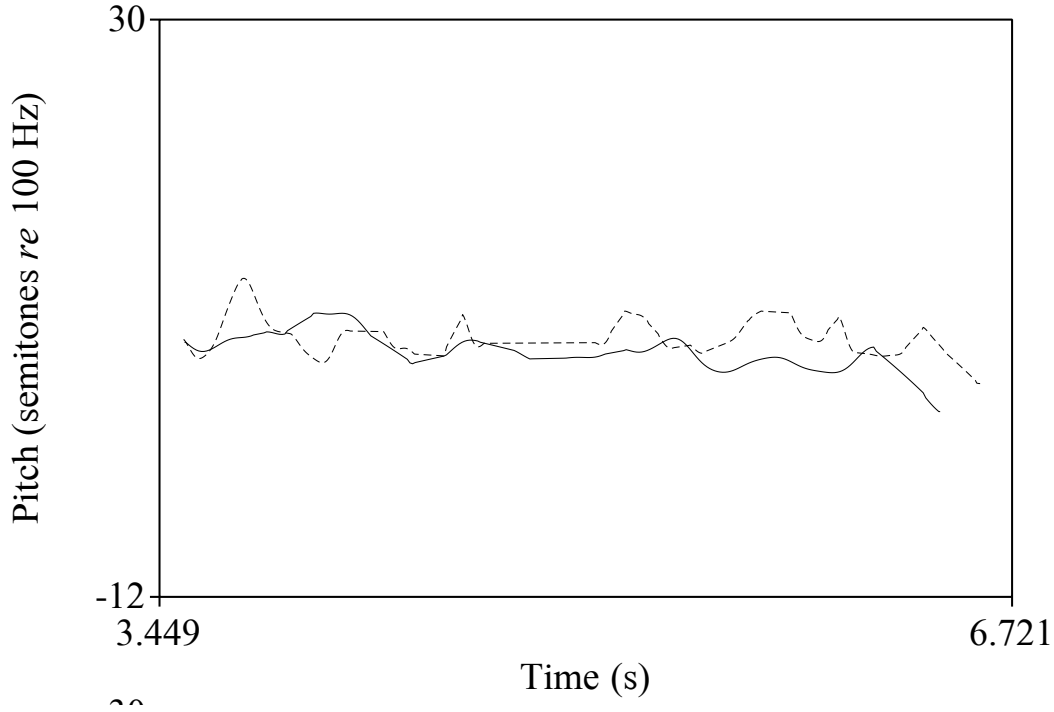

Figura 5.16 – Curva de F0 suavizada (e interpolada) com frequência de corte de 5 Hz de locutor de cerca de 20 anos de trecho de leitura de fábula em PB, "O vento sul e o sol discutiam qual dos dois era o mais forte" (linha cheia), e em francês como língua estrangeira, nível de proficiência básico, *La bise et le soleil se disputaient, chacun assurant qu'il était le plus fort* (linha tracejada). A abscissa se refere ao tempo de leitura em PB; em francês, ela durou 2,3 segundos a mais.

O locutor tinha cerca de 20 anos no momento da leitura e tinha nível de proficiência básico na língua estrangeira. O tempo da leitura em francês está encolhido na figura para caber no mesmo gráfico, tendo sido gastos 2,3 segundos a mais para ler em francês, por conta da inserção de pausas silenciosas e uma articulação mais lenta. No trecho em PB, a taxa de picos de Fo é de 3,5 picos por segundo e, em FLE, de 1,8 picos por segundo. Além disso, o desvio-padrão temporal de sua ocorrência é de 20 ms em PB e 50 ms em FLE, atestando maior lentidão e variabilidade de incidência no tempo em FLE, o que tem a ver mais com a dificuldade de produção na língua.

Outro indicador dessa dificuldade pode ser examinado pelos contornos melódicos que usou nas duas línguas. Utilizando o sistema DaTo que vimos neste capítulo, em francês o tom de fronteira H% foi usado 72% das vezes, indicando não terminalidade de vários trechos de fala (foram 28 tons desse tipo em 39 do total de tipos de con-

torno/tom). As demais proporções importantes foram 13% de tom H para marcar proeminência e 13% de contorno >LH. Já em PB, 21% de todos os tipos de contorno/tom foram do tipo L%, marcador de terminalidade, e as proeminências foram assinaladas a 21% pelo tom H, 21% pelo contorno >LH e 29% pelo contorno HL, muito frequente ao final de trecho assertivo.

Outro descritor interessante para a descrição melódica é o grau de abertura média dos picos da F0, que guardam uma relação com carisma, como mostraram Niebuhr, Thumm e Michalsky (2018), assinalando que aberturas maiores dos picos da F0 tendem a ser associados a uma fala mais carismática, como se pode ver na Figura 5.17. Os picos, possivelmente por serem menos abruptos e talvez, considerados menos contundentes, são interpretados como uma espécie de convite.

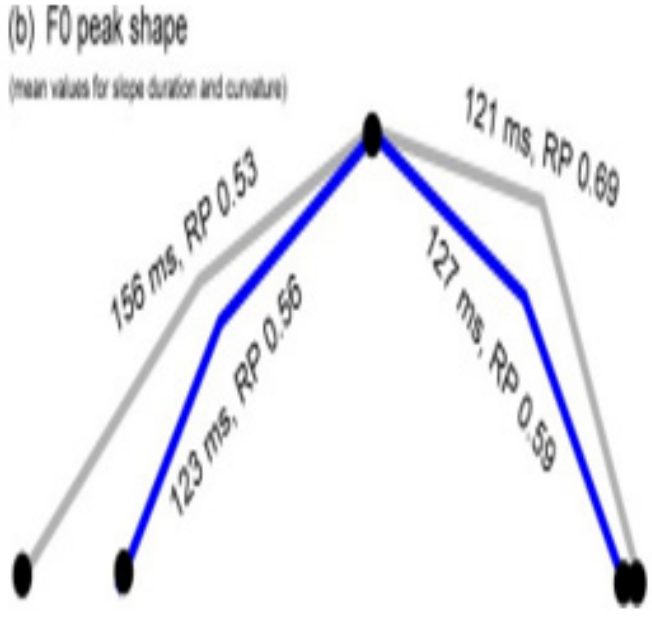

Figura 5.17 – Representação esquemática da abertura de picos da F0, sendo o mais aberto encontrado mais frequentemente em discurso de Steve Jobs, segundo Niebuhr, Thumm e Michalsky (2018). Adaptado da figura 3 do mesmo artigo. As durações são das subidas e descidas, sendo maiores em Jobs. RP é uma medida do grau de convexidade, sendo tanto mais convexa quanto maior a partir de 0,5.

5.2.4 Servindo-se dos descritores melódicos

No que segue, examinaremos descritores melódicos em diferentes tipos de contrastes, para ver o que revelam, em seu conjunto, sobre a entoação da fala. A Figura 5.18 mostra os valores de desvio-padrão da Fo e média das taxas de variação positivas de F0 (média dos trechos positivos da primeira derivada de F0) de uma interpretação da lenda do uirapuru por Camila Pitanga (LIS-PECTOR, 2000). O trecho lido foi dividido em trechos discursivos hierarquizados segundo a proposta de Grosz e Sidner (1986), por isso, embora os trechos apareçam na sequência como foram ditos, da esquerda para a direita, sua numeração reflete seu lugar na hierarquia temático-discursiva. O que importa aqui é a observação das maiores mudanças, que são reflexos de mudanças na interpretação em função do conteúdo.

Selecionamos para ilustrar nas figuras os parâmetros melódicos com mudanças mais bruscas para determinada passagem entre os trechos. De fato, do trecho DS10, em que se relata o lançamento de uma flecha para matar o uirapuru, para o trecho DS5, em que se introduz algo inesperado que será relatado na sequência, enquanto o desvio-padrão da F0 passa de 3 a 3,4 semitons, o valor médio das taxas de subida da curva da F0 passa de 5 a 8 Hertz/quadro, apontando que, embora os trechos tenham extensões temporais muito distintas, como se vê na Figura 5.19, o trecho 5 tem uma subida central da F0 bem mais rápida. Pode-se ver aí também que, por conta dessa maior subida, a variabilidade da F0 aumenta no trecho mais curto. Os dois trechos podem ser ouvidos no repositório do livro como **DS10Pitanga** e **DS-5Pitanga**

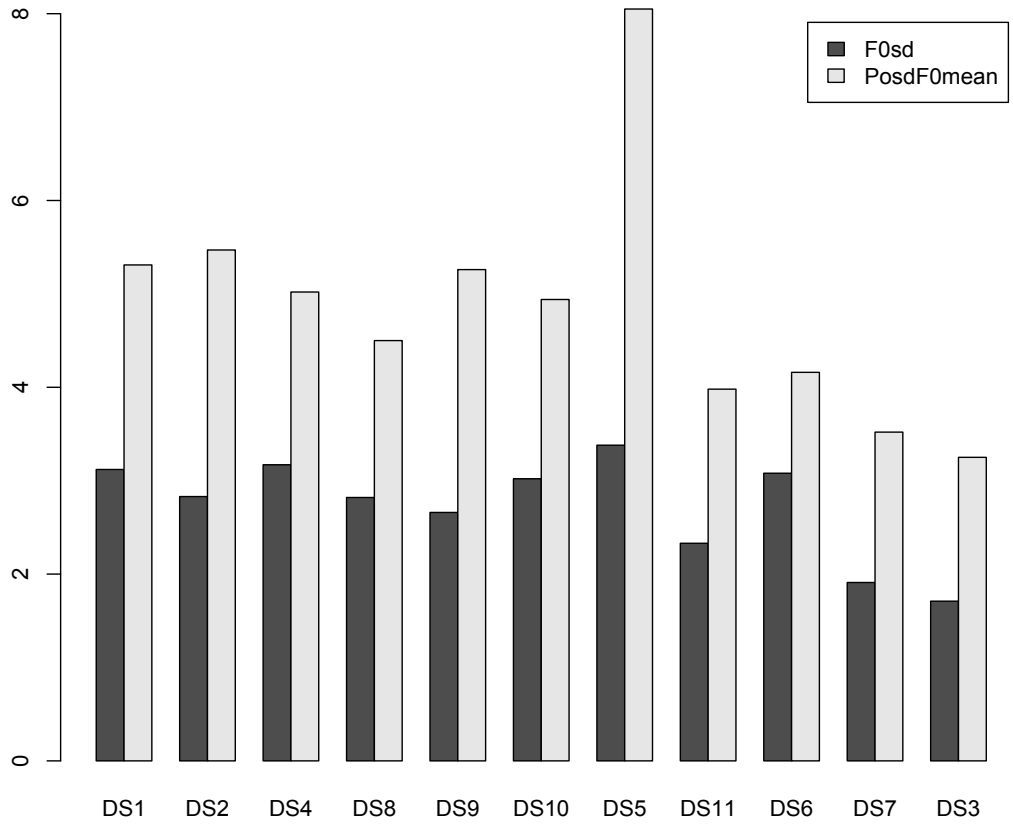

Figura 5.18 – Valores de desvio-padrão de F0 (semitons), barras escuras, e média das taxas de variação positivas de Fo (Hertz/quadro), barras claras, dos trechos discursivos da interpretação da lenda do uirapuru por Camila Pitanga.

Outros descritores melódicos que podem ser comparados estão assinalados na Figura 5.20: enquanto a taxa média de produção dos picos da F0 se mantém praticamente a mesma ao longo de toda a interpretação, a variabilidade dos valores dos picos da Fo vai particularmente aumentando do trecho DS8 até DS5, para indicar justamente, pela vivacidade que essa variação provoca na percepção, o elemento de surpresa que será relatado a partir do trecho DS11, que segue DS5 na linha temporal.

Com o fim de ilustrar a utilidade em conjugar descritores melódicos e duracionais, vistos no capítulo anterior, vamos ver agora o que acontece com a melodia e a pausa numa fala telejornalística. A locutora é uma jornalista de Campinas, cujos dados fazem parte do

trabalho de Mareüil e Barbosa (2018). Ela foi convidada a ler o texto da fábula de Esopo adaptada, "O Vento Sul e o Sol", de duas maneiras: leitura habitual (etiquetada 'normal') e, em sequência, leitura imitando uma locução telejornalística. A leitura foi dividida em 10 trechos de mesmo conteúdo que incluíram ao menos uma pausa silenciosa em um dos estilos. Em outra camada de anotação, as pausas silenciosas produzidas foram segmentadas, não havendo nenhuma pausa preenchida nas duas leituras. O script *Prosody Descriptor* permitiu o cálculo de 12 parâmetros melódicos e dois parâmetros relativos às pausas. Desses, onze parâmetros melódicos apresentaram diferença significativa entre os estilos e um parâmetro, entre os dois relativos às pausas, a duração média. Os diagramas de blocos podem ser vistos na Figura 5.21[4].

4 De cima para baixo e da esquerda para a direita as variáveis são: mediana de F0 em Hertz (f0med), máximo de F0 no trecho em Hertz (f0max), mínimo de F0 no trecho em Hertz (f0min), desvio-padrão de F0 no trecho em Hertz (f0sd), desvio-padrão de máximos de F0 no trecho em Hertz (sdf0peak), grau de abertura do pico de F0 em Hertz (f0peakwidth), desvio-padrão das posições dos picos de F0 no trecho em segundos (sdtf0peak), média da primeira derivada positiva de F0 em Hertz/quadro (df0posmean), desvio-padrão da primeira derivada positiva de F0 no trecho em Hertz/quadro (df0sdpos), média da primeira derivada negativa de F0 em Hertz/quadro (df0negmean), desvio-padrão da primeira derivada negativa de F0 no trecho em Hertz/quadro (df0sdneg) e duração de pausa silenciosa em ms (durSIL).

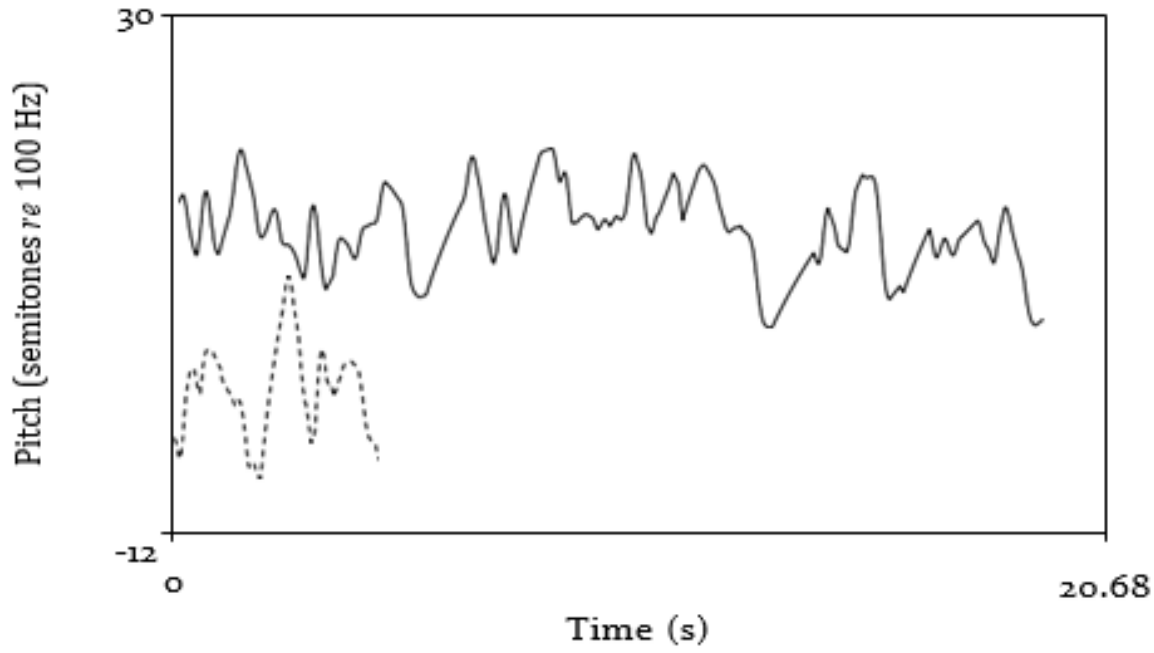

Figura 5.19 – Curvas de F0 suavizada (e interpolada) com frequência de corte de 5 Hz e traçados em semitons rel. 100 Hz dos trechos discursivos 10 (linha cheia) e 5 (linha tracejada) da interpretação da lenda do uirapuru por Camila Pitanga. Escalas verticais deslocadas para facilitar a visualização das duas curvas.

Com exceção das variáveis "desvio-padrão das posições dos picos de F0" e "mínimo da F0", as diferenças aqui encontradas são todas significativas para o nível de significância de 5% em testes de Wilcoxon e revelam o que se vê claramente na Figura 5.21: no estilo telejornalístico, os seguintes parâmetros têm valores maiores do que na leitura habitual, a saber, mediana da F0, máximo da Fo, desvio-padrão da F0, desvio-padrão dos máximos da F0, média e desvio-padrão da taxa de subida da F0, média (em módulo) e desvio-padrão da taxa de descida da F0. São menores no estilo telejornalístico o grau de abertura dos picos da F0 e a duração das pausas silenciosas. Quanto às variáveis não significativas, enquanto o mínimo de Fo deva ser investigado para ver se seria menor na fala telejornalística com mais sujeitos, a variabilidade entre os trechos do próprio desvio-padrão das posições dos picos da F0 deve também merecer um exame num maior volume de dados, pois pode estar relacionada a escolhas variadas sobre em que palavras dar ênfase.

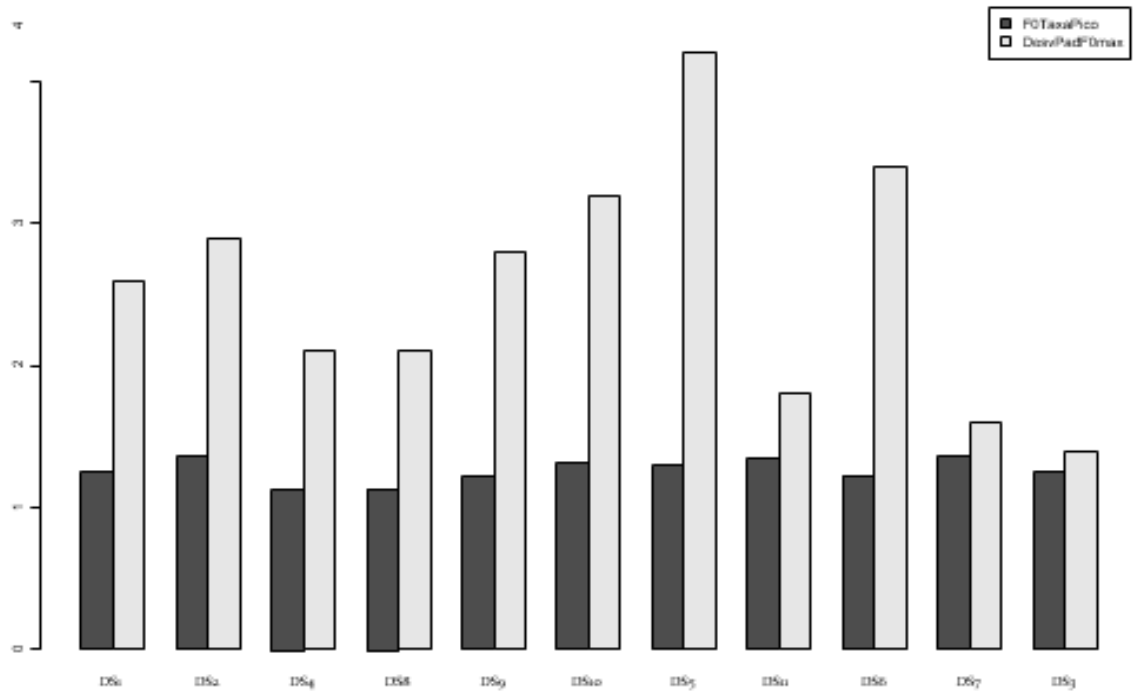

Figura 5.20 – Valores de taxa de picos locais de Fo (picos/segundo), barras escuras, e desvio-padrão de valores máximos de Fo (semitons), barras claras, dos trechos discursivos da interpretação da lenda do uirapuru por Camila Pitanga.

Do ponto de vista da percepção, esses resultados revelam que, comparada com a leitura não profissional, a fala telejornalística da jornalista é mais aguda, com entoação mais variável (evitando a monotonia e criando momentos de surpresa), subidas da curva melódica mais rápidas e variáveis, contribuindo para maior vivacidade. Ela também faz picos da F0 menos abertos, o que pode assinalar maior atratividade na fala, e as pausas são mais curtas, o que a torna mais ágil. Os áudios podem ser ouvidos no repositório nos arquivos **FalaJornal** e **FalaHabitual**.

Além de parâmetros melódicos *stricto sensu*, os parâmetros de qualidade de voz (QV) concernem à atividade vibratória das pregas vocais e, por isso, serão descritos nesse capítulo. Eles dizem respeito à alteração de longo termo do modo de fonação e a suas consequências acústicas. Para uma excelente discussão sobre o tema, ver os trabalhos de Fujimura (1988), Fujimura e Hirano (1995), Titze (2000), Esling e Harris (2005) e Kreiman e Sidtis (2011).

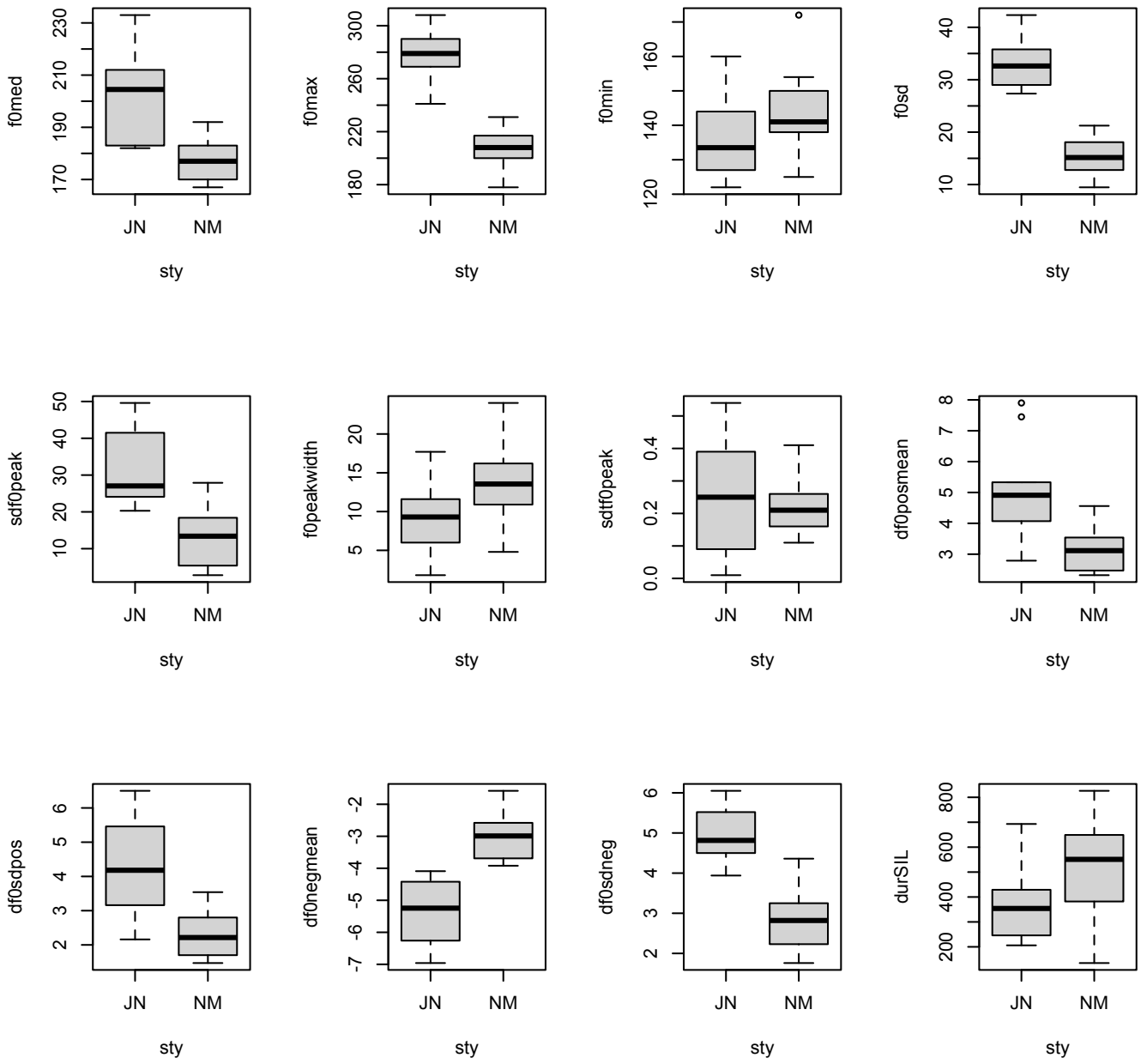

Figura 5.21 – Diagramas de blocos (*boxplots*) de onze variáveis melódicas e uma duracional da leitura em dois estilos e elocução, telejornalístico (JN) e habitual (NM).

5.3 Descritores acústicos de Qualidade de Voz (QV)

Os descritores acústicos da qualidade de voz (QV) são calculados a partir do sinal de fala, mas refletem direta ou indiretamente o que se passa no sinal glotal. Para uma leitura didática,

sem deixar de ser aprofundada, recomendamos o panorama dado por d'Alessandro (2006).

De interesse prosódico são as mudanças no modo de fonação, também chamada de qualidade de voz, como no caso da voz modal (*modal voice*), voz soprosa (*breathy voice*) e voz laringalizada (*creaky voice*). Em termos articulatórios, esses modos de fonação alteram o quociente de abertura do ciclo glotal (OQ, na sigla em inglês para *Open Quoti- ent)*. O OQ é a razão entre o intervalo de tempo em que as pregas vocais estão abertas em relação ao período glotal. Para a fonação modal (normal), esse quociente é cerca de 50%, enquanto para a fonação soprosa, por exemplo, o OQ é bem maior.

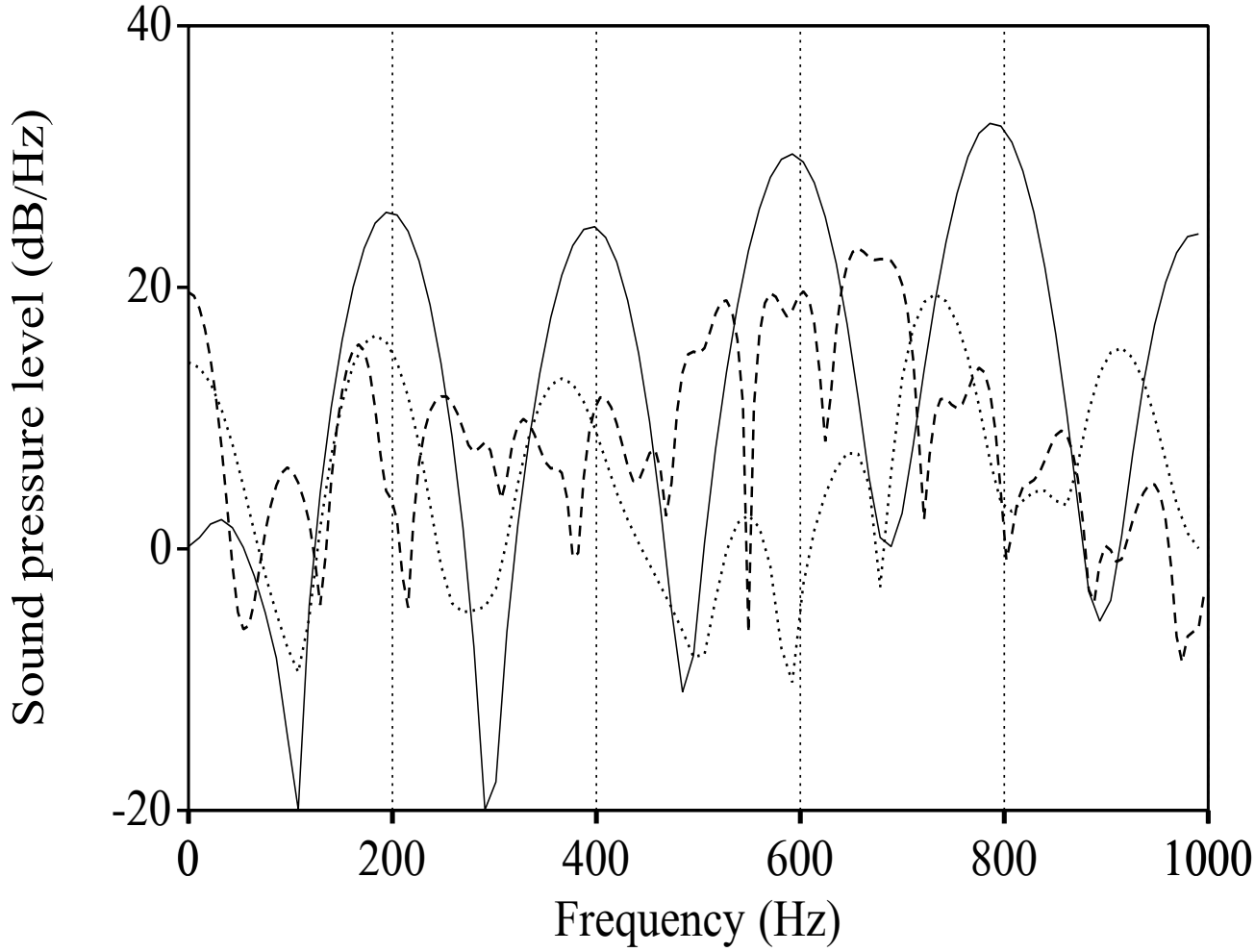

Figura 5.22 – Espectros de Fourier na região média da vogal [ɛ] da palavra "fonética" pronunciada com três modos de fonação na frase "O objeto de estudo da Fonética é essa complexa, variável e poderosa face sonora da linguagem, a fala." por uma fonoaudióloga paulista de cerca de 40 anos na época da gravação. A linha cheia se refere à fonação modal, a linha tracejada à fonação laringalizada e a linha pontilhada à fonação soprosa.

