



UNIVERSIDADE ESTADUAL DE CAMPINAS
Instituto de Estudos da Linguagem

ALEXANDRE MENEZES BARROSO

GradiEnt:
UM MODELO DE ANÁLISE FONOLÓGICA
PARA VARIÁVEIS CONTÍNUAS

CAMPINAS

2025

Alexandre Menezes Barroso

GradiEnt:
Um modelo de análise fonológica para variáveis contínuas

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Linguística no Instituto de Estudos da Linguagem da Universidade Estadual de Campinas, como parte dos requisitos exigidos para a obtenção do título de Mestre em Linguística, na área de Forma e Funcionamento das Línguas Naturais (Fonética/Fonologia).

Orientadora: Prof^ª. Dr^ª. Maria Filomena Spatti Sandalo

Campinas

2025

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Estudos da Linguagem
Ana Lucia Siqueira Silva - CRB 8/7956

B278g Barroso, Alexandre Menezes, 1995-
GradiEnt : um modelo de análise fonológica para variáveis contínuas /
Alexandre Menezes Barroso. – Campinas, SP : [s.n.], 2025.

Orientador: Maria Filomena Spatti Sandalo.
Dissertação (mestrado) – Universidade Estadual de Campinas
(UNICAMP), Instituto de Estudos da Linguagem.

1. Fonologia. 2. Fonética. 3. Teoria da otimalidade (Linguística). I.
Sandalo, Maria Filomena Spatti, 1965-. II. Universidade Estadual de
Campinas (UNICAMP). Instituto de Estudos da Linguagem. III. Título.

Informações complementares

Título em outro idioma: GradiEnt : a phonological analysis model for continuous variables

Palavras-chave em inglês:

Phonology

Phonetics

Optimality theory (Linguistics)

Área de concentração: Linguística

Titulação: Mestre em Linguística

Banca examinadora:

Maria Filomena Spatti Sandalo [Orientador]

Plínio Almeida Barbosa

Michael Becker

Data de defesa: 12-05-2025

Programa de Pós-Graduação: Linguística

Objetivos de Desenvolvimento Sustentável (ODS)

Não se aplica

Identificação e informações acadêmicas do(a) aluno(a)

- ORCID do autor: <https://orcid.org/0009-0005-2794-767X>

- Currículo Lattes do autor: <http://lattes.cnpq.br/2946472563147497>

Maria Filomena Spatti Sândalo

Plínio Almeida Barbosa

Michael Becker

Ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós Graduação do IEL.

Agradecimentos

Agradeço à minha orientadora, profa. Filomena Sandalo, pela paciência e apoio em nosso trabalho conjunto desde a monografia de graduação até aqui e em diante. Agradeço aos membros da banca examinadora, tanto de qualificação quanto de defesa, pelo tempo dedicado à leitura deste trabalho, apontamos e participação em minha jornada acadêmica. Nomeadamente, professores Plínio Barbosa e Michael Becker. Além disso, gostaria de agradecer toda a comunidade do Instituto de Estudos da Linguagem da Unicamp – discentes, docentes e funcionários – que possibilitaram e transformaram o meu tempo na universidade por meio de trocas acadêmicas, aulas e a manutenção do instituto de forma geral.

Agradeço aos meus pais e à minha irmã – Ricardo, Aziza e Luana – pelo apoio familiar que recebi durante meus estudos. Agradeço minha esposa, Ana Carolina, pela paciência e suporte que recebi nos últimos 12 anos juntos. Gostaria também de agradecer aos meus dois gatos, Marina e Telinho, pelo companheirismo em todos os momentos.

Por fim, gostaria de agradecer à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que viabilizou meu trabalho financeiramente. Esta dissertação foi realizada com apoio da CAPES através do Programa de Excelência Acadêmica (PROEX): *número de processo 88887.966937/2024-00 com início em 01/04/2024, Bolsa Mestrado, projeto 88882.329577/2010-01 - UNICAMP - LINGÜÍSTICA.*

Resumo

Esta pesquisa tem como objetivo delinear uma metodologia fonológica cognitivamente plausível (perceptual-articulatória) para análise de variáveis fonéticas contínuas. Essa metodologia, e pesquisa em geral, insere-se no contexto teórico das gramáticas fonológicas baseadas em restrições (como a Teoria da Otimalidade e a Gramática Harmônica), cujas análises são tradicionalmente discretas ou discretizantes. Logo, esta pesquisa pretende investigar a viabilidade de uma abordagem “Otimalizante” e não-discretizante de fenômenos fonéticos. Esse objetivo parte da compreensão de que a metodologia deve adaptar-se ao objeto de estudo e, na interface Fonética-Fonologia, há ainda poucos estudos que contemplam a possibilidade de uma Fonologia contínua e baseada em restrições para lidar com fenômenos fonéticos contínuos. O desenvolvimento dessa metodologia envolve adaptar, onde cabível, os elementos das gramáticas baseadas em restrições para o campo das variáveis contínuas, fazendo uso de uma modelagem matemática e computacional desenvolvida especificamente para esse fim. Enfim, esta pesquisa pretende viabilizar – em termos matemáticos, computacionais e fonológicos – a Otimalidade em ambiente contínuo na interface Fonética-Fonologia.

Palavras-chave: gramática harmônica, variáveis contínuas, otimização, metodologia, linguística computacional

Abstract

This research aims to establish a cognitively plausible (perceptual-articulatory) phonological methodology for analyzing continuous phonetic variables. This methodology, and the research overall, is positioned within the theoretical domain of constraint-based phonological grammars, such as Optimality Theory and Harmonic Grammar, which are traditionally focused on discrete or discretizing analyses. Thus, this study seeks to assess the feasibility of an optimality-theoretic and non-discretizing approach to phonetic phenomena. This objective is based on the notion that the methodology should be adaptable to the subject matter, and in the Phonetics-Phonology interface, there are limited studies that explore the possibility of a continuous, constraint-based Phonology to address continuous phonetic phenomena. The development of this methodology involves adapting, as appropriate, elements of constraint-based grammars to the realm of continuous variables, through the use of mathematical and computational modeling specifically devised for this purpose. Ultimately, this research aims to make possible — in mathematical, computational, and phonological terms — the application of Optimality in a continuous context at the Phonetics-Phonology interface.

Keywords: harmonic grammar, continuous variables, optimization, methodology, computational linguistics

Lista de Figuras

1	Mapeamento de <i>inputs</i> para <i>outputs</i> na OT (KAGER, 2004, p. 8)	17
2	Funcionamento da gramática OT (KAGER, 2004, p. 19)	17
3	Funcionamento da restrição WARP (retirado de BOERSMA; ESCUDERO; HAYES, 2003, p. 2).	22
4	Função de densidade de probabilidade dos dados	26
5	Função heaviside vs. função sigmoide.	33
6	Função sigmoide com limiar em 461 Hz.	35
7	Função sigmoide com limiar em 500 Hz.	35
8	Diferentes taxas de saturação k da função sigmoide (todas com $L = 1$ e $x_0 = 461$ Hz).	37
9	Representação da produção de categorias vocálicas (BOERSMA, 1998, p. 179)	38
10	Representação da produção de categorias vocálicas (BOERSMA, 1998, p. 179)	41
11	Sigmoide, sigmoide complementar e seu produto	43
12	Sigmoide, sigmoide complementar e seu produto normalizado	44
13	Hierarquia contrastiva dos traços das vogais do PB (retirado de LEE, 2010, p. 37)	45
14	Saturação adaptável da restrição perceptual	46
15	Incidência de violações da restrição articulatória	54
16	<i>Output</i> da restrição perceptual	59
17	Relação de multiplicação	61
18	Relação de subtração	62
19	Relação de divisão entre potências da mesma base	64
20	Relação de divisão (gaussiana deslocada)	65
21	Relação de divisão (restrição invertida)	66
22	Violações da restrição perceptual	68
23	Violações da restrição articulatória	71
24	Representação da estabilidade dos três candidatos como retas.	78
25	Fluxograma da otimização.	80
26	Nesta seção discutiremos a parte destacada. (1)	81
27	Sistema com menos e mais entropia.	81
28	Funcionamento de restrições matemáticas (retirado de HOFFMAN; BRADLEY, 2007, p. 613)	83
29	Nesta seção discutiremos a parte destacada. (2)	89

30	Nesta seção discutiremos a parte destacada. (3)	92
31	Demonstração do algoritmo L-BFGS.	97
32	F.d.p. de máxima entropia.	98
33	F.d.p. de máxima entropia <i>versus</i> f.d.p. dos dados.	99
34	Captura de tela do GradiEnt (1).	105
35	Captura de tela do GradiEnt (2).	106
36	Captura de tela do GradiEnt (3).	106
37	Captura de tela do GradiEnt (4).	107
38	Captura de tela do GradiEnt (5).	107
39	Captura de tela do GradiEnt (6).	108
40	Captura de tela do GradiEnt (7).	109

Lista de Tabelas

1	Exemplo de tableau da OT	18
2	Tableau adaptado de FLEMMING, 2021, p. 3	18
3	Tableau adaptado de FLEMMING, 2021, p. 5	19
4	Exemplo de tableau com variáveis fonéticas (1).	20
5	Exemplo de tableau com variáveis fonéticas (2).	21
6	Candidatos retirados de BOERSMA; ESCUDERO; HAYES, 2003, p. 2.	22
7	Gramática de produção (adaptado de BOERSMA, 1998, p. 146)	28
8	Gramática de percepção (adaptado de BOERSMA, 1998, p. 147)	28
9	Domínio dos dados.	29
10	Exemplos de regiões para análise.	29
11	Frequências médias dos dois primeiros formantes de /e/ pretônico (MADRUGA, 2017, p. 56)	53
12	Violações da restrição perceptual.	69
13	Violações da restrição perceptual.	69
14	Violações da restrição articulatória.	71
15	Violações da restrição articulatória.	72
16	Escore harmônico dos candidatos.	74
17	Escore harmônico como soma de violações.	74
18	Escore harmônico $\mathcal{H}$ e estabilidade $\mathcal{E}$	77
19	Valores otimizados dos pesos das restrições e λ_0	97
20	Divergência KL pré- e pós-otimização (decimais dos parâmetros abreviados).	98
21	Probabilidades de formantes, <i>i.e.</i> , o <i>output</i>	101
22	Probabilidades de formantes, <i>i.e.</i> , o <i>output</i>	101
23	Probabilidades de formantes, <i>i.e.</i> , o <i>output</i>	102
24	Sobre os dados.	134

Lista de Abreviaturas e Siglas

Sigla	Descrição
DUR	Duração.
EDK	Estimação de densidade por kernel.
F1	Primeiro formante.
F2	Segundo formante.
F.d.p.	Função de densidade de probabilidade.
HG	Gramática Harmônica (<i>Harmonic Grammar</i>).
KL	Kullback-Leibler.
L-BFGS	Algoritmo de Broyden-Fletcher-Goldfarb-Shanno Limitado.
MaxEnt	Máxima entropia.
OT	Teoria da Otimalidade (<i>Optimality Theory</i>).
PB	Português Brasileiro.
RA	Restrição articulatória.
RP	Restrição perceptual.

Lista de Símbolos

Símbolo	Descrição
$\lim$	Limite: Valor que a função atinge ao aproximar-se de um ponto.
$\log$	Logaritmo: Função inversa da exponenciação.
$\frac{d}{dx}$	Derivada: Taxa de variação funcional em relação a uma variável.
∂	Derivada parcial: Taxa de variação em funções multivariáveis.
$\int$	Integral: Acúmulo de quantidades em um intervalo.
$\iint$	Integral dupla: Integral sobre uma área.
Δ	Diferença: Mudança em uma variável ou função.
∇	Nabla: Operador para gradiente.
$\sum$	Somatório: Adição sequencial de números.
$\prod$	Produto: Multiplicação sequencial de números.
∞	Infinito: Representa um limite não finito.
$f(x)$	Função: Relação de x (ou mais variáveis) a valores.
$\hat{f}(x)$	Estimador: Representa função estimada.
e	Euler: Base do log natural, aprox. 2.718.
$\mathcal{H}$	Escore harmônico: Somatório das integrais das violações ponderadas.
$\mathcal{E}$	Estabilidade: Medida relacionada ao gradiente do escore harmônico.
$\mathcal{L}$	Função lagrangiana: Função objetiva com restrições com multiplicadores.
λ	Multiplicador lagrangiano: Peso das restrições.

Lista de Equações

Conceito	Expressão matemática
Restrição perceptual	$R_P(F1) = L - \frac{L^2}{(1+e^{k_1(F1-F1_{\text{limiar}_1})})(1+e^{-k_2(F1-F1_{\text{limiar}_2})})}$
Restrição articulatória	$R_A(F1, F2) = \sqrt{(C_{F1} - A_{F1})^2 + (C_{F2} - A_{F2})^2} \cdot \frac{E_{\text{realizado}}}{E_{\text{esperado}}}$
Cômputo de violações	$v_i(x) = e^{R_i(x) - \hat{f}(x)}$
Escore harmônico	$\mathcal{H}(x) = \sum_{i=1}^n p_i \cdot v_i(x)$
Estabilidade da harmonia	$\mathcal{E} = (\prod_{i=1}^n (b_i - a_i))^{-1} \cdot \int_{a_i}^{b_i} \left(\sum_{i=1}^n \frac{\partial \mathcal{H}}{\partial x_i} \right) dx_i$
Entropia diferencial	$h(f) = - \int_S f(x) \log(f(x)) dx$
Função lagrangiana	$\mathcal{L}(x, \lambda) = f(x) - \sum_{i=1}^n \lambda_i [g_i(x) - c_i]$
Distribuição de máxima entropia	$f_{\text{MaxEnt}}(F1, F2) = e^{-1 - \lambda_0 - \lambda_1 \cdot v_P(F1) - \lambda_2 \cdot v_A(F1, F2)}$
Divergência KL	$\text{KL}(p q) = \int_{-\infty}^{\infty} p \log \left(\frac{p}{q} \right) dx$

Sumário

1	Introdução	13
2	Gramáticas baseadas em restrições	16
2.1	Componentes da Otimalidade	16
2.2	O problema da discretização	20
3	Fundamentos matemáticos	24
3.1	Dados	24
3.1.1	Normalização	25
3.1.2	Estimação de densidade por kernel	25
3.2	<i>Input</i>	27
3.3	Candidatos	29
3.4	Restrições	30
3.4.1	Restrição perceptual	31
3.4.2	Restrição articulatória	47
3.4.3	Algumas considerações teóricas	56
3.5	Violações	57
3.5.1	Cômputo das violações da restrição perceptual	67
3.5.2	Cômputo das violações da restrição articulatória	70
3.6	Harmonia	72
3.6.1	Escore harmônico	73
3.6.2	Estabilidade	75
3.7	Otimização	79
3.7.1	Maximização da entropia diferencial	81
3.7.2	Divergência de Kullback-Leibler	89
3.7.3	Algoritmo de Broyden-Fletcher-Goldfarb-Shanno	92
3.8	<i>Output</i>	100
3.9	A hipótese e a plausibilidade cognitiva	103
4	Implementação computacional	105
5	Considerações finais	111
6	Referências bibliográficas	113

A	Script para modelagem	119
B	Dados utilizados	134

1 Introdução

Esta dissertação introduz a modelagem fonológica GradiEnt, um quadro teórico baseado na Teoria da Otimalidade e Gramática Harmônica projetado para analisar variáveis fonéticas contínuas. Essencialmente, a modelagem GradiEnt adapta a arquitetura da Gramática Harmônica para que seja capaz de lidar com candidatos contínuos. Ela emprega funções matemáticas como restrições capazes de avaliar qualquer intervalo dentro de um espaço fonético sem realizarmos discretizações. Essas restrições, uma de natureza perceptual e outra articulatória, trabalham conjuntamente para modelar a interface entre as categoria fonológicas abstratas e suas realizações fonéticas. A elaboração e utilização de um modelo e quadro teórico para candidatos/variáveis contínuas na Otimalidade justifica-se pela lacuna metodológica existente na maneira como esses modelos lidam com variáveis fonéticas contínuas. Abordagens tradicionais na Teoria da Otimalidade e Gramática Harmônica discretizam dados fonéticos em intervalos antes da análise em limites artificiais entre fenômenos que, na realidade, podem ser compreendidos de maneira contínua. Isso pode criar fronteiras rígidas entre candidatos acústicos que não representam a realidade do fenômeno, *i.e.*, da variável acústica em análise fonológica. A apresentação deste problema, que justifica a elaboração deste trabalho, será apresentada de maneira mais completa no segundo capítulo desta dissertação.

Os fundamentos matemáticos da modelagem GradiEnt são uma parte alterações à modelagem otimalista tradicional (*e.g.*, escore harmônico) e outra parte arquitetura e contribuições originais (*e.g.*, cômputo das violações, estabilidade e otimização dos pesos), resultando em uma modelagem com as mesmas capacidades que sua abordagem discreta. Entre algumas de suas inovações, podemos citar a introdução das restrições perceptual e articulatória, do emprego de estimação de densidade por kernel para criação de candidatos e do uso da divergência de Kullback-Leibler como parte da elaboração das probabilidades dos candidatos. Outro diferencial é seu tratamento das violações, que são definidas como *interação* entre a densidade estimada dos dados empíricos e a função matemática da restrição fonológica. Além disso, uma interface visual para realização dessa modelagem será apresentada e disponibilizada para fins de reprodutibilidade e acessibilidade, comprovando que a modelagem GradiEnt é não apenas *teoricamente* viável como também *computacionalmente* viável. Uma ressalva importante é a de que apesar das implicações teóricas da modelagem GradiEnt estenderem-se para além de sua implementação matemática e computacional, algo que será devidamente discutido em momentos relevantes, o objetivo e recorte desta pesquisa é primariamente ocupar-se com a viabilidade e praticabilidade da execução da modelagem. É, afinal, um grande empenho em si mesmo realizar a transposição

da abordagem otimalista do ambiente discreto para o contínuo de maneira suficientemente abrangente. Tudo isso será abordado de maneira expansiva nos capítulos seguintes.

Uma vantagem chave da modelagem proposta neste trabalho é sua flexibilidade empírica. Apesar do quadro teórico/modelagem apresentados nesta dissertação serem demonstrados primariamente usando frequências de formantes, especialmente considerando as restrições em uso, seus fundamentos matemáticos foram projetados para lidar com qualquer variável fonética contínua. Essa generalizabilidade permite que o GradiEnt seja potencialmente utilizado em uma grande variedade de análises considerando diferentes fenômenos acústicos, como duração, *onset*, intensidade e *pitch*. Ademais, sua capacidade de lidar com múltiplas variáveis fonéticas contínuas simultaneamente também viabiliza a análise de interações fonéticas mais complexas. As restrições dispostas neste trabalho, portanto, podem ser entendidas como *placeholders* funcionais – *i.e.*, provas de conceito – para futuras restrições mais ou menos complexas eventualmente customizadas para investigações específicas. No entanto, mesmo na condição de *placeholders*, as restrições apresentadas foram pensadas não só em termos de funcionalidade matemática ou computacional como também necessariamente exprimindo aspectos fonológicos teóricos importantes – formando, assim, um quadro teórico coerente e ancorado na literatura relevante, cujos elementos basilares encontram-se prontos para investigações futuras e expansões.

Enfim, no que diz respeito à estrutura deste trabalho, esta dissertação será segmentada em três capítulos principais (capítulos 2, 3 e 4), conforme listados abaixo:

2. **Gramáticas baseadas em restrições**, onde serão realizadas a apresentação e contextualização teórica sobre gramáticas baseadas em restrições (Teoria da Otimalidade, Gramática Harmônica e abordagens de máxima entropia) e o problema da discretização na literatura fonológica atual.
3. **Fundamentos matemáticos**, onde será apresentado o quadro teórico da modelagem GradiEnt, passando por todos os elementos da otimalidade tradicional e de como podem ser reinterpretados para o tratamento de variáveis contínuas, com as devidas demonstrações.
4. **Implementação computacional**, apresentação de uma interface visual escrita em Python3 para a modelagem.

Após esses três capítulos, esta dissertação será concluída no quinto capítulo **Considerações finais**, seguido das **Referências bibliográficas** utilizadas e dos anexos **Script para modelagem** e **Dados utilizados**, que apresenta o código do *script* apresentado no capítulo **Implementação computacional** e os dados utilizados nesta pesquisa, respectivamente. Tendo isso em mente,

podemos prosseguir para a contextualização teórica e apresentação do problema no capítulo seguinte.

2 Gramáticas baseadas em restrições

Antes de entrarmos na modelagem de variáveis contínuas, que é o foco deste trabalho, é preciso contextualizar o quadro teórico da Otimalidade¹ em que esta metodologia se insere.

2.1 Componentes da Otimalidade

A Teoria da Otimalidade (abreviada, OT) é um quadro teórico cuja ideia central é a de que representações superficiais da língua refletem resoluções de conflitos entre restrições que competem entre si (cf. PRINCE; SMOLENSKY, 1993). Essas restrições seriam universais e línguas difeririam umas das outras na disposição do ranqueamento dessas restrições, levando a diferentes resoluções de conflitos. Isso significa que, no cerne da OT, gramáticas são consideradas um sistema de forças conflitantes (KAGER 2004, p. 4). Há, de maneira geral, duas grandes famílias de restrições na OT: marcação e fidelidade. Essas forças podem entendidas como estando em conflito em qualquer gramática (*Ibid.*, p. 4). A marcação é um denominador geral para fatores gramaticais que pressionam-nos em direção a estruturas não marcadas². Essa força é contraposta pela fidelidade, que representa a combinação de fatores gramáticos que preservam contrastes lexicais (*Ibid.*, p. 4-5). É fácil inferir, graças às suas definições, que essas duas forças estão inerentemente em conflito, pois uma preserva contrastes enquanto outra livra-se deles. A OT, por sua vez, é o dispositivo através do qual essas duas forças chocam-se em razão de *inputs* e *outputs*. Ela assume que há uma *forma ótima* para segmentos, resultante de restrições representativas dessas duas forças conflitantes.

¹Uso a palavra *Otimalidade* como termo guarda-chuva para OT clássica, OT estocástica, Gramática Harmônica, gramáticas MaxEnt, e assim em diante para todos os quadros teóricos de gramáticas baseadas em restrições associados ou derivados dos trabalhos de Legendre, Miyata, Smolensky e Prince realizados na década de 1990. É equivalente ao adjetivo *optimality-theoretic*.

²A marcação representa uma forma de universalidade em sentido fraco (*Ibid.*, p. 2). Estruturas *não marcadas* seriam aquelas preferidas e consideradas “básicas” entre gramáticas de várias línguas, enquanto estruturas *marcadas* são aquelas evitadas e usadas apenas para a criação de contraste entre segmentos (*Ibid.*, p. 2-3). Kager (2004, p. 5) oferece alguns exemplos de estruturas favorecidas *versus* desfavorecidas: sílabas abertas *versus* sílabas fechadas, vogais curtas *versus* vogais longas ou obstruintes não vozeadas *versus* obstruintes vozeadas. De maneira geral, em síntese, segmentos, combinação de segmentos ou estruturas prosódicas preferem estruturas não marcadas.

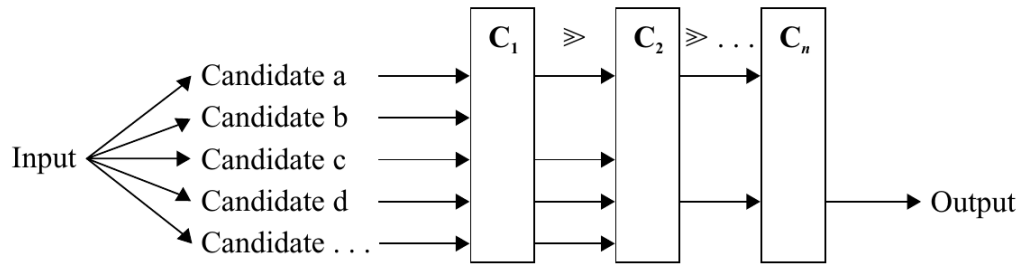


Figura 1: Mapeamento de *inputs* para *outputs* na OT (KAGER, 2004, p. 8)

Como observamos acima, a forma ótima de um segmento é um *output* resultante do choque entre candidatos (possíveis *outputs* de um dado *input*) e restrições da gramática da língua ($C_1, C_2 \dots C_n$). Essas restrições, justapostas umas às outras, são elencadas em forma de *ranking* em que algumas têm prioridades sobre outras. Podemos, então, estabelecer o conceito de *restrição* como o requerimento estrutural que pode ser ou satisfeito ou violado por um output (*Ibid.*, p. 9). Enfim, após realizadas as definições de marcação, fidelidade e restrição, podemos tratar sobre a arquitetura de uma gramática OT.

