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).

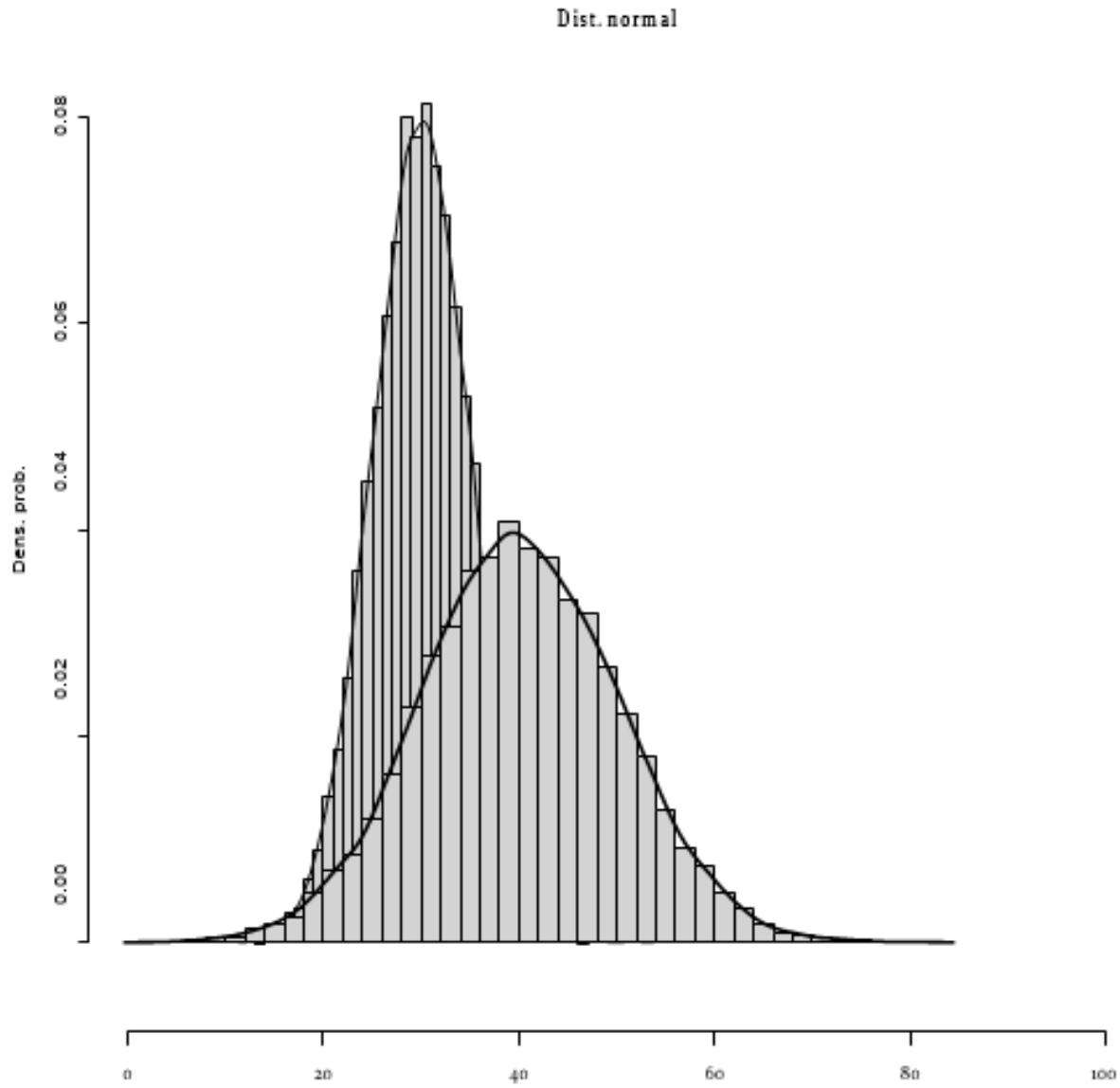

Figura 6.1 – Amostras e populações gaussianas-exemplo: a mais à esquerda com média 30 e desvio-padrão 5 e a mais à direita com média 40 e desvio-padrão 10. As linhas cheias que servem de envoltórias das amostras são traçados das funções de densidade de probabilidade gaussiana que representam as respectivas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