Autores como Shue, Chen e Alwan (2010) mostraram uma correlação de cerca de 65% entre as diferenças entre o primeiro (H1) e segundo harmônicos (H2) do espectro de uma vogal emitida em determinado modo de fonação com o quociente de abertura. Embora haja grande variação interindividual nesse tipo de correlação, como mostraram logo depois Kreiman et al. (2012), vale a pena o cálculo dessa medida, denominada de H1-H2, em vogais com F1 elevada (vogais abertas), para evitar o efeito do primeiro formante sobre a amplitude dos dois primeiros harmônicos.

A Figura 5.22 mostra os espectros de Fourier na região média da vogal [ɛ] da palavra "fonética" pronunciada numa frase-veículo por uma fonoaudióloga paulista. Observe que as amplitudes do primeiro harmônico nas fonações modal e soprosa são maiores do que a do segundo harmônico, mas a relação inversa se dá na fonação laringalizada. Os valores das diferenças de amplitude calculados numa janela de 50 ms centrada na vogal são de 2,0 dB (modal), 6,3 dB (soprosa) e -6,6 dB (laringalizada), o que vai na direção do que tem sido observado na literatura. Os áudios correspondentes podem ser ouvidos no repositório como **VQPBModal**, **VQPBLaringalizada** e **VQPBSoprosa**.

Uma relação no mesmo sentido, maior na fala soprosa e menor na fala laringalizada, foi encontrada no estudo de Shue, Chen e Alwan (2010), embora tenham usado a vogal [i] sustentada[5].

Uma outra medida acústica que reflete o modo de fonação é o pico de proeminência cepstral (CPP, na sigla em inglês para *Cepstral Prominence Peak*). Como vimos em outro lugar (BARBOSA; MADUREIRA, 2015, p. 162-167), o cepstro é técnica de análise espectral que permite separar os sinais da fonte sonora do efeito de filtragem do trato vocal, permitindo observar isoladamente as características sono-

5 Embora haja técnicas para amenizar o efeito dos formantes, recalculando a amplitude dos harmônicos sem uma estimativa desses efeitos, quantidade que tem a sigla H1*-H2*, é recomendável evitar esse artifício, por isso a recomendação de escolher as vogais baixas.

ras do som laríngeo. Um ciclo glotal regular dá maiores picos de proeminência cepstral do que um ciclo irregular. Assim, vozes roucas e soprosas têm valores menores para CPP. Para as mesmas vogais [ɛ] da palavra "fonética" acima, os valores de CPP foram de 26,7 dB (modal), 13,1 dB (soprosa) e 18,1 dB (laringalizada), tendo valor máximo na voz modal e mínimo na voz soprosa, como previsto.

Medidas mais diretas da perturbação da vibração das pregas vocais são as medidas de *jitter* e *shimmer*. O *jitter* é a medida da irregularidade nos períodos glotais. Pode ser calculado ciclo a ciclo (*jitter* local) ou considerando janelas contendo 3 ou 5 ciclos glotais ou ainda a diferença média entre os ciclos numa determinada janela. Pode ser expresso em unidades de tempo ou de modo percentual. Quanto maior seu valor, mais irregular a vibração, sendo menor na voz modal e maior na voz rouca de vibração irregular. Para os trechos lidos pela fonoaudióloga que estamos usando para ilustrar as medidas, o valor do *jitter* local percentual nos três modos de fonação são 1,9% (modal), 2,1% (soprosa) e 6,2% (laringalizada), conforme esperado.

O *shimmer* é a medida da irregularidade nas amplitudes dos ciclos glotais. Pode ser calculado ciclo a ciclo (*shimmer* local) ou considerando janelas contendo 3, 5, 7 ou 11 ciclos. Pode ser expresso em dB ou de modo percentual. Quanto maior seu valor, mais irregular a amplitude vibratória, sendo menor na voz modal e maior na voz rouca de vibração irregular, especialmente com tensão laríngea. Para os trechos lidos pela fonoaudióloga, o valor do *shimmer* local percentual nos três modos de fonação foram: 9,1% (modal), 13,4% (soprosa) e 22,9% (laringalizada). Observe que tanto para o *jitter* quanto para o *shimmer* os maiores valores ocorrem na voz laringalizada, que foi feita de forma bem tensa em sua fala. A voz com vibração mais regular das pregas vocais é mesmo a modal, como esperado.

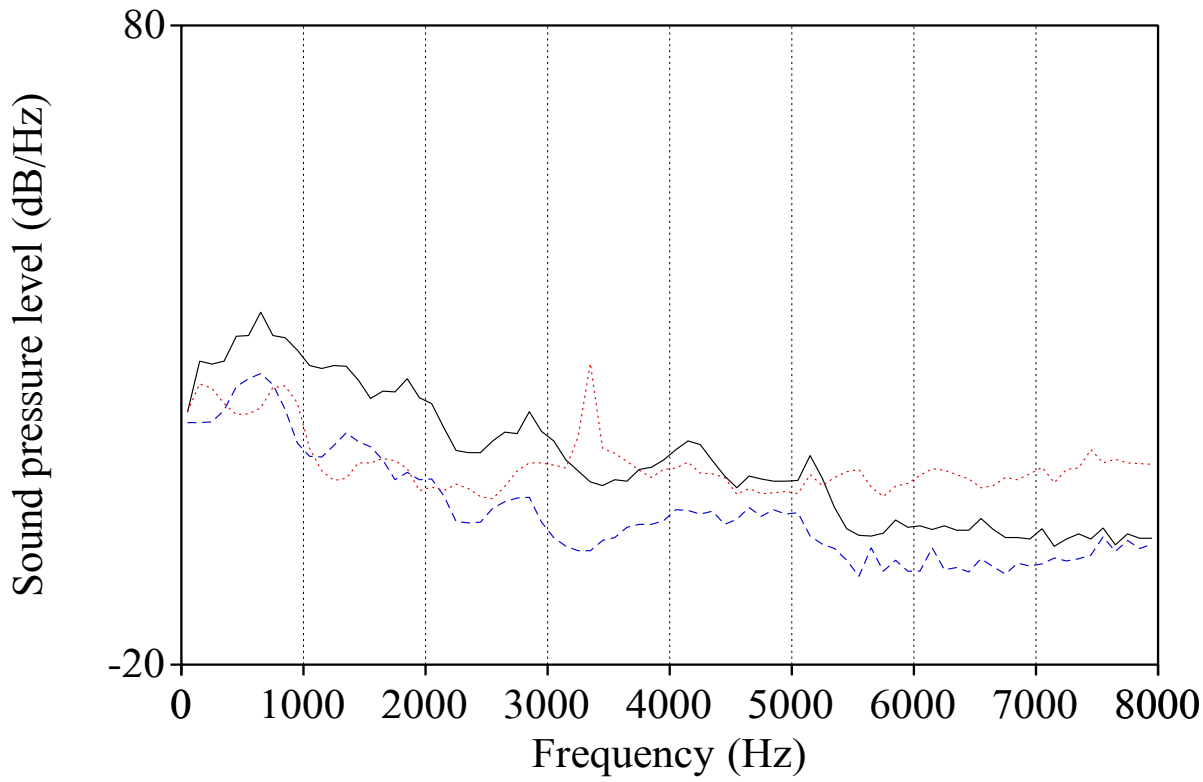

Figura 5.23 – Espectros médios de longo termo (LTAS) da frase "O objeto de estudo da Fonética é essa complexa, variável e poderosa face sonora da linguagem, a fala." pronunciada em três modos de fonação por uma fonoaudióloga paulista de cerca de 40 anos na época da gravação. A linha cheia se refere à fonação modal, a linha tracejada azul à fonação laringalizada e a linha pontilhada vermelha à fonação soprosa. Observe as diferenças de energias de 0 a 1000 Hz e depois nas faixas 1000 a 4000 Hz.

Os dois próximos descritores se referem ao descompasso em energia entre diferentes bandas espectrais. São a ênfase espectral e a inclinação do espectro de longo termo (LTAS, na sua sigla em inglês, *Long-Term Average Spectrum*). A ênfase espectral (*spectral emphasis*, em inglês) foi definida por Traunmüller e Eriksson (2000) como uma medida acústica indireta do esforço vocal. De forma simplificada, pode ser definida como a diferença entre a intensidade total de um som e a intensidade numa faixa de frequência baixa para englobar toda variação da frequência fundamental, segundo a equação 5.2.

$$\text{Ênfase espectral} = I - I_0 \qquad (5.2)$$

Em que I é a intensidade até a frequência máxima do sinal e I_0

é a intensidade do som de 0 a 400 Hz, ambas em dB, com o limiar da banda baixa de 400 Hz fixado para melhor operacionalizar os cálculos[6]. Na prática, uma vez que os sinais de fala são armazenados de forma digital, a frequência máxima disponível para análise é a frequência de Nyquist, que é a metade da frequência de amostragem. Os valores de ênfase espectral para os trechos ilustrados aqui são de: 9,5 dB (modal), 6,6 dB (soprosa) e 8,1 dB (laringalizada), com esforço maior na fala modal.

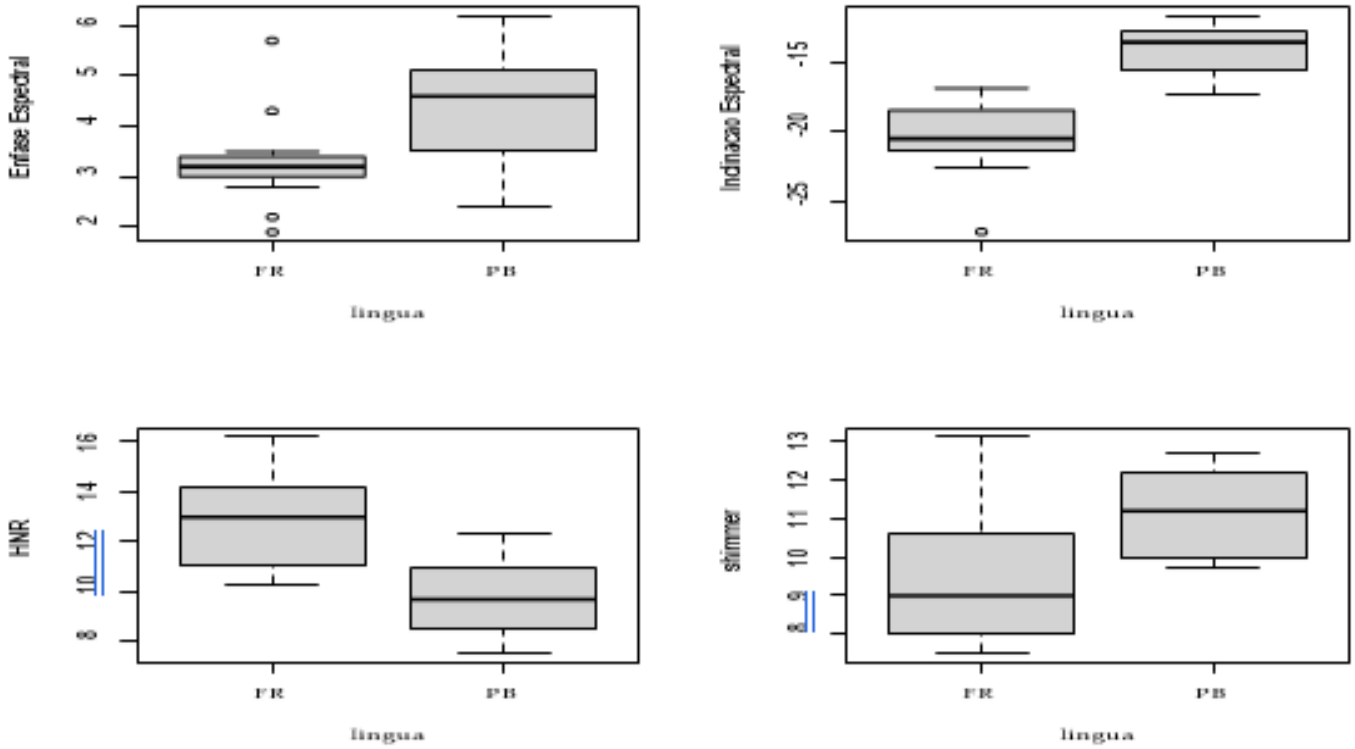

Figura 5.24 – Diagramas de blocos de medidas de QV para estudante de francês em nível básico em leituras em PB e em francês (FR) de fábula de Esopo. As medidas são, de cima para baixo e da esquerda para a direita: ênfase espectral, inclinação espectral, HNR, as três em dB, e *shimmer* em porcentagem.

A inclinação do espectro de longo termo se refere à diferença de energia média entre duas bandas de frequência desse mesmo espetro. Quanto menos inclinado for esse espectro, mais energia se encontra na banda de mais alta frequência, que é um reflexo da produção de

6 Os autores definiram esse limiar exatamente como 1,43 x F0 médio no trecho, isto é, 43% acima da frequência fundamental média no trecho, mas também testaram com valores de 50% e valores fixos como os 400 Hz sugeridos aqui. Os resultados de correlação com o esforço vocal foram praticamente os mesmos.

componentes de frequências altas com amplitude elevada, uma consequência seja de maior esforço vocal, seja de maior soprosidade. Uma medida muito usada da inclinação é a diferença de energia entre a banda de 1000 a 4000 Hz e a banda de 0 a 1000 Hz. Observe na Figura 5.23 que essa diferença é menor na fala soprosa. De fato os valores para o cálculo sobre toda a frase são: -9,6 dB (modal), -7,1 dB (soprosa) e -12,6 dB (laringalizada), revelando a maior energia produzida na banda de 1000 a 4000 Hz por conta da soprosidade.

Outra medida acústica muito útil de qualidade de voz é a razão harmônico-ruído (HNR, da sigla do nome em inglês, *Harmonic to Noise Ratio*), que mede a relação, em dB, da energia dos harmônicos num trecho de fala e e energia do ruído, em toda a faixa espectral. Diferentemente das medidas de ênfase e inclinação espectrais, que consideram bandas diferentes e não separam o que é devido apenas à energia de ruído do que é devido apenas à energia harmônica, a HNR faz isso. Os valores de HNR para os trechos ilustrados aqui são de: 10,2 dB (modal), 7,6 (soprosa) e 2,3 dB (laringalizada), revelando que há bastante ruído na fala laringalizada, afetando todo o espectro de fala. O efeito de ruído em baixa frequência nessa fala não é considerado de forma separada nas medidas de ênfase espectral e inclinação espectral, por isso a diferença com os dois outros modos de fonação não são tão grandes quanto a mostrada com essa medida.

Retomando o exame da fala de um estudante de francês cuja leitura em PB e em francês usamos na seção 5.2.3 para ilustrar a importância de se usar outros descritores melódicos, observa-se em sua fala em língua materna (PB) uma grande frequência de laringalização, como o leitor pode conferir nos áudios nas duas línguas, **MCFR** e **MCPB**. Essa laringalização é acompanhada de maior tensão laríngea e possíveis irregularidades de vibração das pregas vocais. De fato, os diagramas de bloco dos quatro descritores de QV mostrados na Figura 5.24 apontam para essa conclusão para sua leitura em PB, a qual exibe maior ênfase espectral, um espectro de longo termo menos inclinado,

um HNR menor que revela mais ruído na fala e um *shimmer* mais alto, apontando para maior irregularidade de amplitude do ciclo glotal.

Parâmetros melódicos e de qualidade de voz são muito usados para compor personagens na indústria de entretenimento. Os diagramas de bloco da Figura 5.25 mostram claramente os recursos empregados por Camila Pitanga para interpretar três personagens da história de Pedro Malazarte: o personagem principal tem uma fala mais aguda com melodia mais variável e ainda menos ruidosa (maior valor de HNR) e com ciclos glotais mais regulares do que a fala dos outros dois personagens, o marido e sua mulher. Esse dois outros personagens são diferenciados sobretudo pela menor variabilidade melódica na mulher (maior valor de desvio-padrão da F0) e a fala mais ruidosa do marido.

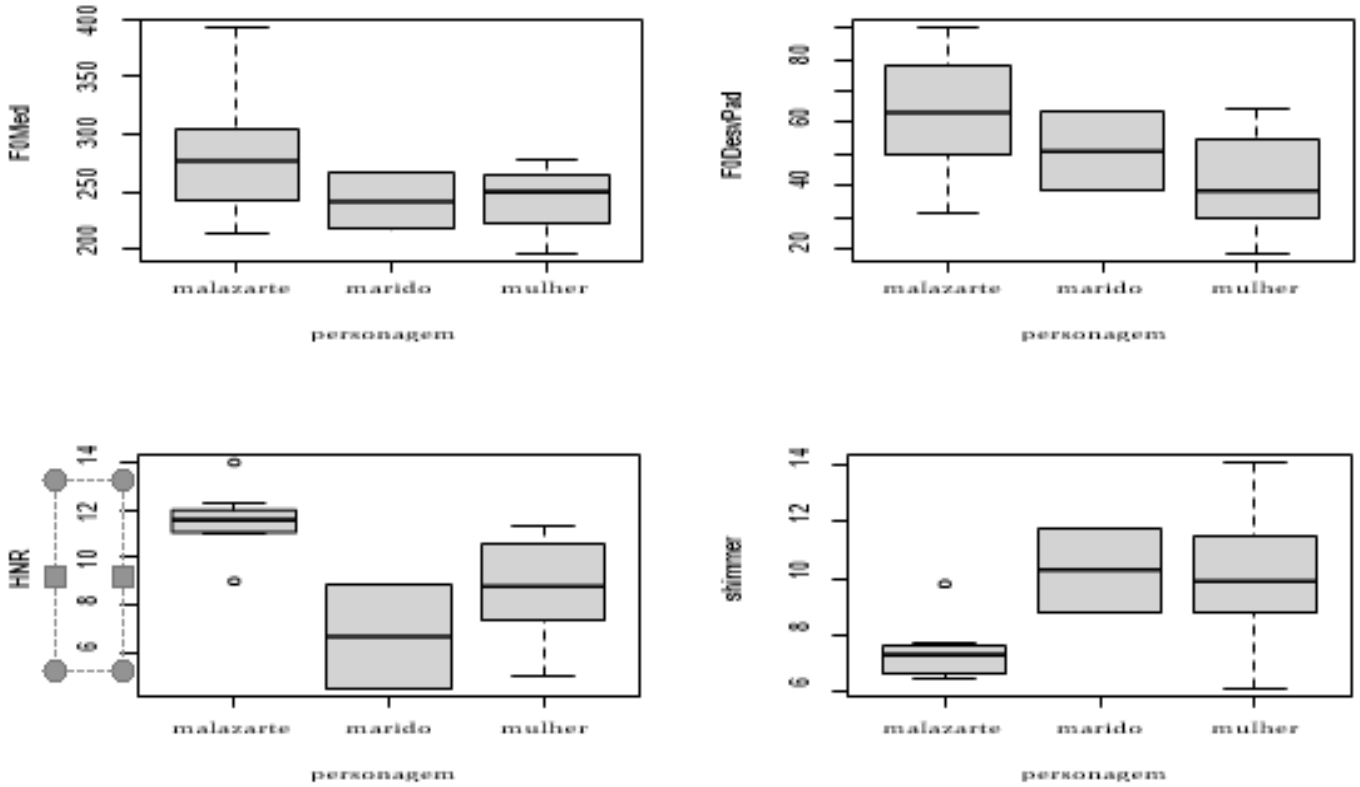

Figura 5.25 – Diagramas de blocos de medidas melódicas e de qualidade de voz na interpretação de três personagens por Camila Pitanga na história de Pedro Malazarte. As medidas são a mediana da Fo em Hertz, o desvio-padrão da Fo em Hertz, a razão harmônico-ruído em dB e o *shimmer* em porcentagem.

5.4 Prelúdio para o próximo capítulo

No próximo capítulo vamos aplicar, no contexto de exemplos de desenho experimental, o que aprendemos com todas as medidas prosódico-acústicas vistas até o momento. Começaremos com uma apresentação sucinta das principais técnicas de análise estatística inferencial usadas na área de prosódia experimental.

Capítulo 6

Elementos de estatística inferencial

Toda investigação experimental envolve a questão da reprodutibilidade dos achados e, portanto, a relação entre a amostra que foi coligida e a população de dados que a subjaz. Assim, este capítulo se dedica a apresentar os conceitos mais relevantes da estatística inferencial. Dentre esses, um componente importante para guiar a interpretação dos resultados fundamentados nas medidas que foram apresentadas neste livro são os testes de hipóteses da estatística inferencial. Enquanto a estatística descritiva se serve de descritores de diferentes ordens como centralidade, dispersão, assimetria e achatamento de uma amostra, além de se servir de histogramas para quantificar as amostras de dados, a estatística inferencial dá um passo a mais por investigar a reprodutibilidade da amostragem.

A estatística inferencial, por meio de testes de hipóteses fundamentados em probabilidade, relaciona as diferenças amostrais dos descritores com as diferenças populacionais, permitindo a continuidade da experimentação. Para bom proveito de um estudo, seu uso não deve se limitar a apontar se diferenças são significativas ou não para um certo nível de significância, mas deve ser completado com a exploração de outros aspectos ainda pouco abordados em nossa área, como intervalo de confiança (*confidence interval*) e tamanho do efeito (*effect size*). Por isso, passaremos a discutir os usos da estatística inferencial na área de prosódia experimental precedidos de considerações gerais sobre estatística. Para uma revisão mais detalhada, refira-se o leitor a obras como as de Crawley (2005), Woods, Fletcher e Hughes (1986), Bunschaft e Kellner (2001), Baayen (2008), Dowdy e Wearden (2001), Johnson (2011), e Rietveld e Hout (1993).

Ao final do capítulo, dois experimentos serão descritos em detalhe, precedidos de uma apresentação sucinta das teorias e observações que os motivaram, para que o leitor possa acompanhar todas as fases do ciclo experimental e possa formar um senso crítico. Em seguida, motivamos o leitor a explorar novas áreas de investigação que serão importantes tanto para a compreensão da fala e sua expressividade quanto para o que podemos aprender com línguas de que temos pouco conhecimento, como as línguas regionais na França. Essa motivação dupla é uma homenagem que faço a dois colegas muito engajados na área.

6.1 Testes estatísticos inferenciais para investigação prosódica

Apontamos em outro lugar (BARBOSA, 2013) que experimentação e estatística inferencial devem estar interligados de tal modo que "one of the first things which the beginner must grasp is that statistics need to be taken into account when the experiment is being planned, or else the results may not be worth treating statistically." (BEVERIDGE, 1957, p. 19).

Isso significa que as hipóteses científicas de um experimento devem ser colocadas de tal maneira que possam ser testáveis por um procedimento estatístico inferencial específico, estabelecendo uma ponte entre a amostra e a população visada. A população estatística não é um conjunto de locutores ou ouvintes, mas um conjunto de dados potencialmente infinito ou tão grande que não se possa medir com os recursos disponíveis e a respeito do qual quer se descobrir algo. Por exemplo, a população que subjaz às durações silábicas da leitura de um texto por um locutor é formada por todas as durações silábicas advindas das leituras de textos similares, lidos de forma semelhante

por esse mesmo locutor. É evidente que, apesar das ressalvas expressas pelos adjetivos “similares” e “semelhante” e o estilo ser leitura, o número potencial de material que se obteria é tão grande que não pode ser medido com os recursos disponíveis pelo experimentador. Por isso, a população deve ter suas características estimadas pela amostra colhida. Qualquer comportamento distinto daquele durante a leitura da amostra não pertence à mesma população de dados.

A estatística inferencial toma, assim, descritores amostrais como média e variância como sendo os mesmos descritores da população para, a partir deles, fazer inferências sobre a probabilidade de os valores estarem numa certa faixa e assim comparar populações de dados. A probabilidade serve como ponte entre o que se conhece e mediu, a amostra, e o que não foi medido mas se deseja conhecer com uma certa precisão, dada pelo valor da probabilidade de ocorrer valores numa determinada faixa. A amostra, que é o conjunto de medidas que se tem de um corpus é, portanto, essencial para a experimentação, mas deve satisfazer determinadas condições. Para tanto, é preciso entender o que é uma amostra.

A amostragem é o procedimento estatístico de seleção aleatória de dados de uma certa população. Se por um lado os locutores ou os ouvintes devem obedecer a critérios de representatividade dos extratos sociolinguísticos a serem investigados em prosódia experimental, por outro lado, mesmo tendo-se obedecido a esses critérios, a seleção que se faz não é completamente aleatória. Tendemos a selecionar o locutor ou ouvinte mais próximo ou o conhecido de um colega de trabalho ou aluno. Por isso, de certa forma, fazemos uma escolha. Por questões econômicas, qualquer outro procedimento que visasse ao cumprimento à risca do caráter aleatório de uma amostragem é simplesmente inviável. Por isso apontaremos alguns cuidados para se evitar que os dados sejam enviesados por algum fator externo.

Outro cuidado que se deve ter é assegurar a independência das amostras, por ser uma condição de aleatoriedade e um pressuposto

básico da maioria dos testes estatísticos, a não ser daqueles justamente que se propõem a inferir aspectos de dependência entre variáveis, como nas séries temporais. Por exemplo, se vamos comparar dados de F0 entre trechos de um enunciado para ver se há diferenças médias de seus valores, não podem fazer parte das amostras todos os valores gerados pelo extrator de F0, pois os valores ao longo de uma vogal são dependentes entre si. Nesse caso, recomenda-se usar apenas os valores médios ou três valores da F0 afastados na vogal. O resultado de um teste inferencial que usasse todos os valores da Fo obtidos de um algoritmo de extração, que gera valores a cada ciclo glotal, seria completamente enviesado. A duração silábica, por sua vez, tende a não gerar valores dependentes, porque é modificável de sílaba para sílaba, caso o locutor o deseje, para fins comunicativos. Os valores da F0, por outro lado, estão atrelados por uma razão de inércia, uma vez que a vibração das pregas vocais não tem sua taxa modificada ciclo a ciclo.

Outro aspecto fundamental para a escolha de um teste estatístico de forma vinculada às hipóteses científicas é o conceito de variável dependente e independente. Essas categorias não são assim nomeadas pelo seu nível de mensuração, a saber, se são categóricas, ordinais ou intervalares (numéricas), mas justamente pelo fato de serem ou não selecionadas pelo experimentador.

A variável dependente é a variável medida pelo pesquisador e supostamente afetada pela manipulação da variável independente, que é a grandeza, em qualquer nível de mensuração, que foi manipulada pelo experimentador em função de suas hipóteses, para ver o efeito sobre a variável dependente, aquela que não manipulou. Por exemplo, se o experimentador tem por hipótese que a sílaba tônica é mais longa do que as átonas, ele escolhe palavras em diversos contextos em que as sílabas variem quanto à tonicidade: tônicas e átonas pré- e pós-tônicas. Assim, tonicidade é a variável independente, nesse caso, categórica. A duração, por sua vez, que é o que ele quer mostrar que varia de forma significativa entre os graus de tonicidade, é a variável dependente.

Nesse exemplo, é preciso estar atento a possíveis influências não manipuladas pelo experimentador, pois a duração depende de muitos outros fatores que acarretariam resultados não devidos diretamente ao grau de tonicidade. Por exemplo, sabe-se que a duração silábica bruta muda com a duração intrínseca e com a posição da sílaba no enunciado. Assim, se uma sílaba átona contém segmentos longos, como [a] e [s], pode vir a durar mais do que uma sílaba tônica com segmentos curtos, como [i] e [R][1]. É por isso que se devem buscar composições similares para a sílaba como em "papa" e "papá". Outra influência a ser mencionada diz respeito ao fato de que, ao final do enunciado e antes de pausas silenciosas, o fenômeno do alongamento final faz com que rimas pós-tônicas sejam bem alongadas e que durem mais do que as tônicas. Essas variáveis não previstas são as variáveis a controlar. Assim, o experimentador deve conhecer bem os fatores que afetam suas variáveis dependentes e controlá-los ou minimizar seus efeitos. Além de fatores intrínsecos a um fenômeno fonético, como os dois citados há pouco, fatores externos como cansaço do participante da pesquisa, horário da gravação ou do teste de percepção com relação à tomada de alimento, condições de saúde gerais bem como problemas fonoaudiológicos e efei tos de aprendizado (que poderiam ocorrer sem estímulos distratores) podem, todos eles, enviesar o resultado de um experimento.

Tendo apontado esses aspectos estatísticos básicos, apresentemos o que é o coração de qualquer análise estatística inferencial, o teste de hipóteses.

1 Foi exatamente essa influência da duração intrínseca que motivou e justificou a normalização da duração no capítulo 4.

6.1.1 Teste de hipóteses

Consideremos, para exemplificar o princípio de um teste de hipóteses, as duas distribuições normais da Figura 6.1. Os retângulos cinzas assinalam a frequência relativa de ocorrência de valores num certo intervalo das distribuições respectivas, enquanto a linha cheia representa a função densidade de probabilidade gaussiana considerando as médias e desvios-padrão das amostras respectivas. Para uma introdução à distribuição gaussiana, ver Bunschaft e Kellner (2001), Baayen (2008) ou ainda Dowdy e Wearden (2001).