Composta por três elementos, chamados LÉXICO (LEXICON ou LEX), GERADOR (GENERATOR ou GEN) e AVALIADOR (EVALUATOR ou EVAL), a arquitetura de uma gramática na OT é relativamente minimalista (*Ibid.*, p. 18-48), mas apresenta alta capacidade analítica. GEN é uma função aplicada sobre *inputs*, sendo estes retirados do LEX, que produz uma série de candidatos, ou seja, *outputs* (representações superficiais) possíveis. Estes candidatos, *outputs* de GEN, são *inputs* do EVAL cujo *output*, por fim, é a representação superficial. Isso pode ser resumido no esquema abaixo:

The grammar as an input–output mechanism

Gen (input) $\Rightarrow \{\text{cand}_1, \text{cand}_2 \dots \text{cand}_n\}$

Eval $\{\text{cand}_1, \text{cand}_2 \dots \text{cand}_n\} \Rightarrow \text{output}$

Figura 2: Funcionamento da gramática OT (KAGER, 2004, p. 19)

Esse tipo de funcionamento de gramática na OT (e gramáticas baseadas em restrições em geral) é aplicado na solução de problemas fonológicos por meio de um tableau com a seguinte estrutura:

/input/	RESTRIÇÃO ₁	RESTRIÇÃO ₂
a. candidato _a	*!	
☞ b. candidato _b		*

Tabela 1: Exemplo de tableau da OT

No tableau acima, temos /input/ como representativo de LEX, temos os candidatos como representativos de GEN e a análise de violações de restrições (onde as violações são marcadas por um asterisco e as restrições estão em ordem de importância) representa EVAL. Percebemos que, conforme a análise do caso acima, o Candidato_b é a forma ótima e *output* da gramática, pois ele viola apenas a Restrição₂ e, conforme ordenado acima, Restrição₁ > Restrição₂.

A Gramática Harmônica (abreviada, HG), apesar de extremamente similar no quesito arquitetura, difere da OT ao utilizar pesos numéricos para suas restrições ao invés de ranqueamentos (FLEMMING, 2021, p. 2). Dentro do quadro teórico disposto pela HG, usamos esses pesos para calcular o *escore harmônico*, ou *harmonia*, dos candidatos à representação superficial e selecionamos aquele cujo escore harmônico é mais próximo de zero, conforme o seguinte tableau:

/input/	R ₁ (P = 15)	R ₂ (P = 8)	R ₃ (P = 8)	Harmonia
☞ a. candidato _a	-1			-15
b. candidato _b		-2		-16
c. candidato _c		-1	-1	-16

Tabela 2: Tableau adaptado de FLEMMING, 2021, p. 3

Nesse tableau, é o Candidato_a que é tido como *mais harmônico* e, assim, torna-se o *output* desse sistema. Trocamos a representação de nossas violações de asteriscos para valores numéricos, mas isso é meramente por convenção, pois a OT tradicional também é capaz de incorrer múltiplas violações em uma restrição em um único candidato. No entanto, mesmo com a adição de pesos, a HG não parece diferir muito da OT em seu *output*, pois ambas oferecem uma representação superficial única e categoricamente superior às outras, visto que foram capazes de serem determinadas como o candidato ótimo ou mais harmônico. E é com o advento da Fonologia probabilística que começa-se a explorar diferentes realizações de *outputs* não mais categóricos. Estaríamos nos distanciando da noção tradicional disposta em *The Sound Pattern of English* (1968) de que deveríamos mapear uma representação superficial para cada uma

interpretação fonética (CHOMSKY; HALLE, 1968, p. 14).³

A HG dispõe de várias versões probabilísticas⁴ para formulação de gramáticas fonológicas estocásticas, sendo uma delas chamada Gramática de Máxima Entropia (abreviada, MaxEnt). Atualmente, MaxEnt é o quadro teórico mais utilizado para criação de gramáticas fonológicas estocásticas (FLEMMING, 2021, p. 2). MaxEnt funciona de tal maneira que, dado um grupo de candidatos, uma probabilidade distinta é assinalada individualmente para cada um, diretamente derivada dos seus escores harmônicos (BEGUŠ, 2018, p. 192). Quanto melhor⁵ o escore harmônico, maior a probabilidade do candidato ser selecionado como *output* – e vice-versa. Vejamos, a fim de exemplo, um tableau MaxEnt:

/input/	R ₁ (P = 15)	R ₂ (P = 8)	R ₃ (P = 8)	H	Probabilidade
☞ a. candidato _a	−1			−15	0.58
b. candidato _b		−2		−16	0.21
c. candidato _c		−1	−1	−16	0.21

Tabela 3: Tableau adaptado de FLEMMING, 2021, p. 5

Observando as probabilidades acima, podemos afirmar sem dúvidas que a gramática MaxEnt é não-categórica. Não assumimos um único candidato vencedor, muito pelo contrário, atribuímos aos candidatos *chances de produção*. Em uma gramática MaxEnt, representações superficiais, então, são entendidas como resultado possível de um cálculo probabilístico derivado do escore harmônico entre candidatos de um grupo. Para mais, uma das competências da gramática MaxEnt é também achar pesos para as restrições de tal maneira que seja maximizada a probabilidade das formas mais observadas e minimizada a probabilidade das formas pouco ou não observadas em dado *corpus* (HAYES; WILSON, 2007, p. 8). Assim, gramáticas estocásticas, de maneira geral, almejam resolver o problema da variação na produção da fala, cujas gramáticas categóricas (como a OT e HG tradicionais) não dão conta. Torna-se necessário, ao pretendermos dar conta da variação, elaborar uma gramática fonológica cuja função vá da representação subjacente, por meio de distribuição probabilística, a uma representação superficial inserida em um grupo de representações superficiais possíveis (MAGRI, 2018, p. 1). E isso

³“The phonological component is a system of rules (...) that relates surface structures (...) to phonetic representations (...), and we will also make specific proposals concerning the system of phonological rules that assign a **phonetic interpretation to each surface structure**.”(CHOMSKY; HALLE, 1968, p. 14, grifo nosso)

⁴As principais implementações estocásticas baseadas em modelos de restrições são OT de ordem parcial, OT estocástica, HG estocástica e MaxEnt (MAGRI, 2018, p. 1).

⁵Por “melhor”, entende-se aqui o escore harmônico com a menor soma de violações para um candidato, *i.e.*, mais próximo de zero.

é de grande importância para o problema que este trabalho pretende adereçar, como explicarei na seção seguinte.

2.2 O problema da discretização

A interface entre fonética e fonologia no contexto de gramáticas baseadas em restrições apresenta desafios interessantes quando se consideram “candidatos contínuos” – isto é, variáveis fonéticas contínuas que atuam como candidatos em análises fonológicas. Um aspecto crítico dessa abordagem envolve a discretização dessas variáveis contínuas em intervalos menores (em *bins*), prática existente na literatura sobre o tema⁶, ainda que esta não seja extensa⁷. Esta discretização, embora metodologicamente necessária, suscita uma questão fundamental sobre a natureza da análise:

Objeção: Em termos de cômputo de violações, sabendo que os candidatos eram um único intervalo antes da discretização, como interpretar valores infinitesimalmente limiares entre dois candidatos? Por exemplo, discretizamos o intervalo “10 Hz até 30 Hz” em dois candidatos “10 Hz até 20 Hz” e “21 Hz até 30 Hz”, como será que devemos interpretar o valor 20.999999999 Hz limiar entre candidatos? Isto é, como adequadamente analisar valores fronteiriços?

Perceba que a pergunta acima tem grandes implicações teóricas. A pergunta traz o seguinte problema: se cada candidato (*i.e.*, *bin* contendo um intervalo de uma variável fonética) é julgado de maneira diferente, como julgar valores limítrofes entre *bins*, sabendo que o fenômeno subjacente era contínuo? Em outras palavras, não é realista lidar com candidatos discretizados de maneira homogênea. Vejamos um exemplo de alguma categoria discreta fonológica imaginária (/input/) que depende dos valores de F1:

/input/	* 3 5 0 H Z	* 4 5 0 H Z
a. 300 - 400 Hz	*!	
☞ b. 401 - 500 Hz		*

Tabela 4: Exemplo de tableau com variáveis fonéticas (1).

Acima, temos uma variável contínua cujo limite inferior é 300 e o superior é 500. Ela foi discretizada em 2 candidatos, candidato *a* (300 - 400 Hz) e candidato *b* (401 - 500 Hz).

⁶Cf. BOERSMA; ESCUDERO; HAYES, 2003

⁷Isto é, análises de restrições na interface Fonética-Fonologia cujos candidatos são fenômenos fonéticos contínuos em comparação com análises mais exclusivamente fonológicas da Otimalidade em geral.

A categoria fonológica imaginária */input/* aparenta preferir valores de F1 mais altos, como, *e.g.*, [-ATR]. Além disso, estamos lidando com duas restrições simples:

*350 Hz: O candidato não pode conter 350 Hz.

*450 Hz: O candidato não pode conter 450 Hz.

Aqui, podemos notar algo possivelmente problemático sobre a abordagem atual da Fonologia otimalista em relação a fenômenos fonéticos. A falta de elementos gradientes. No tableau acima, será que 401 Hz deve ser julgado de maneira totalmente diferente de 400 Hz? Apesar da discretização tê-los separado, eles ainda fazem parte de uma única variável contínua (distribuição) subjacente. Para que o problema fique ainda mais evidente, podemos realizar uma abordagem estocástica. Veja o seguinte tableau adaptado:

<i>/input/</i>	*350 Hz (3)	*450 Hz (1)	H	P
a. 300 - 400 Hz	-1		-3	0.25
 b. 401 - 500 Hz		-1	-1	0.75

Tabela 5: Exemplo de tableau com variáveis fonéticas (2).

Se nossa intenção é prever F1, será que realmente faz sentido fonético que 400 Hz seja agrupado no intervalo com 25% de chance de produção, mas 401 Hz esteja no intervalo assinalado 75%? É uma transição demasiadamente abrupta para candidatos que não estão apenas justapostos, eles são – pré-discretização – um único intervalo contínuo enquanto fenômeno do mundo. Mesmo que as produções individuais sejam discretas, as distribuições dessas produções são contínuas – não diferente de um degradê ao longo de uma gaussiana.⁸ Então, nos resta a pergunta: como podemos assinalar probabilidades discretizantes para fenômenos em distribuição contínua? Perceba que não estou disputando a validade dos resultados desse tipo de análise, pois, estruturalmente, o quadro teórico deste texto é também otimalista. O que questiono aqui é o grau de *precisão* que essa análise alcança. Abaixo, trarei um exemplo de uma publicação que

⁸Há algumas ressalvas importantes a serem feitas aqui. Do ponto de vista da percepção, é possível sim falar em transições abruptas entre categorias fonológicas, especialmente considerando a teoria quântica da fala (cf. STEVENS, 1972) que demonstra que a relação entre os movimentos articulatorios e as produções resultantes não são lineares. Isso explicaria como certas produções, em nosso caso, faixas de formantes, seriam favorecidas em razão de sua estabilidade no momento da articulação. No entanto, a crítica elencada no corpo do texto persiste e não contradiz, de forma alguma, o fato de que a diferenciação entre categorias fonológicas na percepção pode ser potencialmente abrupto. A questão disposta é um apontamento sobre a modelagem na Otimalidade que, mesmo considerando os processos de percepção e produção subjacentes, parece sofrer perda de informações relevantes (especialmente de transição entre categorias) ao realizar essa discretização. É uma investigação, entre outras coisas, sobre os limites da OT no limiar Fonética-Fonologia, na transição entre categorias fonológicas (*e.g.*, fonemas) e sobre a necessidade (ou não) de discretização no aparato matemático da Otimalidade.

posiciona-se em favor da abordagem de intervalos a fim de demonstrar como os problemas elencados acima encontram-se presentes na literatura.

Em *Learning Abstract Phonological from Auditory Phonetic Categories: An Integrated Model for the Acquisition of Language-Specific Sound Categories* (2003), Boersma, Escudero e Hayes apresentam um modelo de aquisição de categorias fonológicas baseado em restrições. Esse modelo baseia-se em parte na existência de uma restrição que os autores chamam de WARP que distorce o espaço perceptual da criança. Veja seu funcionamento:

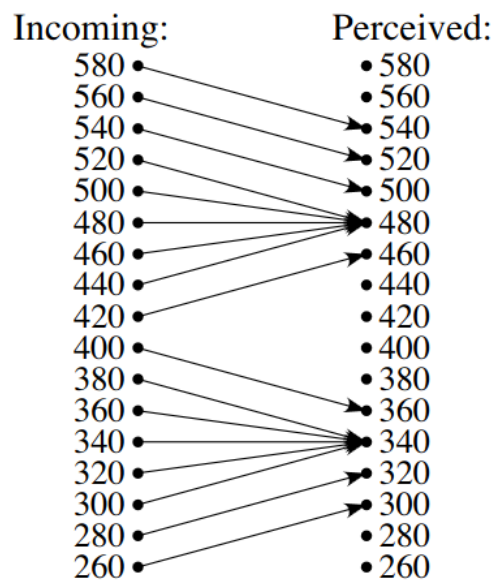


Figura 3: Funcionamento da restrição WARP (retirado de BOERSMA; ESCUDERO; HAYES, 2003, p. 2).

Os autores irão argumentar que os *incoming inputs* são intervalos mapeados a *perceived inputs* de tal maneira que há a conversão entre o fenômeno fonético contínuo infinito e categorias fonológicas discretas finitas (BOERSMA; ESCUDERO; HAYES, 2003, p. 3). Isso traria maior plausibilidade à seguinte modelagem na OT discreta:

[280 Hz] /280/
a. /280 Hz/
b. /320 Hz/
c. /340 Hz/
d. /-/

Tabela 6: Candidatos retirados de BOERSMA; ESCUDERO; HAYES, 2003, p. 2.

O problema desse tipo de análise é justamente o que acabamos de discutir. Por

mais que a discretização tenha sido realizada de maneira não-arbitrária – seguindo simulações e utilizando dados de aquisição (*Ibid.*, p. 3) – a discretização ainda traz, por si só, um aspecto categórico para análises que originam-se de variáveis (distribuições) contínuas. A solução dos autores é condizente com uma Fonologia discreta que, estando a OT no âmbito da Fonologia e não da Fonética, os candidatos devem ser discretos. No entanto, seria tão estranho pensar que, no ponto de contato entre Fonologia e fenômenos fonéticos contínuos, as categorias fonológicas apresentassem comportamento diferenciado? Nesse caso, a discretização poderia estar representando apenas um passo adicional desnecessário para a análise. O distorcer⁹ do espaço perceptual da criança, por exemplo, não precisaria estar atrelado a intervalos discretos específicos e tão homogêneos.¹⁰ Uma maior fluidez das categorias fonológicas poderia representar um maior grau de precisão e flexibilidade nesse tipo de análise. Como disse anteriormente, não coloco em questão os resultados finais desse tipo de análise, mas sim se não poderíamos nos beneficiar com análises ainda mais granulares e precisas. Ao invés de trabalhar com candidatos que homogeneamente em seus intervalos recebem violações, poderíamos realizar operações com candidatos que respeitassem a heterogeneidade dos fenômenos no mundo. Mesmo analisando um candidato dito 300 - 400 Hz, consideraríamos a realidade de que, *e.g.*, produções específicas 348 Hz ou 398 Hz não estariam violando de maneira idêntica a uma mesma restrição. A distribuição de violações deve seguir a distribuição dos dados de maneira fidedigna e granular.

A proposta deste trabalho, enfim, tem aqui seu ponto inicial. Estudar a viabilidade de uma gramática baseada em restrições *contínua*, para uso específico na interface Fonologia-Fonética. A escolha pelo texto de Boersma, Escudero e Hayes apresentado aqui foi pragmática, pois é uma publicação com bastantes citações e que foi influente no estabelecimento desse tipo de discretização. Dito isso, tendo contextualizado as metas deste trabalho, podemos agora passar a explorar novas possibilidades de não-discretização de candidatos.

⁹A palavra “distorcer” aqui é empregada em referência ao fenômeno de *warping* ou *distorted recognition* descrito e nomeado por Boersma em *How we learn variation, optionality, and probability* (1997) em que a percepção é modelada como um mapeamento de um número infinito de valores a um *continuum* auditivo de formantes. Esse fenômeno é empregado em publicações subsequentes, como a citada anteriormente de Boersma, Escudero e Hayes (2003).

¹⁰Perceba que o maior problema parece ser a homogeneidade em como os candidatos são tratados. Uma violação atribuída a um intervalo significa, na OT discreta, que todos os elementos dentro daquele intervalo recebem por igual essa violação. Há uma falta de nuance na distribuição de violações internamente ao candidato discreto quando este é um intervalo.

3 Fundamentos matemáticos

Ao longo deste capítulo, elencarei uma série de expressões matemáticas que conjuntamente formulam um protótipo de modelagem fonológica para variáveis fonéticas contínuas.¹¹ Todas as subseções consistirão na apresentação dos princípios matemáticos seguida de uma discussão sobre sua implementação, demonstrando sua aplicação com exemplos em dados reais.

Com fins demonstrativos e para prova de conceito, irei utilizar neste capítulo dados coletados de 3 falantes de dialetos nordestinos do sexo masculino, todos com cerca de 30 anos de idade e recém-migrados para Campinas – totalizando 114 instâncias de /e/ pretônico. Os dados estão disponíveis na íntegra no **anexo B**. Vale a pena ressaltar que essa modelagem não limita-se a um ou outro tipo específico de dado, mas, considerando que estes são os dados disponíveis para esta pesquisa, serão eles que utilizarei para exemplos de aplicabilidade e implementação – e isso me parece aceitável, pois, em nosso recorte de pesquisa proposto, estamos mais preocupados com as operações da arquitetura matemática do modelo do que com possíveis conclusões fonológicas que tiraríamos dos mesmos.¹² Assim, aqui demonstro a arquitetura do modelo com dados limitados – utilizando apenas formantes como variáveis – para que, ao obter uma quantidade suficiente e mais diversa de dados em uma pesquisa posterior, seja possível realizar uma verdadeira análise com intuito de obter *insights* fonológicos e maior refinamento da modelagem.

Uma última nota é que para aplicação dessa metodologia, foi elaborada uma interface visual (composta por dois *scripts* em Python3) análoga ao MaxEnt Grammar Tool de Hayes e Wilson. Essa interface (ou, melhor dizendo, seus *scripts*), que já está disponível para usuários, será apresentado na seção **Implementação computacional**. Assim, neste capítulo apenas discutirei aspectos matemáticos, e deixarei para discutir os aspectos computacionais em seguida.

3.1 Dados

Primeiramente, é preciso explicitar como os dados devem ser tratados nesse quadro teórico. Afinal, realizaremos uma série de operações matemáticas que requerem consideração

¹¹ Estando este trabalho posicionado nesse ponto de contato entre Fonética e Fonologia, acho importante notar que essa modelagem é fundamentalmente *fonológica* por uma questão tanto de princípios como de finalidades. Digo *princípios* pois parte de um arcabouço teórico majoritariamente pertencente à Fonologia (como a OT e a HG), e digo *finalidades* pois seus objetivos também pertencem às investigações fonológicas, debatendo questões como, por exemplo, modelagem e processos das gramáticas de restrições.

¹² O corpus selecionado serve como base suficiente para demonstrar a funcionalidade da arquitetura matemática proposta, permitindo a análise dos mecanismos computacionais desenvolvidos. O foco central desta pesquisa reside no desenvolvimento e validação do modelo teórico-matemático, sendo essa a contribuição desta pesquisa para a área.

sobre seus *inputs* e seus *outputs* numéricos.

3.1.1 Normalização

Inicialmente, considerei que a normalização dos dados devesse ser realizada com o método de Lobanov (cf. LOBANOV, 1971), mas não obtive sucesso com sua implementação. Houve um problema com a matriz de covariância da estimação de densidade por kernel, utilizada no próximo passo, quando utilizei os dados normalizados por Lobanov. Obtive sucesso, no entanto, com uma normalização por escore Z.¹³ Futuramente, seria interessante pensar em maneiras de como tornar o método de Lobanov compatível com o processo de estimação de densidade por kernel. Enfim, para o funcionamento desse quadro teórico, só é preciso padronizar os dados para uma escala como é feito em cálculos de escore Z, mas devemos levar em conta metodologias vantajosas para a Linguística, como normalização de Lobanov.

3.1.2 Estimação de densidade por kernel

Após a normalização, devemos reunir os dados normalizados e utilizá-los para a criação de uma função de densidade de probabilidade (doravante, f.d.p.). Nós faremos uso dessa f.d.p. para *ponderar* o *output* de funções de restrição, como veremos na seção que discuto as violações. Ela será utilizada para intermediar candidatos com restrições, ajudando a computar violações específicas para cada conjunto de dados coletados. A f.d.p. será capaz de fazer isso pois detém informação probabilística sobre a distribuição dos dados, nos informando a *densidade* de suas regiões. Mas, antes de tudo, devemos pensar em como podemos criar a f.d.p. Podemos fazer isso por estimação de densidade por kernel (doravante, EDK), por meio da seguinte expressão matemática (HASTIE; TIBSHIRANI; FRIEDMAN, 2008, p. 208):

$$\hat{f}_{\mathbf{X}}(x_0) = \frac{1}{N\lambda} \sum_{i=1}^N K_{\lambda}(x_0, x_i) \quad (1)$$

O funcionamento da expressão acima pode ser compreendido da seguinte maneira: para estimarmos o valor de uma f.d.p. $\hat{f}_{\mathbf{X}}$ em um ponto x_0 conhecendo amostras aleatórias

¹³O método de Lobanov, diferentemente do escore Z que utiliza a média e o desvio padrão globais da amostra, calcula esses valores separadamente para cada indivíduo. É, portanto, uma normalização intra-individual, e é amplamente usada na Fonética Acústica. O problema é que, ao aplicar Lobanov e padronizar os dados dentro de cada indivíduo, a matriz de covariância resultante dos dados normalizados dessa maneira ficou não invertível (*i.e.*, seu determinante foi zero). Se essa matriz não for invertível, o cálculo dos pesos do kernel pode falhar (como aconteceu no experimento). É possível que, em uma quantidade limitada de dados ou devido *aos meus dados específicos*, a normalização de Lobanov acabou por reduzir excessivamente a variabilidade, e isso mal condicionou a matriz de covariância. Em pesquisa futura, com uma quantidade maior de dados, o problema encontrado entre a normalização de Lobanov e a estimativa de densidade por kernel deve ser investigado.

$x_1, \dots, x_N$, podemos somar $\sum_{i=1}^N K_\lambda(x_0, x_i)$ as contribuições de cada amostra x_i na vizinhança de x_0 . Por “contribuições”, estamos falando em termos de densidade. Se nosso kernel $K_\lambda(x_0, x_i)$ for gaussiano, por exemplo, podemos pensar que todas as amostras representam uma distribuição gaussiana e, ao somá-las, estimamos a f.d.p. de nossos dados. O parâmetro de largura λ controla a largura do kernel, definindo a amplitude dos picos de cada amostra: valores maiores de λ resultam em uma curva mais suave, enquanto valores menores produzem uma curva menos suave – basicamente, introduzindo a decisão entre mais detalhes mas mais ruído (mais heterogêneo) *versus* menos detalhes mas menos ruído (mais homogêneo). Por fim, a fração $\frac{1}{N\lambda}$ serve para normalizar os dados, fazendo com que sua integral resulte em 1, a fim de satisfazer uma das principais propriedades de f.d.p. Esse é o funcionamento básico de uma EDK.

Neste capítulo, trabalharemos com uma f.d.p. conjunta, *i.e.*, uma f.d.p. de duas ou mais variáveis. Isso introduz algumas diferenças¹⁴ na fórmula da EDK, mas a intuição por trás de uma estimação de densidade por kernel permanece suficientemente similar. Assim, pegaremos as 114 amostras normalizadas e elaboraremos sua f.d.p. por EDK. Faremos uso do kernel gaussiano (distribuição normal bivariada) nos dados normalizados previamente por escore Z (média em 0, desvio padrão de 1).¹⁵ O resultado final pode ser visto aqui:

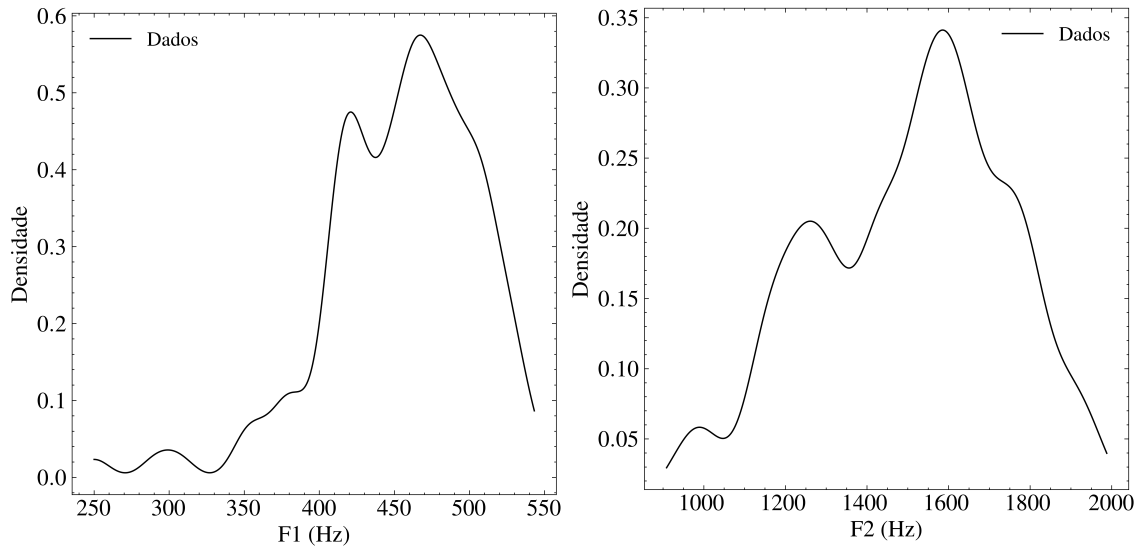


Figura 4: Função de densidade de probabilidade dos dados

Acima, podemos observar dois gráficos, um para o formante F1 e outro para o formante F2. Foi calculada uma f.d.p. conjunta e, para melhor visualização, cada variável foi plotada individualmente. Esses gráficos indicam, de acordo com os nossos dados, quais regiões

¹⁴A estimação agora pode envolver vetores e matrizes, a seleção de largura torna-se diferente também, entre outros elementos afetados pela questão da dimensionalidade.