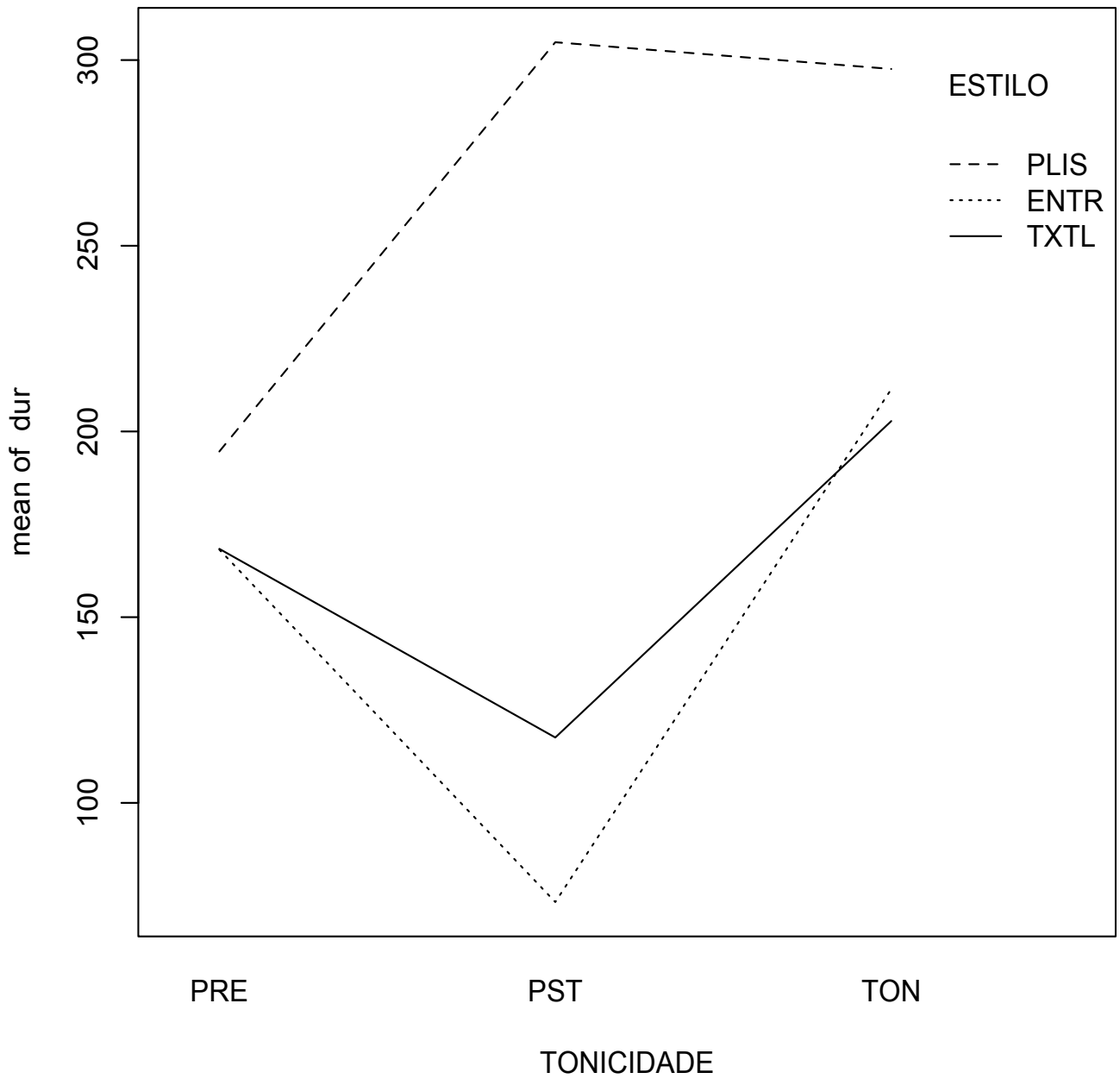

Figura 6.2 – Médias dos três níveis do fator TONICIDADE e dos três níveis do fator ESTILO para o exemplo do texto interligadas por linhas. Observe que as linhas dos estilos entrevista (ENTR) e texto lido (TXTL) são próximas.

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.

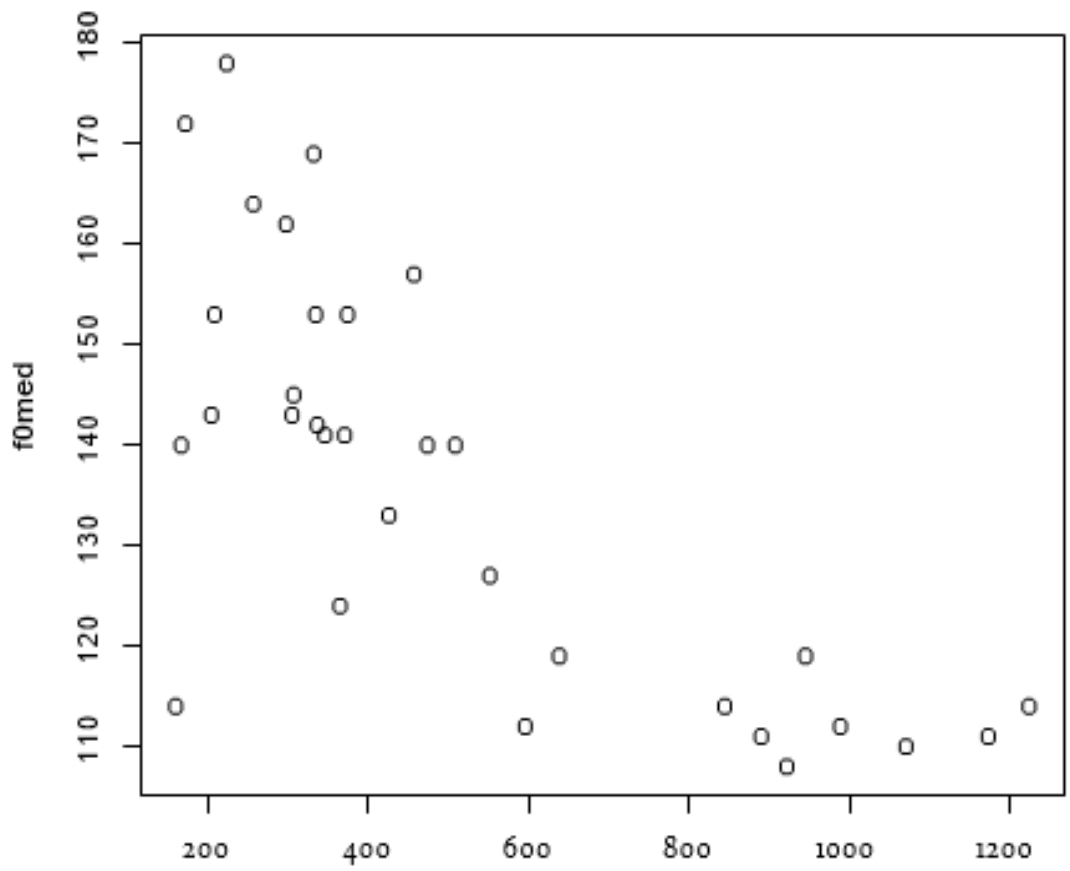

Figura 6.3 – Relação entre duração da unidade VV saliente em ms (abscissa) e mediana da F0 em Hertz na respectiva unidade (ordenada) em trecho de leitura de parágrafo em locutor paulista.

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*.