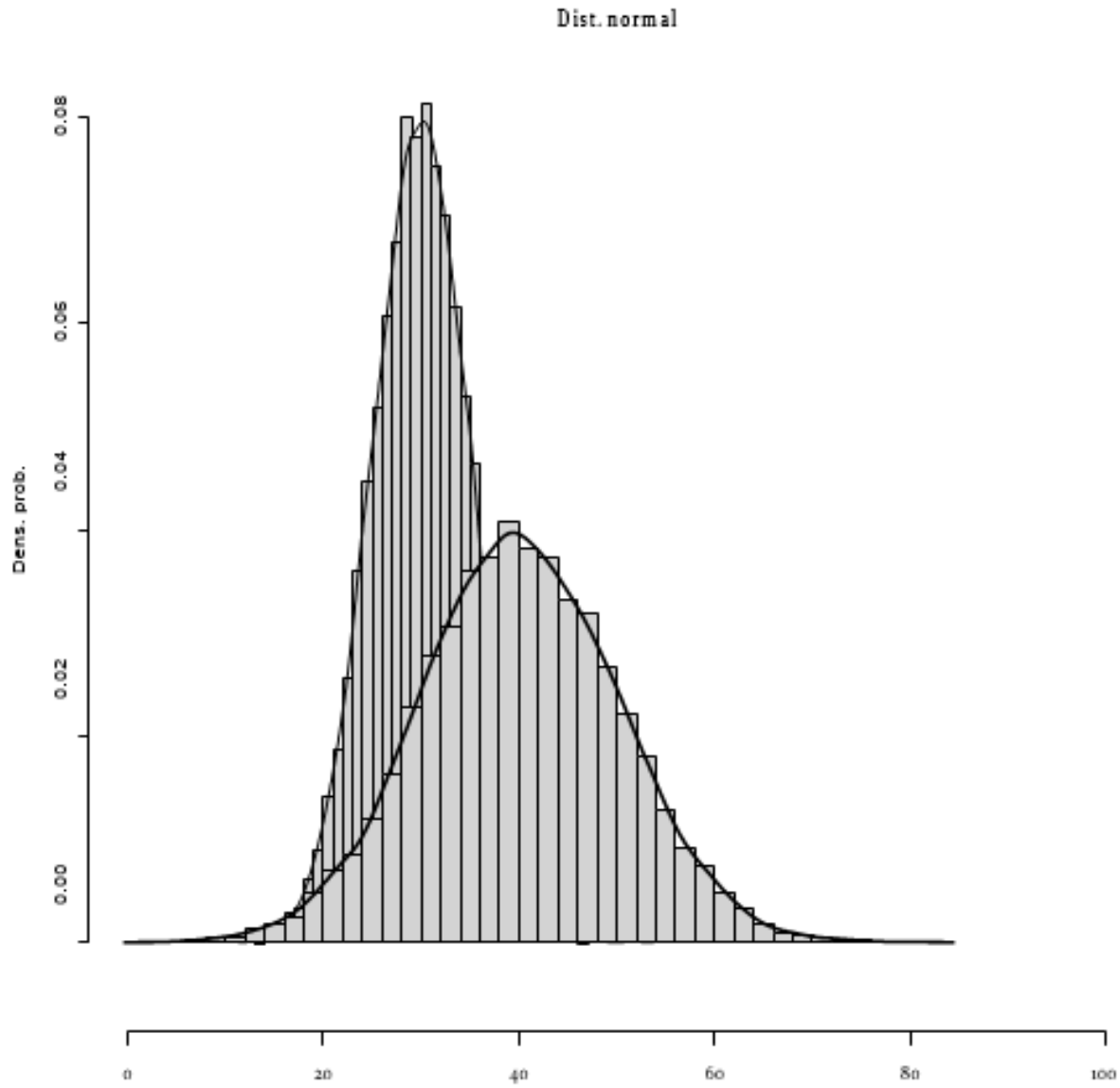

Figura 6.1 – Amostras e populações gaussianas-exemplo: a mais à esquerda com média 30 e desvio-padrão 5 e a mais à direita com média 40 e desvio-padrão 10. As linhas cheias que servem de envoltórias das amostras são traçados das funções de densidade de probabilidade gaussiana que representam as respectivas populações.

Suponhamos que tenhamos feito um estudo piloto e obtivemos da- dos cujo histograma é o da esquerda da Figura 6.1. Suponhamos também que a variável dependente com que trabalhamos, qualquer que seja ela, tem média 30 e desvio-padrão 5 e passou num teste de normalidade, como o de Shapiro-Wilks[2]. Como testar se um novo experimento confirma a hipótese de que a média da variável é 30? É essa a finalidade de um teste de hipóteses. Suponhamos que, com esse novo experimento, grande parte dos valores da variável dependente se encontre entre 45 e 55. Por conta disso, presume-se que a média pode ser maior que 30, senão, como explicar a frequência desses dados? Por isso o experimentador monta o esquema de hipóteses em 6.1.

$$\begin{aligned} H_0 &= \mu_0 \\ H_a &> \mu_0 \\ \alpha &= 0,05 \end{aligned} \tag{6.1}$$

Nesse esquema, H_0 é a hipótese nula, o ponto de partida que foi dado pelo estudo piloto que assume que $\mu_0 = 30$. A hipótese alternativa, H_a, sempre a negação da hipótese nula, é expressa aqui com o sinal de maior (>) por conta da frequência de valores no intervalo de 45 a 55 ser superior a 30. O nível de significância α é o limiar da decisão, a margem de probabilidade que permite ao experimentador escolher entre uma das duas hipóteses. Se a distribuição da hipótese nula gera valores no intervalo do novo experimento maior que α, então essa hipótese é aceita, por essa probabilidade ter sido considerada *a priori* como suficiente para fins do experimento. Se, por outro lado, a probabilidade de valores naquele intervalo for menor que α, considera-se um evento raro que não poderia ser explicado pela hipótese nula, rejeitan-

2 Os testes de normalidade avaliam probabilisticamente se uma distribuição pode ser considerada uma gaussiana. Os mais conhecidos são os de Kolmogorov-Smirnoff, de Lilliefors e o de Shapiro--Wilks. Esse último é dos mais robustos segundo Razali e Wah (2011). Para mais detalhes sobre testes de normalidade, consultar Ghasemi e Zahediasl (2012).

do-a e assumindo uma hipótese alternativa. Observe que em nenhum dos casos se tem certeza de algo, pois a hipótese alternativa continua sendo uma hipótese a ser testada com experimentos sucessivos.

A probabilidade de valores entre 45 e 55 na distribuição gaussiana de média 30 e desvio-padrão 5 é de 0,0013, que é o chamado p-valor. Como essa probabilidade é menor do que 0,05 (α), rejeita-se a hipótese nula, e então assume-se que a média da variável dependente é maior do que 30. Como esse resultado é também uma hipótese, há uma probabilidade de se cometer um erro, chamado em estatística de erro do tipo I, definido como o erro em se rejeitar uma hipótese nula verdadeira. No nosso exemplo esse erro é justamente o nível de significância, pois, considerando todos os experimentos que podem ser feitos, é sempre essa probabilidade, escolhida de antemão, que constitui a probabilidade de erro, pois a hipótese nula pode gerar dados com essa proporção ou menor e não se considerou isso aceitável para fins de experimentação.

Como um novo experimento havia sido realizado, o experimentador calcula a média e desvio-padrão da nova amostra representada à direita na Figura 6.1. Se a população que subjaz essa amostra é a dada pela gaussiana cuja função é desenhada na figura, a probabilidade de valores entre 45 e 55 nessa outra distribuição é de 0,24, superior a α. Até novas descobertas, feitas a partir de novos experimentos, essa é a melhor representação dos dados exemplificados aqui. Mais interessante do que a rejeição da hipótese nula é conhecer mais sobre os dados, supondo que é regido pela distribuição gaussiana de média 40 e desvio-padrão 10. Uma dessas formas é o intervalo de confiança, que define em que limites a maior parte dos dados se encontra. A esse intervalo se associa uma probabilidade. Assim, o intervalo de confiança a 95% é o intervalo em que se encontram 95% dos dados (e, de forma estimada, da população) centrados em torno da média, assumindo uma determinada distribuição estatística. Em nosso exemplo, o intervalo de confiança a 95% é dado pelos limites de 20 a 60, isto é, 95% dos valores da distribuição gaussiana de média 40 e

desvio-padrão 10 se situam entre esses limites. Exemplos concretos nas seções que seguem consolidarão a utilidade de uma melhor exploração das características da amostra, começando por um teste muito usado na área de prosódia experimental, a ANOVA.

6.1.2 ANOVA

A Análise de Variância (ANOVA, da sigla para *Analysis of Variance*) permite testar se existe ao menos uma diferença significativa entre as médias de grupos de amostras. Esses grupos ou níveis estão associados a um ou mais fatores que são, justamente, as variáveis independentes. Vamos tomar um exemplo concreto para explicar como se faz uma ANOVA. Mais detalhes podem ser obtidos em Dowdy e Wearden (2001) e em Baayen (2008), esse último com aplicações para a área da linguagem. Os dados e roteiro para refazer as análises ilustradas aqui se encontram no repositório do livro na pasta **Estatistica/ANOVA**.

Suponha que um experimentador queira iniciar um estudo sobre o papel da duração silábica como correlato do acento lexical em uma pa- lavra extraída de uma leitura. Para tanto, pede a um locutor brasileiro que leia dez vezes um trecho transcrito contendo a palavra "contato". Essas leituras foram intercaladas com leituras de outros trechos sem essa palavra, para desviar a atenção do locutor dos objetivos do experimento, como vimos na seção 3.2.2. Esses dados se encontram no arquivo **contato.txt** na pasta mencionada acima. Ao abrir o arquivo, podem-se ver as durações para cada sílaba da palavra na primeira coluna e uma segunda coluna com a variável independente nominal TONICIDADE, que é o fator da ANOVA. Observe que TONICIDADE tem três níveis, que são as categorias PRE (pré-tônica), PST (pós--tônica) e TON (tô- nica). São três níveis de um único fator, por isso a ANOVA que será realizada se chama de ANOVA de um fator

(*1-Way ANOVA*).

Mesmo sabendo que a diferença na composição das sílabas quanto às consoantes e quanto às vogais tenha um efeito para a duração silábica, o experimentador resolve deixar essa questão para um estudo ulterior. O modelo requer que se testem três suposições para que a ANOVA seja realizável: normalidade dos resíduos[3], homogeneidade das variâncias e independência das amostras. Usamos o programa R (R Development Core Team, 2008) para testar essas suposições respectivamente com testes de Shapiro-Wilks, Fligner Killeen e o gráfico entre resíduos e valores preditos pelo modelo. Os comandos para executar esses passos se encontram no arquivo referente à ANOVA, disponibilizado no repositório.

A saída do modelo de ANOVA do R é mostrada abaixo, cujos as- pectos mais relevantes são os graus de liberdade (Df) do fator e dos resíduos, o valor de F (F value) e o p-valor (Pr (>F)). Além disso, a tabela mostra na primeira coluna o nome do fator (TONICIDADE) e os graus de liberdade dos resíduos (aqui, 27), que tem a ver com o número de dados. O teste de ANOVA pressupõe um esquema de hipóteses em que a hipótese nula é que não há diferenças entre as médias dos níveis (grupos) contra a hipótese alternativa de que existe pelo menos uma diferença entre as médias em cada grupo para um determinado nível de significância (aqui adotado como 5%). Quando se tem mais de dois níveis, é preciso ainda aplicar um teste suplementar, para apontar entre que níveis as médias diferem significativamente, o chamado teste *post hoc*.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
TONICIDADE	2	36743	18372	20.29	4.17e-06	***
Residuals	27	24444	905			

3 Os resíduos ou erros são as diferença entre o valor real da variável e o valor gerado pelo modelo de ANOVA, que é sempre o valor médio da variável em cada nível.

O p-valor da tabela acima é bem menor que 5% ($4,17.10^{-6}$) e, portanto, rejeita-se a hipótese nula: há pelo menos uma média significativamente distinta. O teste *post hoc* de *Tukey Honest Significant Difference* é aplicado e obtém-se a tabela que segue, com p-valores na última coluna, todos abaixo de 5%. Esses passos também foram disponibilizados no repositório.

$TONICIDADE

	diff	lwr	upr	p adj
PST-PRE	-50.8	-84.163569	-17.43643	0.0022347
TON-PRE	34.4	.036431	67.76357	0.0423059
TON-PST	85.2	51.836431	118.56357	0.0000026

Assim, todos os grupos diferem entre si, com média iguais a 168 ms para a pré-tônica, 203 ms para a tônica e 118 ms para a pós-tônica. Observe que há uma diferença de 85 ms entre tônica e pós-tônica, uma queda de duração acima de valores de *Just Noticeable Difference* (JND)[4] para duração que se encontram na literatura e, portanto, passível de ser utilizado como parâmetro revelador do acento lexical, importante aspecto da prosódia lexical.

Uma forma simplificada de apresentar o resultado do modelo de ANOVA e subsequente teste *post hoc* pode ser essa: "Há uma diferença significativa para as médias das durações das sílabas com $F_{2,27} = 20,29$, $p < 5.10^{-6}$, sendo que todos os níveis diferem significativamente entre si para $\alpha = 0,05$". Os valores subscritos ao símbolo do teste F, que é o teste realizado pelo modelo de ANOVA, representam os graus de liberdade respectivamente entre os grupos e dentro dos grupos.

As populações que subjazem às amostras examinadas nesse

4 A mínima quantidade de duração, de frequência e de intensidade necessárias para discriminar esses parâmetros se chama de *Just Noticeable Difference* (JND). Para duração de uma vogal, a JND varia se a vogal é tônica ou átona, com valor menor na sílaba átona, com variação entre 25 e 40 ms. Algo semelhante vale para a duração da sílaba.

exemplo são as durações provenientes das repetições *ad infinitum* das sílabas da palavra "contato", lidas em textos como os usados no experimento pelo mesmo locutor. Não há outra generalização possível. Por isso, um experimento real que queira dizer sobre a duração em diferentes níveis de tonicidade desse locutor deve considerar outras palavras e outros padrões acentuais lexicais além do paroxítono, de forma a ser mais amplamente generalizável e incluir a palavra como fator aleatório num teste que o preveja, como os modelos mistos (*mixed models*) apresentados na seção 6.1.6. Por outro lado, generalizar esses resultados de diferença duracional entre os níveis de tonicidade para qualquer locutor de uma região dialetal requereria uma seleção prévia de locutores dessa região. Especulando ainda, se o experimentador não quiser ficar restrito à frase lida, mas quiser considerar outros estilos de elocução, deverá incluir o estilo como fator fixo[5], se quiser considerar tão somente os estilos estudados. O leitor pode ver que a inclusão de fatores e seus níveis multiplica os dados obtidos, aumentando o tempo de coleta, de medida, de realização de testes e do próprio experimento, exigindo planejamento cuidadoso.

Para ilustrar o conceito de tamanho do efeito e de interação entre fatores, consideremos incluir o fator estilo de elocução no nosso exemplo. Os dados se encontram no arquivo **contatoStiloTon.txt**, na mesma pasta. O leitor poderá ver nesse arquivo que o número de valores de duração foi multiplicado por três, com a inclusão de uma coluna especificando o fator ESTILO com três níveis: texto lido (TXTL), entrevista (ENTR) e palavra isolada (PLIS). O locutor foi submetido a uma entrevista em que apareceu a palavra "contato" dez vezes. A entrevista foi transcrita e trechos dela foram lidos duas semanas depois, todos eles contendo as dez instâncias da palavra "contato". Na mesma ocasião, o locutor leu dez vezes, de forma isolada, a palavra "contato" intercalada com outras de um conjunto de palavras

5 Essa noção será discutida adiante, nos modelos de efeitos mistos.

distratoras. Deseja-se saber se a duração silábica marcaria o acento lexical da mesma forma em qualquer um dos estilos.

Para tanto, criou-se no R um modelo de ANOVA de dois fatores que passou nas três suposições para o teste de ANOVA. O resultado aparece na tabela seguinte, que tem formato semelhante à anterior com exceção do elemento "TONICIDADE:ESTILO" que é o efeito da interação entre fatores.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
TONICIDADE	2	89529	44764	58.72	<2e-16	***
ESTILO	2	238534	119267	156.46	<2e-16	***
TONICIDADE:ESTILO	4	123044	30761	40.35	<2e-16	***
Residuals	81	61747	762			

As médias para cada um dos níveis dos dois fatores podem ser vistas na Figura 6.2 nas posições correspondentes aos níveis PRE, PST e TON assinaladas na abscissa. Observe que as linhas dos estilos entre- vista (ENTR) e texto lido (TXTL) são próximas. Quando a interação não é significativa, essas três linhas que se veem na figura se aproximam de paralelas. Quando não o são, pode indicar que a interação é significativa, isto é, que ao menos um nível de um dos fatores se comporta distintamente em relação ao níveis do outro fator, que é o caso aqui, pois a média da sílaba pós-tônica no estilo palavra isolada (PLIS) não segue o padrão de ser menor do que das duas demais sílabas. E, de fato, o p-valor $< 2.10^{-16}$ da tabela acima, menor do que 0,05, aponta para essa interação.

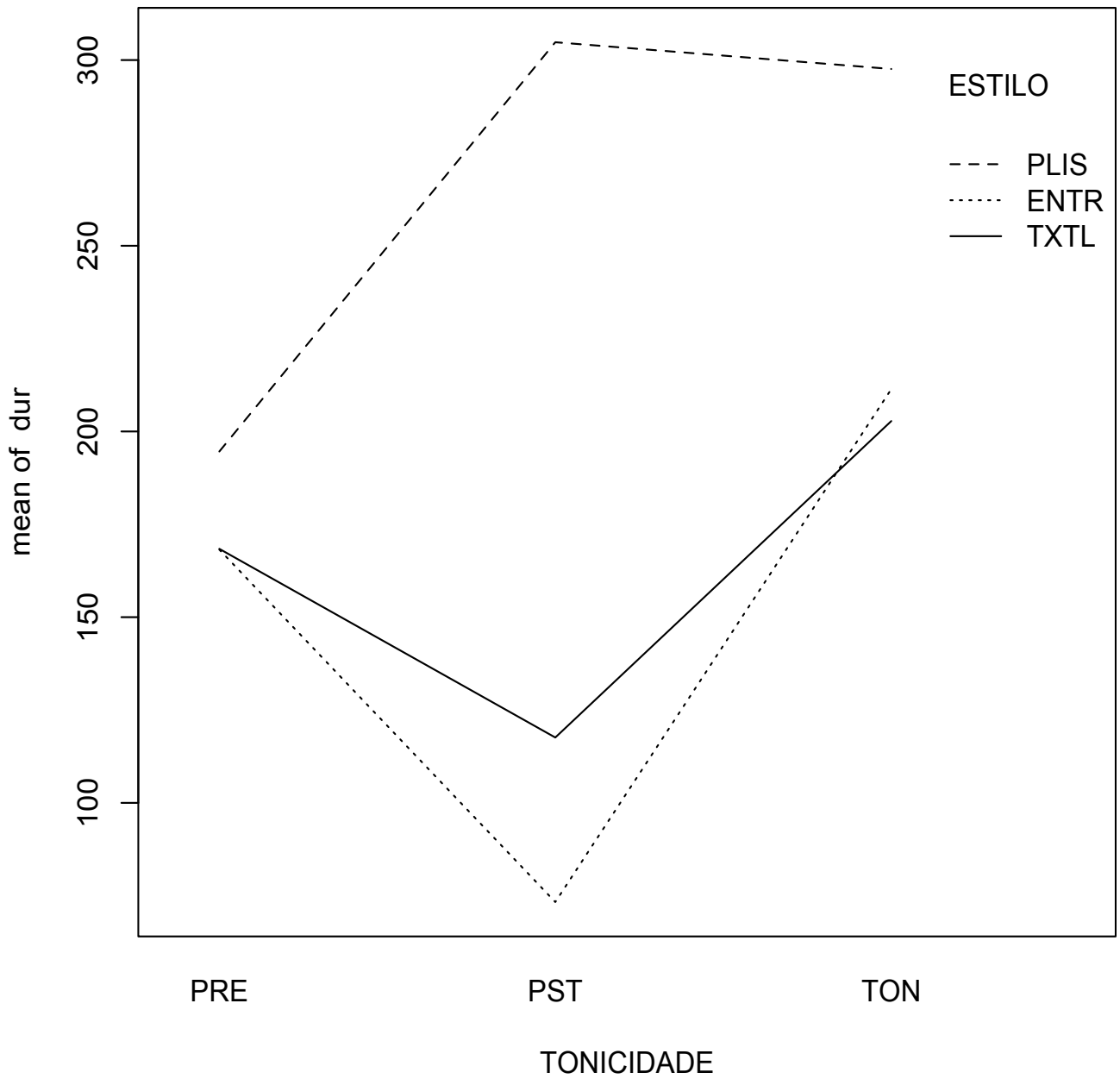

Figura 6.2 – Médias dos três níveis do fator TONICIDADE e dos três níveis do fator ESTILO para o exemplo do texto interligadas por linhas. Observe que as linhas dos estilos entrevista (ENTR) e texto lido (TXTL) são próximas.

Outra informação muito importante é o tamanho do efeito (*effect size*) de um fator sobre a variável dependente. Define-se pela porcentagem da variância da variável dependente explicada pela variância entre os níveis de um fator. Pode ser calculada pela razão entre a soma dos quadrados atribuída a um fator, no cálculo de va-

riâncias da tabela (*Sum Sq*) e a soma dos quadrados total, incluindo a dos resíduos. Calculando essas razões para os fatores TONICIDADE e ESTILO e sua interação, temos as seguintes porcentagens de variância explicada: 17% para o fator TONICIDADE (89529/(89529 + 238534 + 123044 + 61747)), 47% para o fator ESTILO (238534/(89529+238534+123044+61747)) e 24% para a interação entre eles (123044/(89529+238534 +123044 +61747)). Com isso, se descobre que o principal fator responsável em mudar a duração da sílaba é o estilo, como se vê na Figura 6.2 para o estilo de palavra isolada, para o qual todas as durações médias são superiores às médias dos outros dois estilos. Além disso, a sílaba pós-tônica (PST) tem duração média bem maior, que é efeito do fenômeno de alongamento final, uma vez que na palavra isolada a sílaba final está diante da pausa silenciosa final.

Aliar a informação do tamanho do efeito com o intervalo de confiança é muito importante para avançar na compreensão de um fenômeno, levando a pesquisa muito além da mera informação sobre significância. Essas grandezas permitem conhecer o grau do efeito de um fator para explicar a variabilidade da variável dependente, no primeiro caso, e assinalar o grau das diferenças médias, especialmente se são relevantes para a percepção tendo, assim, valor comunicativo. O modelo de ANOVA, se bem prático, nem sempre pode ser aplicado se alguns dos pressupostos acima não for obedecido. Se acontecer, é preciso aplicar os testes não paramétricos equivalentes que são os testes de Kruskal-Wallis, para a ANOVA de um fator e o de Scheirer-Ray-Hare, para a ANOVA de dois fatores. Esse teste foi justamente o teste aplicado no trabalho de Barbosa, Eriksson e Åkesson (2013), que investigou os parâmetros prosódico-acústicos marcadores do acento lexical nos três estilos mencionados nesta seção.

Examinemos agora outro teste estatístico muito comum em prosódia experimental, o teste de Student ou teste t, embora menos usado que a ANOVA, por isso reportado agora.

6.1.3 Teste de Student ou t

Há três tipos de testes de Student, também chamados de testes t, muito úteis em estudos de prosódia experimental: teste t de variáveis independentes, teste t de variáveis dependentes ou teste t pareado e teste t de valor fixo. Os dois primeiros comparam as médias de duas amostras de dados e têm menos suposições do que o modelo de ANOVA, pois requerem apenas a normalidade dos resíduos e a independência das amostras, enquanto o segundo requer o mesmo, mas apenas numa amostra única. Vamos exemplificar cada um dos três com dados de pesquisa que ajudem o leitor a saber quando aplicar um dos três. Mais detalhes também podem ser obtidos em Dowdy e Wearden (2001) e em Baayen (2008), esse último com aplicações para a área da linguagem.

O teste t de variáveis independentes avalia a diferença entre médias de uma variável dependente numa situação em que não é possível relacionar um dado de um grupo com o do outro nas mesmas condições, por exemplo, na situação em que se deseja comparar médias da F0 em duas narrativas para saber se foram feitas pelo mesmo indivíduo. Como em narrativas não é possível comparar valores em cada vogal, na mesma palavra e na mesma sequência, para ver as alterações nesses lugares, simplesmente porque as narrativas não compartilham exatamente o mesmo vocabulário, muito menos a mesma sequência de palavras, o teste t de variáveis independentes deve ser usado. Considere os dados do arquivo **exemploforense.txt**, na pasta **Estatistica/Testet**, com valores de parâmetros segmentais de F2, taxa de movimento de F2 no início da vogal até a estabilidade, tempo para estabilização de F2, frequência de base da F0 (estimativa da frequência mínima em cada vogal) e mediana da F0 por vogal, que fazem parte do trabalho de mestrado de Machado (2014). Por serem parâmetros prosódico-acústicos, verificaremos se tanto a mediana da

F0 quanto a sua frequência de base têm ou não médias significativamente distintas em duas narrativas, assumindo a hipótese nula de que se trata da mesma pessoa e que, portanto, teriam médias idênticas.

Essas narrativas são histórias de vida em entrevistas com dez pessoas, tendo sido uma sorteada como sendo o "criminoso". Trechos da narrativa são comparados entre o criminoso e um suspeito, que são os dados do arquivo de dados no repositório. Considerando tanto a variável *foMedian* (mediana da F0) quanto *Baseline* (frequência de base da F0), os resíduos não passaram no teste de normalidade e, por isso, usaremos o teste t de variáveis independentes não paramétrico, o teste de Wilcoxon para variáveis independentes, também chamado de teste de Mann-Whitney. Para a mediana da F0, o resultado do teste com seu p-valor e a quantidade W, que mede a soma das ordens e é tanto menor quanto mais próximas forem as amostras dos dois grupos, são estes: $W = 248010$, p-valor = $1,57.10^{-5}$. Como o p-valor é menor que 5%, rejeita-se a hipótese nula: as medianas da F0 são significativamente distintas, com o valor de média de 153 Hz para o criminoso e 164 Hz para o suspeito. Repetindo o procedimento para a frequência de base da F0 obtêm-se esses resultados: $W = 239260$, p-valor = $8,14.10^{-5}$. Como o p-valor também é menor que 5%, rejeita-se a hipótese nula: as frequências de base da F0 são significativamente distintas, com o valor de média de 148 Hz para o criminoso e 158 Hz para o suspeito.

Com isso conclui-se que há indícios de que o suspeito não seja o criminoso. Evidentemente, muitos outros parâmetros acústicos devem ser avaliados em casos reais e, em seu trabalho, Machado (2014) faz um estudo amplo quanto aos parâmetros mais relevantes para apontar quem pode, com certa probabilidade, ser o "criminoso" de sua simulação experimental.

Como comentado acima, o teste t de variáveis dependentes ou teste t pareado pode ser usado na situação em que é possível relacionar um dado de um grupo com o do outro nas mesmas condições. O traba- lho de Passetti (2015) é excelente para ilustrar esse teste,

pois foram feitas gravações simultâneas com um celular e um microfone para examinar o efeito do filtro do celular sobre os parâmetros acústicos, tendo sido encontrados efeitos mais importantes nas frequências do primeiro e terceiro formantes. Mas houve também um efeito significativo num parâmetro prosódico, a mediana da frequência fundamental, para o qual fez-se um teste t pareado[6] que foi significativo, com diferenças variando entre 1 e 6 Hz, a depender do indivíduo. É justamente esse tipo de resultado que, se se limitasse à diferença significativa encontrada, não se atentaria para algo crucial: o efeito do celular é em média pouco maior que 2%, com a maior parte das diferenças inferiores a 4 Hz, que não é audível. Para fins forenses, mesmo que consideremos que sejam dados de produção, a variação da mediana da F0, mesmo para um único indivíduo, é tão superior a 4 Hz que o efeito de celular, para esse parâmetro, deve ser ignorado.

O teste t de valor fixo, por sua vez, é usado quando se deseja verificar a hipótese nula de que uma amostra tem uma determinada média, aplicando-se a uma amostra única de dados. Um exemplo que bem o ilustra é o experimento para inferir o p-center em PB (BARBOSA et al., 2005). Nesse trabalho experimental de sincronização fala-metrônomo, seguimos o esteio das pesquisas sobre o *perceptual-center* (*p-center)* que trouxeram evidência de que o momento de ocorrência da sílaba para nosso sistema auditivo é a transição C-V. Por isso, no trabalho hipotetizamos que, ao ser convidado a produzir repetidamente uma sílaba em sincronismo com um metrônomo sonoro, um indivíduo alinharia o início da vogal com cada batida do metrônomo. Foi isso que observamos em linhas gerais, embora a precisão desse sincronismo dependa da composição silábica.