¹⁵Todos os gráficos são transformados de volta para a escala original após as operações, para visualização.

são mais “densas”, *i.e.*, quais regiões acumulam mais dados. Assim, utilizaremos a f.d.p. para transportar informações para o cômputo de violações, notificando-o sobre onde há candidatos passíveis de serem analisados com base na densidade dos dados. E, em regiões de baixa densidade, irá informar nossa restrição que ali não há nada a ser analisado. Mas isso é uma discussão para próximas seções. Podemos denominar essa f.d.p. conjunta dos dados como $\hat{f}_{\text{dados}}(F1, F2)$, utilizando o acento circunflexo acima de f para notar que é uma *estimação*. A largura utilizada acima foi $\lambda = 0.2$. Essa largura foi selecionada partindo da regra de Silverman, que resultou em 0.398, realizando sucessivas reduções para evitar a supersuavização da função resultante. Assim, equilibrou-se maior detalhamento dos dados com suavização suficiente. Agora que elaboramos $\hat{f}_{\text{dados}}(F1, F2)$, podemos seguir em frente com outras explicações matemáticas.

3.2 Input

Quando digo “fenômenos contínuos” ou “variáveis contínuas”, tenho em mente a análise de fenômenos acústicos. Tradicionalmente, em oposição a uma Fonologia que é tida como abstrata, discretizante e categórica, temos a Fonética que é tida como gradiente e contínua. Há muitas abordagens que pretendem realizar o contato entre essas duas áreas em aparente dicotomia na Linguística, e há uma de particular interesse para nós que é a *Fonologia funcional* conforme elaborada por Paul Boersma no livro homônimo *Functional Phonology*:

Este livro trata do uso do som na linguagem. Ele desenvolve uma teoria que utiliza capacidades gerais do comportamento motor humano e da percepção para explicar, bem como descrever os dados das línguas do mundo. Podemos prever, bem como esclarecer generalizações sobre a organização da fala humana e resolver muitas questões fonológicas controversas, apenas separando os papéis da articulação e da audição dos sons da fala. (BOERSMA, 1998, p. 1, tradução nossa)

Ao realizar a interface entre o auditivo e o articulatório, Boersma pretende capturar aquilo que pode eventualmente passar despercebido pelas análises fonológicas. Além disso, uma de suas proposições de interesse para nós é a de que, antes de elaborarmos qualquer tipo de restrição funcional, precisamos compreender quais representações estão sendo avaliadas por essas restrições (*Ibid.*, p. 9). Isto é, quais são nossos candidatos, nossos objetos de análise. Esse é um *insight* particularmente valioso para nós, pois, ao elaborarmos um quadro teórico contínuo, temos agora um pontapé teórico inicial: devemos primeiramente pensar no que são “candidatos contínuos”. Isto é, devemos iniciar nossa modelagem pelo *input*.

Em nossa modelagem, podemos pensar que o *input* é uma especificação fonológica – traços, por exemplo – e os candidatos são intervalos contínuos de valores fonéticos (como formantes) que podem realizar essa especificação. Podemos, assim, nos alinhar à compreensão

de gramática bidirecional apresentada por Boersma (cf. BOERSMA, 1998). Relativizaremos a gramática em termos de *gramática do falante* (gramática de produção) e *gramática do ouvinte* (gramática da percepção), e a conceberemos como duas gramáticas distintas, porém intrinsecamente conectadas. Ainda fazendo uso da nomenclatura de Boersma, chamaremos o *input* (a representação subjacente) de *especificação*.

Tendo estabelecido nossa concepção geral de gramática, devemos agora detalhá-la um pouco mais. Podemos começar nossa discussão apresentando um tableau para cada gramática (produção e percepção). Segundo a gramática de produção, candidatos seriam configurações articulatorias acopladas às suas expressões perceptuais (BOERSMA, 1998, p. 146). Vejamos um exemplo:

<i>especificação</i>	A	B
☞ a. [art ₁] /perc ₁ /		*
b. [art ₂] /perc ₂ /	*!	

Tabela 7: Gramática de produção (adaptado de BOERSMA, 1998, p. 146)

A gramática da percepção, por outro lado, lida com a categorização de *inputs* acústicos. Por exemplo, ao receber um sinal acústico de um primeiro formante F1 de alguma vogal, o ouvinte precisa categorizar esse *input* em alto, médio ou baixo (*Ibid.*, p. 147-148). Por isso, a gramática de percepção apresenta um tableau distinto, vejamos abaixo:

<i>input acústico</i>	A	B
☞ a. /categoria ₁ /		*
b. /categoria ₂ /	*!	

Tabela 8: Gramática de percepção (adaptado de BOERSMA, 1998, p. 147)

Assumindo que todos nós tenhamos uma gramática de produção e uma de percepção¹⁶, o sistema fonológico apresenta um *feedback loop* que compara o resultado perceptual produzido por alguma configuração articulatoria (nosso *output*) com a especificação (nosso *input*) no momento da enunciação (*Ibid.*, p. 27). Similarmente, para esta modelagem, o *input* será uma forma de especificação. Sabendo que parecemos discriminar perceptualmente traços, ao invés de fonemas (cf. CHLÁDKOVÁ *et al*, 2015), poderíamos pensar nosso *input* como sendo, por

¹⁶Conforme Boersma (1998, p. 161), ao observarmos línguas de sinais podemos inferir que a fonologia pode ser baseada na capacidade de qualquer sistema motor e qualquer sistema sensorial que nos permita expressar intenções, desejos e pensamentos.

exemplo, um traço. Mas, em última instância, deixo a questão do *input* em aberto. Elenquei algumas considerações básicas, entretanto, na medida em que a modelagem é utilizada com diferentes tipos de variáveis, é difícil – ou, melhor, impossível – falar sobre *input* de maneira simultaneamente específica e universal. Logo, deixo em aberto a discussão sobre *input*, visto que qualquer afirmação categórica sobre sua natureza limitaria desnecessariamente o escopo de variáveis acústicas que esta modelagem é capaz de analisar.

3.3 Candidatos

Em nosso quadro teórico, a noção de candidatos permanece a mesma, mas a maneira como serão analisados é diferente. Como estamos trabalhando com variáveis contínuas, nossos candidatos podem ser qualquer intervalo dentro do domínio dos dados coletados. Mais especificamente, podemos pensar como intervalos de $\hat{f}_{\text{dados}}$. Em nosso caso:

F1 (Hz)	F2 (Hz)
250 até 543	908 até 1987

Tabela 9: Domínio dos dados.

Desse modo, qualquer intervalo – não importando sua amplitude ou a combinação específica entre intervalos de F1, F2 – é aceitável na condição de que encontre-se dentro dos limites inferiores e superiores de $\hat{f}_{\text{dados}}$. Nossos candidatos, portanto, não precisam ser discretizados previamente. Estamos analisando regiões de uma única variável contínua, um único “candidato contínuo” que podemos observar microscopicamente suas partes sem desmembrá-lo¹⁷ em elementos menores. Podemos selecionar, por exemplo, os seguintes valores como candidatos:

Faixa (F1)	Faixa (F2)
250 até 543	908 até 1987
250 até 350	908 até 1300
350 até 450	1300 até 1700
450 até 543	1700 até 1987
450 até 543	1300 até 1700

Tabela 10: Exemplos de regiões para análise.

¹⁷Pois é fundamentalmente diferente a separação prévia dos dados em *bins* e a análise de diferentes regiões de uma *única* função matemática que descreve *uma* variável contínua.

Ao longo das próximas subseções, irei usar de exemplo para análise as regiões acima.¹⁸ A primeira região é o domínio dos dados, seu alcance completo. As três regiões seguintes são divisões dos menores valores, valores mediais e maiores valores dos dados, divididos em fatias semelhantes. Por fim, a última região analisada é uma demonstração do efeito de modificarmos a combinação entre faixa de F1 e faixa de F2 analisadas, pois realizaremos uma análise conjunta dos dados. Assim, a última região analisada é uma combinação da penúltima região de F1 analisada e da antepenúltima região de F2 analisada.

Talvez aqui surja uma questão: ao separarmos diferentes faixas de análise, não estamos discretizando os candidatos? A resposta é *sim* e é *não*. Por um lado, qualquer análise requer discretização, discriminação e recorte do objeto de estudo. Isso é inevitável. Nesse sentido, sim, realizamos uma discretização ao decidirmos nossos candidatos. Mas não é esse tipo de discretização que tenho em mente ao insistir em “modelagem contínua”, como ficará evidente ao longo deste capítulo. Deve-se notar que apesar estarmos conceitualmente separando nossos candidatos em intervalos específicos, a variável matemática que os origina é contínua e será passível de ser tratada de maneira contínua ao longo de toda a modelagem. Esse é o ponto central dessa proposta de modelo. Ao longo das próximas páginas, ao adentrarmos questões um pouco mais técnicas, será mais fácil de compreender o que significa “contínuo” e “discreto” neste contexto.

3.4 Restrições

Vejamos agora as restrições. Se candidatos agora são contínuos e não estão mais dispostos em unidades discretas, é preciso realizar mudanças fundamentais na maneira como restrições operam e, ao mesmo tempo, preservar a intuição e definição tradicionais de *restrição* em quadros teóricos da OT e da HG.

Restrições “contínuas” que modelam fenômenos fonéticos devem ser fundamentalmente diferentes de restrições de fenômenos discretos. Afinal, como dito, a restrição depende de seu objeto de análise (BOERSMA, 1998, p. 9). Gostaria, então, de compor algumas diretrizes importantes que levam em conta que as restrições neste trabalho são elo entre Fonologia e Fonética, *i.e.*, que representam regras fonológicas agindo sobre produções fonéticas. Estas serão

¹⁸Foram selecionados três intervalos por questões demonstrativas, além disso, há precedentes no uso de intervalos não-logarítmicos em pesquisas da Otimalidade (cf. BOERSMA, 1997; BOERSMA; ESCUDERO; HAYES, 2003). De qualquer modo, para os fins desta pesquisa, essa divisão é pouco significativa, pois trataremos os candidatos (todos eles) como uma única variável contínua subjacente. Seria igualmente aceitável dividir os candidatos de acordo com as bandas críticas da Escala Bark, ou ter 2 ou 4 intervalos ao invés de 3, e assim em diante. Os candidatos na Otimalidade devem ser selecionados conforme a investigação da pesquisa e, aqui tratando-se de uma demonstração e prova de conceito, os candidatos acima foram satisfatórios em razão de serem didáticos o suficiente.

diretrizes para elaboração de restrições. Consigo, pelo menos preliminarmente, pensar em duas que são absolutamente necessárias:

1. Restrições devem representar a interação entre categorias fonológicas abstratas e a realidade da produção fonética. Em outras palavras, conforme a nomenclatura da Fonologia funcional, devem representar contato entre percepção, especificação e articulação.
2. Restrições devem ser capazes de avaliar variáveis contínuas em termos de violações, de maneira análoga a quadros teóricos tradicionais da OT e da HG. Sendo assim, devem receber uma variável de entrada (*i.e.*, candidato) e retornar um valor numérico que representa a quantidade de violações.

A primeira diretriz relaciona-se à conceitualização teórica de nosso modelo enquanto a segunda diretriz diz respeito à arquitetura do mesmo. Aqui, apesar de ainda não muito específico, já começamos nosso trabalho de delimitação do que devemos compreender por “restrição contínua” daqui em diante. As seções seguintes serão norteadas por essas duas diretrizes.

Acho que vale a pena comentar que, considerando a segunda diretriz, é talvez inevitável que nossas restrições tornem-se *funções matemáticas*. Mas isso não é uma ideia original deste trabalho. Edward Flemming trabalhou redução vocálica e funções de custo em *A Phonetically-Based Model of Phonological Vowel Reduction* (2001), e Michael Lefkowitz empregou funções parabólicas como restrições em sua tese de doutorado *Maxent Harmonic Grammars and Phonetic Duration* (2017). Como é possível inferir por seus títulos, ambos são trabalhados no limiar entre Fonologia e Fonética. Gostaria, portanto, de enquadrar meu trabalho como tendo uma proximidade metodológica a abordagens desse tipo.

Enfim, agora que iremos trabalhar nossas restrições, penso que é interessante pensar em pelo menos duas restrições de naturezas distintas. Formularemos nas próximas seções uma *restrição perceptual* e uma *restrição articulatória*. Em outras palavras, como será explicado, uma restrição de *fidelidade* e uma restrição de *marcação*. Vamos começar pela *restrição perceptual* da família das restrições de fidelidade.

3.4.1 Restrição perceptual

Nossa restrição perceptual será uma função sigmoide. Funções sigmóides são aquelas cuja curvatura tem formato de “s”. Podemos fazer uso do comportamento específico desse

tipo de função para modelar a interação entre categorias fonológicas e a produção fonética. Mas, antes de qualquer modelagem, vale a pena contextualizar um pouco as funções sigmóides e sua história na representação de processos cognitivos.

A inspiração por trás de uma restrição perceptual de caráter sigmoide inicia-se pelo *perceptron* de Frank Rosenblatt¹⁹, publicado originalmente em *The Perceptron: A Perceiving and Recognizing Automaton* (1957) e mais amplamente debatido em *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* (1962). Rosenblatt elenca a possibilidade de *reconhecimento de atributos de estímulos* ser essencialmente um *problema de classificação* (HAYEK, 1952 apud ROSENBLATT, 1962, p. 22). Nesse contexto, o perceptron pode ser interpretado como um modelo matemático de um neurônio cujo *input* são estímulos/fenômenos do mundo e seu *output* é uma classificação binária 0 ou 1, em linha com a ideia proposta por McCulloch e Pitts sobre o caráter “tudo-ou-nada” da atividade neuronal ([1943]1990, p. 99). Esse tipo de abordagem, apesar de extremamente significativa na história de aprendizado de máquina e de modelagem de sistemas cognitivos, apresenta algumas limitações que levaram pesquisadores subsequentes a buscar novas modelagens neuronais.²⁰ Dentre os vários desenvolvimentos posteriores ao perceptron original, é de nosso interesse falar sobre as “redes” de perceptrons, também chamadas *redes neurais*²¹, e o dito *perceptron multicamadas*.

Ao empregarem perceptrons multicamadas, muitos pesquisadores fazem uso de funções sigmóides como função de ativação²² (cf. BISHOP, 1995, p. 126-132). Originalmente, Rosenblatt elaborou perceptrons como neurônios binários, cujo estímulo progressivamente saturava-os até, ao alcançarem seu limiar, efetivarem ativação (ROSENBLATT, 1962, p. 72; *Id.*, 1957, p. 4). O valor exato desse limiar está relacionado à sensibilidade do neurônio (*Id.*, 1962, p. 48). Considerando esse funcionamento binário discreto da função de ativação do perceptron de Rosenblatt, podemos falar que essa se trata de uma função de Heaviside ou, mais descritivamente, uma função degrau. A introdução de uma função de ativação sigmoide, portanto, não só introduziu não-linearidade como também fez com que, mesmo respeitando um valor limiar, a transição entre categorias (0 e 1) fosse realizada de maneira contínua. A seguir,

¹⁹Que foi diretamente influenciado por McCulloch e Pitts, e sua obra sobre o primeiro modelo matemático de neurônios *A Logical Calculus of the Ideas Immanent in Nervous Activity* publicado em 1943.

²⁰Perceptrons são inadequados para lidar com problemas que não são linearmente separáveis, como problemas do tipo XOR (BISHOP, 1995, p. 104). Nesse caso, por exemplo, um perceptron não é capaz de dividir (classificar) com uma simples linha reta as coordenadas (0,0) e (1,1) das coordenadas (1,0) e (0,1).

²¹Uma rede neural não necessariamente é uma rede de perceptrons, pois atualmente há muitos outros tipos de neurônios artificiais em uso, mas uma coletânea de perceptrons em rede pode receber o nome de rede neural.

²²Uma função de ativação é uma das operações finais do neurônio, que normalmente pretende introduzir não-linearidade, controlar o fluxo de informação e transformar o *input* de uma maneira ou outra, conforme necessário, controlando seu *output*.

elaborei um gráfico com ambas funções de ativação que discutimos até aqui, para que fique evidente a diferença entre uma ativação discreta de uma ativação contínua. Nesse gráfico, tratei o limiar de ambas funções como sendo 5, para que pudéssemos observar seus comportamentos a partir de um mesmo referencial.

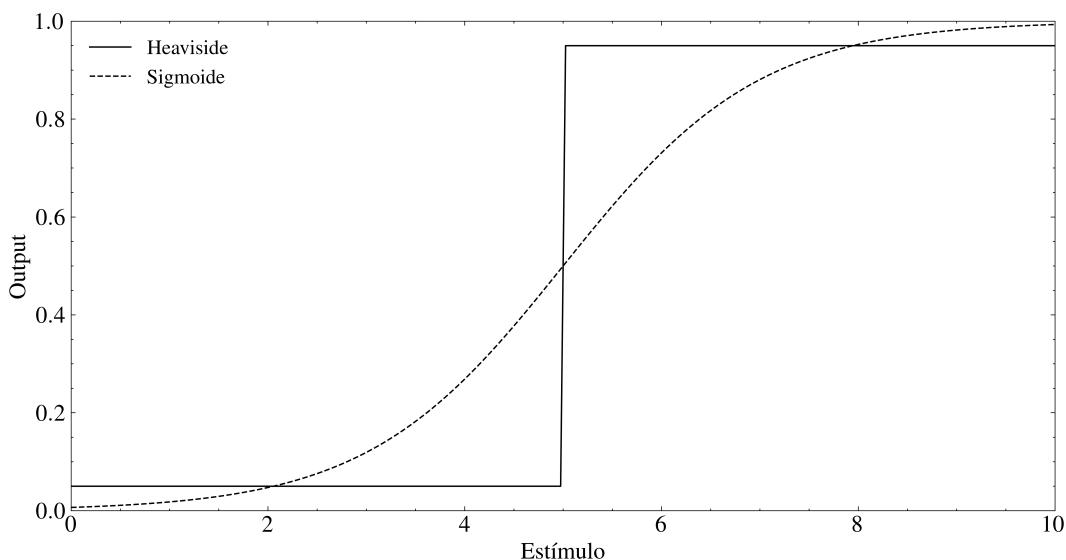


Figura 5: Função heaviside vs. função sigmoide.

Após observarmos esse gráfico e tendo em mente o que foi discutido até aqui, não é difícil notar que estamos falando de uma modelagem do ponto de contato entre entidades cognitivas discretas e, como diz Rosenblatt (1952, p. 1), o mundo dos *fenômenos*. De certo modo, isso assemelha-se às modelagens de HG aplicadas a fenômenos fonéticos. Não é absurdo, conseqüentemente, realizarmos uma “ponte” teórica entre elementos do perceptron e nosso modelo, pois há alinhamento entre nossos objetivos. E é assim que, sob sua inspiração, podemos pensar no conceito de função de ativação sigmoide como uma restrição perceptual para nossa modelagem.

Finalmente, agora podemos afirmar que a restrição perceptual será pensada como reprodução de uma função de ativação sigmoide cujo *input* são estímulos, fenômenos contínuos do mundo, e cujo *output* é uma classificação contínua entre dois polos abstratos. Nossa primeira restrição, sendo assim, satisfaz ambas condições que elencamos anteriormente como diretrizes necessárias de restrições contínuas. Ela representa um ponto de contato entre Fonologia (classificação abstrata da gramática) e Fonética (estímulo do mundo fenomenal), e ela é capaz de delegar violações quando pensamos que ela pode representar um alvo, como veremos a seguir ao discutir seu funcionamento.

Agora, após termos realizado essa contextualização histórica e teórica, podemos

começar a elaborar nossa restrição. A partir deste ponto, a modelagem será demonstrada através da alternância $[e] \sim [\varepsilon]$ e das variáveis F1 e F2 em contexto pretônico. As restrições serão elaboradas de maneira a poderem ser manipuladas por parâmetros customizáveis em uma interface visual, conforme será apresentado em **Implementação computacional**. Vale notar que, apesar desta demonstração específica com formantes, a modelagem não é intrinsecamente limitada a estas variáveis, como já foi dito.

Agora, como me propus a modelar $/e/$ pretônico no PB, em particular a alternância $[e] \sim [\varepsilon]$ não motivada por sândi ou harmonia, podemos pensar que um bom exemplo de restrição perceptual seria modelar a percepção da vogal média pretônica $/e/$. Para tanto, precisamos delimitar frequências²³ de nosso interesse. Isso não é uma tarefa trivial, pois envolve tomar decisões não apenas fonéticas como também cognitivas: quais são os pontos limiares de nossa percepção? Em quais circunstâncias, em termos de produção e percepção, *percebemos* uma categoria fonológica como absolutamente distinta de outra? Quão abrupta, em termos de variáveis acústicas, é a transição entre categorias da percepção? E como correlacionar todas as respostas dessas perguntas, caso um dia sejamos capazes de respondê-las, com valores numéricos de formantes de vogais? Para que não nos afoguemos agora em meio a essas questões, podemos realizar esse protótipo com *placeholders*. Futuramente, podemos tentar apaziguar algumas dessas questões, mas sem dúvidas não seremos capazes de respondê-las completamente, pois atravessam a própria Linguística, estendendo-se por temas que atualmente parecem carregar mais perguntas do que respostas, como Filosofia, Psicologia, Neurologia e Teoria da Mente. Entretanto, neste pedacinho miópico de pesquisa, podemos esboçar nossa restrição de maneira genérica e, futuramente, nos voltar para pesquisas de Aquisição, Psicolinguística, Fonologia e Fonética para encontrar *algumas* respostas. Vejamos o gráfico abaixo:

²³Nessa restrição, trabalharei apenas com F1, pois estamos modelando a percepção da abertura da vogal. Isso é uma decisão que tomei para priorizar simplicidade da modelagem matemática, mas não há nenhum empecilho em trabalharmos a percepção a partir de quantos formantes quisermos.

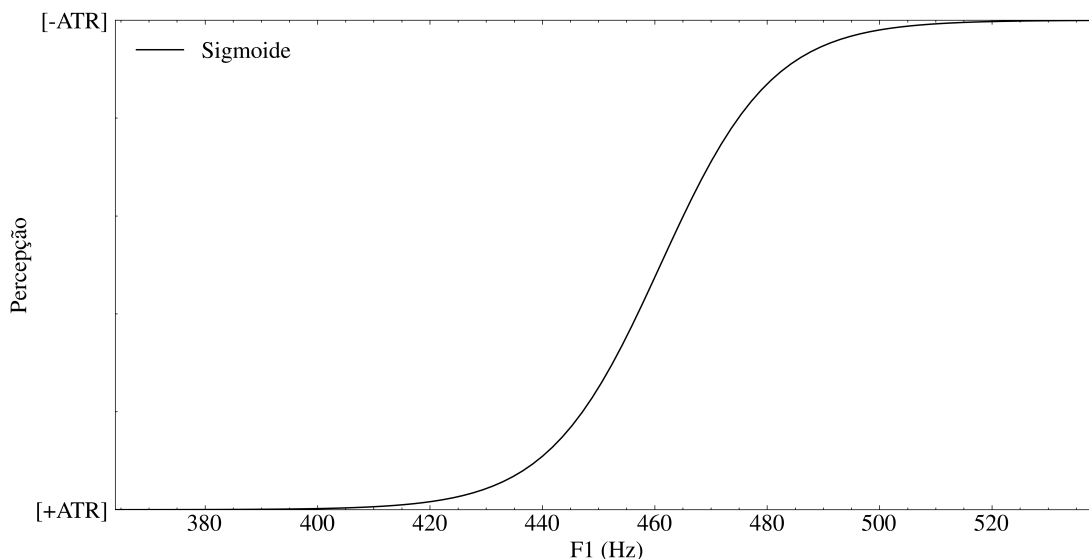


Figura 6: Função sigmoide com limiar em 461 Hz.

Sabendo que, segundo média de 10 falantes masculinos do PB de 5 regiões do Brasil em posição tônica, temos [e] médio em 383 Hz e [ɛ] médio em 539 Hz (FAILS; CLEGG, 1992 apud KENSTOWICZ; SANDALO, 2016, p. 3), o gráfico acima faz uso desses valores para mostrar como eles poderiam estar associados à percepção. Nele, consideramos o limiar em 461 Hz entre as categorias fonológicas [e] e [ɛ] – que é o ponto intermediário entre 383 Hz e 539 Hz. Esse limiar é arbitrário, e não há respaldo em literatura que nos diga que uma simples média ($383 + 539 = 922/2 = 461$) represente a transição entre categorias na percepção. Veja abaixo caso decidíssemos que o limiar de percepção entre categorias fonológicas fosse, dessa vez, 500 Hz:

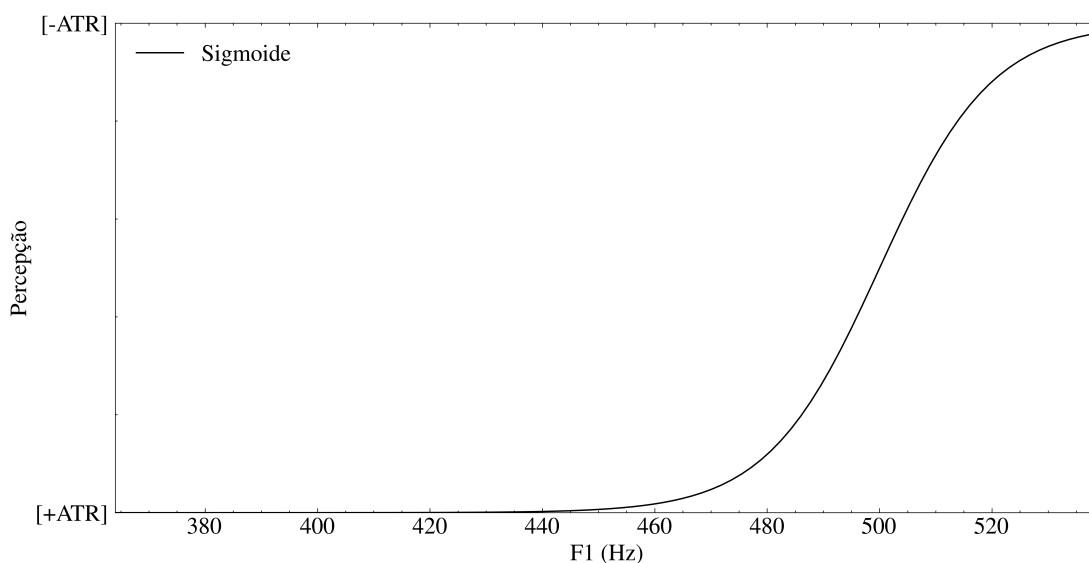


Figura 7: Função sigmoide com limiar em 500 Hz.