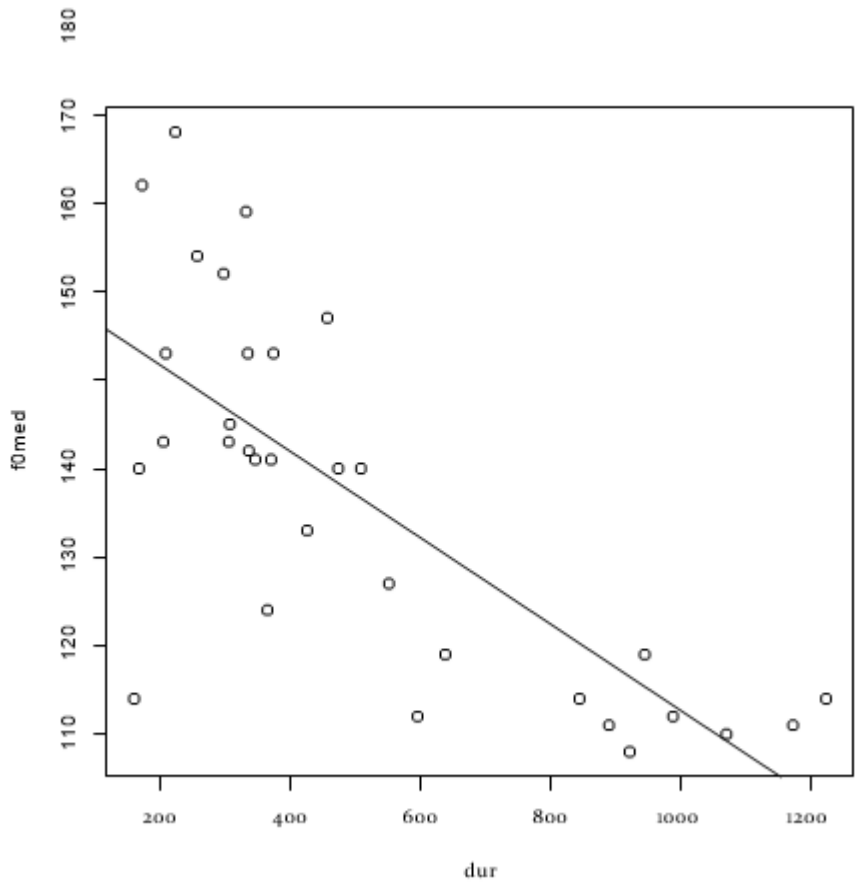

Figura 6.4 – Reta de regressão linear superposta aos dados da Figura 6.3.

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.

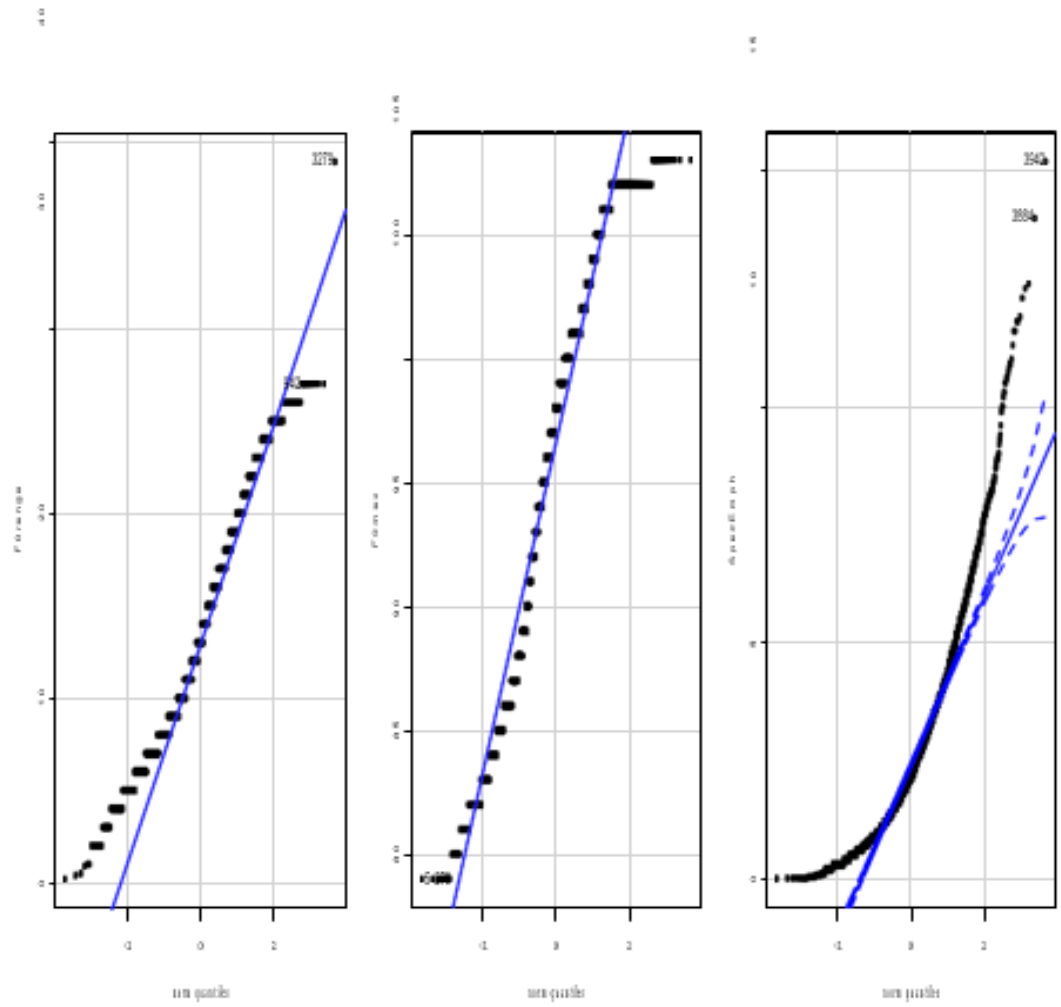

Figura 6.5 – Reta de regressão linear superposta aos dados da Figura 6.3.

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.