Por isso é preciso verificar se a distância temporal entre início de vogal e batida do metrônomo é, em média, nula. Assim, comparamos a média da distribuição com o valor zero da população, contra a

6 A função que fez isso no pacote R é a mesma que faz o teste de variáveis independentes, bastando indicar no argumento da função que o teste é pareado.

hipótese alternativa de que seria diferente de zero. Fizemos esse teste para duas taxas do metrônomo, com a repetição da sílaba [pɛ] a 80 e 108 bpm, com dados que se encontram no arquivo **pcenterpE.txt**, na pasta **Estatistica/Testet**. No arquivo, "delta" é a variável dependente que representa, em milissegundos, a distância entre a batida do metrônomo e o início da vogal, sendo positiva quando a batida do metrônomo ocorre depois do início da vogal e negativa se ocorre antes. A variável independente, "taxa", é a taxa do metrônomo, modificada para aplicar o experimento em sessões distintas.

Usando o teste t de valor fixo com cada conjunto de dados, o de 108 e depois o de 80 bpm, o resultado é a aceitação da hipótese nula, com p-valores maiores do que 5%. Assim, aceita-se a hipótese nula para essas taxas do metrônomo. Claro que pode haver um erro ao se tomar essa decisão, que é o de aceitar uma hipótese nula falsa, o chamado erro do tipo II em estatística, que é representado pela letra β. Estimar de quanto seria esse erro requer experimentos em que se sugerisse de quanto seria o afastamento do sincronismo perfeito (delta = 0). Tanto esse teste quanto o anterior têm equivalentes não paramétricos, que é o já usado teste de Wilcoxon. Os testes não paramétricos não foram aplicados nesses exemplos porque os resíduos passaram no teste de normalidade e os dados foram obtidos de forma independente, não violando pressuposto algum para o teste t.

Se os testes vistos até o momento comparam médias, é preciso também conhecer como comparar inferencialmente as variâncias. O teste F compara duas variâncias, mas há testes que comparam as variâncias entre várias amostras ou níveis, da mesma forma que a ANOVA compara médias entre vários grupos.

6.1.4 Testes para comparação de variâncias

Ao longo deste livro, vimos mais de uma vez a importância de

comparar a variabilidade de algum parâmetro prosódico-acústico entre ao menos duas condições. Por exemplo, a tonicidade de uma sílaba afeta não apenas a duração média, maior na tônica, mas também a variância, também maior na tônica. Reconhecer que uma sílaba se comporta como tônica é verificar se tem essas duas características: mais longa e mais variável.

O teste F é um teste paramétrico que compara variâncias de duas amostras, assumindo-as normais e independentes. As mesmas condições de uso requerem o teste de Levene, que faz o mesmo trabalho, mas para mais de dois conjuntos de dados. Em ambos, a hipótese nula mais usada é a de que as variâncias em todos os conjuntos é a mesma, contra a hipótese alternativa de que são distintas. O teste não paramétrico mais usado é o de Fligner-Killeen, que já mencionamos para verificar uma das assunções da ANOVA, a de que as variâncias dos grupos são estatisticamente iguais.

Ilustremos o uso desse teste com dados do trabalho de Barbosa, Madureira e Mareüil (2017) sobre os parâmetros prosódico-acústicos que distinguem quatro estilos de elocução em quatro línguas. Esses dados se encontram na pasta **Estatistica/TestesVar**, no arquivo **AllLanguagesREST**. Dos resultados desse trabalho apresentaremos apenas aqueles que concernem tão somente as distinções entre as línguas para os estilos de leitura e narração tomados juntos, uma vez que as análises mostraram que são mais distintos dos estilos telejornalístico e político do que entre si. As línguas, *lato sensu*, foram francês (FR) e alemão (AL) padrões e português brasileiro (PB) e europeu (PE). Dez locutores leram e narraram a história dos pasteis de Belém nas quatro línguas. Trechos entre 10 e 20 s foram extraídos para análise sendo pelo menos quatro trechos por locutor em todas as línguas. De cada trecho foram extraídos oito parâmetros prosódico-acústicos para o trecho inteiro, dos quais ilustramos aqui a mediana da F0 e a taxa de picos da Fo para comparação das variâncias ao nível de significância de 5%.

Como para nenhuma das variáveis houve distribuição normal

para os resíduos, considerando a diferença de cada valor com relação à média de cada grupo formado por uma língua, aplicamos o teste de Fligner-Killeen, que assinalou significância para ambas as variáveis. Aplicamos um teste *post hoc* não paramétrico de comparação entre as variâncias[7] com correção de Bonferroni[8] que produziu os seguintes resultados quanto aos p-valores, primeiro para a mediana da Fo, em semitons relativos a 100 Hz:

	PB	PE	FR
PE	0.0094	–	–
FR	0.0051	1.0000	–
AL	0.0340	1.0000	1.0000

Esse resultado revela que o português brasileiro é a única língua que se distingue significativamente em variância das demais. Pode-se então calcular o valor médio do desvio-padrão da F0 para mostrar que seu valor para o PB é de 4,8 semitons, contra 3,5 semitons ou menos para as demais línguas.

Quanto à taxa dos picos da F0, os p-valores revelam que o PB também se distingue das demais línguas para a variância desse parâmetro, com valor de desvio-padrão de 0,14 picos/segundo contra valores médios superiores a 0,21 picos/segundo para as demais línguas. O que significa que o PB é mais regular em suas taxas de picos da F0 entre os trechos de áudio de leitura ou narração.

	PB	PE	FR
PE	7.0e-07	-	-
FR	8.8e-13	0.58	-
AL	2.8e-09	1.00	1.00

O leitor pode treinar esse teste consultando o roteiro no

7 Disponível na biblioteca RVAideMemoire para o software R, função *pairwise.var.test*.

8 Esse método corrige o efeito de não respeito do nível de significância escolhido quando se tem múltiplas comparações.

repositório do livro, pasta **Estatistica/TestesVar**. Recomendamos que realize, como exercício, o teste com outras variáveis dependentes, bem como verifique as diferenças médias de variância dessas variáveis entre os estilos (variável *style*).

6.1.5 Regressão linear e logística

Os modelos estatísticos de regressão mais comumente usados em prosódia experimental permitem investigar a eventual relação entre variáveis intervalares (regressão linear, simples ou múltipla) e entre uma variável categórica e uma proporção (regressão logística). Boas referências sobre regressão linear simples e múltipla e regressão logística podem ser encontradas no grande manual do R (CRAWLEY, 2007) como também em Baayen (2008), esse último com aplicações para a área da linguagem. Um livro mais avançado sobre o assunto, também usando o R, é o de Gelman e Hill (2007).

Observe na Figura 6.3 um gráfico da relação entre entre duas variáveis numéricas, a duração da unidade VV saliente no grupo acentual (a último desse grupo) na abscissa e a mediana da F0 da mesma unidade na ordenada, cujos dados se encontram no repositório do livro, pasta **Estatistca/RegLin**.

Os valores foram medidos a partir da leitura de um texto de 110 palavras d'*A Menina do Nariz Arrebitado*, de Monteiro Lobato, por um locutor paulista de nível universitário. Os picos locais de duração da unidade VV foram determinados pela técnica apresentada na seção 4.3 e representam a duração normalizada da unidade mais à direita do grupo acentual, como vimos na seção 4.7. Observe, na mesma figura, que há um decréscimo da mediana da F0 à medida que a duração da última unidade VV do grupo acentual aumenta.

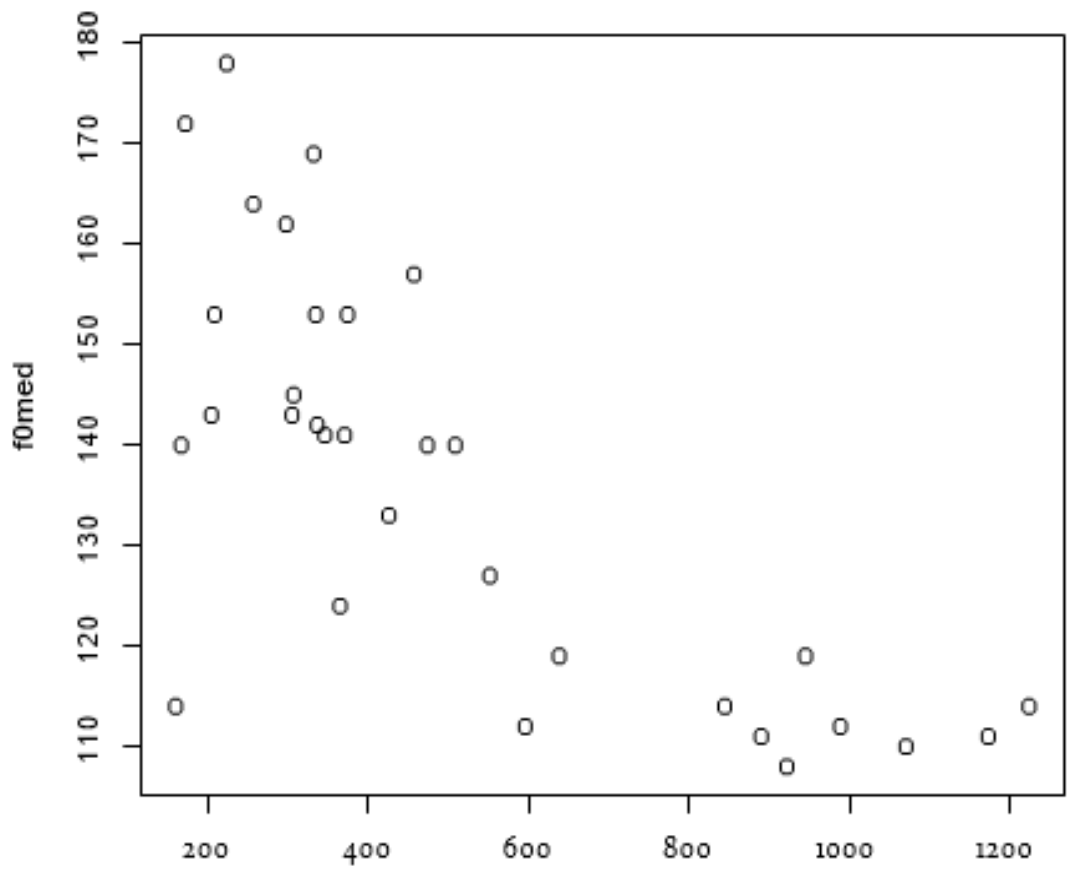

Figura 6.3 – Relação entre duração da unidade VV saliente em ms (abscissa) e mediana da F0 em Hertz na respectiva unidade (ordenada) em trecho de leitura de parágrafo em locutor paulista.

O modelo de regressão que apresentamos em seguida procura responder a questões relacionadas à taxa como se dá esse decréscimo e quão bem segue uma relação linear. No entanto, respondidas essas questões, é conveniente lembrar que se trata aqui apenas de uma relação entre variáveis, e não uma relação de causalidade.

Uma relação linear quer dizer precisamente que as duas variáveis se relacionariam segundo a equação da reta, isto é, $mediana(F0) = a + b.dur$, em que *dur* é a variável preditora e a mediana de F0 é a variável resposta ou a ser explicada. O coeficiente *a* é o de intercepção e *b*, o de inclinação da reta. Como explicado anteriormente, por regressão não ter implicação de causalidade, a escolha diferente, tendo a duração

como variável resposta e F0 como preditora pode ser igualmente concebida. Aqui foi adotado o nível de significância de 5%.

O uso do modelo de regressão linear requer a satisfação das mesmas condições que vimos na seção sobre ANOVA acrescidas da condição de linearidade, embora formulados de forma um pouco diferenciada:

- Os resíduos entre os valores preditos pelo modelo para a variável resposta e os valores medidos da mesma variável devem ser distribuídos normalmente;
- Os valores medidos devem ser independentes;
- A relação entre os valores preditos e os resíduos deve ser de igualdade de variância, condição referida como homocedasticidade;
- A relação entre as variáveis resposta (a variável dependente) e preditora (a variável independente) deve ser linear.

Nesse exemplo, como poderá verificar o leitor, seguindo o roteiro que se encontra na mesma pasta do repositório, os pressupostos foram obedecidos e o modelo apresenta os seguintes resultados:

```
Coefficients:
             Estimate   Std. Error t value  Pr(>|t|)
(Intercept) 161.42881    4.76811    33.86    < 2e-16 ***
dur          -0.04872    0.00792    -6.15    9.1e-07 ***

Residual standard error: 14 on 30 degrees of freedom
Multiple R-squared: 0.558,Adjusted R-squared: 0.543
F-statistic: 37.8 on 1 and 30 DF,  p-value: 9.12e-07
```

O resultado revela um coeficiente de determinação (*Adjusted R-squared* [9]) de 54, 31%, medida que assinala a adequação da função linear (uma reta) com relação aos dados. Esse valor indica, assim, que pouco mais da metade da variância da variável resposta (a mediana da F0) é explicada pela variável preditora (a duração da unidade VV ao fim do grupo acentual). O p-valor desse resultado é o mesmo que para o coeficiente de inclinação, isto é, $9,1 \times 10^{-7}$.

Os coeficientes da reta que aproxima os dados são os coeficiente de intercepção (*Intercept*), de valor aproximado de 161 ms, e o coeficiente de inclinação indicado pelo nome da variável *dur*, de valor aproximado de -0,049 Hz/ms. Assim, a equação da reta que prediz os valores da mediana da F0 a partir da duração da unidade VV saliente correspondente é: *Fomed (Hz)* = 161 -0.049.*dur (ms)*. O intervalo de confiança a 95 % desses coeficientes são os seguintes: de 152 a 171 ms para o coeficiente de intercepção e de -0,065 a -0,033 Hz/ms para o coeficiente de inclinação.

9 O coeficiente de determinação ajustado estima o coeficiente de determinação da população e não das amostras, por isso o tomamos no lugar do *Multiple R-squared*.

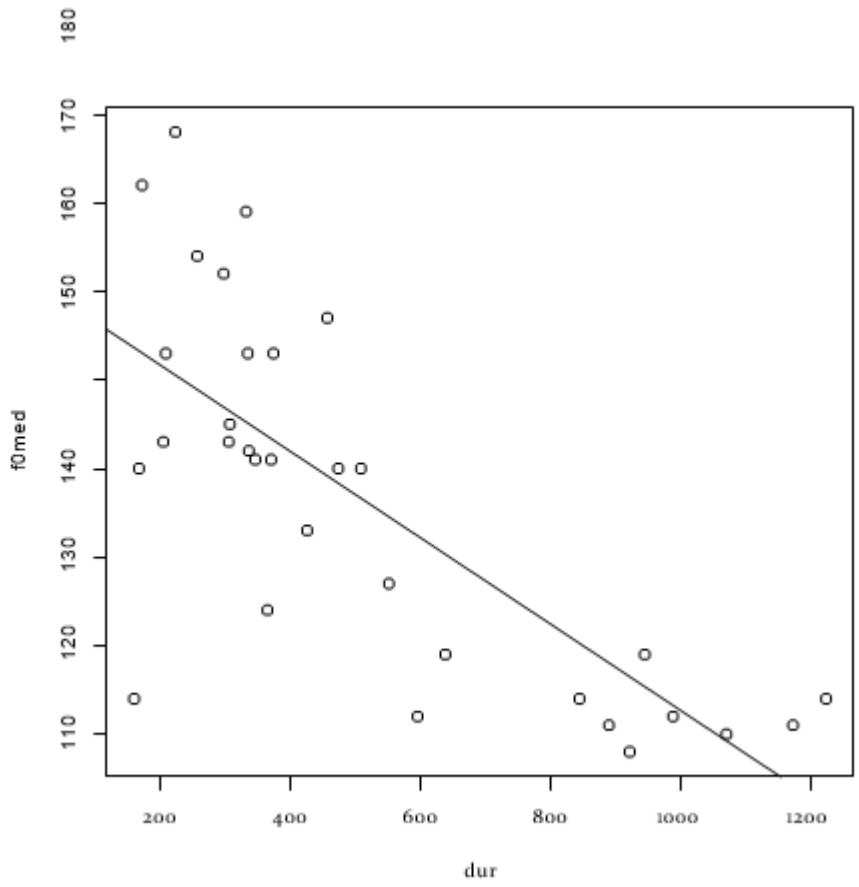

Figura 6.4 – Reta de regressão linear superposta aos dados da Figura 6.3.

A Figura 6.4 mostra a equação da reta que melhor representa os dados, que foi o resultado do modelo criado de regressão linear. O resultado revelou que mais do que a metade da variância dos dados de mediana de F0 na unidade VV saliente é explicada[10] pela duração dessa mesma sílaba, duração essa que inclui a pausa silenciosa no caso das fronteiras fortes, uma vez que esse tipo de sílaba compreende o intervalo entre dois inícios de vogais consecutivas no enunciado.

Conclui-se então que, para esses dados, existe uma tendência a que, à medida que a duração da sílaba fonética saliente aumenta, tanto menor o valor mediano da F0. Isso quer dizer que, quanto mais longa é a duração da sílaba fonética que precede a fronteira do grupo acentual, mais baixo o valor da frequência fundamental a precedendo imediatamente na fronteira. Assim, quantitativamente, a força crescente

10 Este terminologia é estatística, não implica em nada uma relação de causalidade.

da fronteira prosódica seria assinalada pela imbricação do aumento progressivo da duração da sílaba fonética que a precede com o abaixamento progressivo da curva de frequência fundamental. E, mais do que isso, ao menos segundo esses dados, a relação pode ser mais precisamente descrita pela equação da reta de regressão linear.

Os pouco menos de 50% de variância não explicada da duração da unidade VV saliente devem ser buscados em outras variáveis preditoras que podem, em seguida, se combinar linear ou não linearmente para explicar os valores da duração. A combinação de mais de uma variável preditora é um modelo chamado de regressão linear múltipla ou multivariada. A introdução de relações não lineares entre as variáveis, como funções exponenciais, logarítmicas e inversas permite explorar relações não lineares entre variáveis, modelo chamado regressão não linear. Por exemplo, o modelo não linear com a função 1/dur explica cerca de 69% da variância da duração através da mediana da F0.

Além de modelos numéricos como os abordados aqui, o modelo de regressão logística é muito útil na área de prosódia experimental. A regressão logística é uma técnica que permite relacionar dados categóricos a uma proporção, calculada a partir do cálculo da frequência relativa de alguma variável categórica. Por exemplo, no trabalho de Lima-Gregio (2011), cinco fonoaudiólogas avaliaram trechos de final de enunciado quanto à percepção de laringalização, anotando-se quantas delas marcaram que determinado trecho em torno de sílaba CV tem o fenômeno. Todos os trechos foram avaliados quanto a três parâmetros acústicos correlatos de laringalização: (1) ausência de movimento de transição de formantes, característico de um golpe de glote; (2) ausência do ruído de explosão quando a sílaba tinha oclusiva em ataque; (3) alteração típica no espectrograma de banda larga, com estrias verticais de vozeamento mais afastadas do que no contexto. Essas três variáveis foram medidas ao nível categórico, com presença ou não da alteração.

Siga o roteiro e considere os dados na pasta **Estatistica/RegLog**

do repositório do livro para montar o modelo de regressão logística com proporções modeladas por uma distribuição quasi-binomial. Apenas a ausência de movimento de transição de formantes se mostrou significa- tiva (p = 6, 65.10^{-10}) para o nível de significância de 5%, conforme se vê no esquema abaixo:

	Df	Deviance	Resid. Df	Resid. Dev	F	Pr(>F)
NULL			156	99.644		
FormantT	1	17.587	155	82.056	43.374	6.645e-10

Esse resultado sugere, ressalvada a necessidade de uma investigação mais ampla, que a ausência de transição formântica na sílaba CV seja um fator muito saliente para que especialistas percebam que houve laringalização.

O potencial de aplicação da regressão logística é muito amplo, pois concerne a relação entre categorias e proporções. Assim, é a técnica mais apropriada em sociofonética para relacionar a presença de determinados condicionantes sociais e a proporção de algum fenômeno prosódico.

Tendo em vista a impossibilidade prática de considerar um número consideravelmente amplo de locutores ou de enunciados para análise em prosódia experimental, é necessário avaliar o efeito aleatório de outros possíveis locutores sobre as variáveis dependentes, tarefa do modelo de efeitos mistos.

6.1.6 Modelo de efeitos mistos

Na área de fonética experimental, que inclui a de prosódia experimental, é muito importante que examinemos se existe algum efeito de uma variável independente (categórica como na ANOVA, intervalar como na regressão linear) sobre uma variável dependente, descontan-

do especialmente os fatores inerentes à variação na variável dependente advinda: de participantes da pesquisa, sejam eles locutores ou ouvintes; de sentenças usadas no experimento, isto é, de fatores cujos níveis podem ser superiores ou muito superiores aos contidos nos dados, que é justamente o caso de sujeitos e sentenças. Para mais informações sobre modelos de efeitos mistos consulte o leitor o capítulo 7 do livro de Baayen (2008). Para uma ampla consulta sobre cálculo de coeficiente de determinação, poder do teste, comparação entre modelos sem e com efeitos aleatórios do modelo visto, recomendo a página de Ben Bolker: https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html.

Considere o mesmo conjunto de dados usado na seção 6.1.4, que se encontra também na pasta **Estatistica/Modelo Misto**, arquivo **AllLanguagesREST**. Após construir um primeiro modelo com o desvio-padrão da F0 (variável *fosd*) como variável dependente, língua (variável *ling*) e estilo (variável *estilo*) como variáveis independentes da parte fixa do modelo misto e sujeito (variável *suj*) como variável de efeitos aleatórios, verifica-se que o estilo não é variável significativa nem sua interação com língua, assim vamos mostrar os resultados de um novo modelo que considera apenas a variável *ling* como variável independente de efeito fixo.

O modelo misto que montamos considera os sujeitos como variável aleatória sem relação com as línguas respectivas, isto é, sem considerar que em cada língua haveria um comportamento particular para o desvio-padrão da F0 que merecesse incluir no modelo. A razão principal é que não nos importa, no momento, entender o que cabe ao sujeito em cada língua, uma vez que não é um efeito fixo significativo, como visto acima. A variável *fosd* tem distribuição que não passou em teste de normalidade e, por conta disso, foi preciso construir um modelo misto generalizado[11] em que um equivalente não paramétrico

11 Utilizamos a função glmmPQL() do R, conforme roteiro no repositório do livro.

do teste é empregado.

Os resultados desse modelo misto no R revelam o que segue no esquema abaixo, cujos pontos mais relevantes serão destacados. O leitor pode usar o roteiro que se encontra na pasta **Estatistica/ Modelo Misto** para praticar com outras variáveis dependentes.

```
Random effects:
 Formula: ~1 | suj
            (Intercept)  Residual
StdDev:    0.3154051 0.6640363

Fixed effects: f0sd ~ ling
                Value Std.Error  DF   t-value p-value
(Intercep)  1.0376947 0.1099973 345  9.433819  0.0000
lingFR     -0.7620656 0.1511374 345 -5.042205  0.0000
lingPB     -0.3280378 0.1455767 345 -2.253368  0.0249
lingPE     -0.3145618 0.1584690  35 -1.985005  0.0550
```

Na parte dos efeitos aleatórios (*Random effects*), vê-se variável *suj* que representa os locutores explicando cerca de 10% da variância, valor obtido pela razão entre os quadrados dos desvios-padrão do coeficiente de intercepção (*Intercept*)e do total, resíduos e coeficiente, isto é: $\left(\frac{0{,}315}{0{,}315 + 0{,}664}\right)^2$ do desvio-padrão da F0. O que resta de variância a explicar é atribuído aos efeitos fixos, que mostra a distinção da variável dependente do alemão (o R toma a primeira variável da ordem alfabética para compara com as demais) com relação a francês e PB (não é distinto de PE, que tem p-valor maior que 0,05).

É justamente o que é mostrado na forma de valor de t, de um teste de Student, para os dados do PB (*Intercept*) e depois dessa língua para cada uma das outras[12].

12 Embora se possa usar uma função para se obter, a partir dos valores de t, os p-valores, há funções específicas no R que o informam diretamente conforme se vê abaixo pelo uso da função *Anova* do pacote *car*.

	Chisq	Df	Pr(>Chis)
ling	27.893	3	3.825e-06

O resultado revela que há diferença significativa entre os desvios-padrão médios da F0 entre as línguas, a despeito da variabilidade da mesma variável nos sujeitos, pois o p-valor (aprox. 0.00046) é menor do que 5%.

Comparando com a ANOVA simples, de efeitos fixos apenas, cujo resultado é apresentado logo abaixo, percebe-se que o p-valor é bem superior no modelo de efeitos mistos, uma vez que seu baixo valor no modelo simples de ANOVA não considerava outros sujeitos possíveis, como faz o modelo de efeitos mistos. Para concluir, aplicaremos o teste *post hoc* não paramétrico de Wilcoxon para indicar quais línguas são significativamente distintas para o desvio-padrão da F0 e de quanto são distintas. A aplicação desse teste evita a necessidade de se checar as suposições para aplicação de um teste paramétrico.

	Df	Sum Sq	Mean Sq	F value	p-value(>F)
ling	3	102.4	34.14	32.9	<2e-16 ***
Residuals	388	402.6	1.04		

O teste *post hoc* revela que as únicas línguas que não são significativamente distintas para esse parâmetro são PB e PE, com valor médio de cerca de 2,2 semitons, contra 1,5 semitom para o francês e 3,0 semitons para o alemão.

O número de sujeitos para uma análise inferencial confiável é sempre uma questão que norteia qualquer experimento. Se for um dos fatores aleatórios de um modelo de efeitos mistos, o fator sujeito pode dar resultados relevantes, se não for muito restrito. No exemplo que demos acima eram dez por língua. Mas como se pode avaliar, em modelos

menos complexos, o número de sujeitos para que um modelo tenha uma probabilidade razoável, digamos, 80%, de apontar uma diferença significativa, caso ela exista, pode ser apontado por um procedimento de cálculo chamado de poder do teste (*power of the test*).

6.1.7 Poder de um teste

O poder de um teste estatístico é a probabilidade de se rejeitar uma hipótese nula de fato falsa. Como a probabilidade de aceitar uma hi pótese nula falsa é o erro do tipo II, assinalado pela letra β, o poder do teste é seu complemento, isto é, 1-β. Embora raramente reportado nos artigos científicos, o poder do teste só pode ser avaliado se se tem uma estimativa do tamanho do efeito, cujo cálculo difere para cada teste estatístico.

Consideremos o exemplo acima do teste t de variáveis independentes com dados para a pesquisa em fonética forense para calcular o seu poder. Por enquanto não se trata de número de sujeitos, uma vez que é a comparação de dados de parâmetros melódicos de um sujeito etiquetado como "criminoso" e outro como "suspeito". O poder do teste t vai revelar se o número de dados do experimento é de fato suficiente para estatisticamente rejeitar uma hipótese nula falsa. O cálculo exige que se informe o número de dados para cada grupo de amostras, bem como o tamanho do efeito *d* do teste t, definida pela equação 6.2, em que μ_1 e μ_2 são as estimativas das médias das populações referentes às amostras, dadas pelas médias de cada grupo e σ é o desvio-padrão comum dos resíduos.

$$d = \frac{\mu_1 - \mu_2}{\sigma} \tag{6.2}$$

Aplicando uma função do R que calcula o poder do teste[13] obtém-se o resultado abaixo. Como o poder é aproximadamente 1, não é

13 A função *pwr.t2n.test* do pactote *pwr*.

preciso adquirir mais dados, as amostras são suficientes para se tomar uma decisão confiável, o de rejeição da hipótese nula, conforme acima, com valores de média de 148 Hz para o criminoso e 158 Hz para o suspeito.

```
n1 = 287
n2 = 2181
d = 25.48605
sig.level = 0.05
power ~ 1
```

De certa forma já se esperava um resultado assim, afinal o número menor de dados era 287. Mais crucial é testar o poder do teste de regressão linear acima, que tem apenas 32 pares de dados de duração e mediana da Fo. Usando uma função do R para seu cálculo se obtém esse resultado abaixo em que u e v são os graus de liberdade respectivos das variáveis predita e preditora e o tamanho do efeito é definida pela razão entre os coeficientes de determinação e seu complemento ($R^2/(1-R^2)$).

```
u = 31
v = 31
effect size= 1.188184
sig.level = 0.05
power = 0.98
```

Como se vê, o poder do teste também é bem superior a 80%, o que revela que, de fato, podemos ter segurança, sem a necessidade de recolha adicional de dados, do resultado a que se chegou acima, de que existe uma correlação entre as variáveis mediana da Fo e duração na unidade VV saliente, a do final do grupo acentual.

Suponhamos agora que tenhamos gravado narrativas de um gru-

po experimental de 15 sujeitos com uma certa patologia para extrair, de curtas narrativas de cada locutor, um valor mediano da F0 por narrativa cujos valores médios foram 140 Hz. Suponhamos ainda que não temos mais possibilidade de acesso a esses sujeitos e que eram os únicos na região com a alteração que se deseja estudar que afeta o nível da F0. E que na época tenha sido usado um grupo controle também de 15 sujeitos saudáveis, de mesma faixa etária e escolaridade e da mesma região dialetal. Suponhamos ainda que esse grupo tenha produzido narrativas curtas com valor médio das medianas da F0 de 120 Hz. Considerando os dois grupos, pode-se calcular o desvio-padrão dos resíduos, com valor de 25 Hz. Temos, assim, todos os elementos para cálculo do poder do teste t, cujo resultado é o que segue:

```
n1 = 15
n2 = 15
d = 0.8
sig.level = 0.05
power = 0.56
```

Pode-se ver que o poder do teste é menor que 80%. Sendo assim, para se ter confiança na decisão que se tomaria, é preciso gravar mais narrativas. Como não se pode ter acesso ao grupo experimental, conforme explicado acima, decide-se gravar mais sujeitos do grupo controle. A mesma função, quando omitimos o número de sujeitos num dos grupos (o controle, nesse caso) e informamos que queremos um poder do teste de 0,80, informa quantos sujeitos no grupo controle são necessários para se ter esse poder ao nível de significância de 5%:

```
n1 = 15
n2 = 77
d = 0.8
sig.level = 0.05
power = 0.8
```

Observe que o resultado é bastante surpreendente, pois revela que ainda é preciso gravar 62 sujeitos (77-15). Mas isso decorre porque a diferença média entre as medianas dos dois grupos é relativamente pequena no estudo inicial. Assim, o experimentador, se realmente quiser ter alguma segurança em sua decisão, deverá realmente gravar as 62 narrativas faltantes. Observe o leitor que, se os desvios-padrão respectivos dos grupos experimental e controle forem cerca de 30 Hz e 15 Hz, por exemplo, o teste t de variáveis independentes daria um valor de t de 2,3 e um p-valor de 0,014. Assim o experimentador, sem conhecer a relevância do poder do teste, teria rejeitado a hipótese nula sem o cuidado de ver se tem condições adequadas de fazê-lo, pelo cálculo de poder.