Além disso, como veremos mais para frente, até a inclinação dessa curvatura está sujeita a modificações, podendo modelar passagens mais ou menos bruscas entre categorias fonológicas. Poderíamos também, além de F1, incorporar outras informações relevantes para nossa análise que afetam a percepção, como duração ou outros formantes, incluindo formantes adjacentes. Mas não explorarei esses casos, afinal, minha proposta atual é uma demonstração do quadro teórico como um todo e não uma demonstração inesgotável de elaboração de restrições.²⁴

Finalmente, a maneira como traremos a função sigmoide para uma “restrição perceptual” será transformando-a em uma restrição com alvo. Em nosso caso, poderíamos pensar uma restrição que nos informa “vogais médias pretônicas devem ser vogais abertas” e penalizar produções *percebidas* como fechadas. Fundamentalmente, apesar de ainda não ser óbvio, isso deve ser interpretado como uma restrição de fidelidade. Partilho aqui do ponto de vista proposto pela Fonologia funcional de Boersma, que compreende a Fonologia como um sistema em *feedback loop* (BOERSMA, 1998, p. 27). Sob essa concepção de gramática fonológica, pensamos a fidelidade como sendo fidelidade entre *input* e *output*, pois o *output* é compreendido pelo próprio falante da enunciação e é comparado a seu *input* originário (*Ibid.*, p. 30). E essa infidelidade entre *input originário* e *output resultante* se daria por conta do sistema de produção de fala. Consequentemente, no contexto deste trabalho, trato a restrição perceptual como uma restrição de fidelidade, que pode ser entendida tanto em termos do falante consigo mesmo no *feedback loop* fonológico quanto em termos da interação entre falante e ouvinte. De qualquer modo, seria uma restrição relativa à *compreensão*, fazendo parte da gramática de percepção, conforme a nomenclatura da Fonologia funcional.

Trataremos, em nosso caso, “vogais abertas” e “vogais fechadas” como categorias fonológicas discretas [-ATR] e [+ATR], e a penalização se daria de maneira gradiente, mais ou menos abrupta, com base em um limiar específico entre as duas, em dado conjunto de produções fonéticas analisadas (nosso candidato contínuo). Assim, me repito: ao implementar esse tipo de restrição, estamos estabelecendo um ponto de contato entre as categorias fonológicas abstratas e as realizações acústicas fonéticas.

Em termos de HG e quadros teóricos de Otimalidade em geral, precisamos de uma função sigmoide cujo *input* seja o candidato e *output* seja a quantidade de violações incorridas pelo candidato. A quantidade de violações, no caso da sigmoide, dependerá do valor do limiar e da curvatura entre categorias fonológicas ou, em outras palavras, da saturação da percepção de

²⁴Como restrições são expressões matemáticas, a criatividade do pesquisador é o que limita quais variáveis ele deseja implementar, como elas interagem internamente na restrição e quão importante são no cômputo das violações.

categorias fonológicas. Assim, conforme as necessidades de nossa análise, devemos ser capazes de calibrar ambos elementos dessa restrição. Para tanto, optei por utilizar uma função logística – que é um tipo de função sigmoide.

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \quad (2)$$

Na função logística, L representa o valor máximo que a função pode atingir (*i.e.*, a quantidade máxima de violações possível em um ponto), k representa a inclinação da curva (*i.e.*, quão abruptamente passamos de uma categoria para outra) e x_0 representa o valor de x no ponto médio da curva (*i.e.*, o valor limiar). Poderíamos, conforme a análise em questão, experimentar com diversos valores de k . Vejamos alguns exemplos abaixo:

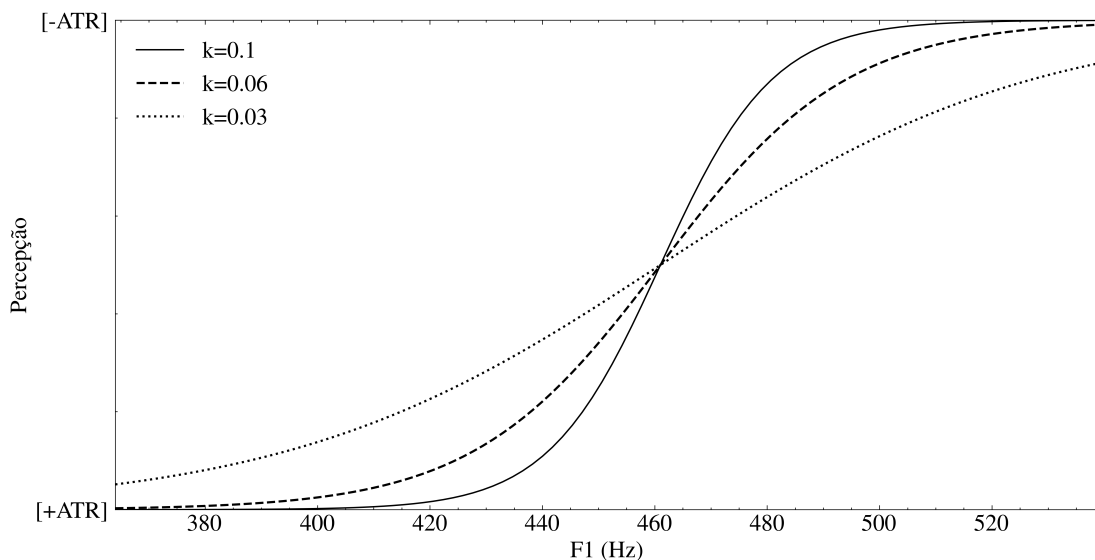


Figura 8: Diferentes taxas de saturação k da função sigmoide (todas com $L = 1$ e $x_0 = 461$ Hz).

No gráfico acima, é possível observar a maneira como diferentes valores de k modificam a taxa de crescimento dessa função. Quanto maior é k , mais abruptamente assumimos que a transição entre a percepção de entidades fonológicas se dá. Veja que com $k = 0.1$, temos uma mudança perceptual da vogal mais abrupta quando comparada com $k = 0.06$ que, por sua vez, é mais abrupto que $k = 0.03$. Apesar de ser um exemplo simplificado que apenas incorpora uma variável (no caso, F1), é possível ver o funcionamento dessa calibragem em ação. Assim, em diferentes análises poderíamos selecionar diferentes pressuposições sobre a percepção (mais especificamente, sobre a transição da percepção) de categorias fonológicas. Dessa maneira, podemos fazer uso da seguinte definição inicial:

$$R_P(F1) = \frac{L}{1 + e^{-k(F1 - F1_{\text{limiar}})}} \quad (3)$$

Vamos agora começar a refinar essa expressão para atender os propósitos de nossa análise²⁵. Podemos iniciar nossa modelagem deixando k positivo²⁶. Em seguida, podemos pensar nosso $F1_{\text{limiar}}$. Ao delimitarmos um limiar perceptual, devemos fazê-lo considerando o hiper-espaço dos alvos fonéticos, que ali encontram-se hiper-articulados (cf. JOHNSON; FLEMING; WRIGHT, 1993; BOERSMA, 2005). Há uma incongruência entre a *produção* fonética e o *parsing* fonético (*i.e.*, percepção). De acordo com o modelo simplificado proposto por Johnson, Flemming e Wright (1993, p. 525), alvos hiper-articulados são um estágio intermediário entre representações categóricas e a produção normal da fala. Assim, escolher um limiar é uma tarefa de equilíbrio que deve levar em conta não apenas médias de produção como também o *stratum* fonológico em que se dá a percepção de categorias. E, além disso, outra complicação é o fato de que categorias vocálicas muitas vezes são produzidas no mesmo espaço acústico, tornando a percepção um esforço probabilístico (BOERSMA, 1998, p. 177-179). A imagem a seguir nos mostra a maneira como categorias vocálicas “intrudem” umas nas outras:

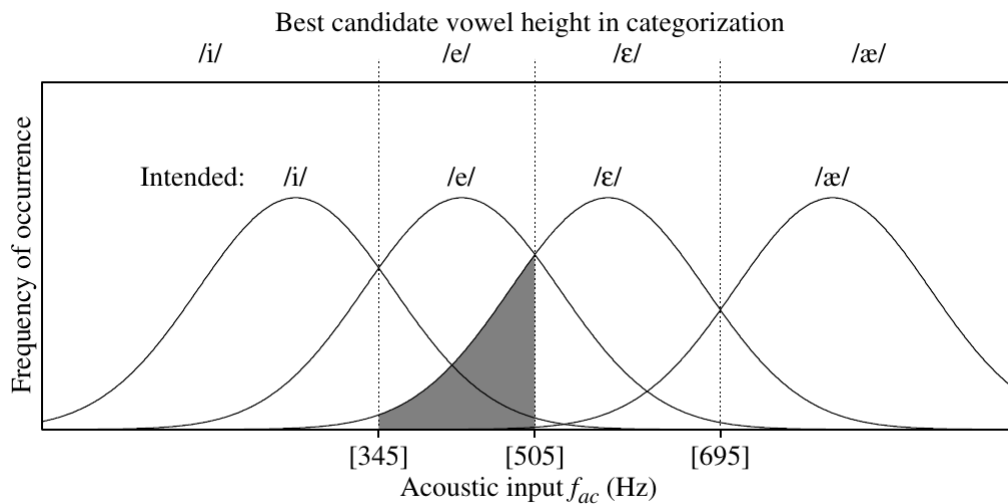


Figura 9: Representação da produção de categorias vocálicas (BOERSMA, 1998, p. 179)

²⁵A fim de exemplo de como análises específicas requerem formulações diferentes, mesmo tendo dito que não abordaria outras formulações, vou excepcionalmente elaborar uma restrição de exemplo que nos informa “a vogal deve ser uma vogal aberta, em virtude de seu F1 e de sua duração”. Poderíamos definir essa restrição da seguinte maneira: $R_P(F1, DUR) = L / (1 + e^{-k(F1 - (F1_{\text{limiar}} \pm \gamma \cdot DUR)))}$. Aqui, o limiar de F1 que afeta nossa percepção é influenciado pela duração, que é multiplicada por uma constante γ que determina quão forte é a influência da duração no limiar de F1. Essa influência pode ser positiva ou negativa, empurrando o limiar para cima ou para baixo, dependendo das pressuposições de nossa análise. Esse é um exemplo de como essas expressões matemáticas podem ser reformuladas para incorporar outras variáveis.

²⁶Mudar o sinal da constante k muda a orientação da sigmoide. Vai fazer com que ela incorra violações em valores mais baixos de F1, e não incorra em valores mais altos. Como estaremos modelando a abertura de /e/ pretônico *aberto*, essa é a orientação que faz mais sentido.

Sendo assim, podemos pensar que os valores limiares de $F1_{[+ATR]}$ e $F1_{[-ATR]}$ em situação hiper-articulada e probabilísticos merecem um cuidado que está fora do escopo desse protótipo. Irei, portanto, realizar uma simplificação para que possamos seguir em frente. Sabendo que nossas violações são dadas pela fórmula de nossa restrição, o limiar perceptual que trabalharemos será dado conforme a fórmula:

$$\begin{aligned}
 \text{violações} &= \frac{L}{1 + e^{k(F1_{\text{médio}} - F1_{\text{limiar}})}} \\
 \text{violações} \cdot (1 + e^{k(F1_{\text{médio}} - F1_{\text{limiar}})}) &= L \\
 \text{violações} + \text{violações} \cdot e^{k(F1_{\text{médio}} - F1_{\text{limiar}})} &= L \\
 e^{k(F1_{\text{médio}} - F1_{\text{limiar}})} &= \frac{L - \text{violações}}{\text{violações}} \\
 \log(e^{k(F1_{\text{médio}} - F1_{\text{limiar}})}) &= \log\left(\frac{L - \text{violações}}{\text{violações}}\right) \\
 k(F1_{\text{médio}} - F1_{\text{limiar}}) &= \log\left(\frac{L - \text{violações}}{\text{violações}}\right) \\
 F1_{\text{limiar}} &= F1_{\text{médio}} - \frac{1}{k} \log\left(\frac{L - \text{violações}}{\text{violações}}\right)
 \end{aligned}$$

Se quisermos resolver nossa restrição em virtude da quantidade de violações, podemos pensá-la da forma acima. Veja que podemos observar uma das propriedades da função logística: ela aproxima-se de 0 sem nunca resultar em 0, pois 0 violações resultaria em uma divisão por 0. A fim de minimizar nosso número de violações na média²⁷, podemos decidir que 0.01 é “próximo o suficiente” de zero e, conseqüentemente, que 0.01 violações é suficientemente satisfatório. Então, ao pensarmos que nossa restrição deve ter como *output* 0.01 na média da categoria, chegaríamos em:

$$F1_{\text{limiar}} = F1_{\text{médio}} - \frac{1}{k} \log\left(\frac{L - 0.01}{0.01}\right)$$

Se adotássemos a convenção²⁸ de que $L = 1$, teríamos:

²⁷Observe que, antes de rearmarmos a expressão acima, o número de violações se referia a $F1_{\text{médio}}$ que ocupa a posição de x em nossa restrição. Assim, “violações” refere-se ao *output* da função em x , que é $F1_{\text{médio}}$ aqui.

²⁸É comum que funções sigmóides sejam empregadas com *output* gradiente entre 0 e 1, como em tarefas de classificação.

$$\begin{aligned}
F1_{\text{limiar}} &= F1_{\text{médio}} - \frac{1}{k} \log \left(\frac{1 - 0.01}{0.01} \right) \\
F1_{\text{limiar}} &= F1_{\text{médio}} - \frac{1}{k} \log(99) \\
F1_{\text{limiar}} &\approx F1_{\text{médio}} - \frac{4.59}{k}
\end{aligned}$$

Quanto maior k , mais abrupta seria a discriminação perceptual, conforme demonstrado anteriormente. Agora, podemos compreender a razão disso. Quanto maior é k , menor é a distância entre $F1_{\text{médio}}$ e o limiar perceptual $F1_{\text{limiar}}$. Podemos trabalhar com $k = 0.1$, um valor relativamente neutro²⁹, e isso significa que, para que a média da categoria de F1 incorra 0.01 violações, seu limiar deve estar aproximadamente 45.9 Hz antes dela. Então, assumindo que a média da vogal pretônica /e/ é 421 Hz³⁰, temos um limiar em cerca de 375 Hz. Isso resulta, enfim, na restrição preliminar:

$$R_P(F1) = \frac{1}{1 + e^{0.1(F1-375)}} \quad (4)$$

Podemos realizar alguns cálculos pontuais, para saber se o funcionamento de nossa restrição está adequado, realizando uma prévia de uma possível modelagem para /ε/. Vejamos:

$$\begin{aligned}
F1 = 300 \text{ Hz} \quad \text{então} \quad R_P(300) &= \frac{1}{1 + e^{0.1(300-375)}} \approx 1 \text{ violação} \\
F1 = 325 \text{ Hz} \quad \text{então} \quad R_P(325) &= \frac{1}{1 + e^{0.1(325-375)}} \approx 1 \text{ violação} \\
F1 = 350 \text{ Hz} \quad \text{então} \quad R_P(350) &= \frac{1}{1 + e^{0.1(350-375)}} \approx 0.92 \text{ violações} \\
F1 = 375 \text{ Hz} \quad \text{então} \quad R_P(375) &= \frac{1}{1 + e^{0.1(375-375)}} \approx 0.50 \text{ violações} \\
F1 = 400 \text{ Hz} \quad \text{então} \quad R_P(400) &= \frac{1}{1 + e^{0.1(400-375)}} \approx 0.07 \text{ violações} \\
F1 = 425 \text{ Hz} \quad \text{então} \quad R_P(425) &= \frac{1}{1 + e^{0.1(425-375)}} \approx 0 \text{ violações} \\
F1 = 450 \text{ Hz} \quad \text{então} \quad R_P(450) &= \frac{1}{1 + e^{0.1(450-375)}} \approx 0 \text{ violações} \\
F1 = 475 \text{ Hz} \quad \text{então} \quad R_P(475) &= \frac{1}{1 + e^{0.1(475-375)}} \approx 0 \text{ violações}
\end{aligned}$$

²⁹Nem pouco e nem muito abrupto. Representa uma transição suave entre dois polos.

³⁰Para o dialeto baiano, conforme Madruga (2017, p. 56). Como pretendemos analisar /e/ pretônico de dialetos que os produzem de maneira aberta, podemos trabalhar esse protótipo com a média do /e/ pretônico do dialeto baiano.

Realizando essa pequena prévia, podemos observar que o funcionamento da restrição está conforme descrito. Ela incorre uma violação para formantes muito anteriores ao limiar perceptual e, após uma breve transição, ela passa a incorrer 0 violações dali em diante. Mas, ao começarmos a calcular nossas violações, nota-se um problema. Se calcularmos R_P para um valor imaginário qualquer muito acima do limiar (digamos, 9000 Hz), ela irá resultar em 0 violações, devido à natureza da sigmoide. Isso não condiz com a realidade linguística. Podemos retomar o gráfico de Boersma (1998, p. 179):

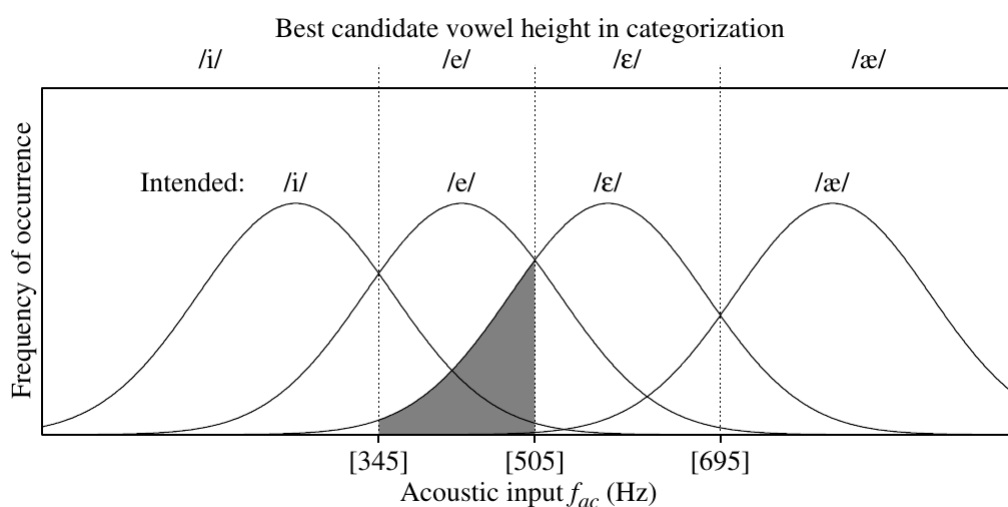


Figura 10: Representação da produção de categorias vocálicas (BOERSMA, 1998, p. 179)

Se estivéssemos, por exemplo, modelando /e/, não seria suficiente dizer que não há violações acima de 345 Hz, como mais ou menos fizemos em nossa restrição. Perceptualmente, acima de 505 Hz nos deparamos a categoria fonológica /ε/ e isso deve ser passível de violação, visto que nossa percepção “adentrou” ou “captou” outra categoria fonológica. Para solucionar essa aparente contradição em nossa restrição, devemos modificar sua expressão. Se uma sigmoide representa passagem entre categorias fonológicas, podemos adicionar uma segunda sigmoide em nossa expressão. Assim, teríamos duas transições perceptuais com dois limiares diferentes – um para valores que antecedem os esperados e outro para valores que sucedem os esperados. Nossa restrição, portanto, pode ser um produto de duas sigmoides. Na segunda sigmoide, podemos utilizar valores negativos para k , sinalizando que ela vai de aprox. 0 violações para aprox. L violações – o contrário da primeira sigmoide, que vai de aprox. L violações para

aprox. 0 violações. Poderíamos, portanto, repensá-la com a seguinte expressão³¹:

$$R_P(F1) = \frac{L^2}{(1 + e^{k_1(F1 - F1_{\text{limiar}_1})})(1 + e^{-k_2(F1 - F1_{\text{limiar}_2})})}$$

O problema é que, fora da área de transição entre a sigmoide e a sigmoide complementar, todos os valores aproximam-se de 0. Como resultado da multiplicação, temos 0 antes da área entre limiares (em virtude da sigmoide complementar) e temos 0 após ela também, criando um “centro” – e esse centro, por sua vez, é a única região cujos valores não são 0, pois é a única parte cujo *output* de uma ou outra sigmoide não era zero originalmente. Precisamos, então, inverter isso. Queremos alcançar o efeito contrário, que esse centro do produto das sigmoides – essa área entre os limiares das sigmoides – incorra aprox. 0 violações e todo o restante aprox. L violações. Para tanto, podemos adicionar o máximo de violações possível L na frente de nossa expressão e subtraí-la de L . Isso faz com que o número de violações seja uma subtração do máximo possível, invertendo essa tendência que descrevi. Onde antes incorria 0 violações, agora incorre L violações ($L - 0 = L$); onde antes incorria L violações, agora incorre 0 violações ($L - L = 0$). Chegamos, finalmente, na seguinte expressão:

$$R_P(F1) = L - \frac{L^2}{(1 + e^{k_1(F1 - F1_{\text{limiar}_1})})(1 + e^{-k_2(F1 - F1_{\text{limiar}_2})})} \quad (5)$$

Resumidamente, o que fizemos aqui foi multiplicar duas funções sigmoides. Nossa restrição perceptual é um teto de violações L subtraído do produto das duas sigmoides. Para que possamos compreender visualmente o que temos feito, vou elaborar o gráfico da sigmoide, sigmoide complementar e produto das sigmoides. Vejamos:

³¹Que é a forma rearranjada da expressão de multiplicação de duas sigmoides descrita anteriormente,

$$\frac{L}{1 + e^{k_1(F1 - F1_{\text{limiar}_1})}} \cdot \frac{L}{1 + e^{-k_2(F1 - F1_{\text{limiar}_2})}} = \frac{L^2}{(1 + e^{k_1(F1 - F1_{\text{limiar}_1})})(1 + e^{-k_2(F1 - F1_{\text{limiar}_2})})}.$$

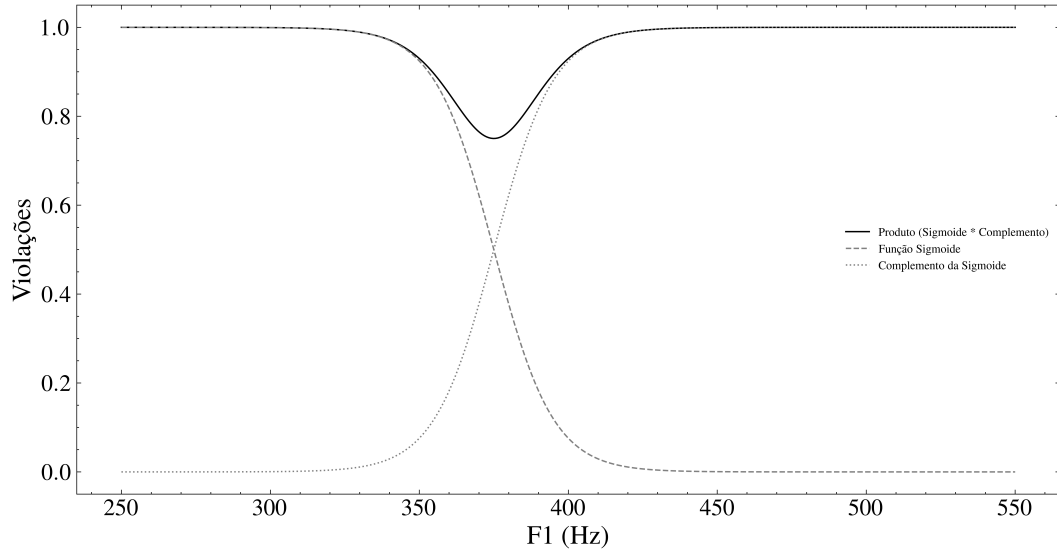


Figura 11: Sigmoide, sigmoide complementar e seu produto

O problema agora é que nossa nova definição parece não alcançar 0 violações quando definimos os limiares muito próximos um do outro nas sigmóides. Isso ocorre em razão de ser um produto de multiplicação: podemos imaginar no gráfico acima a proximidade das sigmóides “espremendo” mais ou menos o total de violações. Se isso for impactar negativamente nossa análise, podemos corrigir isso rapidamente com uma normalização quando necessário, no estilo:

$$R'_P(F1) = \frac{R_P(F1)}{\max_{F1 \in D} R_P(F1)} \quad (6)$$

O denominador de nossa restrição é o que garante sua normalização: estamos dividindo a função por seu máximo em D , seu domínio³². Vejamos agora um gráfico com essa nova definição:

³²Em nosso caso, D é o intervalo de nossos dados

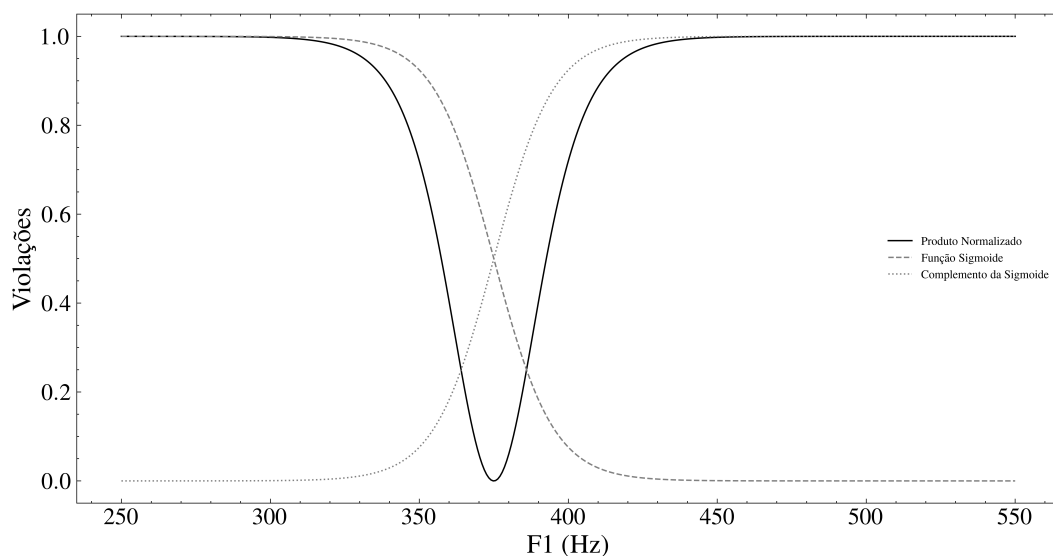


Figura 12: Sigmoide, sigmoide complementar e seu produto normalizado

Podemos observar no gráfico acima que agora “corrigimos” nossa função, atualizando seu valor mínimo para aprox. 0. Apesar disso, considerando como essa normalização acaba adicionando uma camada a mais de cálculo em nossa restrição, esse procedimento deve ser realizado somente quando estritamente necessário. Devemos, então, ajustar nossa estratégia conforme selecionamos nossos limiares.