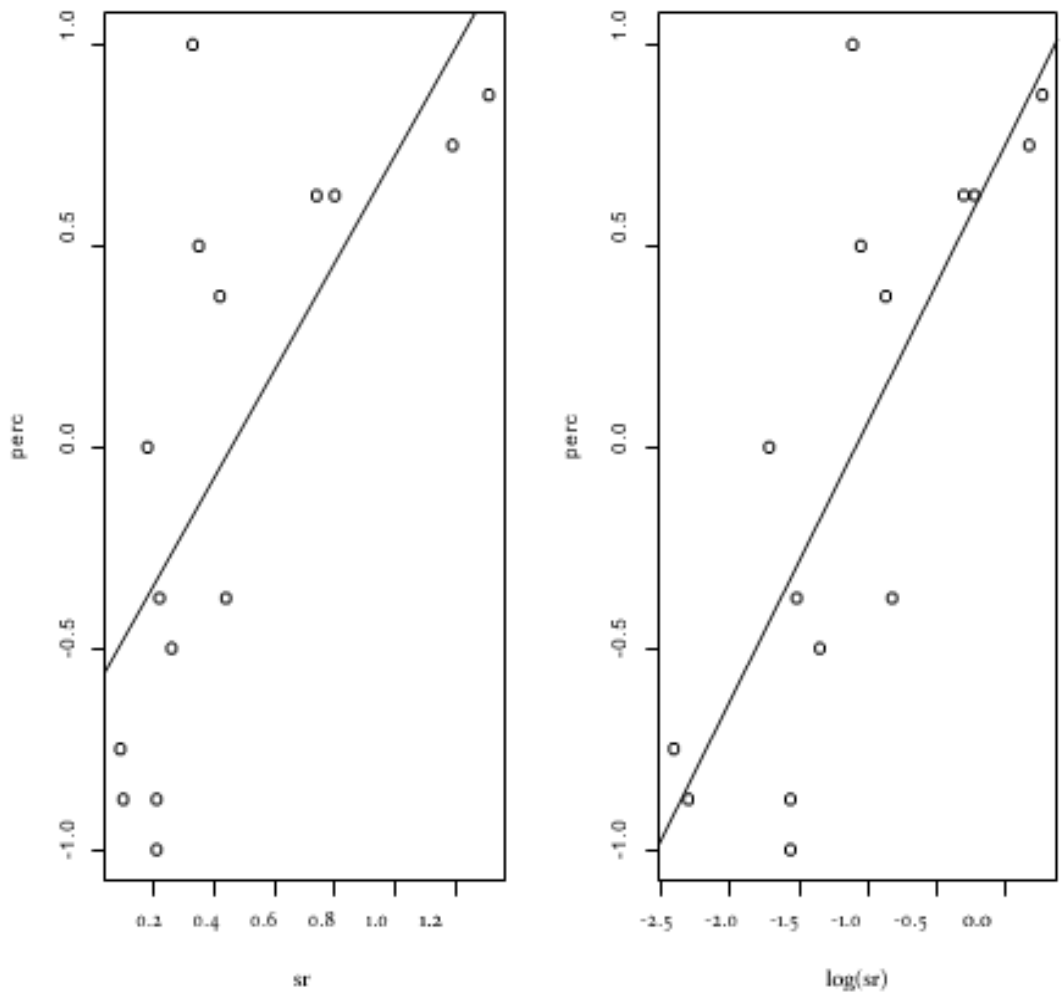

Figura 6.6 – Gráfico de diferença de taxa de elocução (sr) vs. resposta média dos ouvintes no teste de discriminação. A relação da esquerda é linear e a da direita considera o logaritmo da diferença de taxa de elocução. As retas de regressão são dadas para ambos os gráficos.

```
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.

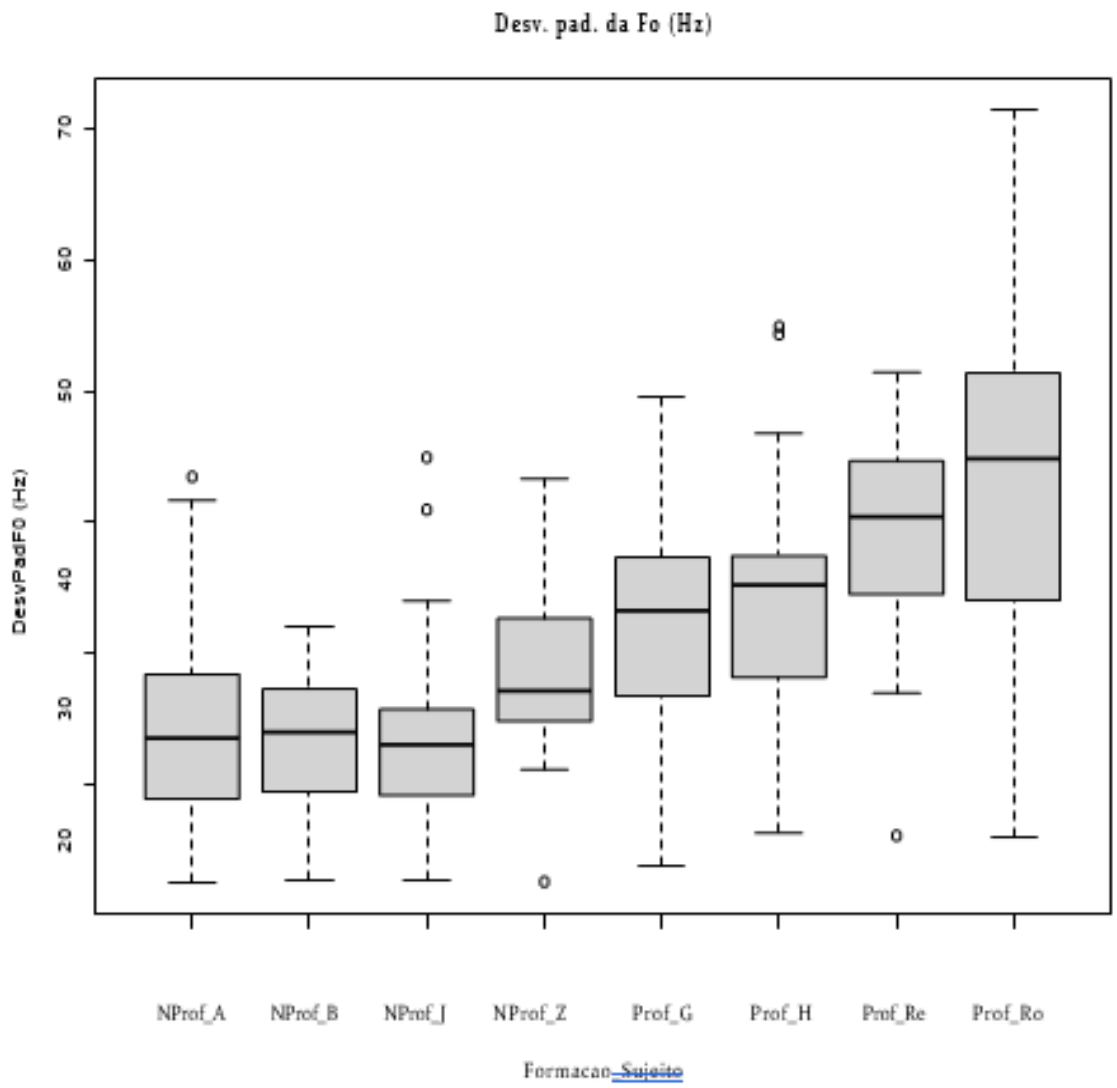

Figura 6.7 – Diagramas de blocos do desvio-padrão da F0 (Hz) para as oito locutoras sendo as