A próxima seção apresenta todos os aspectos de dois experimentos na área de prosódia experimental para que possa servir de modelo para a montagem de um desenho experimental. Ressaltaremos e discutiremos as escolhas e decisões para orientar o pesquisador interessado. Dados e roteiro dos testes estatísticos realizados no R se encontram no repositório do livro.

6.2 Exemplos de desenho experimental em prosódia acústica

O primeiro exemplo é de um estudo recentemente publicado de Barbosa e Niebuhr (2020) sobre as modificações respiratórias e acústi-

cas na fala persuasiva em inglês. O segundo exemplo trata da relação entre a produção e a percepção dos ritmos da leitura e da narração em português brasileiro, publicado por Barbosa e Silva (2012). Embora não sejam estudos experimentais publicados, mas ilustrativos, concluímos com avaliação de dados que nos permitem fazer duas homenagens. O primeiro deles é o exame de diferenças prosódicas de diferentes interpretações profissionais da leitura do "Soneto da Separação" de Vinícius de Moraes, como representativo do imenso e criterioso trabalho sobre expressividade da fala que tem sido conduzido com esmero e delicadeza por minha colega Sandra Madureira, da PUC-SP; o segundo estudo examina diferenças melódicas em leitura de uma curta fábula em diferentes línguas regionais românicas na França, como homenagem ao linguista de campo, Philippe Boula de Mareüil, que tem percorrido o mundo todo gravando e conhecendo línguas minoritárias e as comunidades que militam por sua preservação.

6.2.1 Diferenças melódicas e respiratórias na persuasão

As principais questões que nortearam o estudo sobre fala persuasiva, fruto da colaboração entre as universidades de Campinas (Unicamp) e do sul da Dinamarca (*Southern Denmark University*), se guiaram por alguns pressupostos da Retórica, através de recomendações como "make sure you're breathing deeply into your belly" (CABANE, 2012), em que se coloca o alegado papel primordial da respiração abdominal. Como assinalamos em nosso estudo, há evidência empírica de que a respiração abdominal tem algum benefício para cantores (SALOMONI; HOORN; HODGES, 2016; THORPE et al., 2001) e possa ser útil no tratamento de alterações vocais e respiratórias (XU; IKEDA; KOMIYAMA, 1991). Além disso, a Retórica também aponta que a postura em pé favorece a persuasão. Estudos

prévios mostraram que, no que diz respeito à respiração, a fase expiratória deve ser curta na fala persuasiva (NIEBUHR; NOVÁK-TÓT; BREM, 2016; ROSENBERG; HIRSCHBERG, 2005) e, no que diz respeito aos parâmetros prosódico-acústicos, encontraram- se valores mais elevados da média, amplitude e máximo da F0 e da ênfase espectral (NIEBUHR; NOVÁK-TÓT; BREM, 2016; ROSEN- BERG; HIRSCHBERG, 2005; TOUATI, 1994; D'ERRICO et al., 2013; NIEBUHR; SKARNITZL, 2019).

Como depreende o leitor, esses foram os pontos de partida do estudo que motivaram duas principais hipóteses: (1) uma mudança significativa na atividade respiratória na respiração abdominal, (2) a confirmação de maiores valores dos parâmetros melódicos e de ênfase espectral na fala persuasiva e (3) valores mais elevados de parâmetros prosódico-acústicos e maior expansão de tórax e/ou abdômen na postura em pé.

Para verificar essas hipóteses, gravamos o corpus PERBREATH no laboratório da *Southern Denmark University* com 18 estudantes e professores alemães da universidade que passaram por algumas horas de treinamento formal sobre fala carismática para fins de promover produtos industriais. Todos eles leram, de duas maneiras e em inglês, um trecho de um texto de discurso sobre a venda de um app de controle de horas de trabalho remoto. As leituras foram primeiramente uma leitura habitual e depois persuasiva, como para vender um produto que eles mesmos teriam feito. De forma alternada entre os participantes, eles repetiram as duas leituras uma vez de pé e outra vez sentados. O nível de inglês dos participantes é de pelo menos B2[14] e todos usam a língua diariamente na universidade para falar com colegas e funcionários, uma vez que a universidade é fortemente internacionalizada.

14 No Quadro Europeu Comum de Referência para as Línguas, é o nível intermediário superior com habilidades definidas precisamente e que podem ser conhecidas aqui: https://www.cambridgeenglish.org/br/exams-and-tests/cefr/.

A gravação dos movimentos de expansão do tórax e do abdômen foi feita com o dispositivo Resp Track, conforme mencionamos na seção 4.10, aparelho que é um pletismógrafo respiratório de indutância criado para medir a área da seção transversal do tórax e abdômen por meio de cintas providas de indutores. Simultaneamente, um microfone unidirecional foi usado para capturar o sinal de fala.

Dois scripts foram desenvolvidos para obter os valores de cinco variáveis relacionadas ao ciclo respiratório a partir dos sinais do tórax e do abdômen nas duas posturas e nas duas condições de leitura e sete variáveis prosódico-acústicas calculadas para cada ciclo respiratório a partir do sinal do microfone. No que segue, apresentaremos algumas dessas variáveis, quais sejam, amplitude global dos movimentos de tórax e abdômen e duração da fase expiratória, bem como, entre os parâmetros acústicos, máximo e amplitude da F0 e ênfase espectral.

Quanto à amplitude global, utilizamos o modelo de ANOVA de medidas repetidas[15] que apontou um aumento de cerca de 2 dB na fala persuasiva apenas para o tórax ($F_{1,17}$ =13.9, p < 0.002), mas não para o abdômen, bem como nenhuma diferença significativa para o fator postura, isto é, é o mesmo padrão respiratório global, tanto sentado quanto em pé.

Quanto à duração da fase expiratória, usamos o teste SHR, o equivalente não paramétrico da ANOVA de dois fatores, para o exame dos fatores condição da leitura e sexo. Encontramos, para um p-valor de pelo menos 0,02, que a fala persuasiva tem duração da expiração menor em 400 ms, o que envolve tórax e abdômen.

Para os parâmetros acústicos, modelos de efeitos mistos aponta- ram diferenças significativas para os fatores sexo e condição de leitura, mas não para a postura. Os dados e um roteiro de aplicação de modelos mistos se encontram no repositório do livro, pasta **Esta-**

15 Para esse teste sugerimos a leitura de seu uso em CRAWLEY (2007) e em BAAYEN (2008).

tistica/ModelosEfeitosMistos. Em todos os casos o poder do teste foi de aproximadamente 1, devido ao grau das diferenças médias e ao número de dados. As variáveis amplitude e máximo da F0 pouco se desviaram da gaussiana, mas não a ênfase espectral como se pode ver na Figura 6.5, pelos pontos muito fora do espaço compreendido entre as linhas tracejadas. Por conta disso utilizamos um modelo de efeitos mistos linear paramétrico para as duas primeiras e um generalizado para a ênfase espetral.

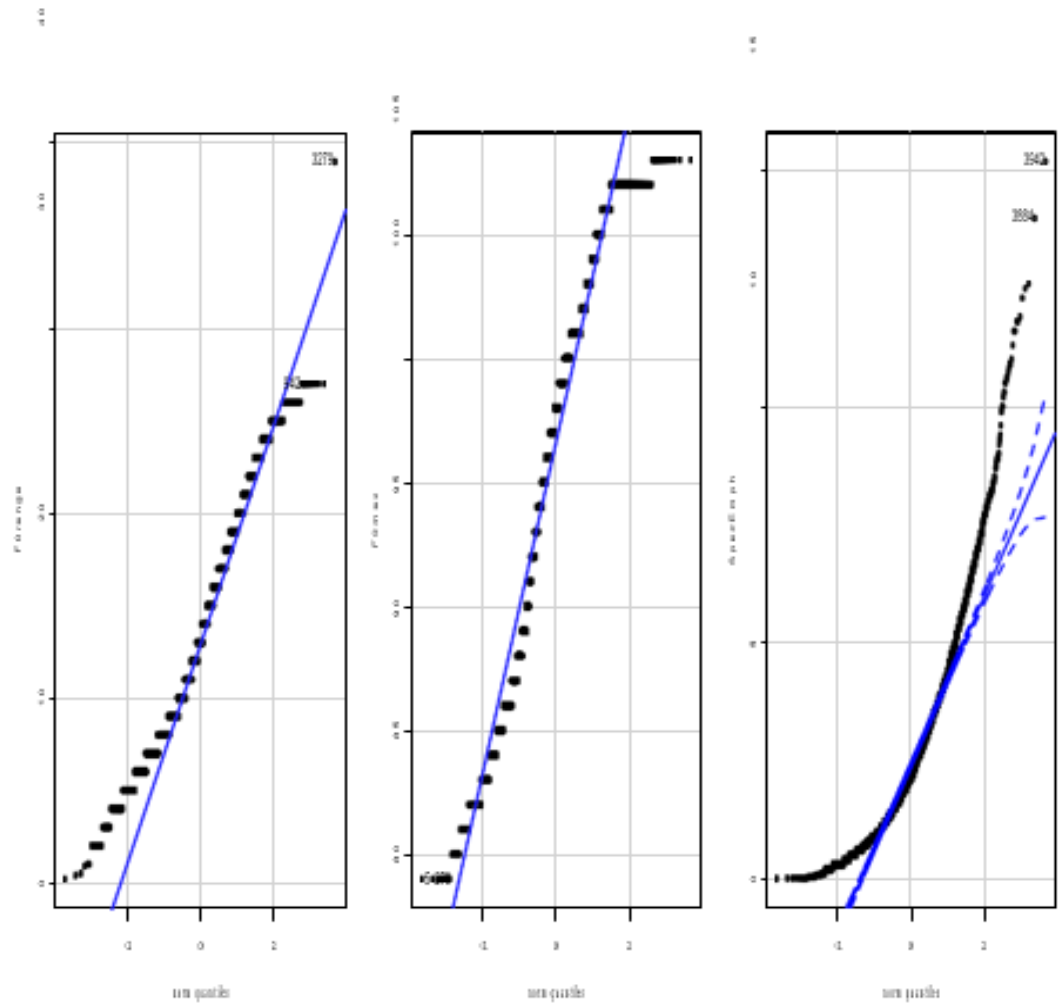

Figura 6.5 – Reta de regressão linear superposta aos dados da Figura 6.3.

A variância explicada pelo fator aleatório, o fator sujeito, é baixa, sendo de 22% (6,388/(6,388+22,352)) para a amplitude da F0 (variável *Forange*), de 27% para o máximo da F0 (variável *Fomax*) e de 9% para ênfase espectral (variável *SpecEmph*). Isso é sinal de que o

número de sujeitos foi suficiente para apontar as diferenças significativas, pois sua variabilidade pouco influencia os resultados, que foram os dos esquemas que seguem.

Para a amplitude da F0, o p-valor indica que o fator sexo não é significativo, tendo as mulheres mesma amplitude que os homens, independentemente da condição de leitura. Entre as condições de leitura, o valor da amplitude da F0 é maior na fala persuasiva: 12,8 semitons (habitual) para 15,3 semitons (persuasão).

	Chisq	Df	Pr(>Chisq)	
task	112.9625	1	<2e-16	***
sex	0.8095	1	0.3683	
task:sex	0.6697	1	0.4132	

Para o máximo da F0, há diferença significativa para sexo e condição de leitura, mas a diferença para sexo era esperada, por as mulheres terem nível superior da F0, mesmo em semitons, visto que usamos a mesma referência de 100 Hz para o cálculo do semitom em ambos os sexos. Entre as condições de leitura, o valor máximo para os sexos respectivos subiu de 99,8 semitons (habitual) para 102,4 semitons (persuasão) nas mulheres e de 90,8 semitons (habitual) para 94,5 semitons (persuasão) nos homens.

	Chisq	Df	Pr(>Chisq)	
task	192.7938	1	< 2.2e-16	***
sex	40.8851	1	1.614e-10	***
task:sex	3.5469	1	0.05966	.

Para a ênfase espectral, o p-valor indica que o fator sexo só é significativo na interação entre os fatores, como se vê em seguida pelos valores médios. Entre as condições de leitura, o valor da ênfase espectral é significativamente maior na fala persuasiva: 1,8 dB (habitual)

para 3,3 dB (persuasão) nas mulheres e 2,1 dB (habitual) para 3,1 dB (persuasão) nos homens, com aumento maior nas mulheres, o que foi acusado pela interação significativa.

	Chisq	Df	Pr(>Chisq)	
task	336.394	1	< 2.2e-16	***
sex	0.0452	1	0.8317	
task:sex	17.4471	1	2.954e-05	***

Com esses resultados dos testes estatísticos, podemos voltar às hipóteses do estudo para concluir que não há papel privilegiado da respiração abdominal na persuasão, mas sim do tórax; que há, de fato, valores mais altos dos parâmetros melódicos e de ênfase espectral na fala persuasiva e que não há efeito da postura em pé ou sentado quanto aos parâmetros respiratórios ou acústicos.

A escolha dos sujeitos falando em língua segunda foi circunstancial, embora não coloque em questão os achados apresentados, pois certamente o grau de fluência elevado que têm em inglês não afeta a habilidade em colocar o aprendizado que receberam sobre fala persuasiva em ação. No entanto, são pessoas que não usam a persuasão como parte de suas atividades diárias, como fazem vendedores e empreendedores. Assim, dois aspectos que poderiam ser estudados ainda são (1) como profissionais que usam a persuasão modificam seus padrões acústicos e respiratórios e (2) como a menor proficiência numa língua afetaria esses padrões. Quanto ao segundo aspecto, o estudo de (ISEI- JAAKKOLA; NAGANO-MADSEN; OCHI, 2018) mostra que locutores suecos ou japoneses lendo na língua segunda (aprendizes suecos do japonês e aprendizes japoneses do sueco) usam mais os músculos do tórax e que os picos dos movimentos musculares respiratórios são, em língua segunda, mais frequentes, irregulares e de menor amplitude.

6.2.2 Vínculo entre produção e percepção do ritmo da fala

Há alguns anos, investigamos a relação entre a capacidade de discriminar o ritmo da fala em diferentes trechos com os parâmetros acústicos que poderiam explicar a discriminação feita pelos ouvintes (BARBOSA; SILVA, 2012). Tínhamos a intuição, por experiência diária, de que os ouvintes tenderiam a dizer que dois trechos de fala são tanto mais distintos no ritmo da fala quanto mais distintas forem as taxas de elocução, as variações, níveis ou taxa de picos da F0 e o esforço vocal medido pela ênfase espectral. Assim, foi essa nossa hipótese, de que haveria uma relação direta e crescente entre diferenças nos valores médios desses parâmetros nos trechos e a proporção de respostas "diferente" quanto ao ritmo. Por os testes terem sido feitos com ouvintes leigos, usamos o termo "modo de falar".

O corpus é um subconjunto do corpus BELÉM de leitura da história da origem dos pastéis de Belém seguida da narração consecutiva com as próprias palavras. Do corpus, para garantir um teste de percepção que durasse cerca de 25 minutos, escolhemos leitura e narração de três locutores paulistas, duas mulheres e um homem entre 30 e 45 anos, todos de nível universitário. Trechos de áudio entre 9 e 18 segundos foram retirados para montar um teste de discriminação no Praat com os áudios oriundos aleatoriamente de qualquer locutor e qualquer um do dois estilos que foram avaliados por dez ouvintes universitários em seus vinte anos.

Os trechos de áudio de cada par foram separados por um tom puro de cerca de 1000 Hz para que se soubesse quando se passava de um trecho para outro. Após escutar, o ouvinte tinha que clicar numa escala de cinco graus variando de 1 (mesmo modo de falar) a 5 (modos de falar completamente diferentes), segundo sua percepção da distinção. Os testes foram feitos também com os áudios deslexica-

lizados, mas não foram encontradas diferenças significativas entre as respostas dadas com ou sem a deslexicalização. Onze parâmetros acústicos foram calculados para todos os trechos e, em cada par apresentado aos ouvintes, calculou-se a diferença média entre os parâmetros. Avaliou-se a correlação entre os parâmetros prosódico-acústicos e as respostas dos ouvintes. Três modelos de regressão com maior variância explicada são mostrados aqui para fins de aprendizado, considerando apenas variáveis com correlação superior a 50%, o que descartou a mediana da F0 e a ênfase espectral que faziam parte das hipóteses. Sobraram apenas parâmetros duracionais. Os dados e roteiro das análises feitas se encontram na pasta **Estatistica/RegLin** do repositório do livro. Para as análises, as respostas dos ouvintes foram transformadas linearmente de 1 a 5 para -1 a 1, sendo 0 a resposta neutra. As regressões consideram o nível de significância de 5%.

O primeiro modelo considera a correlação entre diferença na taxa de elocução (variável *sr*) e resposta do teste de discriminação (variável *perc*). Tanto o coeficiente de intercepção quanto o de inclinação foram significativos, sendo o coeficiente de determinação (R^2), ou seja, a porcentagem da variância da resposta explicada pela diferença média da taxa de elocução, de cerca de 48%, próximo ao desejável de pelo menos 50%. No entanto, o poder do teste foi de apenas 65%. Como para um modelo simples esse foi o parâmetro com mais correlação, era preciso mudar algo e resolvemos fazer uma regressão não linear com a função logarítmica, após ter notado que, à medida que a diferença em taxa de elocução aumenta, a resposta média dos ouvintes aumenta numa taxa menor, como se pode ver no lado esquerdo da Figura 6.6.

Esse segundo modelo também teve coeficiente de intercepção e de inclinação significativos, com coeficiente de determinação de cerca de 59% e poder do teste de 81%, que foi um grande ganho. Em seguida, pensamos em combinar linearmente as variáveis dois a dois e o modelo com maior coeficiente de determinação e poder foi o que con-

sidera as diferenças de taxa de elocução (variável *sr*), as diferenças na taxa de picos locais de duração normalizada da unidade VV (variável *pr*) e a interação entre ambas. Esse modelo produziu um coeficiente de determinação de cerca de 64% e poder do teste de 89%, conforme tabela abaixo.

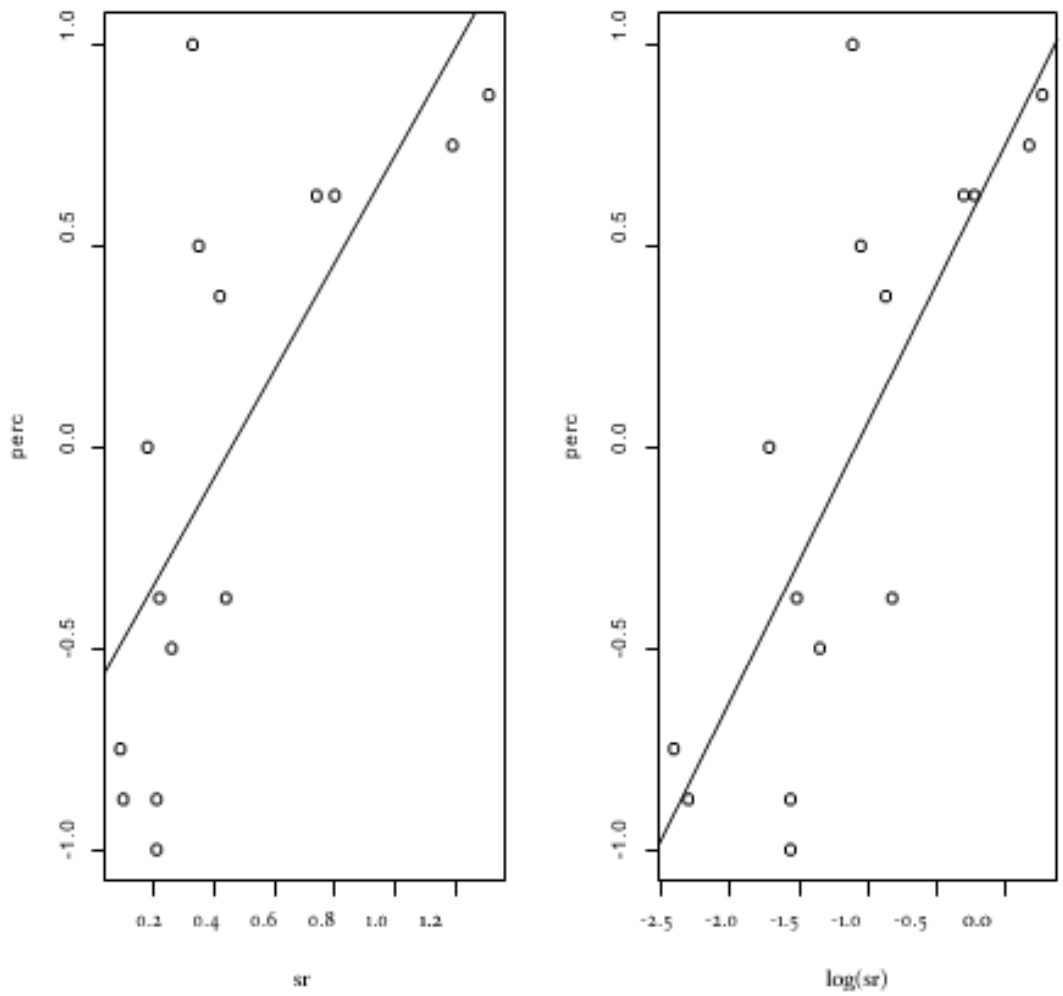

Figura 6.6 – Gráfico de diferença de taxa de elocução (sr) vs. resposta média dos ouvintes no teste de discriminação. A relação da esquerda é linear e a da direita considera o logaritmo da diferença de taxa de elocução. As retas de regressão são dadas para ambos os gráficos.