Podemos agora voltar a definir limiares e inclinações. E é esse caráter adaptável que torna o produto de duas sigmóides mais atraente para nós do que, por exemplo, uma função gaussiana. Digo isso pois *visualmente*, até agora, pode parecer que apenas recriamos uma curva gaussiana. No entanto, quando definirmos os valores para nossa restrição, veremos a razão desse não ser o caso. E para começar a estabelecer valores específicos para nossa restrição, precisamos fazer uma reflexão sobre o que pretendemos analisar. Para fins de exemplo – e é um bom exemplo – vamos imaginar que estamos preocupados em modelar /e/ pretônico fechado, e usá-remos como referência [e]. Poderíamos pensar que, nesse experimento imaginário, ao modelar /e/ desejamos ser mais tolerantes com [ɛ] do que com [i] ao assinalar violações para frequências de formantes. Isso pois – *em situação que não seja de harmonia ou sândi e em posição pretônica* – a alternância [e] ~ [ɛ] é, para falantes do PB, menos contrastiva que [e] ~ [i]. Para melhor expressar esse argumento, vejamos a seguinte hierarquia contrastiva dos traços das vogais do PB (LEE, 2010, p. 37):

sigmóides para alcançar esse efeito. Vide o gráfico abaixo:

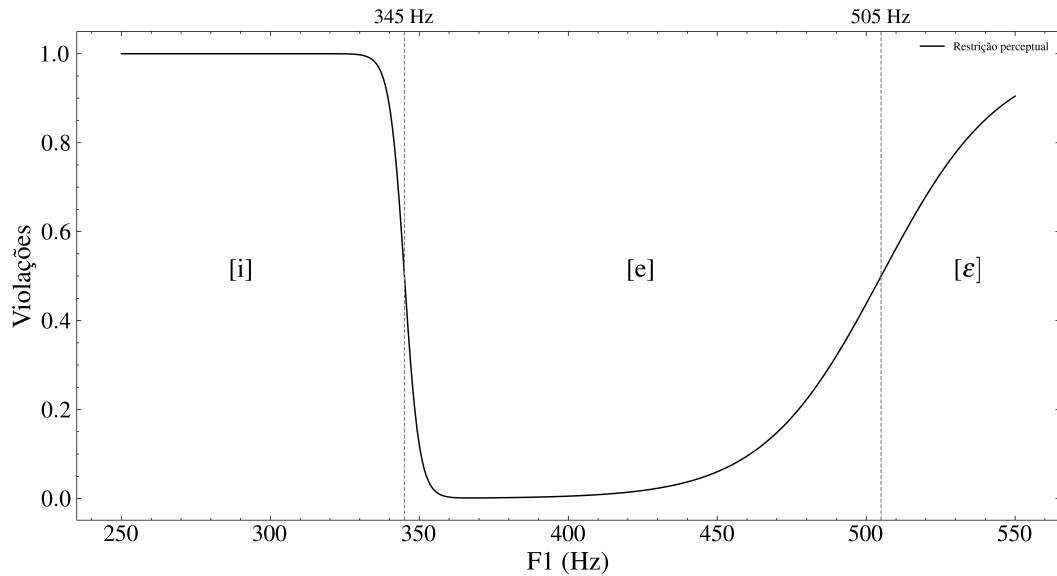


Figura 14: Saturação adaptável da restrição perceptual

O gráfico acima representa nossa restrição perceptual modelando um possível /e/ pretônico fechado. Ele demonstra como podemos manipulá-la³⁴ para alcançar relações fonológicas relevantes para nossa análise. No caso acima, *contraste*. E utilizando as propriedades do produto de duas sigmóides, podemos fazer com que a transição entre [e] ~ [i] seja mais abrupta (em termos de violações) do que entre [e] ~ [ε]. Assim, retomando o que havia dito anteriormente, não estamos recriando uma função gaussiana: estamos fazendo uso das propriedades específicas das sigmóides para nossa modelagem.

Por fim, para encerrar nossa discussão sobre restrição perceptual, é importante fazermos alguns cálculos pontuais para checar se seu novo funcionamento matemático está de acordo com nossas expectativas fonológicas. Vamos, para isso, usar a definição matemática de nossa restrição R_P com $k_1 = 0.05$, $k_2 = 0.4$, $F1_{\text{limiar}_1} = 505$ e $F1_{\text{limiar}_2} = 345$. Vejamos:

$$F1 = 300 \text{ Hz} \quad \text{então} \quad R_P(300) \approx 0.99 \text{ violações}$$

$$F1 = 325 \text{ Hz} \quad \text{então} \quad R_P(325) \approx 0.99 \text{ violações}$$

$$F1 = 350 \text{ Hz} \quad \text{então} \quad R_P(350) \approx 0.11 \text{ violações}$$

$$F1 = 375 \text{ Hz} \quad \text{então} \quad R_P(375) \approx 0 \text{ violações}$$

³⁴No caso do gráfico, foi representada R_P com $k_1 = 0.05$, $k_2 = 0.4$, $F1_{\text{limiar}_1} = 505$ e $F1_{\text{limiar}_2} = 345$. Os valores dos limiares foram retirados de Boersma (1998, p. 179) e os valores de $k_{1,2}$ foram decididos arbitrariamente. Poderíamos também ter calculado valores limiares com nossa função $F1_{\text{limiar}} \approx F1 - (4.59/k)$ definindo a partir de qual ponto específico a restrição começaria a incidir violações.

$F1 = 400 \text{ Hz}$	então	$R_P(400) \approx 0$ violações
$F1 = 425 \text{ Hz}$	então	$R_P(425) \approx 0.01$ violação
$F1 = 450 \text{ Hz}$	então	$R_P(450) \approx 0.06$ violações
$F1 = 475 \text{ Hz}$	então	$R_P(475) \approx 0.18$ violações
$F1 = 500 \text{ Hz}$	então	$R_P(500) \approx 0.43$ violações
$F1 = 525 \text{ Hz}$	então	$R_P(525) \approx 0.73$ violações

A restrição agora avalia e assinala violações de maneira linguisticamente razoável. Espero, portanto, que tenha sido possível demonstrar a modelagem de uma restrição de percepção e sua adaptabilidade. Podemos tornar nossa restrição o quão simples ou complexa desejarmos, tanto no quesito comportamento da função como na diversidade de variáveis que pode vir a receber como *input*. Dito isso, podemos formalizar a seguinte definição:

Definição:

A **restrição perceptual** é formulada como **a subtração do teto de violações pelo produto de duas funções sigmóides**. As sigmóides emulam a **saturação da percepção** que resulta na transição entre categorias percebidas. Essa restrição pretende descrever o **processo perceptual de discriminação** entre **categorias fonológicas**.

Com essa definição realizada e, por enquanto, suficientemente fundamentada, podemos seguir em frente e tratar da restrição articulatória.

3.4.2 Restrição articulatória

Se a restrição perceptual é uma restrição de fidelidade, a restrição articulatória pode ser entendida como uma restrição de marcação. Estaríamos considerando a marcação relativa entre gestos adjacentes durante a produção (BOERSMA, 1998, p. 152), mediada pelo conceito de *esforço*. A restrição articulatória, portanto, deve ser pensada como uma restrição de marcação que busca simultaneamente uma configuração específica do trato vocal e a minimização do esforço articulatório. Assim, ela opõe-se à restrição perceptual (que modela fidelidade entre *output* e *input* mediada pela percepção em *feedback loop* fonológico), em virtude de preferir formas que minimizam o esforço articulatório em relação às configurações adjacentes. Fundamentalmente, é uma restrição de coarticulações.

A elaboração dessa restrição foi motivada, pelo menos em parte, pelo entendimento

que a *Teoria Motora* oferece sobre a percepção da fala, afirmando que conhecimento anatômico do aparelho fonador está essencialmente atrelado tanto à produção quanto à percepção (cf. BALL, RAHILLY, 2013; LIEBERMAN, BLUMSTEIN, 1988; LIBERMAN, MATTINGLY, 1985). Isso vem ao encontro dos conceitos de especificação, articulação e percepção em *feedback loop* fonológico. Assim, ao aplicar essa restrição, estaremos modelando, através das frequências dos formantes, configurações do trato vocal supralaríngeo (LIEBERMAN, BLUMSTEIN, 1988, p. 147). E essas configurações do trato vocal – essas configurações articulatórias – são parte também da gramática fonológica, passível de ser modelada como restrição:

A primeira afirmação da teoria motora, conforme revisada, é que os objetos da percepção da fala são os gestos fonéticos pretendidos pelo falante, representados no cérebro como comandos motores invariantes que solicitam movimentos dos articuladores através de certas configurações linguisticamente significativas. Esses comandos gestuais são a realidade física subjacente às noções fonéticas tradicionais, como 'retroflexão da língua', 'arredondamento dos lábios' e 'elevação da mandíbula', que servem de base para as categorias fonéticas. Eles são os eventos elementares da produção e percepção da fala. (...) Na perspectiva fonológica, é claro que os gestos em si devem ser vistos como grupos de traços, como 'labial', 'oclusivo', 'nasal', mas esses traços são atributos dos eventos gestuais, não eventos em si. Portanto, perceber um enunciado é perceber um padrão específico de gestos pretendidos. (LIBERMAN, MATTINGLY, 1985, p. 2-3, tradução nossa)

Nossa restrição articulatória, então, tratará essas combinações como configurações articulatórias que fazem parte *tanto da produção quanto da percepção do falante* (LIBERMAN, MATTINGLY, 1985, p. 30). E, assim, estando no ponto de encontro entre produção e percepção – entre o contínuo e o categórico – e sendo capaz de calcular violações em termos de *desvio* de um alvo, podemos afirmar que satisfizemos as duas diretrizes que delimitamos para restrições contínuas. Além disso, como dito anteriormente, sendo uma restrição de marcação *entre gestos adjacentes*, ela deve incorporar a minimização do esforço articulatório. Vejamos o que Flemming nos diz sobre minimização de esforço:

Propusemos que uma classe de restrições de minimização de esforço penaliza o movimento articulatório entre segmentos, com movimentos maiores violando restrições de classificação mais alta, por exemplo, mover o corpo da língua da frente para trás é mais trabalhoso do que movê-lo da frente para o centro ou do centro para trás. A magnitude do movimento ao longo de uma dimensão é mais facilmente representada com dimensões escalares, assim como a distinção de contrastes é mais facilmente representada com dimensões escalares. Não exploramos totalmente esse potencial, formulando restrições contra mover *n* etapas em uma dimensão articulatória, com restrições classificadas de acordo com o tamanho do movimento, em vez disso, formulamos restrições bastante específicas, como *FRONT-BACK, *FRONT - CENTRAL. Adotamos a última abordagem para manter os nomes das restrições interpretáveis e em reconhecimento ao fato de que as dimensões articulatórias são mais tentativas e ad hoc do que as dimensões auditivas propostas aqui. Ou seja, não faremos nenhuma

tentativa de construir um conjunto geral de dimensões articulatórias, e algumas das dimensões propostas podem ser decompostas em múltiplas dimensões em um quadro teórico mais geral. (FLEMMING, 1995, p. 73, tradução nossa)

Apesar da concepção de *marcação* entre autores parecer ser divergente³⁵, há uma compreensão similar sobre a minimização do esforço. Vejamos agora a compreensão de Boersma de minimização de esforço:

A fonologia do falante também avaliará até que ponto o princípio funcional de minimização do esforço articulatório é respeitado. Essa avaliação ocorre em um espaço de características articulatórias, que incluem as possíveis posições, formas, movimentos e tensões dos lábios, bochechas, ponta da língua, corpo da língua, véu palatino, raiz da língua, paredes faríngeas, epiglote, estruturas laríngeas, pregas vocais e pulmões. A trajetória da implementação da enunciação através deste espaço articulatório é uma viagem ao longo de muitas posições, cada uma das quais é caracterizada como um vetor medido ao longo de escalas de grau de fechamento ou tensão. Embora essas escalas sejam contínuas, as línguas discretizam a maioria delas. Por exemplo, graus de fechamento supralaríngeo podem ser: completo (geralmente provocado por um movimento balístico: oclusivas e nasais); crítico (geralmente provocado por um movimento controlado, que o torna preciso o suficiente para manter o ruído de fricção ou vibração: fricativas); aproximante (articulação secundária forte, faringealização); estreito (0,3 - 1 cm²; vogais altas, semivogais, líquidas, raiz da língua retraída); aberto (1 - 4 cm²; vogais neutras); ou largo (4 - 15 cm²; lábios abertos, raiz da língua avançada). (BOERSMA, 1998, p. 15, tradução nossa)

Considerando que a minimização do esforço articulatório deve ser, conforme entendido por ambos autores, um processo multidimensional e passível de ser modelado, podemos tentar construir nossa restrição observando os princípios postulados pelos autores. Dando foco em nossa modelagem de maneira mais macroscópica e holística, parece-me melhor que, por enquanto, elaboremos essa restrição de maneira que capture os princípios dessa discussão sem nos apegarmos a modelagens granulares do trato vocal. Isso, no entanto, deverá ser pauta de um refinamento em pesquisa subsequente. E, ainda no que tange a aspectos para serem trabalhados futuramente, gostaria de falar sobre gestos adjacentes.

Há uma limitação metodológica importante a ser considerada: embora a restrição articulatória possa incorporar dados adjacentes (e.g., formantes das vogais imediatamente anteriores e posteriores às vogais médias pretônicas /e/), esta demonstração focará em uma versão mais simplificada do modelo. Para este protótipo, que visa demonstrar o método de forma generalizada, optou-se por uma abordagem operacionalmente análoga à concepção original, preservando o rigor metodológico sem depender de dados adjacentes ainda não estabelecidos no escopo atual

³⁵Boersma escreve sobre *marcação* entre gestos adjacentes, enquanto Flemming escreve sobre *marcação* entre representações auditivas (cf. BOERSMA, 1998; FLEMMING, 1995).

da pesquisa.³⁶ Ao invés de falar em esforço relativo à articulação adjacente, usarei o conceito de “esforço relativo ao ponto neutro da língua antes da fala”. Em termos articulatórios, estamos tratando como se cada produção fosse um deslocamento do ponto neutro da língua até o ponto da produção. Isso só seria verdade se assumíssemos que todas as produções analisadas correspondem ao início da produção da fala e que a língua necessariamente encontra-se em descanso antes de todas as vogais médias pretônicas analisadas. Isso não corresponde com a realidade, mas nos permite demonstrar como a minimização do esforço pode ser contemplada em nossa restrição – não relativa às articulações adjacentes, mas relativa a um ponto neutro.

Para nossa restrição articulatória, podemos pensar em três expressões matemáticas diferentes que irão, conjuntamente, defini-la. Vamos começar pelas seguintes definições para calcularmos um fator de esforço relativo:

$$E_{\text{Realizado}} = \sqrt{(C_{F1} - N_{F1})^2 + (C_{F2} - N_{F2})^2}$$

$$E_{\text{Esperado}} = \sqrt{(A_{F1} - N_{F1})^2 + (A_{F2} - N_{F2})^2}$$

Podemos compreender as duas expressões acima da seguinte maneira:

A é a coordenada articulatória F1,F2 esperada (alvo) para a vogal analisada. Pode ser representada pelos valores de F1,F2 médios de determinado dialeto para determinada vogal analisada.

C_{F1}, C_{F2} é a coordenada articulatória F1,F2 julgada pela restrição, a qual iremos delegar violações, produzida pelo falante. São as coordenadas que recebem os valores das variáveis acústicas F1 e F2, portanto $C_{F1} = F1$ e $C_{F2} = F2$.

N é a coordenada articulatória F1,F2 em posição neutra, onde a língua está em posição de descanso.

$\sqrt{(x - N_{F1})^2 + (y - N_{F2})^2}$ é a distância euclidiana entre alguma coordenada articulatória F1,F2 e a coordenada articulatória F1,F2 neutra. Em nosso protótipo, essa é nossa definição rudimentar de “esforço”.

$E_{\text{Realizado}}$ é o esforço (distância euclidiana) entre a produção F1,F2 analisada e a coordenada neutra.

³⁶ Isso deverá ser realizado em uma pesquisa subsequente.

E_{Esperado} é o esforço (distância euclidiana) entre a produção F1,F2 prototípica/média/esperada e a coordenada neutra.

Em síntese, E representa o esforço de um gesto articulatório, A representa o gesto articulatório alvo da análise e C_{F1}, C_{F2} (ou, seu correlato acústico originário $F1, F2$) representam o gesto articulatório produzido. Efetivamente, estamos calculando a distância euclidiana³⁷ (ou seja, a menor distância possível dentro de uma geometria euclidiana) entre um ponto articulatório neutro, um alvo e uma produção. Esse “ponto neutro” é a coordenada de descanso da língua antes da produção da fala. Sabendo que C_{F1}, C_{F2} são representações de configurações articulatórias da língua, iremos calcular o esforço com base na movimentação da língua indo de seu ponto de descanso pré-fala até o ponto da produção da fala. Como observado, usaremos F1 e F2 como coordenadas C_{F1}, C_{F2} , tendo em mente que são representativos da altura e anterioridade/posterioridade da língua, e realizaremos um cálculo de distância euclidiana entre essas diferentes coordenadas. Assim, as duas expressões acima representam dois cálculos: o esforço *esperado* e o esforço *realizado*. O *esforço esperado* é a distância entre o ponto neutro da língua e a configuração articulatória que esperamos que o falante vá realizar em nossa análise, segundo F1,F2 médios de seu dialeto ou outras informações que temos *a priori*. O *esforço realizado*, por outro lado, é a distância entre o ponto neutro da língua e a configuração articulatória que o falante, no ato da fala, acabou realizando e estamos analisando. Assim, vejamos como ficaria nosso fator de minimização de esforço:

$$\frac{E_{\text{Realizado}}}{E_{\text{Esperado}}} \quad (7)$$

Esse fator acima ponderará nossa restrição articulatória³⁸ e pode ser compreendido da seguinte maneira: quando o esforço realizado $E_{\text{Realizado}}$ for menor que o esperado E_{Esperado} diremos que houve uma minimização do esforço. Em contrapartida, quando o esforço realizado $E_{\text{Realizado}}$ for maior que o esforço esperado E_{Esperado} diremos que não houve minimização do esforço. Desse modo, ao pensarmos em *esforço*, estamos computando a distância entre o ponto neutro da língua antes da produção da fala e o ponto da língua nas configurações articulatórias em análise. Assim, definimos o esforço em virtude da distância entre a língua descansando e o ponto médio que a língua alcança (em termos de altura e anterioridade/posterioridade) na produção

³⁷Gostaria de fazer menção à tese de doutorado de Madruga, *A Fonética e Fonologia da Harmonia de [ATR] no Português Brasileiro* (2017), que também fez uso do conceito de distância euclidiana entre formantes, inclusive no debate de harmonia vocálica no PB.

³⁸Podemos adicionar uma constante ϵ em seu denominador caso o alvo da análise coincida com o ponto neutro da língua, para evitarmos uma divisão por 0. Vou omitir isso dessa equação, pois essa é uma questão mais específica da implementação e menos conceitual.

de uma categoria fonológica específica. Como F1 e F2 nos contam sobre a posição da língua, utilizamos F1 e F2 como um sistema de coordenadas geométricas do trato vocal. Isso é uma versão simplificada da proposta do trato vocal como espaço multidimensional/geométrico. Em pesquisa subsequente, como já havia dito anteriormente, será necessário substituir a “coordenada neutra da língua” por “coordenadas adjacentes posteriores e anteriores”. Na interface visual do programa apresentado no capítulo seguinte **Implementação computacional**, me certifiquei que, pelo menos, o usuário pudesse alterar os valores da coordenada neutra para adequá-lo à realidade de sua pesquisa. Isso é uma solução temporária, mas talvez suficiente para um primeiro protótipo de modelo. Gostaria agora de apresentar na íntegra a expressão matemática de nossa restrição articulatória. Vejamos:

$$R_A(F1, F2) = \sqrt{(C_{F1} - A_{F1})^2 + (C_{F2} - A_{F2})^2} \cdot \frac{E_{\text{realizado}}}{E_{\text{esperado}}} \quad (8)$$

Podemos compreender a fórmula acima da seguinte maneira: a restrição articulatória incorre violações com base no desvio entre a produção analisada e a produção alvo. Em nosso caso, esse alvo pode ser os valores médios de F1, F2 para a vogal média pretônica /e/ no dialeto estudado.³⁹ Assim, quanto maior o desvio entre a produção e a média dialetal⁴⁰, mais violações são incorridas. Mas essa violação é ponderada pelo esforço: se o desvio ocorre de maneira que minimiza o movimento da língua, nosso fator de esforço tem valor menor que 1 e diminui a quantidade de violações. Por outro lado, se o desvio ocorre de maneira que aumenta a movimentação da língua, as violações são acentuadas, pois nosso fator será maior que 1. Resumidamente, a restrição articulatória observa a distância entre a produção e um alvo, e busca compreender se esse desvio do esperado (o desvio da “média”) se deu em razão de minimizar o esforço coarticulatório entre categorias fonológicas. Se sim, ela incorre menos violações; se não, ela incorre mais violações.

Agora, para que comecemos a trabalhar com elementos menos abstratos, vejamos a tabela abaixo que nos fornece informações sobre algumas médias de F1 e F2 em um contexto pretônico no PB:

³⁹Vale a pena notar que seremos um tanto liberais na seleção de valores dessa restrição, em termos dialetais. Nossos dados das regiões do Ceará e do Maranhão e nosso alvo que será da região da Bahia devem ser – em uma pesquisa cujo intuito é a avaliação fonológica e que almeja extrair *insights* relevantes – pensados de maneira mais dialetalmente coerente. Aqui, isso não é um problema, pois estamos tratando as restrições como *placeholders* funcionais, mas é algo a se pensar em pesquisas futuras e/ou que empreguem essa modelagem.

⁴⁰Por exemplo, o valor médio dos formantes produzidos por falantes do dialeto pesquisado. É o valor esperado que aquele falante naquela comunidade com aquele dialeto produza.

Dialeto	Gênero	F1 (Hz)	F2 (Hz)
Gaúcho	Masc.	410	2054
Baiano	Masc.	421	1887

Tabela 11: Frequências médias dos dois primeiros formantes de /e/ pretônico (MADRUGA, 2017, p. 56)

Se estivéssemos modelando a gramática de falantes do dialeto baiano, poderíamos estabelecer um alvo para a configuração articulatória como $F1 = 421$, $F2 = 1887$. Para um gaúcho, por outro lado, poderíamos estabelecer o alvo $F1 = 410$, $F2 = 2054$. Podemos trabalhar em nosso protótipo com o alvo baiano $F1 = 421$ e $F2 = 1887$. Em relação ao ponto neutro, pode-se dizer que a posição neutra da língua – ou seja, a posição da língua antes da fala – é o ponto de referência para seus movimentos e corresponde aproximadamente a [ɛ], conforme David Odden (2013, p. 325). Então, usaremos a frequência média de [ɛ] como posição neutra da língua, considerando que essa posição pode ser “usada como ponto de referência para definir movimentos relativos da língua” (Ibid., p. 325, tradução nossa). Mais especificamente, utilizaremos $F1 = 610$ e $F2 = 1900$. Essa seria, de acordo com J. C. Catford (2001, p. 154), uma frequência média razoável para [ɛ] para falantes masculinos.⁴¹ Isso tornaria nossa restrição:

$$R_A(F1, F2) = \sqrt{(C_{F1} - 421)^2 + (C_{F2} - 1887)^2} \cdot \frac{\sqrt{(C_{F1} - 610)^2 + (C_{F2} - 1900)^2}}{\sqrt{35890}} \quad (9)$$

E essa restrição apresenta o seguinte comportamento:

⁴¹É importante notar que essas frequências podem estar fazendo referência a um [ɛ] da língua inglesa, mas seu uso é proposital nesta pesquisa por questões de coerência. Odden descreve a posição neutra da língua no trato vocal como sendo a mesma da produção de [ɛ], assim, assumindo que esse [ɛ] seja referente à língua inglesa, devemos buscar os formantes a que Odden se refere e não necessariamente ao [ɛ] de algum dialeto do PB. Não estamos especificamente atrás de [ɛ] em uma ou outra língua, mas buscando a posição “neutra” da língua que, ao considerarmos Odden e Catford, seria equivalente a uma vogal produzida em $F1 = 610$ e $F2 = 1900$.