profissionais precedidas de "Prof" na abscissa e as não profissionais de "NProf".

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.

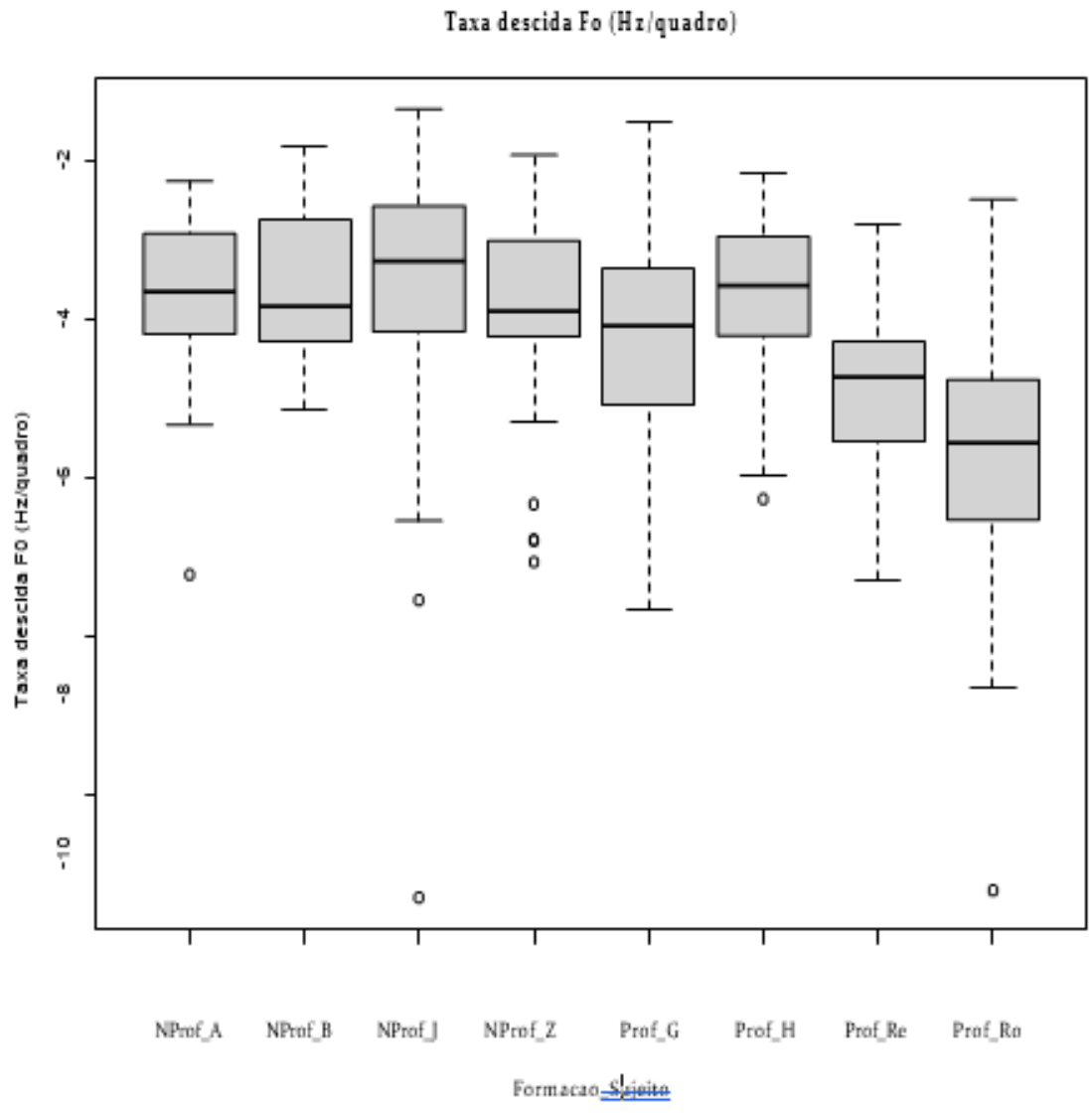

Figura 6.8 – Diagramas de blocos da taxa de descida média da F0 (Hz/quadro) para as oito locutoras sendo as profissionais precedidas de “Prof” na abscissa e as não profissionais de “NProf”.

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.

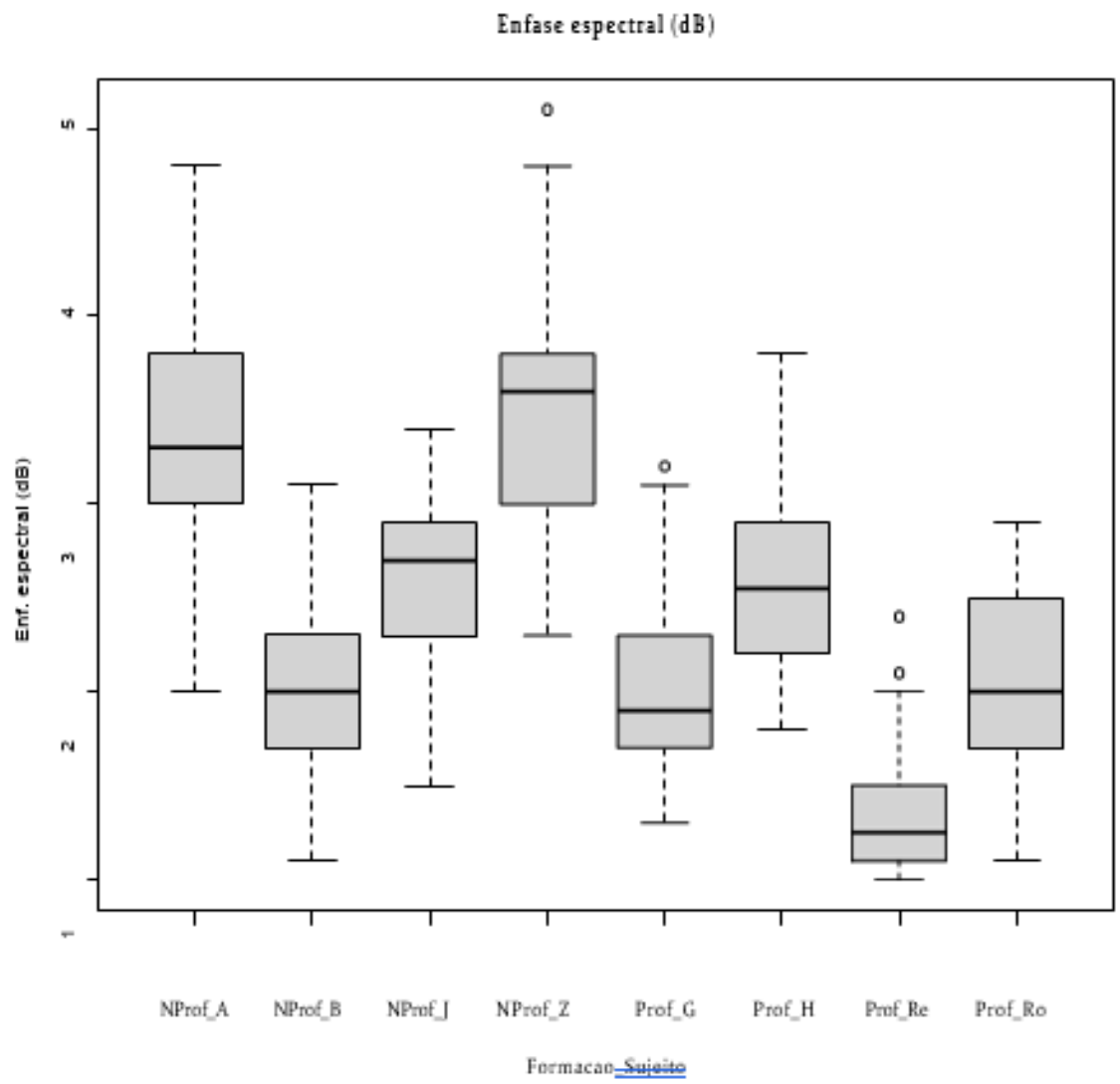