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -1.5216     0.3674    -4.141  0.00164 **
sr            2.7128     0.7896     3.436  0.00556 **
pr            9.6812     3.6902     2.623  0.02368 *
sr:pr       -10.3152     3.7138    -2.778  0.01798 *
---
Residual standard error: 0.427 on 11 degrees of freedom
Multiple R-squared: 0.7185,Adjusted R-squared: 0.6417 F-
statistic: 9.359 on 3 and 11 DF, p-value: 0.002309
```

Quanto às hipóteses do estudo, apenas taxa de elocução e taxa de picos locais de duração normalizada da unidade VV foram relevantes para explicar a discriminação de modo de falar feita pelos ouvintes. Esse resultado é interessante pelo fato de dizer respeito à sucessão silábica dada pela taxa de elocução e à sucessão de sílabas proeminentes dada pela segunda taxa, que são parâmetros classicamente associados ao conceito de ritmo silábico e acentual.

Desenvolvimentos desse estudos podem ser muitos, como o exame de outros parâmetros melódicos como desvio-padrão da F0 e taxas de subida e descida da taxa da F0. Outros estilos podem ser incluídos para confirmar se os parâmetros encontrados aqui ainda são válidos, ampliando a gama de estilos. A isso se pode juntar locutores de diferentes regiões do país, como também com e sem experiência musical, para ver se essa influi na capacidade de discriminação.

6.3 Motivando a investigação em áreas sub-exploradas da prosódia experimental

Conforme comentei na introdução a este capítulo, passo a apresentar o trabalho de dois colegas entusiasmados por seu traba-

lho, para incentivar o leitor a se embrenhar em áreas ainda pouco exploradas da prosódia.

6.3.1 Diferenças de expressividade na fala profissional

Há alguns anos, Sandra Madureira, pesquisadora da PUC-SP, se debruça sobre a relação entre som e sentido (MADUREIRA, 2011, 2016; MADUREIRA; FONTES; CAMARGO, 2019), tocando questões muito atuais da área de prosódia, como a integração da informação de movimentos da face e do som na veiculação da expressividade da fala (MADUREIRA; FONTES, 2015, 2019), incluindo as emoções (MADUREIRA, 2004); como o papel de ajustes articulatórios na descrição da voz e da fala; como a coordenação entre movimento respiratório e som na produção da fala e do canto (BARBOSA et al., 2020), sem contar seus interesses diversos na variação de pronúncia e expressão entre diferentes localidades, diferentes interpretações de textos literários e também na pronúncia de língua estrangeira.

A Prof*a* Sandra também se dedicou ao estudo da fala de locutores profissionais, jornalistas, atores e atrizes, tanto brasileiros quanto portugueses. Escolhemos assim esse campo de sua investigação para motivar o leitor a se inteirar de alguns aspectos da variabilidade prosódica. Vamos ilustrar diferenças entre leitura por profissionais e não profissionais da fala, bem como diferenças individuais na declamação do Soneto da Fidelidade de Vinícius de Moraes. Os arquivos de áudio e de anotação se encontram na pasta **Audios/Capitulo6/Expressividade** e as tabelas de dados, texto lido e roteiros de teste estatístico de Análise Discriminante Linear na pasta **Estatistica/Expressividade**.

Quatro locutoras profissionais da voz, do Clube da Voz em São Paulo e quatro locutoras não profissionais leram o texto que se encontra na pasta mencionada acima. As gravações foram feitas no es-

túdio de Rádio da PUC-SP. Segmentamos toda a leitura em 30 trechos, sendo cada trecho de mesmo conteúdo nas oito locutoras. Em seguida, utilizamos o script *Prosody Descriptor* para gerar parâmetros melódicos e de qualidade de voz para examinar diferenças entre os grupos de locutoras quanto ao uso profissional da voz. Entre as profissionais, todas se encontram na faixa de 40 a 50 anos e entre as não profissionais, duas se encontram entre 20 e 25 anos e duas entre 40 e 45 anos de idade.

As diferenças significativas para um teste de Wilcoxon ao nível de significância de 5% foram para as seguintes variáveis: mediana da Fo (Hz), desvio-padrão da Fo (Hz), taxa de descida da Fo (Hz/quadro), ênfase espectral (dB) e razão harmônico-ruído (dB). Se a mediana da F0 revela, sobretudo nesse caso, diferenças individuais, incluindo as relacionadas à faixa etária, as demais podem revelar aspectos interessantes da fala profissional. Observe na Figura 6.7 que as locutoras não profissionais variam menos a Fo em relação às profissionais, um contrate de 25 Hz vs. 37 Hz, o que tem um efeito de chamar mais a atenção do ouvinte, de fazer o conteúdo do texto evocar modos distintos de uso da melodia, possivelmente prendendo mais a atenção de quem escuta, como o leitor poderá ouvir nos áudios disponibilizados.

Se observar agora a Figura 6.8, verá que as locutoras profissionais têm em média descidas melódicas mais íngremes com valor médio de 4,6 Hz/quadro contra 3,8 Hz/quadro nas não profissionais, causando um efeito mais enfático nas terminações de enunciados.

Quanto ao correlato do esforço vocal, com diagramas de blocos na Figura 6.9, o uso profissional da voz faz com que as locutoras com essa prática façam, em média, menor esforço: 2,0 dB vs. 2,9 dB nas não profissionais. Quanto à razão harmônico-ruído, são as não profissionais que têm valor médio maior, apontando, sobretudo, pregas vocais menos desgastadas, gerando menos ruído, mas que certamente não pode ser desvinculado da faixa etária, porque entre as não profissionais há duas jovens e entre as profissionais (locutora H), há uma

fumante. Os valores médios são de 13,6 dB nas não profissionais e 12,2 dB nas profissionais.

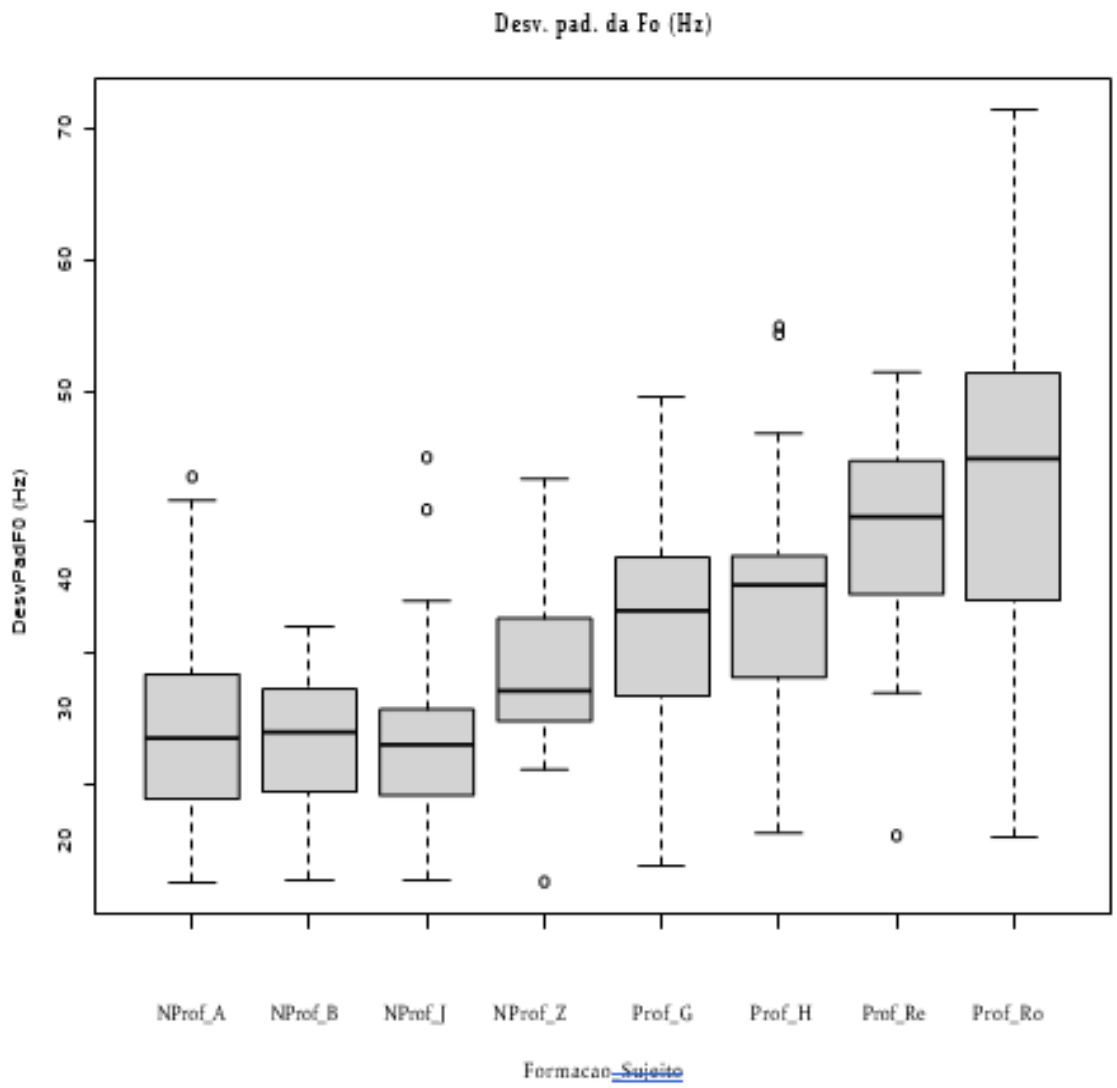

Figura 6.7 – Diagramas de blocos do desvio-padrão da F0 (Hz) para as oito locutoras sendo as profissionais precedidas de "Prof" na abscissa e as não profissionais de "NProf".

Outro campo de investigação caro à Prof*a* Sandra é a criatividade através da expressão vocal, que pode ser ilustrada comparando diferentes declamações do Soneto da Fidelidade retiradas do YouTube, feitas por sete locutores de faixas etárias distintas, a julgar pela própria voz, como o locutor CV, que aparenta ser o mais velho. Essa avaliação pode ser feita pelo próprio leitor, ouvindo os áudios disponibilizados na pasta remota.

As leituras foram segmentadas por versos numa camada de anotação e, na segunda camada, delimitaram-se as pausas silenciosas de

cada um dos sete locutores. Em seguida, utilizamos o script *Prosody Descriptor*, desta vez para obter, além dos parâmetros melódicos e de qualidade de voz, aqueles relacionados ao uso da pausa silenciosa, sua duração média (variável durSIL) e o período médio de sua recorrência (variável IPI, por *Inter Pausal Interval)*, pois a pausa pode ser usada para criar efeitos dramáticos, como se vê na Figura 6.10.

Observe na figura que o locutor YR faz uma pausa de mais de quatro segundos, que tem um efeito dramático ao final da declamação. Observe também que o locutor CV tem as pausas que duram mais e recorrem com menor frequência na declamação.

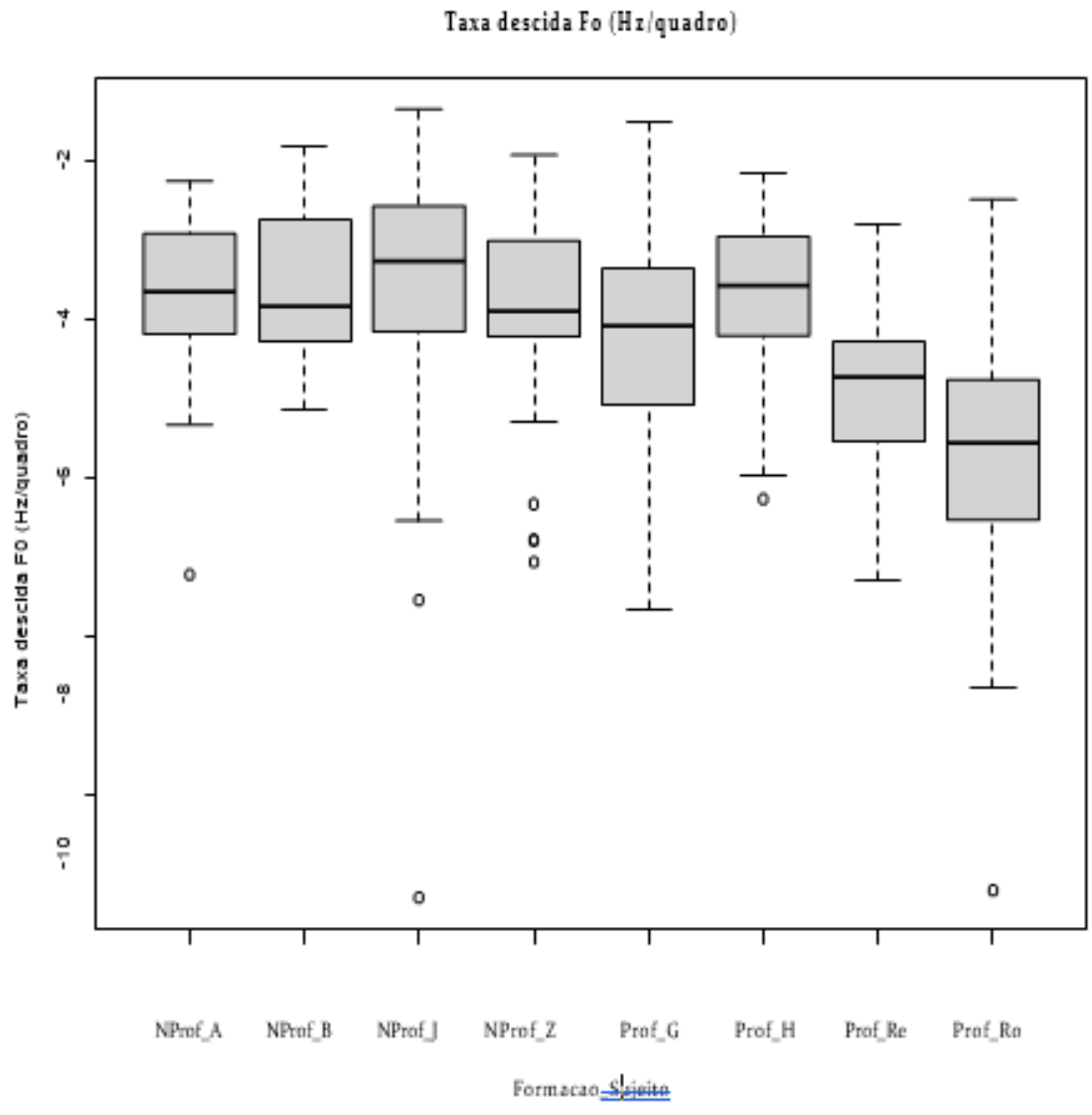

Figura 6.8 – Diagramas de blocos da taxa de descida média da F0 (Hz/quadro) para as oito locutoras sendo as profissionais precedidas de “Prof” na abscissa e as não profissionais de “NProf”.

Quanto aos demais parâmetros, observe aqueles mais diferenciados entre os locutores na Figura 6.11: coeficiente de variação[16] da intensidade (variável *cvint*) em porcentagem, ênfase espectral (variável *emph*) em dB, máximo da F0 (variável *fomax*) em Hz e taxa média de subida da F0 (variável *dfoposmean*) em Hz/quadro.

Observe que o coeficiente de variação da intensidade é maior em ME, que justamente usou o recurso de diminuir a intensidade em alguns versículos para provocar algum efeito no ouvinte. O locutor ML se destaca por ter esforço vocal maior e bem mais variável que os demais, enquanto o locutor SC tem os máximos da F0 mais elevados, usando também o recurso de os variar mais. Juntamente com SM, esse locutor tem as mais altas taxas de subida melódica. YR, o que usou uma pausa extremamente longa para efeito dramático, é o locutor com valores mais baixos e menos variáveis para os quatro parâmetros mostrados aqui. Todos esses aspectos parecem sugerir que, em seu conjunto, esses parâmetros poderiam diferencias os sete locutores, revelando assim que cada um tem características prosódicas singulares.

16 O coeficiente de variacão é a razão entre o desvio-padrão e a média de uma variável, expressando assim uma variabilidade relativa.

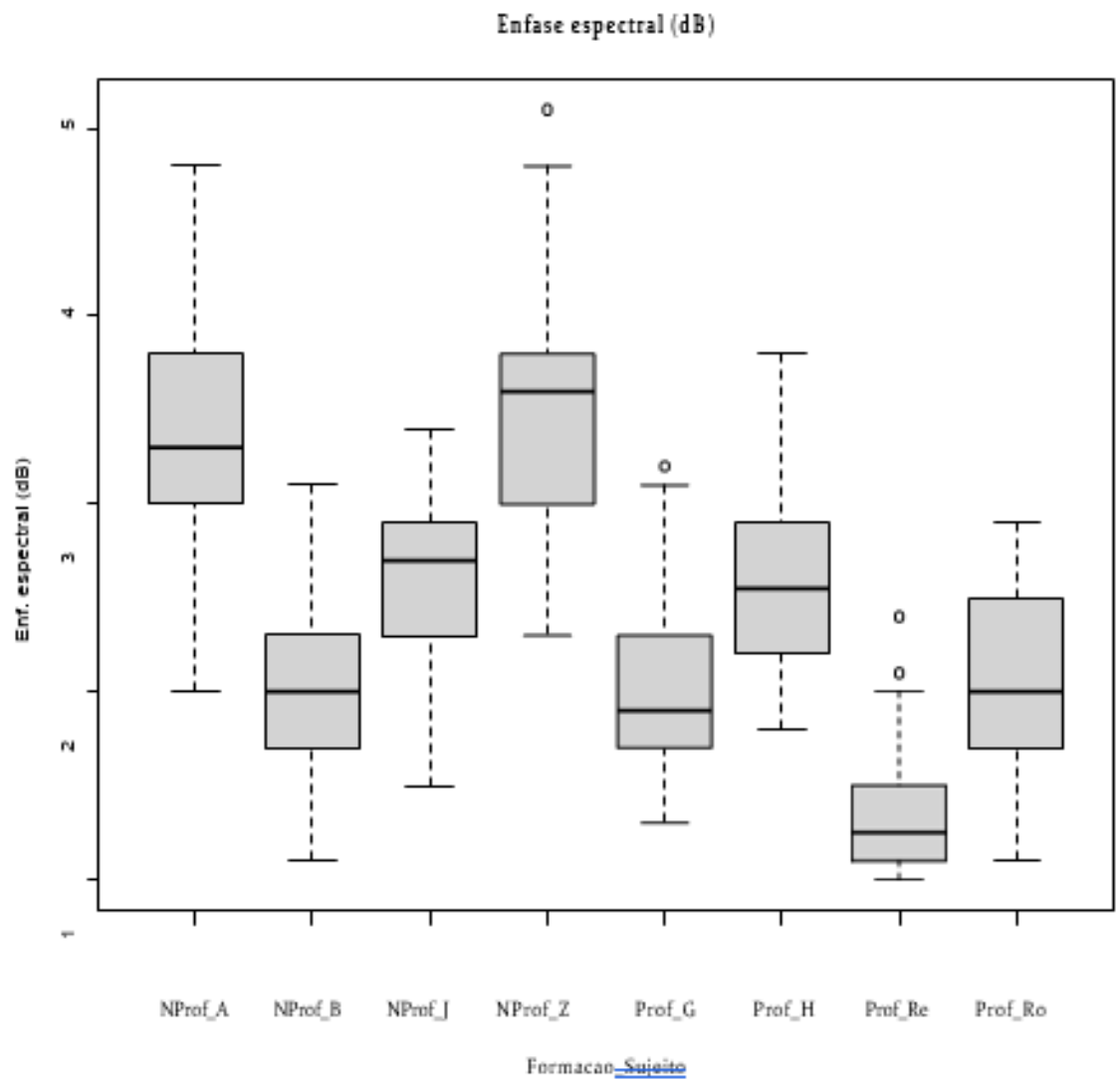

Figura 6.9 – Diagramas de blocos da ênfase espectral (dB) para as oito locutoras, sendo as profissionais precedidas de "Prof" na abscissa e as não profissionais de "NProf".

Suj	CV	JM	ME	ML	SC	SM	YR
CV	86	0	0	0	0	0	14
JM	0	90	0	0	0	10	0
ME	0	0	90	0	0	10	0
ML	0	0	0	82	18	0	0
SC	0	0	0	9	82	9	0
SM	0	15	0	0	0	85	0
YR	0	0	0	0	0	0	100

Utilizamos a técnica da Análise Discriminante Linear (LDA, na sigla em inglês), para classificar os sete locutores no espaço paramétrico formado por oito parâmetros: os quatro mencionados acima acres-

cidos da taxa de picos da F0, da taxa média de descida da F0 e dos desvios-padrão das taxas de subida e descida da F0. O roteiro para realizar o teste no R se encontra na pasta remota mencionada acima. O resultado em porcentagem de reconhecimento pode ser visto acima, revelando que todos os locutores são reconhecidos numa proporção superior a 80%. O locutor CV é confundido em 14% dos versículos declamados com YR, ML confundido 18% dos versículos com SC e SM 15% com JM.

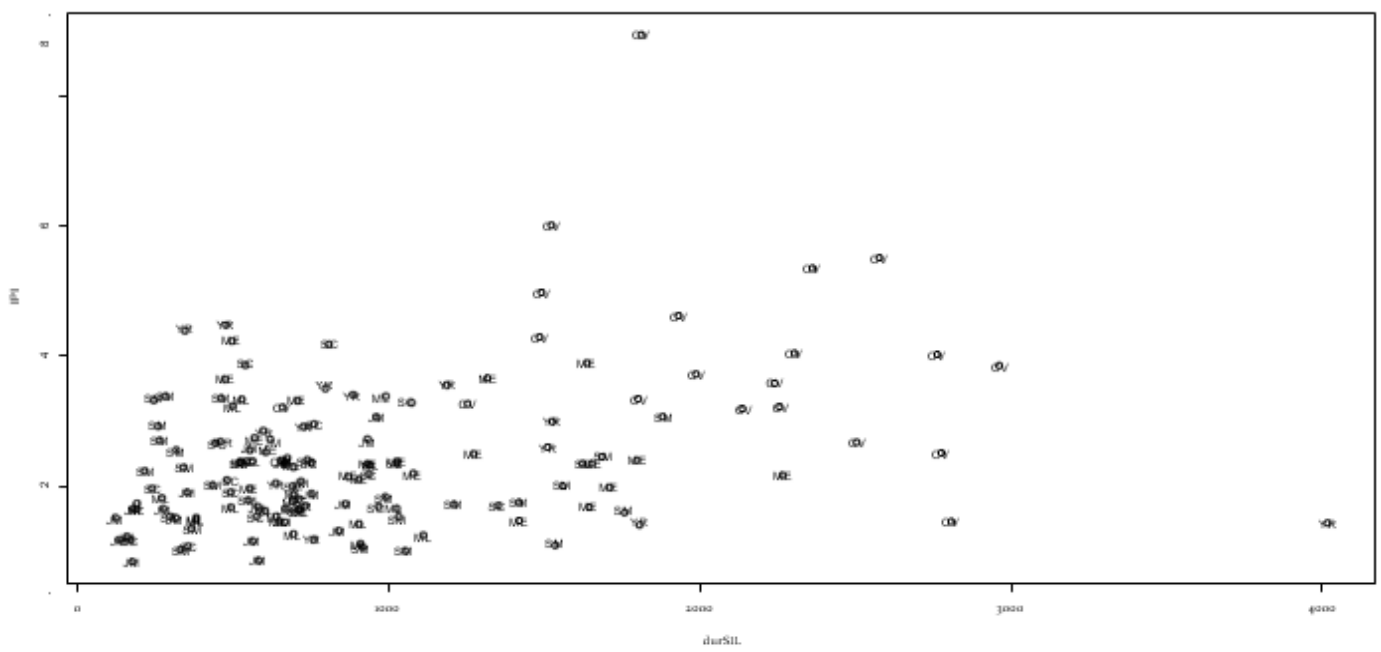

Figura 6.10 – Gráfico de dispersão considerando a duração das pausas silenciosas na abscissa (em milissegundos) e o intervalo de ocorrência com relação à pausa anterior (em segundos) na ordenada, para sete locutores declamando o Soneto da Fidelidade.

Os resultados para as duas ilustrações desta seção revelam, de modo claro, a importância da investigação da variabilidade prosódica para indicar uso profissional da voz, bem como para distinguir a performance dos indivíduos para o uso da fala para fins estéticos.

6.3.2 Diferenças melódicas entre línguas românicas regionais na França

Há mais de dez anos, Philippe Boula de Mareüil, pesquisador do LIMSI (atual LISEN) em Orsay, França, viaja o mundo para gravar línguas minoritárias, tendo começado pelas línguas regionais da França dentro do projeto *Atlas sonore des langues régionales de France* (MAREÜIL et al., 2008). Philippe é também pesquisador de sotaques, com inúmeros trabalhos na área (VAISSIÈRE; MAREÜIL, 2004; WOEHRLING; MAREÜIL, 2006; MAREÜIL et al., 2008; MAREÜIL; BARDIAUX, 2011; MAREÜIL, 2012a) e um livro sobre o assunto (MAREÜIL, 2010), tendo abordado aspectos diversos do sotaque regional e estrangeiro.

No trabalho de Woehrling e Mareüil (2006), a questão da diferenciação acústica de sotaques regionais no território francês é abordada aliando a sua percepção com a análise dos parâmetros acústicos segmentais de realização da vogal neutra (*schwa*), dos valores das frequências dos dois primeiros formantes e da presença de uma consoante de travamento de vogais nasais, muito comum no sul da França. A partir de análises de classificação multidimensional, os autores mostraram que se pode separar o sul do norte da França, como também a região da Suíça românica.

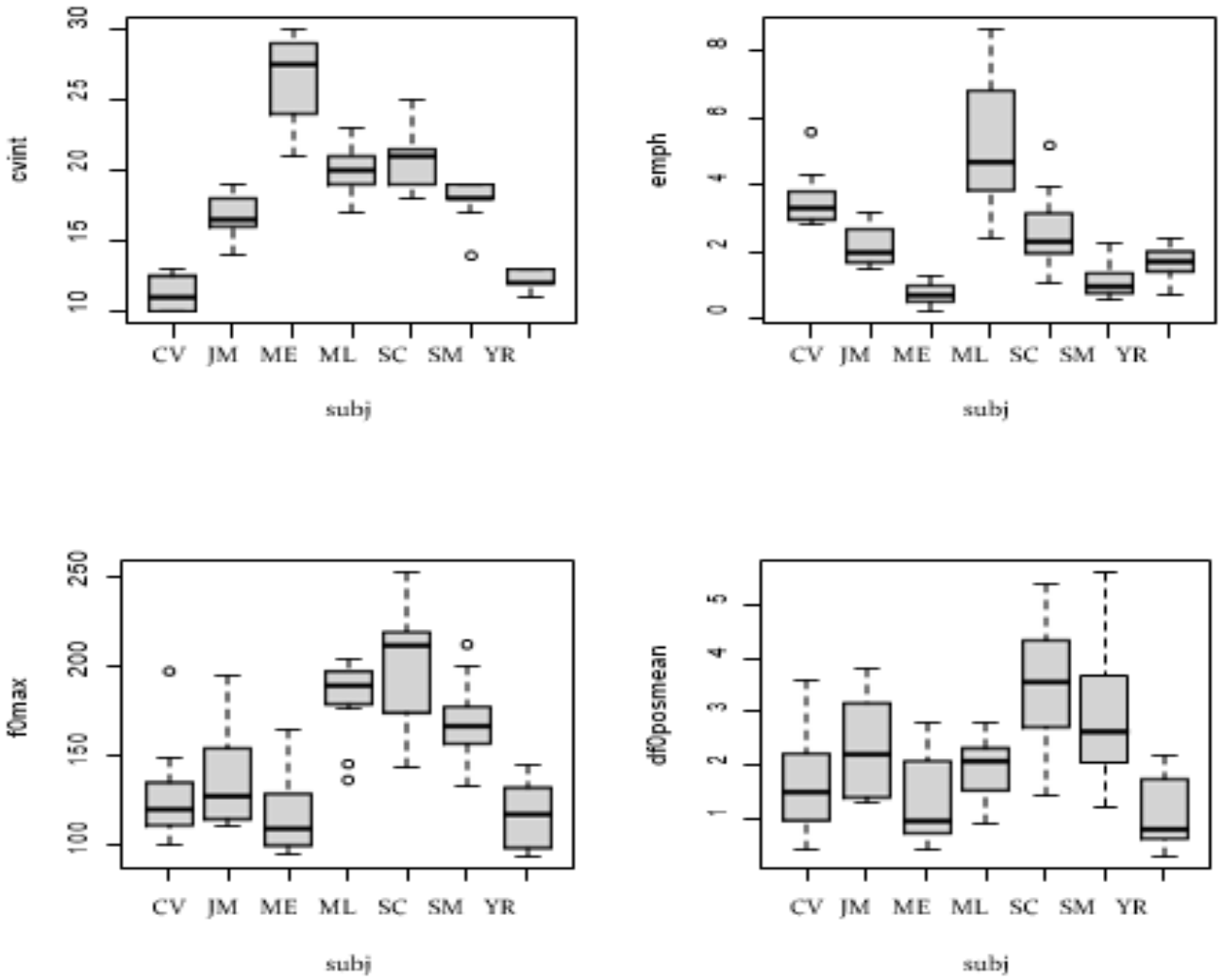

Figura 6.11 – Diagramas de blocos de coeficiente de variação da intensidade em porcentagem, acima à esquerda; ênfase espectral em dB, acima à direita; máximo da F0 em Hz , abaixo à esquerda e taxa média de subida da F0 em Hz/quadro, abaixo à direita para sete locutores que declamaram o Soneto da Fidelidade.

Em um belo estudo sobre mudanças prosódicas diacrônicas no estilo radiofônico (MAREÜIL, 2012b), Philippe estudou como mudou, ao longo de 50 anos, a locução de rádio em Paris, abordando aspectos como proeminência inicial em nomes próprios e o alongamento final. Por conta de seu interesse em prosódia e variação linguística, damos aqui um *aperçu* de características notadamente melódicas de locutores dos dialetos românicos de 21 cidades francesas de norte a sul e de leste a oeste da França, comparando-as com as características melódicas do locutor de Paris. Conforme metodologia do projeto do atlas sonoro, todos os locutores, um por cidade, leram a tradução da fábula de Esopo

que apresentamos neste livro sobre a disputa entre o vento e o sol para ver quem tirava o casaco de um viajante.

Baixamos todos os áudios do endereço https://atlas.limsi.fr/liste.html e segmentamos cada um em dez trechos de mesmo conteúdo equivalente nas 22 línguas. No mesmo site, os trechos estão todos transcritos ortograficamente. As cidades que selecionamos foram: Amiens (norte), Angers (noroeste), Arsac-en--Velay (sudeste), Arvillard (centro-leste), Aubigny-Les-Clouzeaux (centro-oeste), Banvillars (leste), Bélis (sudoeste), Caraman (sul), Gap (sudeste), Harau-court (nordeste), Labaroche (nordeste), Lignières (centro), Montgaillard (sul), Montsauche-lès-Settons (centro), Naves (centro), Neufchâtel- en-Saosnois (noroeste), Nice (sul), Pancheraccia (centro), Paris (centro- norte), Plerneuf (noroeste), Réville (noroeste) e Sanary-sur-Mer (sudeste). Os locutores de todas essas cidades são homens com idade superior a 40 anos, embora a maior parte seja formada por pessoas com mais de 60 anos. A razão da escolha de apenas homens é dupla: é o sexo da maior parte dos locutores do atlas e permite eliminar um fator de variação.

Utilizamos o script *Prosody Descriptor* para calcular os valores médios de nove parâmetros melódicos, com o fim de caracterizar prosodicamente os locutores dos dialetos de cada lugar. O leitor entenda que, do ponto de vista experimental, seria fundamental ter um número amplo de locutores, mas não é viável com esse corpus, pois o atlas sonoro tem apenas um locutor por cidade, limitação parcialmente contornada pela escolha de locutor representativo da língua regional. Evitamos, assim, o uso de parâmetros mais diretamente ligados a aspectos individuais como mediana da F0, mínimo da F0, bem como parâmetros de qualidade de voz, que poderiam inclusive refletir mais a faixa etária do que a prosódia da língua regional. Os parâmetros foram estes: desvio-padrão da F0, taxa de picos da F0, abertura dos picos da F0, desvio-padrão dos valores e da ocorrência no tempo de picos da F0, valores médios e de desvio-padrão das subidas e descidas da F0.

Para cada locutor, calculamos a média de cada um dos nove parâmetros considerando os dez trechos segmentados, organizando-as num vetor, conforme se vê nos dados no repositório do livro, pasta **Estatistica/Linguas Regionais Franca**, com os áudios e arquivos TextGrid de anotação do Praat na pasta **Audios/Capitulo6/Linguas Regionais Franca**. Assim, cada língua regional é representada por um vetor de nove parâmetros melódicos médios identificado pela cidade. Com essa tabela de vetores, realizamos uma análise de classificação hierárquica cujo roteiro também se encontra na pasta e cujo resultado se pode ver na Figura 6.12.

Pode-se ver uma distribuição equitativa das cidades nos grupos, com Paris agrupada com cidades do centro e do norte, perto de sua região geográfica. Não parece haver grandes agrupamentos claros, sendo necessário para um real experimento sobre variação prosódica interdialetal uma investigação com mais locutores e parâmetros que meçam diferenças prosódicas relacionadas à posição da sílaba tônica, por exemplo. Além disso, é importante salientar que parte fundamental da caracterização fonética de uma língua é seu aspecto segmental, não observado aqui.

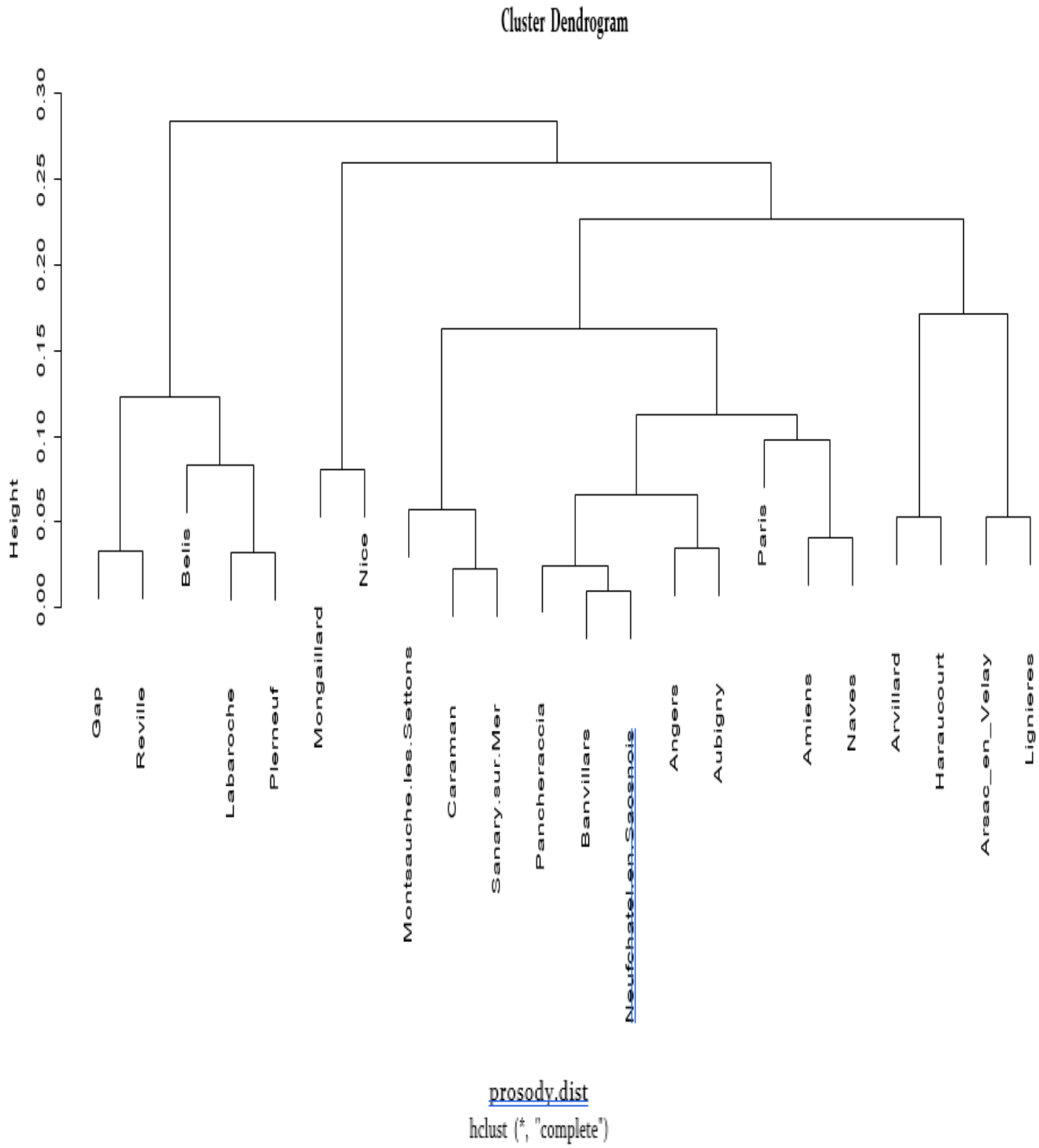

Figura 6.12 – Dendrograma de classificação hierárquica com o método que favorece o encontro de amostras similares. Pode se ver nas folhas, na parte baixa da árvore, as cidades que se agrupam por proximidade maior dos vetores contendo os nove parâmetros melódicos médios.

Quanto à abertura média dos picos da F0, a Figura 6.13 mostra os diagramas de blocos para as 22 línguas, para que se vejam línguas próximas segundo a hierarquização feita e mostrada no dendrogra-

ma. Embora a classificação seja feita com base nos nove parâmetros melódicos médios, a figura aponta a proximidade da mediana desse parâmetro para as cidades de Paris, Amiens e Naves, próximas geograficamente.

Esse curto *aperçu* visou a dar ao leitor uma visão do potencial experimental da comparação de parâmetros prosódico-acústicos entre línguas e variedades linguísticas, área que podemos referir como prosódia comparada, área de pesquisa ainda em sua infância[17]. Seus limites certamente esbarram no volume de dados necessário para que se permita uma estimativa apropriada da variação intra-sujeito e da variação inter-sujeito, bem como da variação no seio da própria língua e entre línguas e variedades distintas. Considerando a possibilidade de variação ampla quando um locutor muda seu estilo de elocução, o leitor pode ter uma ideia da enormidade da tarefa experimental.

17 Recomendamos a leitura do excelente artigo de Goldman et al. (2014) para avaliação de diferenças prosódicas num grande número de estilos com imediata aplicação das técnicas por ele usadas para a investigação de diferenças entre línguas regionais.

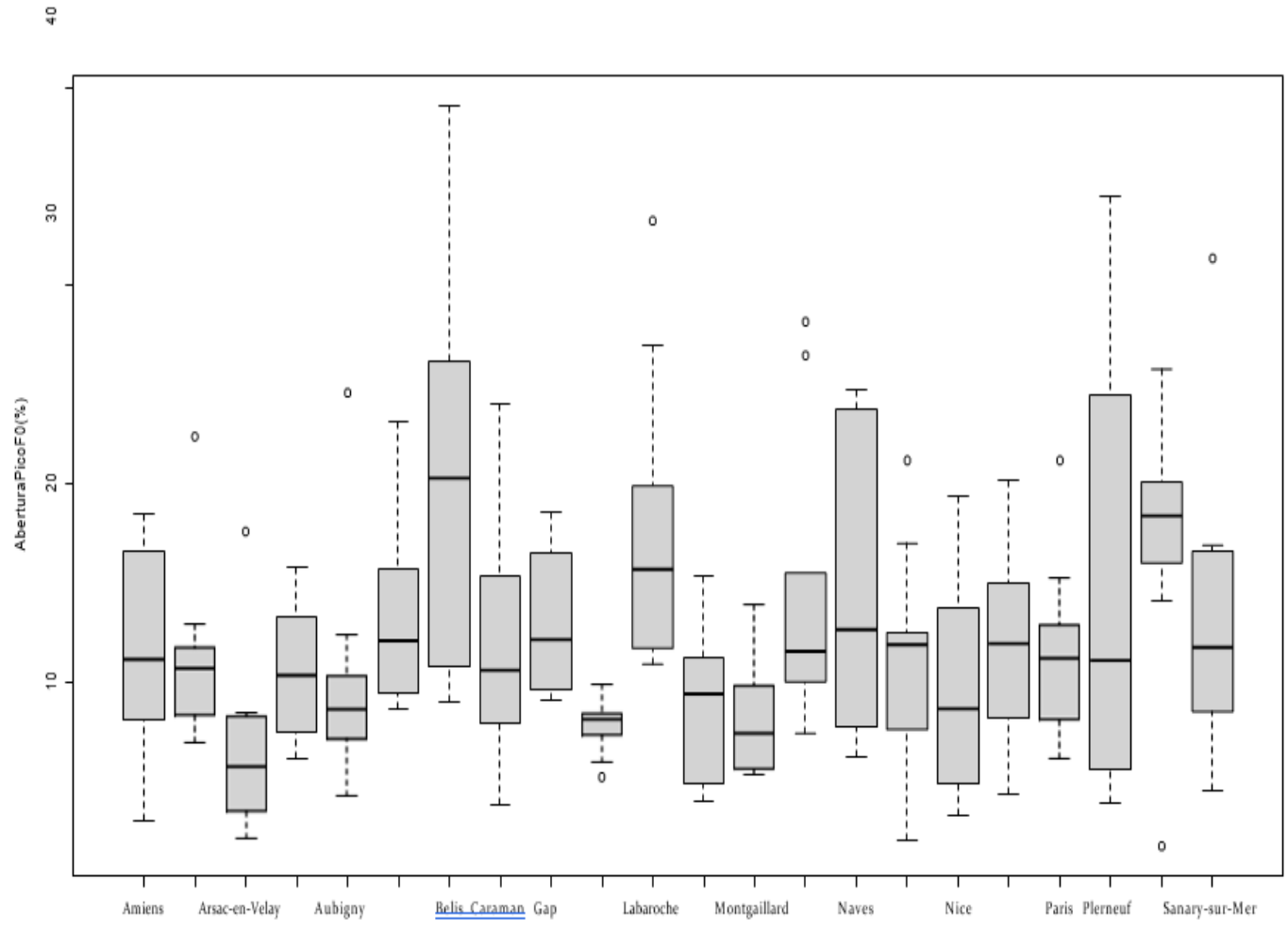

Figura 6.13 – Diagramas de blocos da abertura média dos picos da F0 para as 21 línguas regionais e o francês padrão (Paris), assinalando a proximidade do valor mediano de Paris com Amiens e Naves.

Capítulo 7

Exercícios propostos

7.1 Aprendendo a segmentar e etiquetar unidades VV e a refletir sobre grupos acentuais

7.1.1 Finalidade

Aprender a: (1) marcar os inícios de vogais, (2) etiquetar a unidade VV, (3) gerar dados de duração normalizada e de grupo acentual com o script *SGDetector*, (4) refletir sobre as relações entre produção de duração e percepção de fronteiras e proeminências.

7.1.2 Material

Na pasta **Exercicio/Material-Exercicio1** do repositório do livro se encontram:

1. Arquivo DOCX **Exercicio 1 - Percepção de Funções Básicas**;
2. Arquivo de áudio WAV **LeituraJornalista**;
3. Arquivo TextGrid **LeituraJornalistaExemplo**;
4. Script com extensão PSC **SGDetector**;
5. Arquivo TableOfReal **BP**;

6. Arquivo de tabela de correspondência IPA-Notação do script **TabelaIPAMarcacaoSGDetector**;
7. Pasta **Correção** com o TextGrid completo para conferir o resultado.

7.1.3 Procedimentos e questões

1. Ouça tantas vezes quanto quiser o áudio e utilize o arquivo DOCX com o que se diz para marcar, no próprio texto, primeiramente as fronteiras com graus forte (//) e menos forte (/) e depois as palavras que parecem se destacar do fundo, que mais chamam atenção. Observe que são duas tarefas distintas; faça cada uma concentrando-se apenas no que é pedido. Reserve o resultado para comparar com o que encontrará na análise de duração;
2. Abra no Praat os arquivos **LeituraJornalista.wav** e **Leitura- JornalistaExemplo.TextGrid** e selecione-os simultaneamente. Observe que cerca de metade do áudio já está segmentado e eti- quetado em unidades VV. Continue a segmentação e etiquetagem, conforme aprendeu na seção 4.3, para todo o resto. A tabela de correspondência entre símbolos do IPA e os símbolos ASCII usados no script se encontram no arquivo **TabelaIPAMarcacaoSGDetector.pdf**. Confira os resultados com o TextGrid completo na pasta **Correção** antes de rodar o script. Se for o caso, corrija;
3. Rode o script **SGDetector** com seu arquivo TextGrid. Ele deve estar numa mesma pasta que o próprio script e o arquivo de referência **BP.TableOfReal**. Recorra ao Manual do script, em caso de erro;
4. Examine os dois arquivos TXT gerados. O que tem o termo **dur** ao final do nome original contém, na penúltima coluna, as

durações normalizadas da unidade VV e o que tem o termo **SG** ao final do nome original contém as durações e número de unidades VV em cada grupo acentual;

5. Observe os lugares de fronteira de duração normalizada das unidades VV (na última coluna do arquivo com o termo **dur**, os lugares marcados com 1) e compare com os lugares em que marcou proeminências e fronteiras no arquivo DOCX. Quais são as coincidências? Onde difere, é possível atribuir à melodia uma proeminência/fronteira percebida e não marcada pela duração? E a razão de uma proeminência/fronteira marcada pela duração, mas não percebida, pode ser atribuída a um valor baixo de duração normalizada? Comente;

6. Abra o arquivo com o termo **SG** ao final do nome original. Esse arquivo contém as durações e número de unidades VV em cada grupo acentual. Entre os grupos acentuais, o que varia mais: a duração ou o número de unidades VV?

7. Calcule as taxas de elocução e de articulação do áudio. Compare com as das taxas de entrevistas medidas na seção 4.6 e comente se é maior ou menor e por que, segundo sua opinião.

7.2 Aprendendo a comparar parâmetros melódicos e respiratórios

7.2.1 Finalidade

Aprender a comparar semelhanças e diferenças entre valores médios de parâmetros melódicos e respiratórios entre dois locutores e entre dois estilos de elocução.

7.2.2 Material

Na pasta **Exercicio/Material-Exercicio2** do repositório do livro se encontra o arquivo TXT **Exercicio2-Dados**, contendo um arquivo de dados reais com valores melódicos e respiratórios calculados por ciclo respiratório em dois locutores (JN e AL) em duas posturas (SENT, sentado e EMPE, em pé), tendo ambos sido lidos nessas posturas em dois estilos de elocução, habitual (HB) e persuasivo (PS). As variáveis res- piratórias medidas são: a duração da inalação em segundos (durinh), a amplitude da inalação em unidades arbitrárias relativas ao vale anterior (ampinh), duração da fase expiratória em segundos (durexp) e duração do ciclo respiratório em segundos (durBG), enquanto as variáveis melódicas são: a mediana da F0 em semitons rel. 100 Hz (Fomedian), o desvio-padrão da F0 em semitons (Fosd), o máximo da Fo no ciclo em semitons rel. a 100 Hz (Fomax), o mínimo da F0 no ciclo em semitons rel. a 100 Hz (Fomin), a amplitude melódica no ciclo em semitons rel. a 100 Hz (Forange) e a taxa de picos da F0 em picos por segundo (Forate).