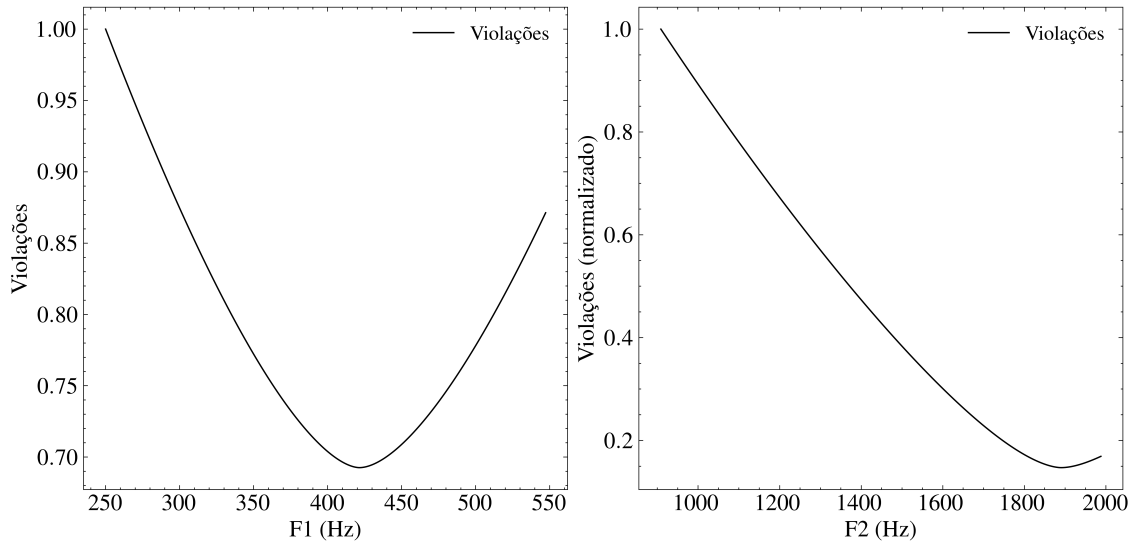


Figura 15: Incidência de violações da restrição articulatória

As violações de nossa restrição parecem calibradas o suficiente, considerando que queremos modelar /e/ pretônico aberto. Há maior incidência de violações fora da região de [ε], progressivamente aumentando conforme aumenta a distância. É importante notar que esses dois gráficos, na realidade, representam uma única função matemática de duas variáveis (F1,F2). Dessa maneira, uma maneira mais fidedigna de apresentá-la seria com um gráfico tridimensional. Senti, porém, que seria mais intuitivo se observássemos individualmente cada eixo dessa função separadamente, realizando um “recorte” dela. No mais, para de fato decidir se nossa restrição está em concordância com nossas expectativas fonológicas, devemos realizar alguns cálculos de demonstração. Vamos lá:

$$R_A(250, 1606) = \sqrt{(250 - 421)^2 + (1606 - 1887)^2} \cdot \frac{\sqrt{(250 - 610)^2 + (1606 - 1900)^2}}{\sqrt{35890}}$$

$$\approx 804.5$$

$$R_A(334, 1541) = \sqrt{(334 - 421)^2 + (1541 - 1887)^2} \cdot \frac{\sqrt{(334 - 610)^2 + (1541 - 1900)^2}}{\sqrt{35890}}$$

$$\approx 850.2$$

$$R_A(401, 1870) = \sqrt{(401 - 421)^2 + (1870 - 1887)^2} \cdot \frac{\sqrt{(401 - 610)^2 + (1870 - 1900)^2}}{\sqrt{35890}}$$

$$\approx 29.2$$

$$R_A(472, 1708) = \sqrt{(472 - 421)^2 + (1708 - 1887)^2} \cdot \frac{\sqrt{(472 - 610)^2 + (1708 - 1900)^2}}{\sqrt{35890}}$$

$$\approx 232$$

$$R_A(513, 1549) = \sqrt{(513 - 421)^2 + (1549 - 1887)^2} \cdot \frac{\sqrt{(513 - 610)^2 + (1549 - 1900)^2}}{\sqrt{35890}}$$

$$\approx 671.6$$

Aqui, após nossa primeira demonstração, surge um questionamento sobre as magnitudes relativas de F1 e F2. Há cerca de 1000 Hz de diferença entre F1 e F2, e essa diferença pode fazer com que o número de violações seja muito mais influenciado por F2 do que por F1, em razão de seus valores serem muito mais altos. Em outras palavras, divergência de F2 é desproporcionalmente representada no resultado final. Apesar dessa ser uma observação importante, ela não deve nos preocupar muito, pois isso é minimizado ao trabalharmos com valores normalizados. Além disso, os valores das violações também passarão por um procedimento de normalização entre as restrições perceptual e articulatória.⁴²

Enfim, essa é a restrição articulatória. Tomando combinações de formantes como representantes de configurações articulatórias, calcularemos a distância entre diferentes configurações e avaliaremos, em termos de violações, quão distante determinada produção acústica (enquanto produto de determinada configuração articulatória) está de um alvo determinado *a priori* para nossa análise. Similarmente à restrição perceptual, a restrição articulatória pode ser dita *bidirecional*, pois faz parte tanto da produção quanto da compreensão do falante. Além disso, podemos dizer que nossas restrições são complementares, pois pertencem a *strata* diferentes: percepção e articulação. E, diferentemente da restrição perceptual, podemos dizer que restrições articulatórias pertencem à família das restrições de marcação e priorizam formas menos marcadas em relação aos gestos vizinhos (BOERSMA, 1998, p. 152). Além disso, o falante tende a minimizar seus gestos articulatórios, isto é, o número e complexidade de seus gestos e coordenações (*Ibid.*, p. 3). Consequentemente, para incorporar a minimização dos gestos articulatórios, foi elaborado um “fator de esforço” que alivia ou agrava a quantidade de violações incorridas. Logo, podemos realizar a seguinte formalização dessa restrição:

Definição:

A **restrição articulatória** é caracterizada pela **distância euclidiana** entre um **alvo** e a **produção efetivada**, que deve ser **multiplicada** por um **fator de minimização de esforço** que **agrava ou atenua** a quantidade de violações.

Assim, encerro minha apresentação formal sobre restrições, restando apenas algumas considerações que farei a seguir.

⁴²O problema da ordem de grandeza das violações das restrições (perceptual vs. articulatória) foi considerado um problema *computacional* e será tratado adiante, durante a implementação computacional da modelagem.

3.4.3 Algumas considerações teóricas

Gostaria de encerrar essa seção realizando algumas repetições do que já foi discutido. Primeiramente afirmando que a restrição perceptual opõe-se à articulatória. Ela o faz no sentido de que a minimização do esforço – a marcação relativa entre configurações articulatórias adjacentes – vai na via contrária à fidelidade entre *input* e *output* priorizada pela restrição perceptual. Assim, tentei preservar nesse protótipo relações tradicionais da literatura de Fonologia otimalista: a interação entre marcação e fidelidade. A escolha por elaborar esses dois tipos específicos de restrições (uma articulatória, uma perceptual) teve como intenção, portanto, mostrar como poderíamos trazer a Fonologia otimalista de maneira *prática* para esse tipo de análise contínua. Não bastaria neste protótipo que tivéssemos “apenas” ou “quaisquer” restrições, pois nossa tarefa também deve ser demonstrar restrições *teoricamente plausíveis* em uma interface Fonologia-Fonética. Simultaneamente, e por *design*, um esquema de restrições perceptuais-articulatórias também nos permite explorar possíveis conciliações entre abordagens gesturalistas (cf. DIEHL; KLUENDER, 1989; FOWLER, 1986, 1996; LIBERMAN; MATTINGLY, 1985; NEAREY, 1997; OHALA, 1996) e gramáticas baseadas em restrições.

Vale a pena dizer que, sendo percepção e articulação consideradas *famílias* de restrições, é válido quebrar essas duas restrições em restrições menores e mais específicas. A organização da quantidade e especificidade das restrições depende da análise pretendida. Demonstrativamente neste trabalho, a restrição da percepção foi apenas *uma* – que poderia, por exemplo, ter sido dividida em outras menores⁴³ ou acrescida a outras⁴⁴. O mesmo vale para a restrição articulatória.

Mais adiante – bem mais para o final deste texto – veremos que a plausibilidade cognitiva é um fator fundamental para a elaboração desse tipo de modelagem. Já adiantando um pouco do que será exposto futuramente, pode-se dizer que é a plausibilidade cognitiva (mais especificamente, fonológica) que *justifica* esse tipo de modelagem em primeiro lugar. A incorporação de pressupostos fonológicos na *base* da análise diferencia essa modelagem de outros modelos contínuos que são puramente estatísticos ou matemáticos. Nós não desejamos “apenas” a modelagem de fenômenos fonéticos – pois poderíamos fazê-la por outros métodos estatísticos, como ANOVA, modelos de efeitos fixos, *etc* – o que nós almejamos é a modelagem de fenômenos fonéticos *sob pressupostos fonológicos*. É claro que a escolha metodológica específica de cada trabalho envolve o conhecimento do objeto de estudo e seu recorte teórico,

⁴³Como duas restrições de uma sigmoide.

⁴⁴Como uma restrição perceptual acerca duração.

mas, como pretendo argumentar neste trabalho, sinto que há mais méritos do que deméritos em expandir nosso leque de metodologias disponíveis.

3.5 Violações

Admito que até aqui nossa modelagem não foi *contínua*. Nossos cálculos, por exemplo, foram realizados com variáveis discretas. Mas, começando nesta seção, passaremos a trabalhar exclusivamente com variáveis contínuas, mobilizando os conceitos que vimos até aqui. Espero também que fique mais evidente quais são os benefícios dessa abordagem trabalhada. Também vale notar que, a partir deste ponto e por questões de complexidade de cálculo, começaremos a usar nossos *scripts* e deixaremos de fazer cálculos manualmente nas demonstrações. Dito isso, podemos discutir como serão calculadas nossas violações.

Primeiramente, precisamos pensar no que é uma violação. Para tanto, gostaria de realizar duas citações. Já no primeiro parágrafo do prefácio de *Optimality Theory* (KAGER, 2004), obra dedicada a introduzir o leitor ao quadro teórico da OT, René Kager escreve:

Este livro apresenta uma introdução à Teoria da Otimalidade, um quadro teórico gramatical de origem recente (Prince e Smolensky 1993, McCarthy e Prince 1993a, b). A ideia central da Teoria da Otimalidade (OT) é que as formas superficiais da língua refletem resoluções de conflitos entre demandas ou *restrições* concorrentes. Uma forma superficial é “ótima” no sentido de que incorre as violações menos graves em um conjunto de restrições violáveis, classificadas em uma hierarquia específica da língua. As restrições são universais e codificam diretamente marcação e princípios que garantem a preservação dos contrastes. As línguas diferem na classificação das restrições, dando prioridade a algumas restrições sobre outras. Tais classificações são baseadas na “dominação estrita”: se uma restrição supera outra, a restrição de classificação mais alta tem prioridade, independentemente das violações daquela de classificação mais baixa. No entanto, tais violações devem ser mínimas, o que prevê a propriedade de *economia* dos processos gramaticais. (KAGER, 2004, p. xi, tradução nossa)

A próxima menção mais explicativa de *violação* ou *violações* encontra-se no primeiro capítulo:

A violação de uma restrição não é causa direta de ingramaticalidade, nem é a satisfação absoluta de todas as restrições essencial para os *outputs* da gramática. Ao invés disso, o que determina o melhor *output* de uma gramática é a violação menos custosa das restrições. As restrições estão intrinsecamente em *conflito*, portanto, cada *output* logicamente possível de qualquer gramática necessariamente violará pelo menos alguma restrição. As gramáticas devem ser capazes de regular os conflitos entre restrições universais, a fim de selecionar a forma de *output* mais “harmoniosa” ou “ótima”. (KAGER, 2004, p. 3, tradução nossa)

Gostaria de chamar atenção para o fato de que, em ambas citações (pertencendo ao prefácio e ao primeiro capítulo, respectivamente), o conceito de *violação* é apresentado como in-

tuitivo, não dispondo de maiores explicações sobre seu funcionamento. Isso é recorrente durante a obra. Kager futuramente irá discutir sobre propriedades de *violabilidade*, *grau de violações*, *economia de violações* e assim em diante sem nunca explicitamente nos informar o que é, conceitualmente, uma violação em termos cognitivos ou da gramática fonológica (cf. KAGER, 2004, p. 12, p. 20, p. 21). Dessa maneira, parece ser suficiente saber que se restrições são violáveis, o candidato que não atende às expectativas da restrição *incorre violações*. A violação, da maneira como é geralmente concebida, é um *subproduto* da interação entre candidatos e restrições. E, sendo a violação um tipo de subproduto, é inevitável que devemos modelar uma *interação* ao modelar a violação.

Na última seção, ao discutirmos ambas restrições, chamei o *output* da função da restrição de “violação”. Mas, rigorosamente, não é assim que gostaria que concebêssemos nossas *violações*, pois “violações como *output* da função da restrição” é uma concepção que, especificamente no caso de nossa modelagem contínua, não sustenta-se sob escrutínio. Gostaria de realizar esse argumento partindo dos trabalhos funcionais-computacionais de Flemming (cf. FLEMMING, 1997, 2001, 2004, 2011, 2017), notando semelhanças e diferenças com meu projeto. Sendo assim, acho que devemos ver o exemplo de “violação” em *Scalar and categorical phenomena in a unified model of phonetics and phonology* (FLEMMING, 2001), sua segunda publicação sobre o tema e artigo em que expõe os princípios de sua metodologia. A modelagem de Flemming é colocada em termos de um problema de otimização tradicional⁴⁵, e isso significa que a forma ótima será pensada como aquela que minimiza a função de custo. A função de custo, por sua vez, é a soma das funções de restrições, cujos *outputs* são considerados violações (FLEMMING, 2001, p. 19). Isso é compatível com a lógica de quadros teóricos da OT e HG, pois também apresentam-se como problemas de minimização de violações entre restrições concorrentes. O problema é que esse tipo de “violação” só funciona ao tratarmos de variáveis discretas. Podemos pensar esse problema tanto por uma abordagem teórica quanto por uma prática. Ao tratarmos de variáveis contínuas, estamos trabalhando com um número *literalmente* infinito de variáveis e, sob uma perspectiva teórica, isso significaria que todos os candidatos teriam infinitas violações. E, sob uma perspectiva prática, seria difícil conciliar dados empíricos à análise contínua. Para exemplificar, poderíamos pensar que o *input* de nosso quadro teórico seriam intervalos. Se nossos dados coletados tivessem sido, por exemplo, 101, 104.3, 109.1 e 110 Hz, não seria a mesma coisa colocar como *input* o intervalo 101 e 110 Hz que os contém, pois

⁴⁵“Tradicional” pois até mesmo sua nomenclatura é análoga à da literatura de problemas de otimização (função de custo, minimização, etc). Assim, não é um modelo de otimização apenas em substância ou conteúdo, mas em forma também.

isso também incluiria 102.5, 103.9, 105.1 Hz e infinitos outros. Em modelagem contínua, não há intervalo que irá perfeitamente corresponder com os elementos discretos (dados coletados) de uma lista. Sendo assim, como dito anteriormente, devemos repensar o conceito de violação oriundo de abordagens discretas.

Ao lidarmos com variáveis contínuas, portanto, estamos tanto abrindo novas possibilidades de análise quanto nos deparando com barreiras metodológicas antes inexistentes. Assim, como falei, precisamos repensar as violações, pois devemos ser capazes de restringir nossa análise apenas aos dados coletados, mesmo que não possamos selecionar valores discretos para análise de violações. Sendo assim, como nosso quadro teórico deve ser genérico o suficiente para poder ser trabalhado com diferentes *corpora*, devemos refletir sobre o modo como nossos dados devem interagir como as restrições para gerar as violações. Isto é, não basta dizer que “violações são os *outputs* das funções de restrições” como nas análises discretas, pois nossos candidatos são infinitos dentro de qualquer intervalo analisado, significando que, mesmo respeitando o domínio dos dados, nossa análise perderia seu elemento empírico.⁴⁶ Vamos ver no gráfico abaixo exatamente o que quero dizer com isso:

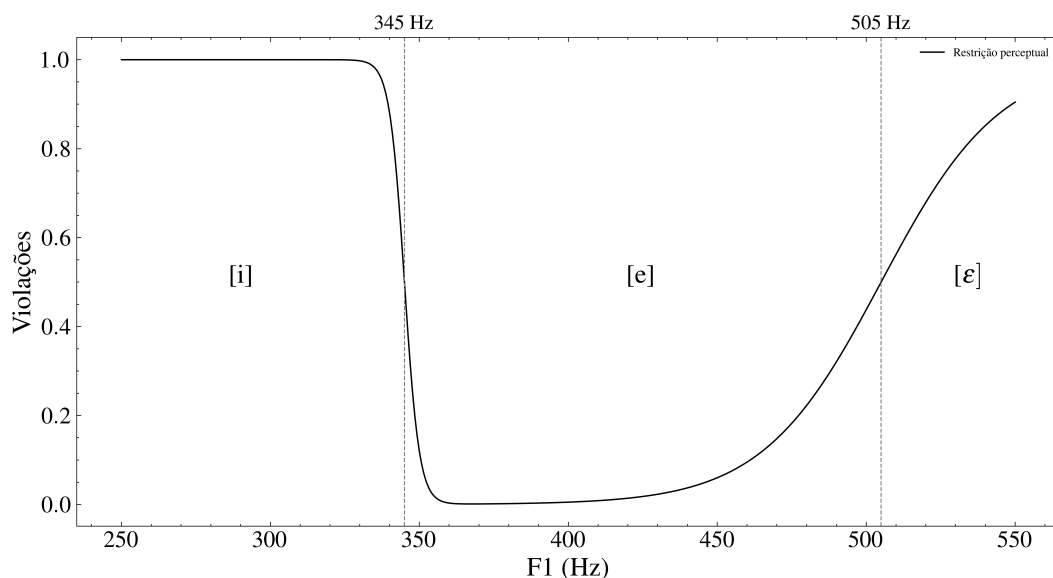


Figura 16: *Output* da restrição perceptual

O gráfico acima representa o *output* da função matemática que descreve a restrição perceptual. Como falei, se considerássemos apenas *output* de função, nossa análise sempre incluiria valores pertencentes ao nosso *corpus* e infinitos outros valores não-pertencentes também. Perceba que ele, de maneira homogênea, considera infinitamente todos os valores entre 250

⁴⁶Pois ela incluiria os dados e infinitos outros valores. Seria mais correto pensar que, na infinitude do candidato contínuo, poderíamos encontrar amostras dos dados coletados.

e 550 Hz. O subir e descer em seu *output* tem a ver com a modelagem fonológica, mas não com a quantidade de amostras coletadas em cada região. Como poderíamos, então, comparar dois grupos de dados diferentes, se nossa restrição não “reage” aos dados? Isso significa que, se quiséssemos modelar dois dialetos distintos, teríamos que recalibrar a restrição para cada dialeto, pois sozinha ela é incapaz de “considerar” ou “julgar” amostras coletadas dos dialetos distintos (a menos que os considere discretamente, como conversamos). Perceba que nós almejamos o contrário: queremos que dialetos distintos sujeitem-se às mesmas restrições *universais* para observarmos a quantidade distinta de violações que eles incorrem por restrição. A universalidade das restrições, afinal, é uma das premissas da OT e HG. Isso significa voltar ao conceito de “violação é um subproduto da interação entre candidato e restrição” e tentar descobrir como se dá essa interação. Perceba que já temos os dois elementos necessários para pensar alguns testes: restrições (R_P, R_A) e candidatos ($\hat{f}_{\text{dados}}$), conforme estabelecido nas seções anteriores deste capítulo. Vamos agora finalmente começar nossa investigação pela seguinte pergunta: “se cada *output* da função da restrição deve estar *atrelado a um candidato real*, qual é a interação em jogo entre eles?”. Em outras palavras, devemos refletir sobre a relação (matemática) entre R_P, R_A e $\hat{f}_{\text{dados}}$.

Anteriormente, ao falar sobre a criação da f.d.p. $\hat{f}_{\text{dados}}$, comentei que ela transportaria informações durante o cômputo de violações. Mais especificamente, transportaria informações sobre a densidade dos dados coletados em diferentes regiões do domínio. Isso significa que, mais corretamente em nossa modelagem, podemos dizer que as violações são dadas pela interação entre funções de restrição e a f.d.p., pois a f.d.p. é o que informa às restrições onde ela deve ou não deve contabilizar violações. Com isso estabelecido, podemos pensar nas quatro operações aritméticas básicas (adição, subtração, divisão e multiplicação) e decidir, em termos matemáticos e fonológicos, qual é a interação entre candidato e restrição em jogo aqui. Vale a pena comentar que não estamos buscando uma interação *categoricamente correta*, pois não acho que devemos pensar estes elementos binariamente como *certos* ou *errados*. Há interações – relações entre restrição e dados – mais ou menos *fonologicamente plausíveis*. Assim, devemos buscar a relação entre partes que melhor parece representar elementos fonológicos conhecidos, deixando aberta a possibilidade de diferentes pesquisas requererem cômputos de violações ajustáveis. No mais, queremos que nossos dados ajam como um contrabalanço às restrições no cômputo de violações. Ou seja, se há concentração de dados (alta densidade) em alguma região, devemos assumir que essa densidade nos conta algo sobre a gramática dos falantes.⁴⁷ Assim, considerando cada *corpus*

⁴⁷Afinal, esse é um dos pontos mais importantes desse quadro teórico proposto: a análise de dados fonéticos como variáveis contínuas, de maneira empírica.

e sua f.d.p. estimada por EDK, violações em regiões de alta densidade devem ser atenuadas. Dito isso, podemos agora começar a investigar algumas relações aritméticas básicas entre candidatos e restrição relevantes para este trabalho:

Adição: $R_i(x) + \hat{f}_{\text{dados}}(x)$. Podemos descartar de antemão uma relação de adição. Caso contrário, estaríamos considerando que valores altos da f.d.p. (regiões de alta densidade de dados coletados) deveriam representar um *aumento* da quantidade de violações. Isso não condiz com a realidade fonológica.

Multiplicação: $R_i(x) \cdot \hat{f}_{\text{dados}}(x)$. A relação de multiplicação também não é ideal. Quando f.d.p. ou restrição aproximam-se de 0, elas também arrastam para 0 o resultado da violação naquela região em razão da multiplicação. Isso significa, em casos como a restrição perceptual que esperamos uma relação complementar entre valores de f.d.p. e restrição, que as violações são demasiadamente estáticas. Vejamos:

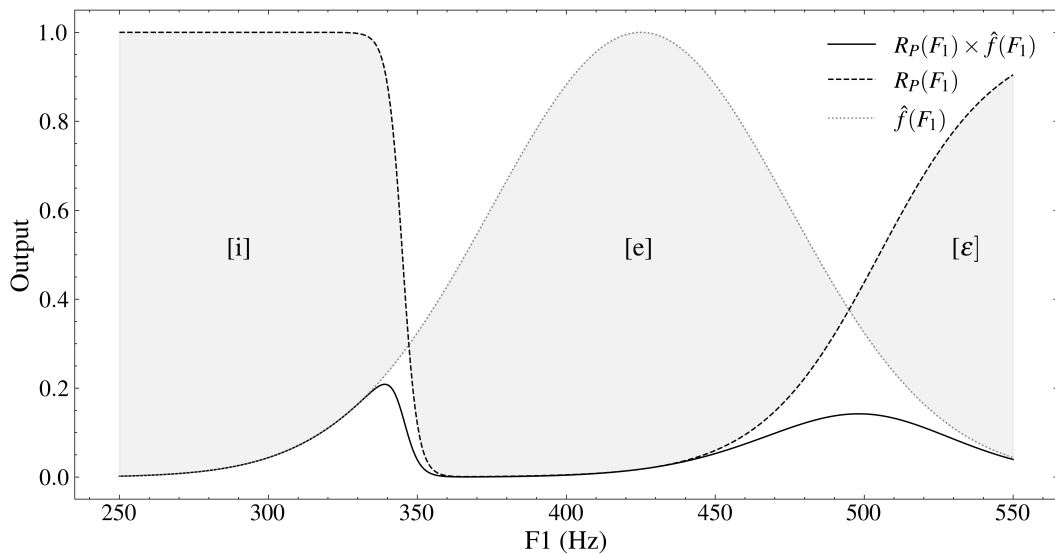


Figura 17: Relação de multiplicação

No gráfico acima e nos subsequentes, plotei a restrição conjuntamente com uma f.d.p. gaussiana imaginária $\hat{f}$ que representa nossos dados nestes exemplos. Assim, temos a curva da restrição R_P , a curva da densidade dos dados $\hat{f}$ e, finalmente, a curva da quantidade de violações. O sombreamento representa a divergência entre restrição e dados, para apoio visual. Perceba que há maiores aumentos das violações nas interseções entre limiares, onde ambas f.d.p. e restrição distanciam-se de 0. Essa é uma relação interessante, pois podemos pensar que estamos analisando exclusivamente a f.d.p. dos dados pela ótica da restrição. No entanto, apesar

de não haver nada intrinsecamente problemático com a relação de multiplicação, ela não atende perfeitamente ao quadro teórico matemático que estamos elaborando. O fato das violações tenderem a 0 em todo o intervalo, salvo em curtos limiares entre restrição e f.d.p., nos dificulta realizar generalizações.⁴⁸ Assim, como futuramente utilizaremos comportamento das violações para gerar uma distribuição “prototípica” de máxima entropia, esse tipo de comportamento não é útil.⁴⁹ Ao trazer quase tudo para 0, ela demasiadamente descaracteriza o comportamento da restrição para modelagem. Podemos, dessa maneira, descartá-la em razão de não ser manipulável.