Figura 6.9 – Diagramas de blocos da ênfase espectral (dB) para as oito locutoras, sendo as profissionais precedidas de "Prof" na abscissa e as não profissionais de "NProf".

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.

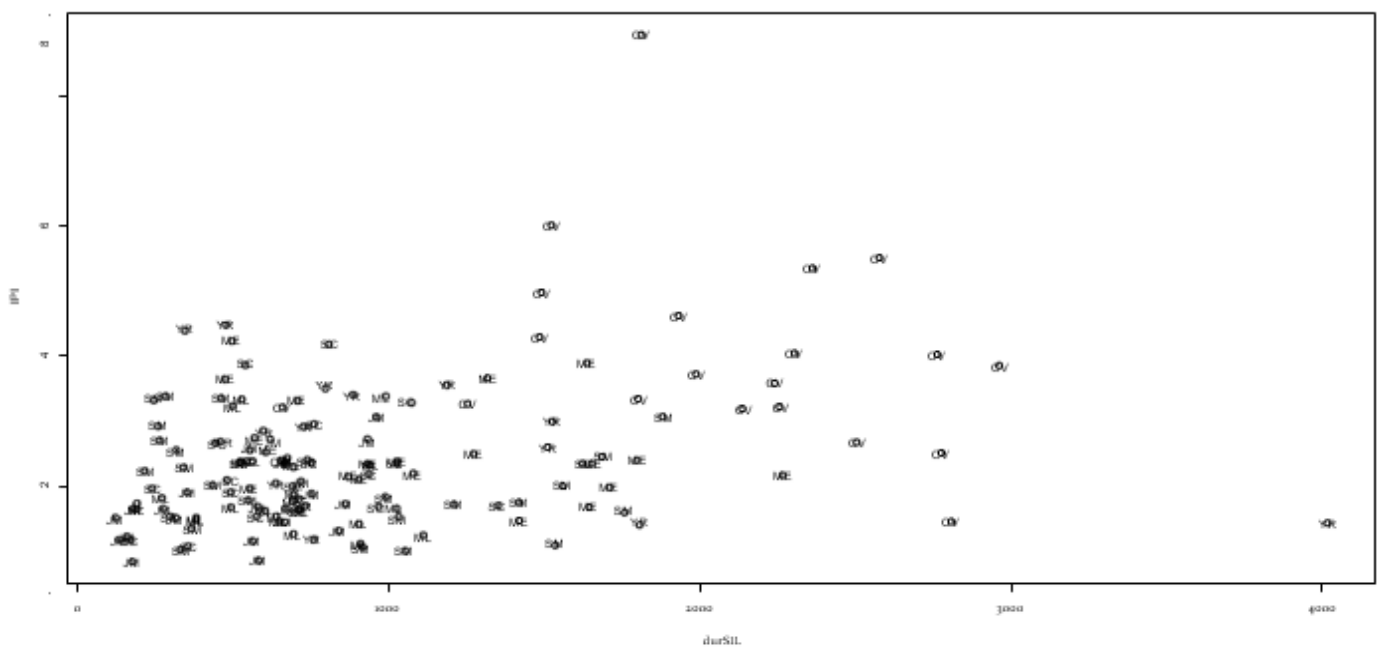

Figura 6.10 – Gráfico de dispersão considerando a duração das pausas silenciosas na abscissa (em milissegundos) e o intervalo de ocorrência com relação à pausa anterior (em segundos) na ordenada, para sete locutores declamando o Soneto da Fidelidade.

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.