7.2.3 Procedimentos e questões

Leia o arquivo **Exercicio2-Dados.txt** num programa de estatística e faça o seguinte:

1. Calcule descritivamente as médias e desvios-padrão das variáveis acima para os dois locutores em ambas as condições de leitura, independentemente da postura;
2. Para cada variável, utilize um teste t de variáveis independentes ou equivalente não paramétrico, após testar a normalidade dos resíduos como visto no capítulo anterior, para apontar em qual condição de leitura o valor médio de cada variável é maior ou

menor e de quanto. Apenas aponte diferenças quando o teste t for significativo para o nível de significância de 5%.

3. O que se conclui sobre o efeito da persuasão? Qual locutor foi mais efetivo em realizar mudanças e em quais tipos de variáveis, melódicas ou respiratórias?

7.3 Aprendendo a montar um desenho ex- perimental

7.3.1 Finalidade

Aprender a refletir sobre questões relativas à montagem de um desenho experimental cruzando dois tipos de ilocução (instrução e convite) e duas atitudes sociais (hostilidade e gentileza).

7.3.2 Procedimentos e questões

Considere os textos abaixo como exemplos de cenários para a produção da ilocução de convite nas duas atitudes mencionadas acima, com a sentença-chave em negrito, retiradas do trabalho de Siqueira (2018).

Para o convite gentil:

Você e um amigo aproveitaram a manhã livre para começar a assistir a nova temporada de um seriado que vocês estavam esperando há quase um ano. Às 12 h vocês já haviam assistido quatro episódios e, por não estarem com fome ainda, decidiram assistir mais um. Agora são 12:50 h, vocês terminaram o episódio e você percebe que está com fome. Então diz:

Estou com fome, vamos almoçar agora?

Para o convite hostil:

Imagine que você e mais dois colegas de sala combinaram

de se reunir hoje de manhã para fazer um trabalho da faculdade que deve ser entregue até amanhã. Agora é meio dia e vocês estão fazendo o trabalho desde as 8 h. Você percebeu que ainda falta muito para finalizar o trabalho e que seus colegas estão conversando e não fizeram quase nada. Você não comeu nada durante a manhã toda e está com fome, tendo já convidado duas vezes os colegas para almoçar, mas vocês decidiram terminar a introdução do trabalho antes de ir. Você finalizou a introdução e, quando seus colegas leram, os dois concordaram que não ficou boa e que você deveria apagar e escrever novamente. Você fica bravo por estar fazendo o trabalho praticamente sozinho e ainda ter que refazer uma parte e diz: Vocês deveriam fazer a introdução do jeito que acharem melhor, já que ainda não fizeram nada. **Estou com fome, vamos almoçar agora?**

Tendo entendido a ideia envolvida na criação de tais cenários para a produção de uma sentença final, monte um desenho experimental para investigar as diferenças melódicas e de duração entre as duas ilocuções e as duas atitudes, realizando as seguintes etapas:

1. Crie dois cenários possíveis para as mesmas atitudes e sentença final interrogativa acima, mas para uma ilocução de instrução;
2. Repita o procedimento de criação de cenários para ter mais quatro frases interrogativas distintas, cruzando as duas ilocucões com as duas atitudes.

Procure responder às seguintes questões relacionadas ao desenho experimental:

1. Como você segmentaria as frases em unidades menores? Que tipo de unidade de segmentação usaria e por quê?
2. Que hipóteses sobre diferenças melódicas, de qualidade de voz e

de duração entre as duas atitudes e duas ilocuções você faria?

3. Como escolheria os locutores, gravaria o corpus e validaria os enunciados obtidos, no sentido de verificar se realmente veiculam as duas atitudes e ilocucões?

4. Que tipo de teste estatístico usaria para apontar diferenças significativas?

7.4 Aprendendo a variar condições experimentais: fronteira prosódica

7.4.1 Finalidade

Aprender a refletir sobre manipulação de níveis da variável independente "fronteira prosódica".

7.4.2 Procedimentos e questões

Considere sentenças como "Foi bem difícil fazer a **prova sá**bado?" e "Foi bem difícil fazer a **prova, sa**be?" com os trechos em negrito contendo sílabas idênticas com mesma tonicidade. Certamente a fronteira prosódica entre "prova" e "sabe" é mais forte do que a entre "prova" e "sábado". A partir dessa ideia e para investigar as diferenças melódicas e rítmicas com a variação da força da fronteira, crie uma sentença adicional imaginando que teria força diferente das duas exemplificadas e responda e faça o que segue.

1. Para ajudar a raciocinar, grave a fala de um colega apresentando as três frases isoladamente num meio de um slide. Pode ser com um celular com bom microfone, tomando o cuidado de converter o formato de áudio para WAV ou MP3, que podem ser lidos pelo

Praat;

2. Como você segmentaria os enunciados obtidos? Que tipo de unidades linguísticas consideraria?
3. Que variáveis prosódico-acústicas podem variar com a força da fronteira?
4. Que domínios seriam mais afetados, antes ou depois da fronteira? Até que porção dos enunciados a fronteira mais forte teria efeito?
5. Quais os eventuais limites e dificuldades de um desenho experimental para avaliar o efeito dessa função prosódica?

7.5 Aprendendo a variar condições experimentais: proeminência

7.5.1 Finalidade

Aprender a refletir sobre manipulação de níveis da variável independente "fronteira prosódica".

7.5.2 Procedimentos e questões

Considere a sentença "Foi bem difícil fazer a **prova** sábado."

Para investigar as diferenças melódicas e rítmicas em enunciados gerados a partir da sentença com diferentes níveis de saliência na palavra "prova", faça o que segue e responda às questões levantadas.

1. Como você instruiria um locutor e que tipo de procedimento adotaria para variar o nível de saliência na palavra "prova"?
2. Como você segmentaria os enunciados? Que tipo de unidades

linguísticas consideraria?

3. Que variáveis prosódico-acústicas podem variar com a mudança da saliência em “prova”?
4. Que domínios seriam mais afetados, apenas durante a palavra saliente ou em sua vizinhança também? E qual vizinhança, mais antes ou depois da palavra saliente?
5. Quais os eventuais limites e dificuldades de um desenho experimental para avaliar o efeito dessa função prosódica?

7.6 Aprendendo a investigar a melodia com taxas crescentes de elocução

7.6.1 Finalidade

Aprender a investigar diferenças melódicas por conta da aceleração da fala.

7.6.2 Material

Na pasta **Exercicio/Material-Exercicio6** do repositório do livro se encontram:

1. Arquivo DOCX **Narizinho**;
2. Arquivos de áudio WAV de leitura do texto por locutor de Brasília em três taxas de elocucão: **PALT** (lenta), **PANM** (nromal) e **PRPT** (rápida).

7.6.3 Procedimentos e questões

1. Abra os arquivos de áudio no Praat e crie, para cada um, um

objeto de anotação TextGrid com uma camada de intervalos separando as sentenças segundo o texto dado, dando uma etiqueta a cada sentença;

2. Salve os objetos TextGrid na mesma pasta dos áudios;
3. Rode o script *Prosody Descriptor Extractor* considerando apenas a camada que foi segmentada, o que implica desabilitar qualquer outra camada passível de análise. Para tanto leia cuidadosamente o manual do script e indique o "chunk tier" como sendo a camada 1, a que você segmentou e etiquetou as sentenças;
4. Use a saída do script para investigar, por meio de teste de ANOVA, ou seu equivalente não paramétrico, seguido de teste *post hoc*, unicamente as diferenças melódicas das três leituras, uma por taxa de elocução;
5. Que diferenças chamam a atenção? Em que parâmetros melódicos e por quê?

7.7 Aprendendo a investigar efeitos de imitação

7.7.1 Finalidade

Aprender a investigar diferenças melódicas, de qualidade de voz e de pausa em diferentes imitações de jornalistas locutores de TV e rádio.

7.7.2 Material

1. Arquivo DOCX **PrimoBasilio**;
2. Arquivos de áudio WAV de leitura do texto por locutor profissional de rádio de Minas Gerais nos seguintes estilos: leitura normal,

imitando os estilos de locução da rádio CBN e do canal BandNews e imitando o estilo de Sandra Annenberg.

7.7.3 Procedimentos e questões

1. Abra os arquivos de áudio no Praat e crie, para cada um, um objeto de anotação TextGrid com uma camada de intervalos separando as sentenças segundo o texto dado, dando uma etiqueta a cada sentença e com uma camada de intervalos segmentando as pausas silenciosas conforme instruções da seção 4.5;
2. Salve os objetos TextGrid na mesma pasta dos áudios;
3. Rode o script *Prosody Descriptor Extractor* considerando apenas a camada que foi segmentada, o que implica desabilitar qualquer outra camada passível de análise. Para tanto leia cuidadosamente o manual do script e indique o "chunk tier" como sendo a camada 1, a que você segmentou e etiquetou as sentenças;
4. Use a saída do script para investigar, por meio de teste de ANOVA, ou seu equivalente não paramétrico, seguido de teste *post hoc*, as diferenças melódicas e de qualidade de voz entre os quatro trechos lidos;
5. Para quais parâmetros melódicos e de qualidade de voz há diferenças significativas?
6. Para quais das duas medidas relativas à pausa (taxa de produção e duração) há diferenças significativas? Esses resultados batem com a sua percepção das imitações?
7. Como procederia para avaliar a qualidade das imitações?

Referências Bibliográficas

ARANTES, P. *Fo_extrema*. 2008. Programa de software para a plataforma Praat.

ARANTES, P.; ERIKSSON, A.; LIMA, V. Minimum sample length for the estimation of long-term speaking rate. In: *Proc. 9th International Conference on Speech Prosody 2018*. Poznan, Polônia: [s.n.], 2018. p. 661–665. DOI: 10.21437/SpeechProsody.2018-134.

BAAYEN, R. H. *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press, 2008.

BARBOSA, P. A. *Caractérisation et génération automatique de la structuration rythmique du français*. Tese (Doutorado) — Institut National Polytechnique de Grenoble, França, 1994.

BARBOSA, P. A. At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration. In: *Proceedings of the First ESCA Tutorial Research Workshop on Speech Production Modeling and Fourth Speech Production Seminar*. Autrans, França: [s.n.], 1996. p. 85–88.

BARBOSA, P. A. Explaining Brazilian Portuguese resistance to stress shift with a coupled-oscillator model of speech rhythm production. *Cadernos de Estudos Lingüísticos*, v. 43, p. 71–92, 2002.

BARBOSA, P. A. *Incursões em torno do ritmo da fala*. Campinas: Pontes/Fapesp, 2006.

BARBOSA, P. A. How prosodic variability can be handled by a dynamical speech rhythm model. In: *Proceedings of the 16th International Conference of Phonetic Sciences*. Saarbrücken: [s.n.], 2007. p. 331–336.

BARBOSA, P. A. Automatic duration-related salience detection in Brazilian Portuguese read and spontaneous speech. In: *Proc.of the Speech Prosody 2010 Conference*. Chicago, Estados Unidos: [s.n.], 2010. p. 100067: 1–4. Disponível em http://www.speechprosody2010.illinois.edu/papers/100067.pdf.

BARBOSA, P. A. Conhecendo melhor a prosódia: aspectos teóricos e metodológicos daquilo que molda nossa enunciação. *Revista de Estudos da Linguagem (UFMG)*, v. 20, n. 1, p. 11–27, 2012.

BARBOSA, P. A. Elementos essenciais para um entendimento dos limites e vantagens da estatística inferencial na pesquisa fonética. *ReVEL*, n. 7, p. 51–67, 2013.

BARBOSA, P. A. Intonation modeling in cross-linguistic research. In: ARMSTRONG, M. E.; HENRIKSEN, N.; VANRELL, M. del M. (Ed.). *Intonational Grammar in Ibero-*

Romance: Approaches across linguistic subfields. Londres: John Benjamins, 2016. p. 115–134.

BARBOSA, P. A. *Prosódia*. São Paulo: Parábola, 2019.

BARBOSA, P. A. Cross-linguistic comparison of automatic detection of speech breaks in read and narrated speech in four languages. In: RASO, T.; IZ'RAEL, S. (Ed.). *In Search of Basic Units of Spoken Language: A corpus-driven approach*. Amsterdam: John Benjamins, 2020. p. 285–299.

BARBOSA, P. A.; ARANTES, P. Investigation of non-pitch-accented phrases in Brazilian Portuguese: no evidence favoring stress shift. In: SOLÉ, M. J.; RECASENS, D.; ROMERO, J. (Ed.). *Proceedings of the XVth International Congress of Phonetic Sciences*. Barcelona, Espanha: The 15th Organizing Commitee, 2003. p. 135–143.

BARBOSA, P. A. et al. Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors. In: *Proc. Ninth European Conference on Speech Communication and Technology*. Lisboa, Portugal: [s.n.], 2005. p. 1441–1444.

BARBOSA, P. A.; ARANTES, P.; SILVEIRA, L. S. Unifying stress shift and secondary stress phenomena with a dynamical systems rhythm rule. In: *Proceedings of the Speech Prosody 2004 Conference*. Nara, Japão: [s.n.], 2004. p. 49–52.

BARBOSA, P. A.; ERIKSSON, A.; ÅKESSON, J. Cross-linguistic similarities and differences of lexical stress realisation in Swedish and Brazilian Portuguese. In: ASU, E. L.; LIPPUS, P. (Ed.). *Nordic Prosody. Proceedings of the XIth conference, Tartu 2012, Estonia*. Frankfurt am Main, Alemanha: Peter Lang, 2013. p. 97–106.

BARBOSA, P. A.; MADUREIRA, S. *Manual de Fonética Acústica Experimental: Aplicações a dados do português*. São Paulo: Cortez, 2015.

BARBOSA, P. A.; MADUREIRA, S. The interplay between speech and breathing across three Brazilian Portuguese speaking styles. In: *Proceedings of Speech Prosody 2018*. Poznan, Polônia: [s.n.], 2018. p. 369–373.

BARBOSA, P. A. et al. Speech breathing and expressivity: An experimental study in reading and singing stylesn. In: QUARESMA, P. et al. (Ed.). *Lecture Notes in Computer Science 12037*. Amsterdam: Springer International Publishing, 2020. p. 393–398.

BARBOSA, P. A.; MADUREIRA, S.; MAREÜIL, P. B. de. Cross-linguistic distinctions between professional and non-professional speaking styles. In: *Proceedings of the 18th Annual Conference of the International Speech Communication Association*. Estocolmo, Suécia: [s.n.], 2017. p. 3021–3025.

BARBOSA, P. A.; MAREÜIL, P. B. de. Imitating broadcast news style: Commonalities and differences between French and Brazilian professionals. In: MELLO, H.; PANUNZI, A.; RASO, T. (Ed.). *Lecture Notes in Computer Science*. [S.l.]: Springer International Publishing, 2018. v. 11122, p. 419–428.

BARBOSA, P. A.; MIXDORFF, H.; MADUREIRA, S. Applying the quantitative target approximation model (qTA) to German and Brazilian Portuguese. In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association*. Florença: Casual Productions, 2011. p. 2025–2028.

BARBOSA, P. A.; NIEBUHR, O. Persuasive speech is a matter of acoustics and chest breathing only. In: SOMEONE (Ed.). *At the edges of language*. Heidelberg: Springer, 2020. p. 1–20.

BARBOSA, P. A.; SILVA, W. da. A new methodology for comparing speech rhythm structure between utterances: Beyond typological approaches. In: CASELI, H. et al. (Ed.). *PROPOR 2012, LNAI 7243*. Heidelberg: Springer, 2012. p. 329–337.

BARTKOVA, K.; SORIN, C. A model of segmental duration for speech synthesis in French. *Speech Communication*, v. 6, n. 3, p. 245–260, 1987.

BECKMAN, M. E.; ELAM, G. A. *Guidelines for ToBI Labelling*. 1993. The Ohio State University Research Foundation.

BEVERIDGE, W. *The Art of Scientific Investigation*. New Jersey, Estados Unidos: The Blackburn Press, 1957.

BOLINGER, D. Intonation: Levels versus configurations. *Word*, v. 7, n. 3, p. 199–210, 1951.

BOLINGER, D. Accent is predictable if you are a mind-reader. *Language*, v. 48, p. 633–644, 1972.

BOLINGER, D. *Intonation and its parts: Melody in Spoken English*. Stanford: Stanford University Press, 1986.

BOLINGER, D. *Intonation and its uses*. London: Edward Arnold, 1989.

BROWMAN, C. P.; GOLDSTEIN, L. Tiers in articulatory phonology with some implications for casual speech. In: KINGSTON, J.; BECKMAN, M. E. (Ed.). *Papers in Laboratory Phonology I*. Cambridge, Reino Unido: Cambridge University Press, 1990. p. 341–376.

BROWMAN, C. P.; GOLDSTEIN, L. M. Articulatory phonology: an overview. *Phonetica*, v. 49, p. 155–180, 1992.

BUNSCHAFT, G. G.; KELLNER, S. R. *Estatística sem mistérios*. Petrópolis: Vozes, 2001.

BYRD, D. Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica*, v. 57, n. 1, p. 3–169, 2000.

BYRD, D. et al. Phrasal signatures in articulation. In: BROE, M. B.; PIERREHUMBERT, J. B. (Ed.). *Papers in Laboratory Phonology V: Acquisition and the Lexicon*. Cambridge, Reino Unido: Cambridge University Press, 2000. p. 70–87.

BYRD, D.; SALTZMAN, E. The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, v. 31, p. 149–180, 2003.

CABANE, O. F. *The Charisma Myth: How Anyone Can Master the Art and Science of Personal Magnetism*. Nova Iorque: Penguin, 2012.

CAMPBELL, W. N. *Multi-level Timing in Speech*. Tese (Doutorado) — University of Sussex, 1992.

CAMPBELL, W. N. Automatic detection of prosodic boundaries in speech. *Speech Communication*, v. 13, p. 343–354, 1993.

CAMPOS, L. C. P. *Radialista: análise acústica da variação entoacional na fala profissional e na fala coloquial*. Dissertação (Mestrado) — Universidade Estadual de Campinas, 2012.

CASTRO, L. *O comportamento dos parâmetros duração e frequência fundamental nos fonoestilos político, sermonário e telejornalístico*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro, 2008.

CAVALCANTI, J. C. O. *Análise de parâmetros fonético-acústicos em gêmeos idênticos: Implicações para a comparação forense de locutor*. Tese (Doutorado) — Universidade Estadual de Campinas, 2021.

CINTRA, G. Distribuição de padrões acentuais no vocábulo em português. *Confluência*, v. 5, n. 3, p. 82–93, 1997.

CLASSE, A. *The Rhythm of English Prose*. Oxford: Blackwell, 1939.

COLE, J.; MO, Y.; BAEK, S. The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes*, v. 25, n. 7, p. 1141–1177, 2010.

COLE, J.; SHATTUCK-HUFNAGEL, S. The phonology and phonetics of perceived prosody: What do listeners imitate? In: *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Florence: [s.n.], 2011. p. 969–972.

COWAN, N. *Attention and Memory. An Integrated Framework*. Nova York: Oxford University Press, 1997.

CRAWLEY, M. J. *Statistics: An Introduction Using R*. Hoboken, NJ, Estados Unidos: John Wiley & Sons, 2005.

CRAWLEY, M. J. *The R book*. Hoboken, NJ, Estados Unidos: John Wiley & Sons, 2007.

CRESTI, E. *Corpus di italiano parlato*. Florença: Accademia della Crusca, 2000.

CRYSTAL, D. *Prosodic systems and intonation in English*. Cambridge, Reino Unido: Cambridge University Press, 1969.

CRYSTAL, D. *The Cambridge Encyclopedia of Language*. Cambridge, Reino Unido: Cambridge University Press, 1997.

D'ALESSANDRO, C. Voice source parameters and prosodic analysis. In: SUDHOFF, S. et al. (Ed.). *Methods in empirical prosody research*. Berlim: Mouton de Gruyter, 2006. p. 63–88.

D'ERRICO, F. et al. The perception of charisma from voice: a crosscultural study. In: *Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction*. Genebra: [s.n.], 2013. p. 552–557.

DINGA, N. et al. Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews*, v. 81, p. 181–187, 2017.

DOGIL, G.; BRAUN, G. *The PIVOT model of speech parsing*. Viena, Áustria: Verlag, 1988.

DOWDY, S.; WEARDEN, S. *Statistics for research*. Nova York: John Wiley & Sons, 2001.

ESLING, J.; HARRIS, J. G. States of the glottis: an articulatory pho- netic model based on laryngoscopic observations. In: HARDCASTLE, W. J.; BECK, J. (Ed.). *A festschrift for John Laver*. Mahwah, NJ, Estados Unidos: Lawrence Erlbaum Associates, 2005. p. 347–383.

FERNANDES, N. H. *Contribuição para uma análise instrumental da acentuação e intonação do português*. Dissertação (Mestrado) — Universidade de São Paulo, 1976.

FLECK, L. Observation scientifique et perception en général. In: BRAUNSTEIN, J.-F. (Ed.). *L'Histoire des sciences*. Paris: Librairie Philosophique J. Vrin, 1992. p. 245–272. [1935].

FÓNAGY, I.; MAGDICS, K. Emotional patterns in intonation and music. *Sprachtypologie und Universalienforschung*, v. 16, p. 293–326, 1963.

FRAISSE, P. Rhythm and tempo. In: DEUTSCH, D. (Ed.). *The Psychology of Music*. Nova York, Estados Unidos: Academic Press, 1982. p. 149–180.

FREIRE, B. F. A. *Influência do português brasileiro sobre a prosódia do neerlandês falado por imigrantes holandeses no Brasil*. Dissertação (Mestrado) — Universidade Estadual de Campinas, 2020.

FRY, D. B. Experiments in the perception of stress. *Language and Speech*, v. 1, p. 126–152, 1958.

FUJIMURA, O. *Vocal physiology: voice production, mechanisms, and functions*. Nova York: Raven Press, 1988.

FUJIMURA, O.; HIRANO, M. *Vocal fold physiology: voice quality control*. San Diego, Estados Unidos: Singular Publishing Group, 1995.

GAITENBY, J. H. *The elastic word*. [S.l.], 1965.

GELMAN, A.; HILL, J. *Data analysis using regression and multilevel/hierarchical models*. Nova Yorque: Cambridge University Press, 2007.

GERRITS, E. *The categorisation of speech sounds by adults and children: a study of the categorical perception hypothesis and the development weighting of acoustic speech cues*. Tese (Doutorado) — Utrecht University, 2001.

GERRITS, E.; SCHOUTEN, B. Categorical perception depends on the discrimination task. *Perception & psychophysics*, v. 66, n. 3, p. 363–376, 2004.

GHASEMI, A.; ZAHEDIASL, S. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, v. 10, n. 2, p. 486–489, 2012.

GOBBO, O. *Marcadores discursivos em uma perspectiva informacional: análise prosódica e estatística*. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, 2019.

GOLDMAN, J.-P. et al. Speaking style prosodic variation: an 8-hours 9-style corpus study. In: *Proceedings of the 7th International Conference on Speech Prosody*. Dublin, Irlanda: [s.n.], 2014. p. 105–109.