Subtração: $R_i(x) - \hat{f}_{\text{dados}}(x)$. A relação de subtração é interessante, e aqui começamos a observar as primeiras possibilidades funcionais de *violações* para nossa modelagem. Quando os valores da f.d.p. são altos (*i.e.*, alta densidade de dados coletados), nós proporcionalmente reduzimos os valores de violações naquela região. Conceitualmente, isso faz sentido – com ressalvas. Vejamos o gráfico das violações:

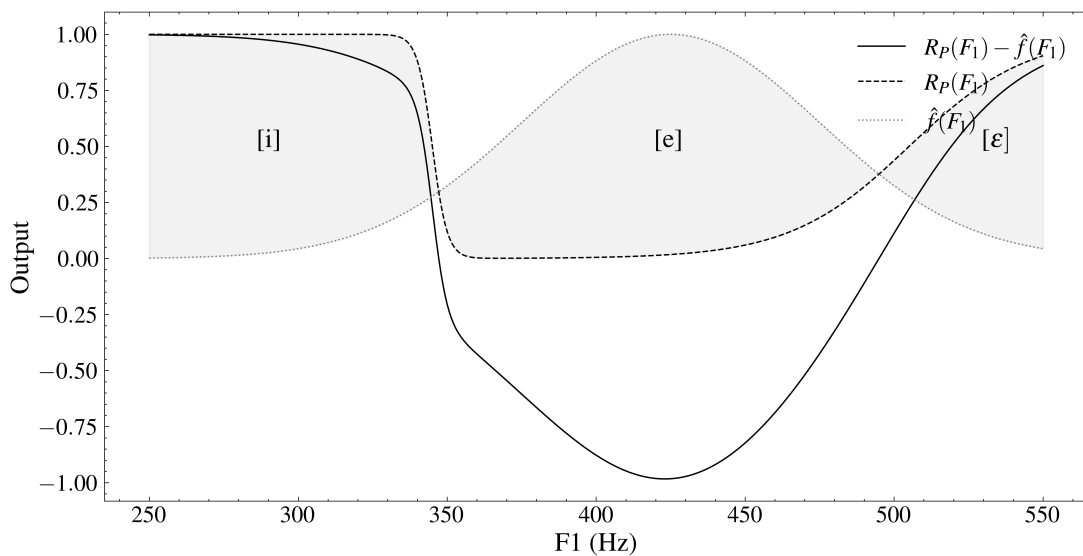


Figura 18: Relação de subtração

Em termos de comportamento, podemos falar que essa relação nos interessa. Onde cresce a frequência dos dados, diminuem as violações. Porém, em termos operacionais, há algumas barreiras para sua adoção. Podemos perceber, logo de antemão, que as violações ficaram

⁴⁸Não é possível realizar generalizações nem para dentro do intervalo dos dados e nem para fora, pois as violações são 0 se ou densidade de probabilidade dos dados ou a restrição são 0.

⁴⁹Digo isso por experiência própria. A relação $R_i(x) \cdot \hat{f}_{\text{dados}}(x)$ foi minha primeira hipótese de violação, pois é uma relação matematicamente intuitiva em que os dados “ponderam” o *output* da restrição, e realizei uma série de experimentos que fracassaram. Não só tive problemas computacionais com *underflow* como também tive que aceitar a realidade que, apesar de fonologicamente plausível, a descaracterização completa da restrição no resultado dessa interação inviabiliza uma distribuição baseada em restrições, dados e violações, pois a maioria da curva das violações torna-se 0.

negativas. Isso deve nos direcionar para o fato de que a subtração pode resultar em números negativos de violações, incompatível com análises otimalistas. Além disso, sabendo que os valores de uma f.d.p. costumam ser baixos, funções de restrições cujo *output* é um valor numérico de alta magnitude acarretam o fato de uma subtração não fazer diferença significativa durante o cômputo de violações. Isso, infelizmente, desqualifica a subtração. Porém, não precisamos ainda abrir mão de seu comportamento, como veremos a seguir.

Divisão: $R_i(x) \div \hat{f}_{\text{dados}}(x)$ ou $\hat{f}_{\text{dados}}(x) \div R_i(x)$. Conceitualmente, a relação de divisão parece promissora. Ao dividir o *output* da função da restrição pela f.d.p., acabamos com um comportamento similar à da subtração. Ou seja, regiões de maior densidade aumentam o valor do denominador e, conseqüentemente, resultam em um valor reduzido de violações. O problema, no entanto, é o seguinte: regiões de baixa densidade da f.d.p., onde não houve muitas ou nenhuma amostra coletada, resultaram em uma divisão por 0. Mesmo se invertêssemos o numerador e o denominador (não é matematicamente relevante qual é restrição e qual é f.d.p.), acabaríamos nessa mesma situação, pois ambos podem apresentar o valor 0. Entretanto, há uma solução relativamente simples para esse problema, que também integra a relação anterior (de subtração). Podemos realizar uma transformação exponencial. Ao invés de diretamente dividir a restrição pela f.d.p., podemos realizar a divisão como se fossem potências da mesma base:

$$v_i(x) = \frac{y^{R_i(x)}}{y^{\hat{f}(x)}}$$

Dessa maneira, aplicamos uma transformação exponencial à divisão original. Isso nos garante algumas propriedades importantes. Primeiramente, garante que nunca tenhamos que realizar uma divisão por 0, pois $y^0 = 1$. Garante também uma maior acentuação das violações fora das regiões de maior densidade da f.d.p. Essa é uma operação que modula o componente determinista (a restrição) com a informação probabilística (a f.d.p. dos dados coletados). Isso significa que, apesar da restrição modelar todo o escopo do intervalo analisado, sua influência no sistema não é constante e é dependente dos dados coletados – que é nossa intenção, pois, sendo assim, diferentes *corpora* irão resultar em diferentes quantidades de violações. Agora, nos resta estabelecer uma base para nossos expoentes. E, de maneira pragmática, sinto que seja apropriado usarmos e como base. A função exponencial e^x apresenta taxa de crescimento proporcional ao seu próprio valor (*i.e.*, a derivada de e^x é e^x), isso reflete uma propriedade de crescimento ou decaimento contínuo e uniforme onde a velocidade de mudança em qualquer ponto é exatamente igual ao seu valor naquele ponto. Essa base, então, captura uma relação direta entre o estado atual do sistema e sua taxa de variação. Isso resulta, finalmente, na seguinte expressão:

$$v_i(x) = \frac{e^{R_i(x)}}{e^{\hat{f}(x)}}$$

Que, pela propriedade de divisão de potências da mesma base, pode ser simplificado da seguinte maneira:

$$v_i(x) = e^{R_i(x) - \hat{f}(x)} \quad (10)$$

Assim, resgatamos o comportamento da relação de subtração por meio da transformação exponencial da relação de divisão, sem agora preocuparmo-nos com violações negativas e com a magnitude dos fatores.⁵⁰ Observemos o gráfico abaixo:

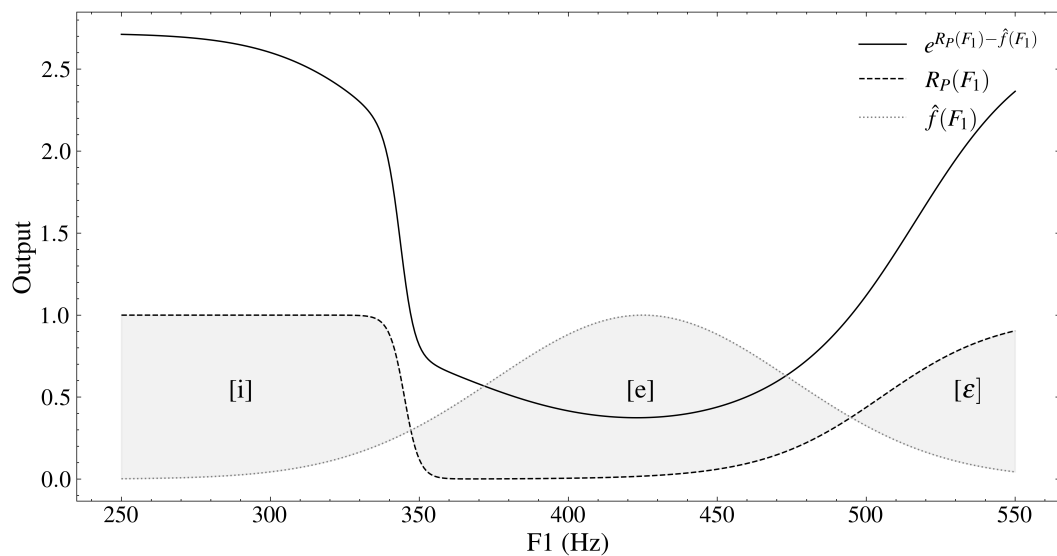


Figura 19: Relação de divisão entre potências da mesma base

O comportamento parece promissor. Conceitualmente, alinha-se com nossas expectativas fonológicas: ele acentua regiões de altas violações, mas atenua regiões de alta densidade de dados. Mas talvez apenas pelo gráfico acima, não sejamos capazes de efetivamente julgar a eficácia dessa definição de violação. As regiões com maiores violações estão relativamente bem separadas da região de alta densidade de dados, a gaussiana $\hat{f}$. Podemos, a fim de avaliar seu comportamento em situações de contraposição, mover a gaussiana (*e.g.*, um *corpus* com F1 mais alto do que a gaussiana do gráfico acima) e manter a mesma restrição para observar a produção de violações enquanto interação dos dados e restrição:

⁵⁰Pois e^x sempre será positivo e acentuará exponencialmente mesmo pequenas subtrações.

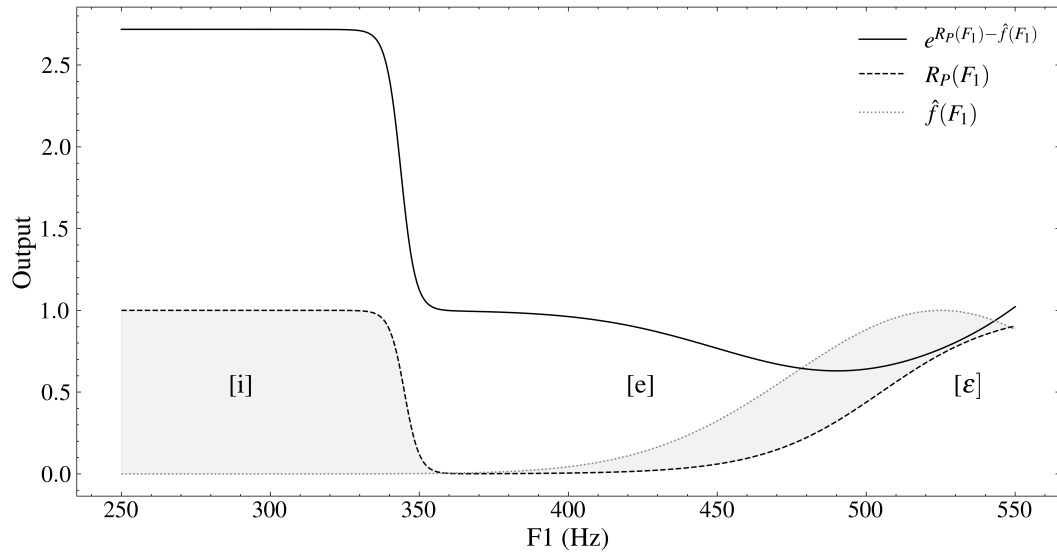


Figura 20: Relação de divisão (gaussiana deslocada)

Ao compararmos este gráfico com o anterior, percebemos que o deslocamento da densidade na f.d.p. realmente representa um deslocamento significativo nas violações. Podemos pensar que o exemplo acima é representativo de uma situação em que usamos parâmetros de [e] na modelagem da restrição perceptual, mas nossos dados coletados demonstraram-se [ɛ]. Poderíamos estar estudando a mudança na produção dessa vogal média pretônica em razão de contato dialetal. Nossos falantes hipotéticos poderiam ser paulistanos migrantes no Nordeste que, ao produzir mais [ɛ] em situação pretônica graças ao contato dialetal, estariam reduzindo a quantidade de violações incorridas naquela região de formantes em sua gramática fonológica. Estaríamos modelando o panorama de violações de suas gramáticas. Mas isso aplica-se a outras situações também, vejamos:

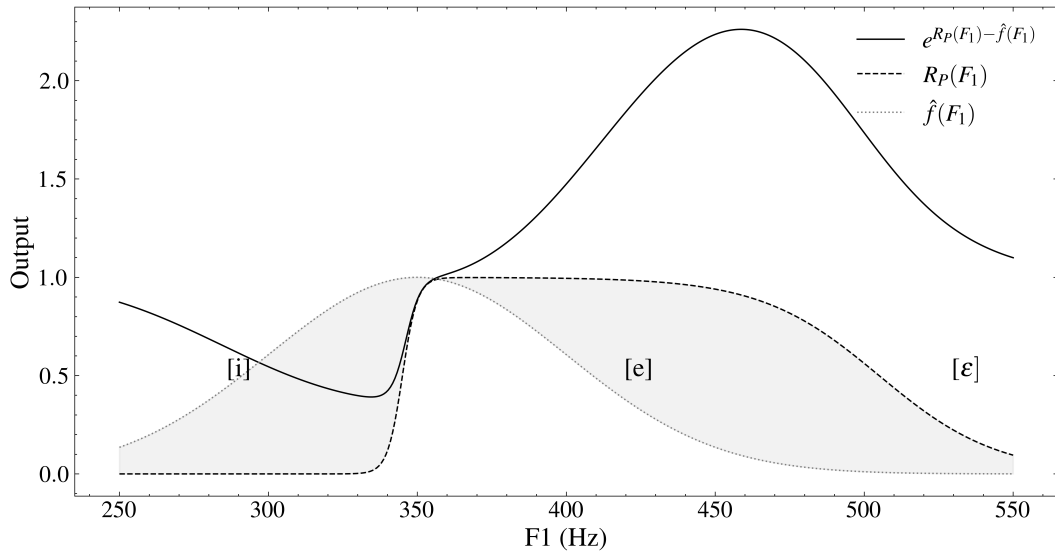


Figura 21: Relação de divisão (restrição invertida)

Acima, a categoria [e] está incorrendo violações enquanto [i] e [ε] não estão (ou, estão incorrendo *menos* violações). É uma inversão da restrição perceptual apresentada até agora. Nesse caso hipotético, poderíamos ter coletado os dados e percebido que sua densidade ficou mais ou menos entre [i] e [e]. Como observado no gráfico, a interação entre candidatos e restrição origina uma curva de violações (dada por $e^{R_i(x) - \hat{f}(x)}$) que acentua regiões de mais violações, mas atenua regiões de maior densidade. Isso significa que as violações não são contadas apenas deterministicamente pela modelagem das restrições, mas pela interação entre o *corpus* (candidatos) e a restrição. Essa relação adéqua a idealização teórica da restrição à realidade empírica dos dados. Estamos, dessa forma, atendendo também à nossa diretriz de que nossos dados ajam como um contrabalanço às restrições no cômputo de violações. No entanto, antes de seguirmos em frente, gostaria de repetir que a seleção dessa relação específica não configura uma afirmação universal. É possível pensar que, em diferentes pesquisas com diferentes necessidades, poderíamos chegar em diferentes expressões matemáticas. Assim, apenas investiguei algumas – dentre infinitas – possibilidades fonologicamente plausíveis de violações.

Após essa pequena investigação de relações entre *corpus*/candidatos e restrição, pode-se dizer que encontramos uma definição de nosso interesse ($e^{R_i(x) - \hat{f}(x)}$). Sendo assim, finalmente, gostaria de apresentar a definição formal de quantidade de violações de nosso quadro teórico:

$$v_i = \int_a^b e^{R_i(x) - \hat{f}(x)} dx \quad (11)$$

Em outras palavras, a quantidade de violações de determinada região/intervalo (can-

didato analisado) é o acúmulo de violações incididas entre seus limites. Para calcular esse acúmulo, podemos realizar uma integração⁵¹ e selecionar limites inferior e superior que correspondam com os limites do intervalo analisado. A expressão acima é uma fórmula genérica, para que possamos adequá-la à realidade de cada análise. Poderíamos estar lidando com uma restrição de duas variáveis, F1 e F2, por exemplo:

$$v_i = \int_{a_{F2}}^{b_{F2}} \int_{a_{F1}}^{b_{F1}} e^{R_i(F1, F2) - \hat{f}(F1, F2)} dF1 dF2 \quad (12)$$

Assim, em virtude de seu caráter genérico, vale a pena realizarmos uma definição formal que norteie sua aplicação:

Definição:

As **violações** originam-se da interação entre **candidato e restrição**. Em ambiente de variáveis contínuas, podemos pensar que **a quantidade de violações de um intervalo analisado é a integral** de $e^{R_i(x) - \hat{f}(x)}$ sob os **mesmos limites do intervalo analisado**.

Dessa forma, a fim de demonstrar seu uso nas restrições concebidas nesse protótipo, vale a pena apresentar o cômputo das violações de nossas restrições.

3.5.1 Cômputo das violações da restrição perceptual

Seguindo nossa definição, devemos utilizar a seguinte expressão para calcular as violações de nossa restrição perceptual:

$$v_P = \int_{a_{F1}}^{b_{F1}} e^{R_P(F1) - \hat{f}(F1)} dF1 \quad (13)$$

O problema é que a distribuição de nossos dados está configurada em uma f.d.p. conjunta de duas variáveis $\hat{f}(F1, F2)$, e a restrição perceptual só deve fazer uso de uma variável, F1. Uma saída para este problema é a *marginalização* da f.d.p., onde integramos (ou, “marginalizamos”) a variável que não precisamos. Em nosso caso, poderíamos pensar da seguinte maneira:

$$\hat{f}_{\text{marginal}}(F1) = \int_{-\infty}^{\infty} \hat{f}_{\text{dados}}(F1, F2) dF2$$

O que resultaria na seguinte equação para violações:

⁵¹No contexto de uma integração definida (que poderíamos pensar como um somatório contínuo), quando integramos uma função entre dois limites a e b , estamos calculando o *valor acumulado* da função, que é muitas vezes representado como a área sob a curva da função entre os limites da integral.

$$v_P = \int_{a_{F1}}^{b_{F1}} e^{R_P(F1) - \hat{f}_{\text{marginal}}(F1)} dF1 \quad (14)$$

Para realizar nosso primeiro cálculo de violações, é importante definir também como trabalharemos com R_P daqui em diante. Assim, penso que seja interessante usar os seguintes parâmetros:

$$R_P(F1) = 1 - \frac{1^2}{(1 + e^{1(F1-600)})(1 + e^{-7(F1-345)})} \quad (15)$$

O teto $L = 1$ é por convenção e simplicidade. Os valores de $k_{1,2} = 1, -7$ foram selecionados arbitrariamente, emulando as relações hierárquicas de contraste entre [e], [ɛ] e [i].⁵² Os limiares 600 Hz e 345 Hz, que representam os pontos de inflexão durante a saturação da percepção, marcam pontos de transição entre a percepção de categorias fonológicas. Podemos agora observar como fica o cômputo de violações da restrição perceptual relativa à f.d.p. marginalizada $\hat{f}_{\text{marginal}}(F1)$ dos dados de minha monografia (sombreei a distribuição dos dados para apoio visual):

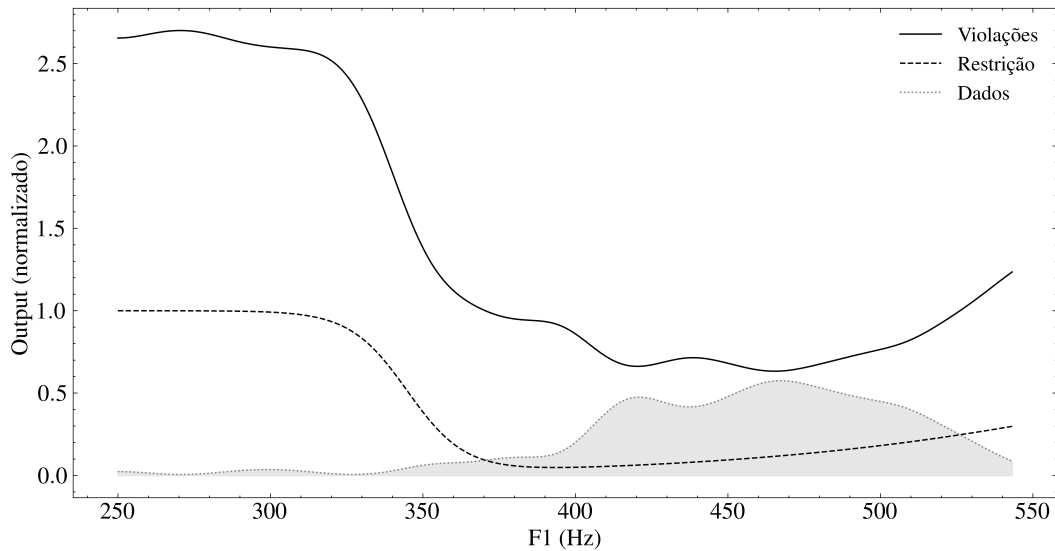


Figura 22: Violações da restrição perceptual

Como nossos dados, neste caso, não estão em aparente relação de contraposição com a restrição, podemos dizer que as violações seguem a restrição. Onde a restrição assinala poucas violações há maior densidade de dados, e onde a restrição assinala mais violações há menor densidade. Podemos observar o comportamento das violações acima e nos perguntar que tipo de relação queremos que nossas amostras tenham em relação à modelagem das restrições.

⁵²Conforme explicitado anteriormente ao discutir a restrição perceptual, seguindo a hierarquia contrastiva dos traços das vogais do PB apresentada por Lee (2010, p. 37).

Queremos que elas, como acima, complementem-se? Seria isso sinal de uma restrição bem formulada para determinado objeto de estudo? Por outro lado, será que não seria melhor modelar uma relação de oposição para que, ao jogarmos restrições *contra* dados, capturássemos alguma sutileza ou relação de alteridade entre categorias? Esses tipos de decisões não são triviais. Dessa maneira, o gráfico acima pode ser interpretado como uma modelagem demonstrativa, não sendo particularmente crítica ou, por assim dizer, “café com leite”. Dito isso, podemos realizar alguns cálculos⁵³ e tabelá-los abaixo:

F1 (Hz)	Violações	%
250 até 543	6.142	99.9
250 até 350	3.701	60.2
350 até 450	1.306	21.2
450 até 543	1.134	18.4

Tabela 12: Violações da restrição perceptual.

Podemos pensar cada faixa acima como um candidato, como discuti anteriormente na seção sobre candidatos. Coloquei uma última coluna à direita das violações intitulada “%” que nos indica, em relação a todas as violações dos dados, qual porcentagem de violações o candidato detém. Isso é importante para sabermos não apenas um valor numérico descontextualizado, mas um valor que nos diz a concentração de violações do candidato. Assim, sabemos que a faixa 250 Hz até 350 Hz percorre cerca de um terço de nossos limites inferior e superior, mas agrega mais da metade (60%) de todas as violações possíveis entre o domínio dos dados coletados. Isso pode ser observado também no gráfico acima, pois corresponde ao espectro perceptual de [i]. Mas esse tipo de análise é apenas uma confirmação do que podemos tranquilamente inferir visualmente e conceitualmente. Vejamos agora um exemplo menos óbvio e mais granular:

F1 (Hz)	Violações	%
440 até 450	0.105	1.7
440 até 445	0.053	0.8
445 até 450	0.052	0.8
449.8 até 449.9	0.001	0.01

Tabela 13: Violações da restrição perceptual.

⁵³Que serão realizados computacionalmente, conforme apresentado no capítulo seguinte.

Dada a natureza contínua de nossa análise, podemos selecionar como candidatos qualquer intervalo de formantes⁵⁴. Isso significa que podemos computar violações que vão desde candidatos estreitos (candidatos como 440 Hz até 445 Hz) até maiores extensões (de 250 Hz até 500 Hz). Sob a perspectiva de uma análise contínua, o pesquisador tem uma maior autonomia para selecionar e analisar diferentes candidatos, deixá-los ou não sobrepostos, faixas maiores ou menores, e assim em diante sem a necessidade de passar por um processo de discretização *a priori*. Além disso, as violações são incorridas de maneira *gradiente* ao longo do candidato, ao invés de maneira homogênea em cima de intervalos discretizantes, de maneira facilmente observável e manipulável.