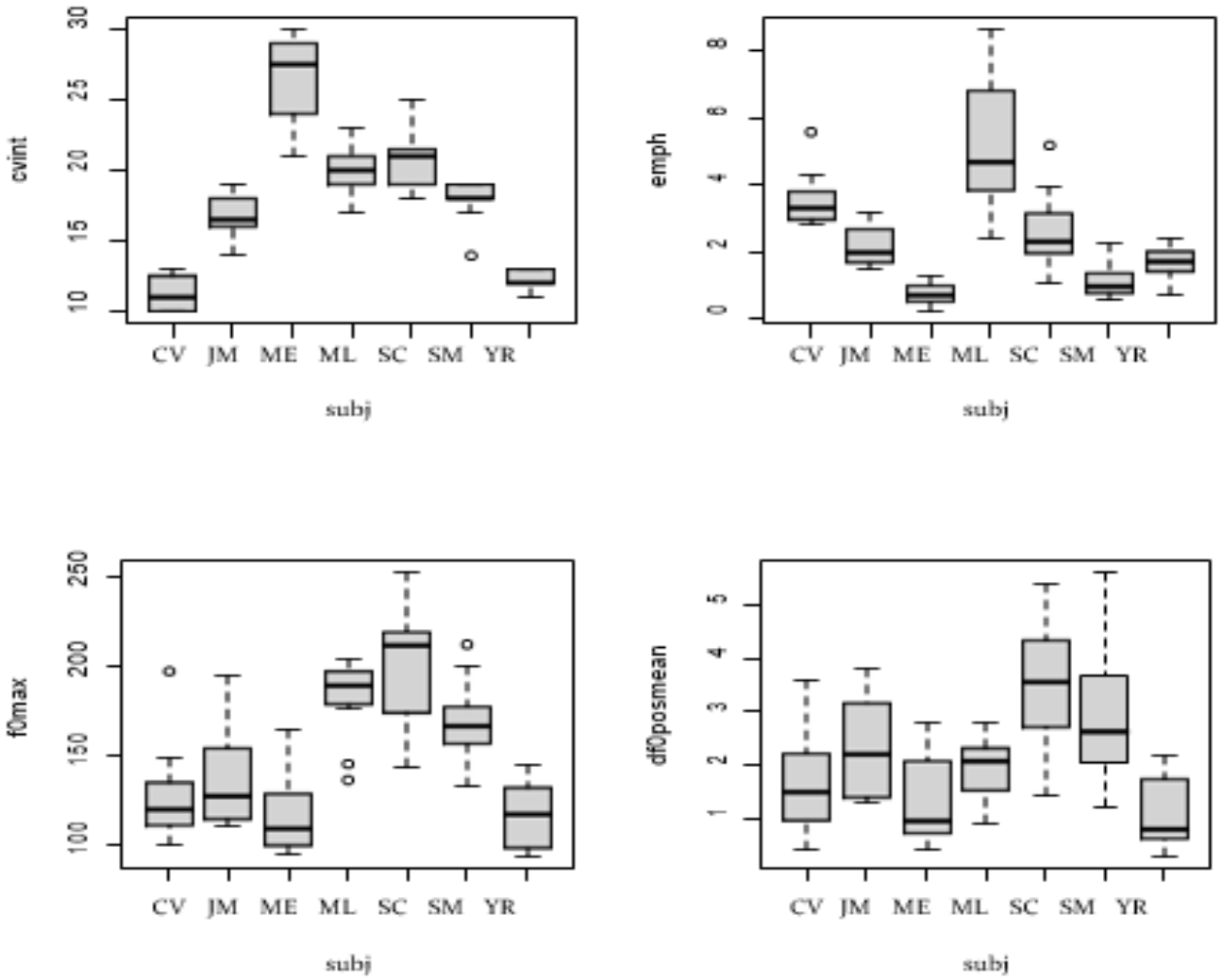

Figura 6.11 – Diagramas de blocos de coeficiente de variação da intensidade em porcentagem, acima à esquerda; ênfase espectral em dB, acima à direita; máximo da F0 em Hz , abaixo à esquerda e taxa média de subida da F0 em Hz/quadro, abaixo à direita para sete locutores que declamaram o Soneto da Fidelidade.

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.

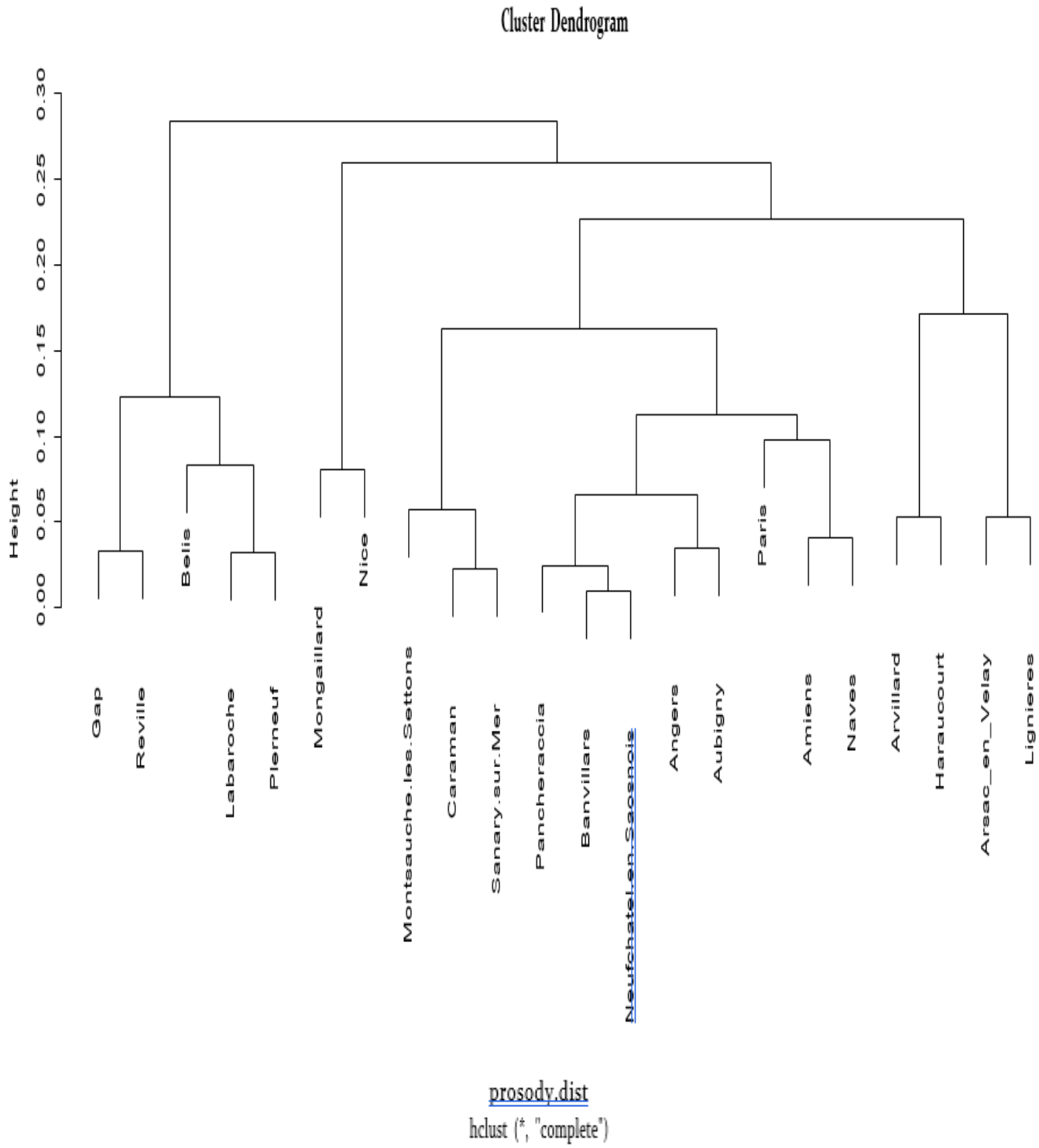