GRABE, E.; WARREN, P. Stress shift: do speakers do it or do listeners hear it? In: CONNELL, B.; ARVANITI, A. (Ed.). *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*. Cambridge, Reino Unido: Cambridge University Press, 1995. p. 95–110.

GREEN, D. M.; SWETS, J. A. *Signal detection theory and psychophysics*. Nova Iorque: Wiley, 1966.

GRICE, M.; LADD, D. R.; ARVANITI, A. On the place of phrase accents in intonational phonology. *Phonology*, v. 17, n. 2, p. 143–185, 2000.

GROSJEAN, F.; COLLINS, M. Breathing, pausing and reading. *Phonetica*, v. 36, p. 98–1146, 1979.

GROSZ, B. J.; SIDNER, C. L. Attention, intention, and the structure of discourse. *Comp. Ling.*, v. 12, p. 175–204, 1986.

HACKING, I. Statistical language, statistical truth and statistical reason: The self-authentification of a style of scientific reasoning. In: MCMULLIN, E. (Ed.). *The Social Dimensions of Science*. Notre Dame, Estados Unidos: University of Notre Dame Press, 1992. p. 130–157.

HART, J. 't; COLLIER, R.; COHEN, A. *A perceptual Study of Intonation. An experimental-phonetic approach to speech melody,*. Cambridge: Cambridge University Press, 1990.

HIROSE, K.; FUJISAKI, H. Analysis and synthesis of voice fundamental frequency contours of spoken sentences. In: *ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing, v. 7*. Paris: [s.n.], 1982. p. 950–953.

HIRST, D. Form and function in the representation of speech prosody. *Speech Communication*, v. 46, p. 334–347, 2005.

ISEI-JAAKKOLA, T.; NAGANO-MADSEN, Y.; OCHI, K. Respiratory control, pauses, and tonal control in L1's and L2's text reading – a pilot study on Swedish and Japanese. In: *Proceedings of Speech Prosody 2018*. Poznan, Polônia: [s.n.], 2018. p. 873–877.

JOHNSON, K. *Quantitative methods in linguistics*. Nova York: John Wiley & Sons, 2011.

KELSO, J. A. S.; SALTZMAN, E. L.; TULLER, B. The dynamical perspective on speech production: data and theory. *Journal of Phonetics*, v. 14, p. 29–59, 1986.

KIMBALL, A. E.; COLE, J. Avoidance of stress clash in perception of conversational American English. In: *Proceedings of VII th Speech Prosody Conference*. Dublin, Irlanda: [s.n.], 2014. p. 497–501.

KLATT, D. H. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, v. 3, p. 129–140, 1975.

KLATT, D. H. Synthesis by rule of segmental durations in English sentences. In: LINDBLOM, B.; OHMAN, S. (Ed.). *Frontiers of Speech Communications Research*.

Nova York: Academic Press, 1979. p. 287–299.

KLATT, D. H. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.*, v. 82, n. 3, p. 737–793, 1987.

KOHLER, K. J. Invariability and variability in speech timing: from utterance to segment in German. In: PERKELL, J.; KLATT, D. H. (Ed.). *Invariance and Variability in Speech Processes*. Ann Arbor: Erlbaum Hillsdale, 1986. p. 268–298.

KOHLER, K. J. Prosody in speech synthesis: The interplay between basic research and TTS application. *Journal of Phonetics*, v. 19, n. 1, p. 121–138, 1991.

KOHLER, K. J. Paradigms of experimental prosodic analysis: from measurement to function. In: SUDHOFF, S. et al. (Ed.). *Methods in Empirical Prosody Research*. Berlim: de Gruyter, 2006. p. 123–152.

KOHLER, K. J. What is emphasis and how is it coded? In: *Proceedings of Speech Prosody 2006*. Dresden, Alemanha: [s.n.], 2006. p. 748–751.

KREIMAN, J. et al. Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *J. Acoust. Soc. Am.*, v. 132, n. 4, p. 2625—-2632, 2012.

KREIMAN, J.; SIDTIS, D. *Foundations of Voice Studies*. Sussex, Reino Unido: Wiley-Blackwell, 2011.

LADD, D. R. Levels vs. configurations, revisited. In: AGARD, F. B. et al. (Ed.). *Essays in Honor of Charles F. Hockett*. Leiden: E. J. Brill, 1983. p. 93–131.

LADD, D. R. Phonological features of intonational peaks. *Language*, v. 59, p. 721–759, 1983.

LADD, D. R. *Intonational Phonology*. Cambridge: Cambridge University Press, 1996.

LAUF, R. Aspekte der Sprechatmung: Zur Verteilung, Dauer und Struktur von Atemgeräuschen in abgelesenen Texten. In: BRAUN, A. (Ed.). *Beiträge zu Linguistik und Phonetik*. Stuttgart: Franz Steiner Verlag, 2001. p. 406–420.

LEHISTE, I. *Suprasegmentals*. Cambridge, Massachussets: MIT Press, 1970.

LEINER, H. C.; LEINER, A. L.; DOW, R. S. The human cerebro- cerebellar system: its computing, cognitive, and language skills. *Behavioural Brain Research*, v. 44, p. 113–128, 1991.

LEVELT, W. J. M. *Speaking: from Intention to Articulation*. Cambridge, MA: M.I.T. Press, 1989.

LIBERMAN, A. et al. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, v. 53, p. 358–368, 1957.

LIBERMAN, M.; PRINCE, A. On stress and linguistic rhythm. *Linguistic Inquiry*, v. 8, n. 2, p. 249–336, 1977.

LIMA-GREGIO, A. M. *Oclusiva glotal e laringalização em sujeitos com fissura palatina segundo abordagem dinamicista*. Tese (Doutorado) — Universidade de Campinas, 2011.

LINK, L. *Individualtypische Aspekte des Atemgeräusches. Eine experimentalphonetische Untersuchung*. Dissertação (Mestrado) — Marburg University, 2012.

LISPECTOR, C. *Doze lendas brasileiras*. Rio de Janeiro: Luz da Cidade, 2000. CD.

LÖFQVIST, A. Stability and change. *Journal of Phonetics*, v. 14, p. 139–144, 1986.

LUCENTE, L. *Aspectos dinâmicos da fala e da entoação no português brasileiro*. Tese (Doutorado) — Universidade Estadual de Campinas, 2012.

LUCENTE, L. Introdução à análise entoacional. In: *Prosódia da fala: pesquisa e ensino*. São Paulo: Blucher, 2017. p. 7–26.

LUCENTE, L.; BARBOSA, P. A. Sistema DaTo de notação entoacional do português brasileiro: teoria e funcionamento. *Cadernos de Pesquisas em Linguística (PUCRS)*, v. 4, p. 41–66, 2009.

MACHADO, A. de P. *Uso de técnicas acústicas para verificação de locutor em simulação experimental*. Dissertação (Mestrado) — Universidade Estadual de Campinas, 2014.

MACMILLAN, N. A.; CREELMAN, C. D. *Detection theory: A user's guide*. Nova Iorque: Psychology Press, 2004.

MACNEILAGE, P. F. The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, v. 21, p. 499–511, 1998.

MADUREIRA, S. A expressão de atitudes e emoções na fala. In: KIRILLOS, L. (Ed.). *Expressividade*. São Paulo: Revinter, 2004. p. 15–25.

MADUREIRA, S. The investigation of speech expressivity. In: MELLO, H.; PANUNZI, A.; RASO, T. (Ed.). *Illocution, modality, attitude, information patterning and speech annotation*. Florença: Firenze University Press, 2011. p. 101–118.

MADUREIRA, S. Intonation and variation: the multiplicity of forms and senses. *Dialectologia*, VI, p. 54–74, 2016.

MADUREIRA, S. et al. *Brazilian Portuguese and European Portuguese contrasted: an experimental acoustic study of speech segments in clash and non-clash conditions*. 2004. Trabalho apresentado na International Conference on Tone and Intonation. 9-11 de setembro. Massaria, Santorini, Grécia.

MADUREIRA, S.; FONTES, M. Gestural prosody and the expression of emotions: a perceptual and acoustic experiment. In: *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: [s.n.], 2015. p. 1–5.

MADUREIRA, S.; FONTES, M. Vocal and facial expressions and meaning effects in speech expressivity. In: *Proceedings of the 10th International Conference of Experimental Linguistics*. Lisboa: [s.n.], 2019. p. 81–84.

MADUREIRA, S.; FONTES, M. . A. S.; CAMARGO, Z. Sound symbolism, speech expressivity and crossmodality. *Signifians (Signifying)*, v. 3, p. 98–113, 2019.

MARCUS, S. M. *Perceptual centres*. Tese (Doutorado) — Cambridge University, 1976.

MAREÜIL, P. B. de. *D'où viennent les accents régionaux ?* Paris: Le Pommier, 2010.

MAREÜIL, P. B. de. *Accents et styles - une étude à base de perception et d'analyses acoustiques à travers le traitement automatique de la parole*. 2012. Thèse d'habilitation à diriger des recherches (HDR) en sciences du langage. Université Sorbonne Nouvelle.

MAREÜIL, P. B. de. A diachronic study of initial stress and other prosodic features in the French news announcer style: corpus-based measurements and perceptual experiments. *Language & Speech*, v. 55, n. 2, p. 263–293, 2012.

MAREÜIL, P. B. de; BARBOSA, P. A. Caractérisation de styles de parole et d'accents étrangers à travers l'imitation : comparaisons entre français et portugais brésilien. *Revue française de linguistique appliquée*, v. 23, n. 1, p. 31–44, 2018.

MAREÜIL, P. B. de; BARDIAUX, A. Perception of French, Belgian and Swiss accents by French and Belgian listeners. In: *Proceedings of Fourth ISCA Workshop on Experimental Linguistics*. Paris: [s.n.], 2011. p. 47–50.

MAREÜIL, P. B. de et al. Accents étrangers et régionaux en français: caractérisation et identification. *Traitement Automatique des Langues*, v. 49, n. 3, p. 135–163, 2008.

MASSINI, G. *A duração no estudo do acento e do ritmo em português*. Dissertação (Mestrado) — Universidade de Campinas, 1991.

MCGUIRE, G. A brief primer on experimental designs for speech perception research. *Laboratory Report*, v. 77, n. 1, p. 2–19, 2010.

MELO, E. B. de. *Perceptual-Center: Para o Estudo de Fatores Fonéticos e Rítmicos na Sincronização Fala-Metrônomo*. Dissertação (Mestrado) — Universidade Nacional de Brasília, 2016.

MILROY, L. *Language and social networks*. Nova York: Basil Blackwell, 1987. 2ª edição.

MIXDORFF, H.; BARBOSA, P. A. Alignment of intonational events in German and Brazilian Portuguese–a quantitative study. In: *Proc. of Speech Prosody 2012 Conference*. Xangai, China: [s.n.], 2012. p. 83–86.

MORAES, J. A. de. Corrélats acoustiques de l'accent de mot en portugais brésilien. In: *Proceedings of the XIth International Congress of Phonetic Sciences*. Talinn, Estônia: [s.n.], 1987. p. 313–316.

MORTON, J.; MARCUS, S.; FRANKISH, C. Perceptual centers (p-centers). *Psychological Revue*, v. 83, n. 5, p. 405–408, 1976.

NIEBUHR, O.; NOVÁK-TÓT, E.; BREM, A. Prosodic constructions of charisma in business speeches a contrastive acoustic analysis of Steve Jobs and Mark Zuckerberg. In: *Proc. 8th International Conference of Speech Prosody*. Boston, EUA: [s.n.], 2016. p. 1–3.

NIEBUHR, O.; SKARNITZL, R. Measuring a speaker's acoustic correlates of pitch - but which? a contrastive analysis based on perceived speaker charisma. In: *Proc. 19th International Congress of Phonetic Sciences*. Melbourne, Austrália: [s.n.], 2019. p. 1774–1778.

NIEBUHR, O.; THUMM, J.; MICHALSKY, J. Shapes and timing in charismatic speech–evidence from sounds and melodies. In: *Proc. 9th International Conference of Speech Prosody*. Poznan, Polônia: [s.n.], 2018. p. 984–989.

OLLER, K. D. The effect of position in utterance on speech segment duration in English. *J. Acoust. Soc. Am.*, v. 54, p. 1235–1247, 1973.

O'SHAUGHNESSY, D. A study of French vowel and consonant durations. *Journal of Phonetics*, v. 9, p. 385–406, 1981.

O'SHAUGHNESSY, D. A multispeaker analysis of durations in read French paragraphs. *J. Acoust. Soc. Am.*, v. 76, p. 1664–1672, 1984.

PASSETTI, R. R. *O efeito do telefone celular no sinal da fala: uma análise fonético-acústica com implicações para a verificação de locutor em português brasileiro*.

Dissertação (Mestrado) — Universidade Estadual de Campinas, 2015.

PIERREHUMBERT, J. B. *The phonology and phonetics of English intonation*. Tese (Doutorado) — MIT, 1980.

PIERREHUMBERT, J. B. Synthesizing intonation. *Journal of the Acoustical Society of America*, v. 70, p. 985–995, 1981.

PIERREHUMBERT, J. B.; HIRSCHBERG, J. The meaning of intonation contours in the interpretation of discourse. In: COHEN, P. R.; MORGAN, J.; POLLACK, M. E. (Ed.). *Plans and Intentions in Communication and Discourse (SDF Benchmark Series in Computational Linguistics)*. Cambridge, EUA: MIT Press, 1990. p. 271–311.

PIKE, K. L. *The intonation of American English*. Ann Arbor: University of Michigan Press, 1945.

POEPPEL, D.; ASSANEO, M. F. Speech rhythms and their neural foundations. *Nature Reviews*, v. 21, p. 322–334, 2020.

POLLACK, I.; PISONI, D. On the comparison between identification and discrimination tests in speech perception. *Psychonomic Science*, v. 24, n. 6, p. 299–300, 1971.

POMPINO-MARSCHALL, B. On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics*, v. 17, p. 175–192, 1989.

POMPINO-MARSCHALL, B. *The syllable as a prosodic unit and the so-called P-centre effect*. Munique, Alemanha, 1991. 66-124 p. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2008. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

RASO, T. O C-ORAL-BRASIL e a teoria da língua em ato. In: RASO, T.; MELLO, H. (Ed.). *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG, 2012. p. 55–90.

RASO, T.; MELLO, H. *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG, 2012.

RAZALI, N. M.; WAH, Y. B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of statistical modeling and analytics*, v. 2, n. 1, p. 21–33, 2011.

REPP, B. H. Categorical perception: Issues, methods, findings. In: LASS, N. J. (Ed.). *Speech and Language: Advanced and Basic Research and Practice*. Nova York: Academic Press, 1984. p. 243–335.

RIETVELD, T.; CHEN, A. How to obtain and process perceptual judgements of intonational meaning. In: SUDHOFF, S. et al. (Ed.). *Methods in empirical prosody research*. Berlim: Mouton de Gruyter, 2006. p. 283–319.

RIETVELD, T.; HOUT, R. van. *Statistical techniques for the study of language and language behaviour*. Berlin: Mouton de Gruyter, 1993.

ROSE, R. L. *The communicative value of filled pauses in spontaneous speech*. Tese (Doutorado) — Birmingham University., 1998.

ROSENBERG, A.; HIRSCHBERG, J. Acoustic/prosodic and lexical correlates of charismatic speech. In: *Proc. 9th European Conference on Speech Communication and Technology*. Lisboa: [s.n.], 2005. p. 513–516.

SALOMONI, S.; HOORN, W. van den; HODGES, P. Breathing and singing: objective characterization of breathing patterns in classical singers. *PLoS ONE*, v. 11, p. e0155084, 2016.

SANTEN, J. P. V. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech & Language*, v. 8, n. 2, p. 95–128, 1994.

SCHOUTEN, B.; GERRITS, E.; HESSEN, A. V. The end of categorical perception as we know it. *Speech communication*, v. 41, n. 1, p. 71–80, 2003.

SHATTUCK-HUFNAGEL, S. Speech errors as evidence for a serial order mechanism in sentence production. In: COOPER, W. E.; WALKE, E. C. T. (Ed.). *Sentence processing: Psycholinguistic studies presented to Merril Garrett*. Hillsdale, EUA: Lawrence Erlbaum, 1979. p. 295–342.

SHATTUCK-HUFNAGEL, S.; KLATT, D. The limited use of distinctive features and markedness in speech production: evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, v. 18, p. 41–55, 1979.

SHUE, Y.-L.; CHEN, G.; ALWAN, A. On the interdependencies between voice quality, glottal gaps, and voice-source related acoustic measures. In: *Proc. Interspeech 2010*. Makuhari, Chiba, Japão: [s.n.], 2010. p. 34–37.

SILVA, W. da. *Correlatos prosódicos da expressão da ironia sarcástica no português brasileiro*. Tese (Doutorado) — Universidade Estadual de Campinas, 2019.

SILVERMAN, K. et al. ToBI: a standard for labeling English prosody. In: *Proceedings of the Second International Conference on Spoken Language Processing*. Banff, Canadá: [s.n.], 1992. v. 2, p. 867–870.

SIQUEIRA, J. *Diferenças prosódicas em atos diretivos combinados a atitudes distintas*. 2018. Trabalho de Conclusão de Curso.

TAYLOR, P. A. *A phonetic model of English intonation*. Tese (Doutorado) — University of Edinburgh, 1992.

THORPE, C. et al. Patterns of breath support in projection of the singing voice. *J. Voice*, v. 15, p. 86–104, 2001.

TITZE, I. *Principles of Voice Production*. Iowa City, Estados Unidos: National Center for Voice and Speech, 2000.

TOUATI, P. Prosodic aspects of political rhetoric. *Working Papers Dep. of Linguistics and Phonetics, Lund, Sweden*, v. 41, p. 168–171, 1994.

TRAUNMÜLLER, H.; ERIKSSON, A. Acoustic effects of variation in vocal effort by men, women, and children. *J. Acoust. Soc. Am.*, v. 107, p. 3438–3451, 2000.

TROUVAIN, J. Laughing, breathing, clicking - the prosody of nonverbal vocalisations. In: CAMPBELL, N.; GIBBON, D.; HIRST, D. (Ed.). *Proc. Speech Prosody 2014*. Dublin, Irlanda: [s.n.], 2014. p. 598–602.

TROUVAIN, J.; TRUONG, K. Comparing non-verbal vocalisations in conversational speech corpora. In: *Proc. 4th Int. Workshop on Corpora for Research on Emotion Sentiment & Social Signals*. Istanbul: [s.n.], 2012. p. 36–39.

TULLER, B.; KELSO, J. A. S. Phase transitions in speech production and their perceptual consequences. In: JEANNEROD, M. (Ed.). *Attention and Performance XIII*. Hillsdale, Estados Unidos: Erlbaum, 1990. p. 429–452.

TULLER, B.; KELSO, J. A. S. The production and perception of syllable structure. *Journal of Speech and Hearing Research*, v. 34, p. 501–508, 1991.

UMEDA, N. Vowel duration in American English. *J. Acoust. Soc. Am.*, v. 58, p. 434–445, 1975.

VAINIO, M. et al. New method for delexicalization and its application to prosodic tagging for text-to-speech synthesis. In: *Proc. of Interspeech 2009 - Speech and Intelligence*. [S.l.: s.n.], 2009. p. 1703–1706.

VAISSIÈRE, J.; MAREÜIL, P. B. de. Divers aspects de l'identification d'une langue ou d'un accent : du segmental à la prosodie. In: *Actes du colloque MIDL 2004. Identification des langues et des varié tés dialectales par les humains et par les machines*. Paris: [s.n.], 2004. p. 1–5.

VALLE-BARBOSA, T. S. do. *Sobreposição de fala em diálogos: um estudo fonético-acústico*. Dissertação (Mestrado) — Universidade Estadual de Campinas, 2013.

VIEIRA, J. M. *Para um estudo da estruturação rítmica na fala disártrica*. Tese (Doutorado) — Universidade Estadual de Campinas, 2007.

WARD, N. G. *Prosodic patterns in English conversation*. Cambridge, Reino Unido: Cambridge University Press, 2019.

WIGHTMAN, C. W. ToBI or not ToBI? In: *Proc. Speech Prosody Conf.* Aix-en-Provence: [s.n.], 2002. p. 25–30.

WITTEN, I. H. A flexible scheme for assigning timing and pitch to synthetic speech. *Language and Speech*, v. 20, p. 240–260, 1977.

WOEHRLING, C.; MAREÜIL, P. B. de. Identification d'accents régionaux en français: perception et catégorisation. *Bulletin du Programme de Phonologie du Français Contemporain*, n. 6, p. 89–103, 2006.

WOODS, A.; FLETCHER, P.; HUGHES, A. *Statistics in language studies*. Cambridge: Cambridge University Press, 1986.

XU, J.; IKEDA, Y.; KOMIYAMA, S. Bio-feedback and the yawning breath pattern in voice therapy: a clinical trial. *Auris Nasus Larynx*, v. 18, p. 67–77, 1991.

XU, Y. Speech melody as articulatorily implemented communicative functions. *Speech Communication*, v. 46, p. 220–251, 2005.

XU, Y. In defense of lab speech. *Journal of Phonetics*, v. 38, p. 329–336, 2010.

XU, Y. Speech prosody: a methodological review. *Journal of Speech Sciences*, v. 1, n. 1, p. 85–115, 2011.

XU, Y.; WANG, Q. E. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, v. 33, p. 319–337, 2001.

SOBRE O AUTOR

Plinio Almeida Barbosa

Plinio Almeida Barbosa é linguista com formação inicial em Engenharia Eletrônica na Graduação e no Mestrado pelo Instituto Tecnológico de Aeronáutica. Seu doutorado em Signal-Image-Parole/ Option Parole foi defendido no Institut de la Communication Parlée e Institut National Polytechnique de Grenoble, França. Tem título de livre-docente em Fonética e Fonologia pela Universidade Estadual de Campinas, onde é Professor Associado III. Tem formação em Engenharia Eletrônica e Linguística, com ênfase na área de Fonética experimental, atuando principalmente nos seguintes temas: análise e modelamento dinâmicos da prosódia da fala, prosódia experimental, teoria de sistemas dinâmicos e de osciladores acoplados, ciências da fala e da linguagem, estilos de elocução, emoção na fala, relações entre atividade respiratória e fala. É bolsista P-Q do CNPq.

OBRAS JÁ PUBLICADAS

COLEÇÃO ALTOS ESTUDOS EM LINGUÍSTICA

A aventura de Saussure
Eliane Silveira

"Ai, se seu te pego...": aspectos prosódicos de estruturas desgarradas em língua portuguesa
Aline Ponciano dos Santos Silvestre

Aquisição atípica da linguagem: modelos linguísticos e prática clínica
Cristiane Lazzarotto-Volcão, Marian Oliveira e Maria João Freitas

Educação intercultural, letramentos de resistência e formação docente
Rodriana Dias Coelho Costa, Kléber Aparecido da Silva e Edinei Carvalho dos Santos

Formas de tratamento e "cordialidade": mudança linguística e conceptualizações culturais
Geisa Mara Batista

Gramaticalização e gramática gerativa
Lorenzo Teixeira Vitral

Linguagem, cognição e ensino: reflexão sobre a linguagem em crianças com e sem diagnósticos

Thalita Cristina Souza Cruz e Fernanda Moraes D'Olivo

Manual de Prosódia Experimental
Plinio A. Barbosa

Monotongação de ditongos orais no português brasileiro: uma revisão sistemática da literatura
Nancy Mendes Torres Vieira

O caso mais grosseiro da semiologia: o que Saussure pode nos dizer sobre os nomes próprios?
Stefania Montes Henriques

Uma abordagem da cena genérica como embreante paratópico: em pauta as cartas privadas de Mário, Drummond, Freud, Sêneca e John Wesley
Manuel Veronez

COLEÇÃO LINGUÍSTICA EM AÇÃO
Introdução à estatística para linguistas
Livia Oushiro

Investigando os sons de línguas não nativas: uma introdução
Felipe Flores Kupske, Ubiratã Kickhöfel Alves e Ronaldo Mangueira Lima Jr.

Linguística no feminino. Vozes femininas que fizeram a linguística no Brasil
Danniel Carvalho e Raquel Freitag

Manual de Morfologia Distribuída
Ana Paula Scher, Indaiá de Santana Bassani e Paula Roberta Gabbai Armelin

REVISÃO
Sandra Madureira

CAPA E PROJETO GRÁFICO
Ad&a Studio

FICHA CATALOGRÁFICA

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Barbosa, Plínio A.

Manual de prosódia experimental [livro eletrônico] Plínio A. Barbosa. -- 1. ed. -- Campinas, SP : Editora da Abralin, 2022. -- (Linguística em Ação)

PDF

Bibliografia.

ISBN 978-85-68990-23-0

1. Língua e linguagem 2. Linguística - Estudo e ensino 3. Prosódia - Estudo e ensino I. Título. II. Série.

23-144516 CDD-410.7

Índices para catálogo sistemático:

1. Linguística : Prosódia : Estudo e ensino 410.7

Henrique Ribeiro Soares - Bibliotecário - CRB-8/9314

DOI 10.25189/9788568990230

EDITORA DA **ABRALIN**

editora.abralin.org