Figura 6.12 – Dendrograma de classificação hierárquica com o método que favorece o encontro de amostras similares. Pode se ver nas folhas, na parte baixa da árvore, as cidades que se agrupam por proximidade maior dos vetores contendo os nove parâmetros melódicos médios.

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.

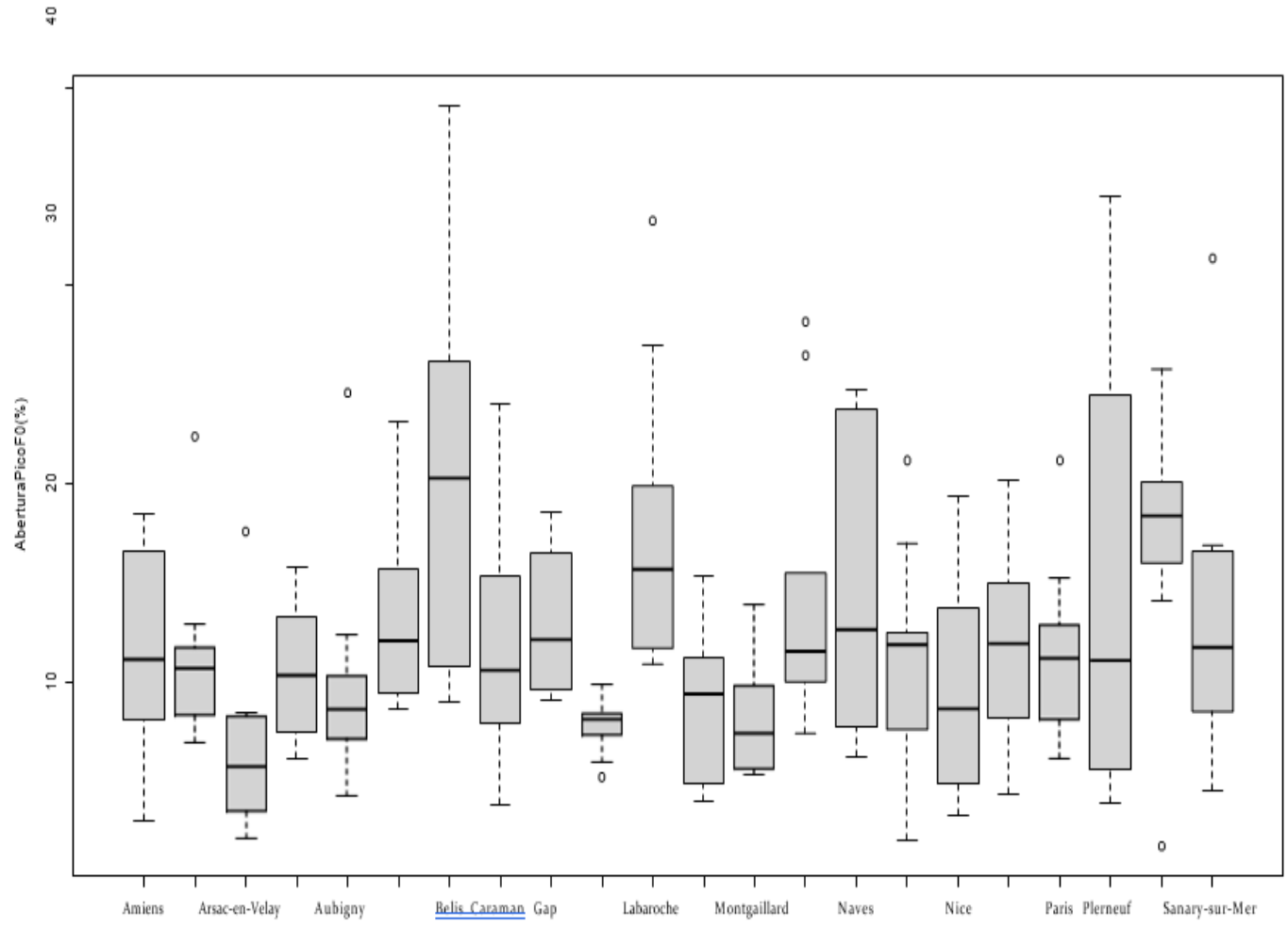

Figura 6.13 – Diagramas de blocos da abertura média dos picos da F0 para as 21 línguas regionais e o francês padrão (Paris), assinalando a proximidade do valor mediano de Paris com Amiens e Naves.