3.5.2 Cômputo das violações da restrição articulatória

De maneira similar, podemos definir o cômputo das violações da restrição articulatória da seguinte maneira:

$$v_A = \int_{a_{F2}}^{b_{F2}} \int_{a_{F1}}^{b_{F1}} e^{R_A(F1, F2) - \hat{f}_{\text{dados}}(F1, F2)} dF1 dF2 \quad (16)$$

Como a restrição articulatória tem como variáveis F1 e F2, não há necessidade de marginalizar ou realizar qualquer outra operação na f.d.p., resultando na expressão acima. Para nossa demonstração nesse protótipo, usarei os seguintes parâmetros para nossa restrição articulatória:

$$R_A(F1, F2) = \sqrt{(C_{F1} - 421)^2 + (C_{F2} - 1887)^2} \cdot \frac{\sqrt{(C_{F1} - 610)^2 + (C_{F2} - 1900)^2}}{\sqrt{35890}} \quad (17)$$

Esses são os mesmos parâmetros discutidos durante sua elaboração, sem mudanças. Temos uma multiplicação entre as coordenadas F1, F2 e a coordenada alvo 412, 1887 e temos a posição neutra da língua como sendo a coordenada 610, 1900. Vejamos abaixo seu funcionamento atual:

⁵⁴Dentro do domínio (limite) dos dados e respeitando, como veremos ao fim deste capítulo, um tamanho de intervalo mínimo.

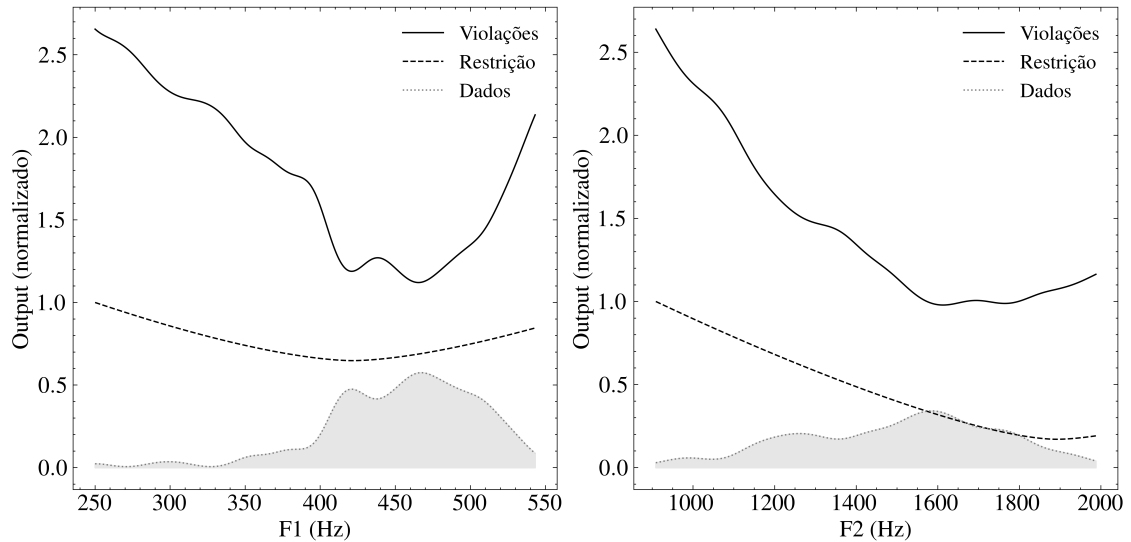


Figura 23: Violações da restrição articulatória

O comportamento das violações é o que espera-se de uma restrição que, na prática, computa distâncias entre formantes. Quanto mais distante do alvo, maior o número de violações incorridas. Podemos demonstrar esse comportamento ao calcular as seguintes violações:

F1 (Hz)	F2 (Hz)	Violações	%
250 até 543	908 até 1987	1094.795	100
250 até 350	908 até 1300	570.273	52.2
350 até 450	1300 até 1700	24.235	2.2
450 até 543	1700 até 1987	5.620	0.5

Tabela 14: Violações da restrição articulatória.

Os valores acima são esperados, pois estão em conformidade com o gráfico e os aspectos conceituais dessa restrição. Mas há um aspecto que podemos explorar na restrição articulatória que não estava presente na restrição perceptual que, como podemos ver acima, é o fato de analisarmos duas dimensões simultaneamente. Veja que não somamos 100% de nossas violações ao dividir em 3 seções o domínio total dos dados – apenas quando o analisamos por inteiro. Isso se dá pois, ao calcular violações dessa maneira, estamos levando em consideração as violações conjuntamente entre dimensões e não separadamente. Isso é de interesse para nós, pois abre a possibilidade de realizar análises combinatórias. Podemos parear F1 com diferentes F2 (ou vice versa) e checar como isso aumenta ou diminui violações da configuração articulatória. Vejamos:

F1 (Hz)	F2 (Hz)	Violações	%
250 até 350	908 até 1300	570.273	52.2
250 até 350	1300 até 1700	130.568	11.9
250 até 350	1700 até 1987	80.995	7.4
250 até 350	1700 até 1987	80.995	7.4
350 até 450	1700 até 1987	13.651	1.2
450 até 543	1700 até 1987	5.620	0.5
400 até 400.9	908 até 1987	1.726	0.1
500 até 500.9	1900 até 1900.9	0.0001	0.00001

Tabela 15: Violações da restrição articulatória.

A tabela acima apresenta 8 candidatos – 8 intervalos de F1,F2 – e foi dividida em três seções. A primeira e a segunda seções demonstram que podemos realizar análises mantendo o intervalo de uma das variáveis fixo enquanto o combinamos com diferentes intervalos da outra. A terceira seção demonstra que também podemos trabalhar com maior granularidade dos dados. No último candidato da tabela, temos efetivamente apenas quantidade de violações incorridas em cerca de 500, 1900 Hz. Isso demonstra que, em termos de análise e metodologia, esse é um quadro teórico versátil, pois os candidatos – apesar de contínuos – podem flexibilizar-se em recortes maiores ou menores.

Sinto que nossa discussão sobre violações foi o primeiro passo para verdadeiramente realizar a composição de uma metodologia contínua. Durante a definição de nossas restrições, ainda estávamos presos a demonstrações e exemplos discretos que, nesta seção sobre violações, foram descartados e devidamente repensados. Agora, devemos aproveitar este ímpeto e começar nossa discussão sobre o *escore harmônico* e reimaginá-lo de maneira fonologicamente bem fundamentada.

3.6 Harmonia

Gostaria de separar essa seção em duas subseções: *escore harmônico* e *estabilidade*. Na subseção *escore harmônico*, pretendo brevemente trabalhar com a definição tradicional de escore harmônico como soma de violações multiplicadas por seus pesos. Como não é necessária nenhuma reimaginação ou inovação, não é preciso escrever longamente sobre este processo que, por si só, já é um processo simples de soma. Na subseção *estabilidade*, por outro lado, desejo introduzir uma segunda referência de harmonia para candidatos contínuos, que torna-se possível

apenas em um quadro teórico completamente contínuo. Essa segunda referência, que apelidei de *estabilidade*, pode ser observada junto do escore harmônico ao realizarmos juízos em nossa pesquisa sobre fenômenos fonológicos, a fim de termos uma visão mais holística dos candidatos.

3.6.1 Escore harmônico

Diferentemente do conceito um tanto difuso e “implícito” de violações, o conceito de escore harmônico é explicitamente definido ao longo da literatura fonológica. Isso inclui também definições matemáticas, o que nos economiza tempo. Vejamos:

$$\mathcal{H}(x) = \sum_{i=1}^n p_i R_i(x) \quad (18)$$

A expressão acima, encontrada sem muitas variações ao longo da literatura (cf. GOLDWATER; JOHNSON, 2003; HAYES; WILSON, 2007; POTTS *et al*, 2010), determina como calcula-se o escore harmônico em um quadro teórico de HG. Em outras palavras, o escore harmônico $\mathcal{H}(x)$ de um candidato x pode ser definido como a soma $\sum_{i=1}^n$ de suas violações $R_i(x)$ multiplicadas por seus devidos pesos p_i , que resulta na expressão $\mathcal{H}(x) = \sum_{i=1}^n p_i R_i(x)$ comumente apresentada em discussões sobre HG. Essa é uma definição que também nos é inteiramente adequada, mas requer modificações. No âmbito deste protótipo, a expressão acima poderia ser traduzida para a seguinte equação:

$$\mathcal{H}(x) = \sum_{i=1}^n p_i v_i(x) \quad (19)$$

A expressão acima, por sua vez, pode ser escrita da seguinte maneira a fim de melhor explicitar todos os seus elementos:

$$\mathcal{H}(x) = \sum_{i=1}^n p_i \cdot e^{R_i(x) - \hat{f}(x)} \quad (20)$$

Conforme discutimos na seção anterior, nossas violações não se dão apenas por $R(x)$ (pelo *output* da função da restrição) e sim por $e^{R(x) - \hat{f}(x)}$. Essa é a única alteração necessária. Vejamos abaixo como essa expressão ocorre neste protótipo:

$$\mathcal{H}(F1, F2) = p_P \cdot e^{R_P(F1) - \hat{f}_{\text{marginal}}(F1)} + p_A \cdot e^{R_A(F1, F2) - \hat{f}_{\text{dados}}(F1, F2)}$$

Como ainda não discutimos os pesos p_P, p_A de nossas restrições, podemos considerá-los como tendo valor igual a 1. Para alcançar os resultados dos candidatos, similarmente às

violações, devemos integrar o escore harmônico, da seguinte maneira:

$$\mathcal{H} = \int_{a_{F2}}^{b_{F2}} \int_{a_{F1}}^{b_{F1}} p_P \cdot e^{R_P(F1) - \hat{f}_{\text{marginal}}(F1)} + p_A \cdot e^{R_A(F1, F2) - \hat{f}_{\text{dados}}(F1, F2)} dF1 dF2$$

Finalmente, ao calcularmos essa expressão para nossos candidatos, podemos obter os seguintes resultados:

F1 (Hz)	F2 (Hz)	$\mathcal{H}$
250 até 543	908 até 1987	1100.938
250 até 350	908 até 1300	573.975
350 até 450	1300 até 1700	25.542
450 até 543	1700 até 1987	6.755

Tabela 16: Escore harmônico dos candidatos.

Note que esses mesmos valores poderiam ser obtidos se somássemos manualmente os valores discutidos na seção anterior. Veja:

F1 (Hz)	F2 (Hz)	v_P	v_A	$v_P + v_A = \mathcal{H}$
250 até 543	908 até 1987	6.142	1094.795	1100.938
250 até 350	908 até 1300	3.701	570.273	573.975
350 até 450	1300 até 1700	1.306	24.235	25.542
450 até 543	1700 até 1987	1.134	5.620	6.755

Tabela 17: Escore harmônico como soma de violações.

As violações, em nossa implementação computacional, serão normalizadas entre restrições e ponderadas por seus pesos, fazendo com que essa grande disparidade de contribuição para o escore total se reduza. Para nossos fins, dessa maneira, basta usar a definição existente de escore harmônico. Futuramente, na seção seguinte, iremos fazer uso dessa definição de harmonia na maximização da entropia diferencial.

Definição:

O **escore harmônico** de um candidato é a **soma** de suas **violações** multiplicadas por seus respectivos **pesos** de suas restrições.

Gostaria, antes de entrarmos na próxima seção e discutirmos otimização, de fazer um pequeno *detour* em nossa rota. Desde o começo deste capítulo, viemos reimaginando elementos da OT e HG para trabalhar com candidatos contínuos. Nos próximos parágrafos que compõem a subseção seguinte, gostaria de falar não de algo existente a ser transformado, mas de como podemos fazer uso criativo das propriedades de variáveis contínuas. A seguir, pretendo apresentar um conceito que apelidei de *estabilidade* que surge no âmbito contínuo e que pode gerar informações interessantes para nossa pesquisa.

3.6.2 Estabilidade

A noção de “estabilidade” é minha tentativa experimental de apropriar para a Fonologia um conceito já amplamente usado e conhecido em outras áreas, o uso de gradientes. O gradiente de uma função de múltiplas variáveis é um vetor que indica taxa de mudança e direção do aumento e diminuição do *output* da função. Ele é composto pelas derivadas parciais da função em relação a cada variável. Desse modo, ao observarmos o gradiente da função do escore harmônico, podemos encontrar *plateaus* de violações locais ou globais e podemos nos nortear ao saber em qual direção aumentam ou diminuem as violações de um dado candidato. A maneira específica como trabalho “estabilidade” nesta subseção, no entanto, não é uma fórmula ou abordagem que encontrei na literatura matemática – apesar de, por eu não ser muito versado nela, é possível que algo tenha me escapado. Isso se dá pois, de maneira geral, funções matemáticas podem conter variáveis de diferentes unidades e magnitudes e, ao somarmos as derivadas parciais do gradiente (como farei), estamos misturando elementos que, teoricamente, podem não ser misturáveis em primeiro lugar. No entanto, as variáveis do escore harmônico estão todas intermediadas pela relação $e^{R_i(x) - \hat{f}(x)}$ e estruturadas em termos de violações. Isso me leva a crer que, pelo menos *a priori*, a soma das derivadas pode nos oferecer informações interessantes, pois estaremos sempre realizando essa operação no escore harmônico que é sempre uma função de violações, ou seja, da mesma unidade e na mesma magnitude⁵⁵. Assim, é necessário realizar rigorosamente alguns testes de validação para observarmos se os resultados da estabilidade alinham-se com as expectativas fonológicas, *i.e.*, se o conceito captura adequadamente fenômenos linguísticos. Para realizar uma demonstração, podemos começar retomando a função de escore harmônico:

⁵⁵É discutível o fato de ser ou não da mesma magnitude, pois envolveria conhecer os pesos de cada restrição e de julgar o que é (ou deixa de ser) uma mesma magnitude em termos numéricos para violações. No entanto, de maneira geral e com as devidas ressalvas, não acho que seja errado dizer que é da mesma magnitude.

$$\mathcal{H}(x) = \sum_{i=1}^n p_i v_i(x)$$

Que especificamente em nosso protótipo se dá como:

$$\mathcal{H}(F1, F2) = p_P \cdot v_P(F1) + p_A \cdot v_A(F1, F2)$$

E que podemos agora explicitar seu gradiente, o vetor composto por suas derivadas parciais:

$$\nabla \mathcal{H} = \left(\frac{\partial \mathcal{H}}{\partial F1}, \frac{\partial \mathcal{H}}{\partial F2} \right)$$

Como queremos saber a influência conjunta das derivadas parciais, que nos dão a taxa e direção da mudança, irei somá-las e dividi-las por sua extensão.⁵⁶ O valor resultante é, discutivelmente, uma medida rudimentar da taxa e direção de mudança do escore harmônico naquele ponto. Como nossos candidatos são contínuos, devemos então realizar a integração dessa soma sob o intervalo do candidato, para que tenhamos em mãos *a média do valor acumulado da taxa e direção da mudança no escore harmônico*. Podemos então definir a medida de estabilidade do escore harmônico pela seguinte fórmula genérica para um número i de variáveis analisadas:

$$\mathcal{E} = \left(\prod_{i=1}^n (b_i - a_i) \right)^{-1} \cdot \int_{a_i}^{b_i} \left(\sum_{i=1}^n \frac{\partial \mathcal{H}}{\partial x_i} \right) dx_i \quad (21)$$

Utilizei a letra $\mathcal{E}$ para denotar “estabilidade”. A equação acima nos informa o equivalente a uma estabilidade ou instabilidade das violações de um candidato, pois indica uma média de *quanto* variam as violações do candidato em seu intervalo e se seu comportamento tende a um *aumento* ou *diminuição* das violações. Ou seja, se sua tendência é o aumento ou diminuição do escore harmônico e se esse aumento ou diminuição está ocorrendo em maior ou menor escala. É por isso que apelidei esse conceito, definido pela fórmula acima, de “estabilidade”. Em nosso caso, conforme as definições elencadas neste capítulo, a expressão não-genérica seria a seguinte:

$$\mathcal{E} = ((b_{F1} - a_{F1})(b_{F2} - a_{F2}))^{-1} \cdot \left(\int_{a_{F2}}^{b_{F2}} \int_{a_{F1}}^{b_{F1}} \frac{\partial \mathcal{H}}{\partial F1} + \frac{\partial \mathcal{H}}{\partial F2} dF1 dF2 \right)$$

E que podemos multiplicar seus termos para compreender melhor a lógica por trás

⁵⁶Comprimento (se unidimensional, uma variável), área (se bidimensional, duas variáveis) ou volume (se tridimensional ou mais, três ou mais variáveis).

de sua definição genérica:

$$\mathcal{E} = \frac{\int_{a_{F2}}^{b_{F2}} \int_{a_{F1}}^{b_{F1}} \frac{\partial \mathcal{H}}{\partial F1} + \frac{\partial \mathcal{H}}{\partial F2} dF1 dF2}{(b_{F1} - a_{F1})(b_{F2} - a_{F2})} = \frac{\text{Quantidade de mudança}}{\text{Área analisada}}$$

Isso é um exemplo de como podemos obter novos tipos de informação sobre dados e candidatos em uma abordagem contínua, requerendo criatividade e um trabalho matemático mais artesanal. Agora, com essa definição formulada, podemos tabelar os resultados dessa medida de estabilidade para cada candidato. Mas, antes, algumas ressalvas. Primeiramente, essa não é uma medida convencional em termos matemáticos e a emprego aqui de maneira experimental e específica para nossa modelagem, sabendo que requer testagem e considerações cuidadosas em seu uso. Além disso, em relação a seu funcionamento, o sinal do valor resultante tem a ver com a direção da integração. Um sinal positivo indica que, de a para b na integração, a quantidade de violações têm aumentado e, em contrapartida, um sinal negativo indica que elas têm decrescido. É assim que podemos aferir em que direção⁵⁷ diminuem as violações, além de saber a quantidade de mudança da região analisada (que é a própria grandeza escalar, *i.e.*, o valor absoluto). Além disso, coloquei como N/A a estabilidade do “candidato” (com fortes aspas) que inclui todo o domínio dos dados. Como ele não é realmente um candidato na prática, pois inclui todos os valores do domínio e não se contrapõe a nenhum outro candidato (já que inclui tudo), não acho que faça sentido que tenha uma métrica de estabilidade. Sempre o coloquei na tabela por questão de referência, mas aqui não se aplica. Observe:

F1 (Hz)	F2 (Hz)	$\mathcal{H}$	$\mathcal{E}$
250 até 543	908 até 1987	1100.938	N/A
250 até 350	908 até 1300	573.975	-351.426
350 até 450	1300 até 1700	25.542	-13.233
450 até 543	1700 até 1987	6.755	+2.853

Tabela 18: Escore harmônico $\mathcal{H}$ e estabilidade $\mathcal{E}$.

Ignorando o “candidato” de referência e focando-se apenas no três abaixo, podemos dizer que há uma tendência de diminuição de violações entre 250,450 Hz e 908,1700 Hz. Por

⁵⁷“Direção” em sentido amplo e não formal, pois o resultado dessa equação é uma grandeza escalar e não um vetor. Assim, ele não apresenta a *direcionalidade* de um vetor, contendo apenas uma indicação relativa a se o valor do escore harmônico está aumentando ou diminuindo ao longo do domínio integrado.

outro lado, de 450 Hz e 1700 Hz para cima temos uma tendência ao aumento de violações, apesar de relativamente pequena. Disso, pode-se inferir que a faixa ótima de produção do /e/ pretônico⁵⁸ está dentro do último candidato. Sabemos que, nele, há algum *plateau* em que violações, após reduzirem e alcançarem uma produção mais harmônica, voltam a subir. Isso está alinhado com nossas restrições, suas violações resultantes e com a realidade empírica dos dados. Vejamos o gráfico abaixo:

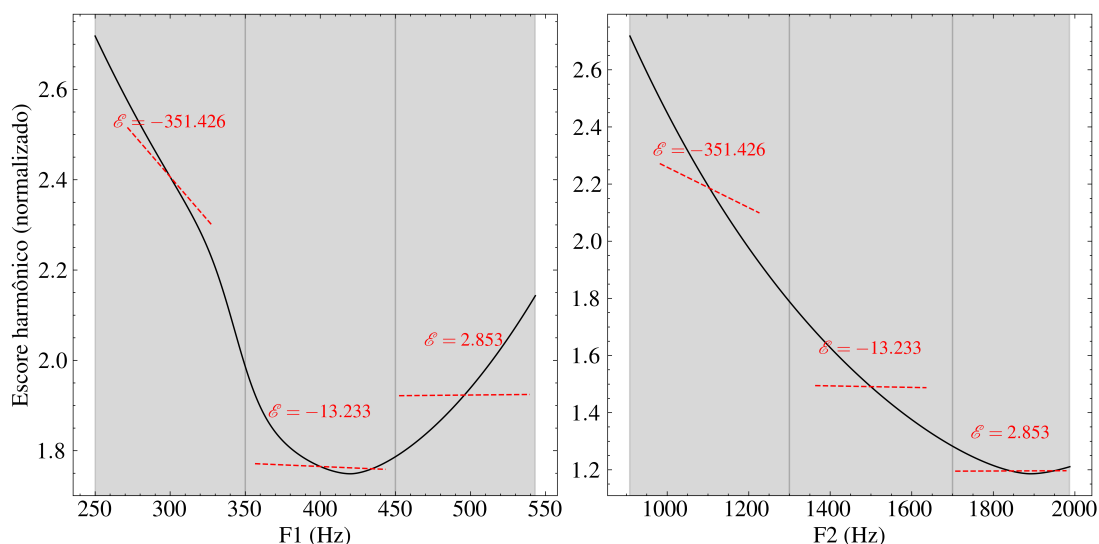


Figura 24: Representação da estabilidade dos três candidatos como retas.

O gráfico acima é uma representação bidimensional de um gráfico originalmente tridimensional, para fins de clareza. A inclinação das retas pontilhadas vermelhas é o valor da estabilidade $\mathcal{E}$ convertido em um ângulo em radianos usando a função arco-tangente para fins de visualização.⁵⁹ Podemos, dessa maneira, observar visualmente a *tendência* das violações. Pelos valores absolutos, podemos também perceber que o primeiro candidato apresenta grande taxa de mudança, rapidamente diminuindo suas violações conforme sobem as frequências de F1 e F2. O segundo candidato, por sua vez, apresenta uma curva do escore harmônico que aparenta ser menos íngreme, pois sua taxa de mudança é muito menor, indicando que, apesar de suas violações diminuírem ao aumentar de frequência, é uma mudança significativamente menos brusca. O terceiro candidato, enfim, nos diz que continuar aumentando F1 e F2 significa incorrer mais violações. A taxa ou velocidade de mudança nas violações do terceiro candidato é

⁵⁸Lembrando que estamos lidando com os dados da alternância [e] ~ [ɛ] em posição pretônica em dialetos nordestinos no PB.

⁵⁹A possível disparidade visual entre ângulos referentes à mesma reta nos gráficos de F1 e F2 se dá pelo fato de, originalmente, isso ser um gráfico tridimensional único. Para fazer essa separação, os ângulos das retas foram corrigidos pela escala dos eixos F1 e F2 (para fins de visualização), o que pode causar uma alteração visual – mas os valores foram calculados de forma idêntica.

a menor de todas e isso, conjuntamente com a mudança do sinal negativo para positivo, indica que alcançamos um valor mínimo para o escore harmônico ali dentro e agora, ao alcançar frequências mais altas, só nos distanciaremos da produção ótima. Podemos agora, com alguma confiança, lançar mão da seguinte definição:

Definição:

A **estabilidade** mede a **direção e a taxa de mudança** de um **candidato**. Ela é dada pela **média da integração da soma das derivadas parciais da função do escore harmônico** sobre a extensão do **candidato**.

Considerando que nossos candidatos contínuos podem ser reduzidos, ampliados e recortados durante a análise – e essa é sua maior diferença com candidatos que foram *discretizados* aprioristicamente – sinto que essa noção de estabilidade pode ser muito útil como ferramenta de busca por *plateaus* do escore harmônico. Ela deve, é claro, ser sempre utilizada conjuntamente com o próprio escore harmônico, pois não distingue entre um *plateau* de muitas violações e um *plateau* de baixas violações, daí que apenas denota *estabilidade* de uma região. Além disso, mesmo em caráter complementar aos elementos tradicionais de análise, é preciso rigorosamente testá-la, pois a expressão matemática da estabilidade é, de certa maneira, não convencional.

Agora que discutimos os elementos básicos da HG (candidatos, restrições, violações e escore harmônico), restam apenas os elementos que dependem de algoritmos de otimização. Isto é, pesos de restrições e probabilidades. Sendo assim, a seção seguinte tratará das otimizações necessárias em nosso quadro teórico.

3.7 Otimização

Antes de procedermos com explicações técnicas, gostaria de esboçar um panorama geral da *otimização* em nosso projeto. Sendo a otimização um processo iterativo e recursivo, ela inescapavelmente sofre com algum grau de lógica circular em que o *produto final do algoritmo* deve ser considerado *durante sua calibragem* que é procedimentalmente anterior a ele. Assim, falar sobre “pesos de restrições” é antes falar sobre nosso produto final e seu formato, visto que os pesos são ajustados em virtude desse produto final. O problema, é claro, é que o produto final depende ele mesmo dos pesos das restrições – como falei, há uma certa lógica circular em jogo aqui. Assim, a fim de tornar tudo isso um pouco menos confuso, faço questão de usar um apoio visual:

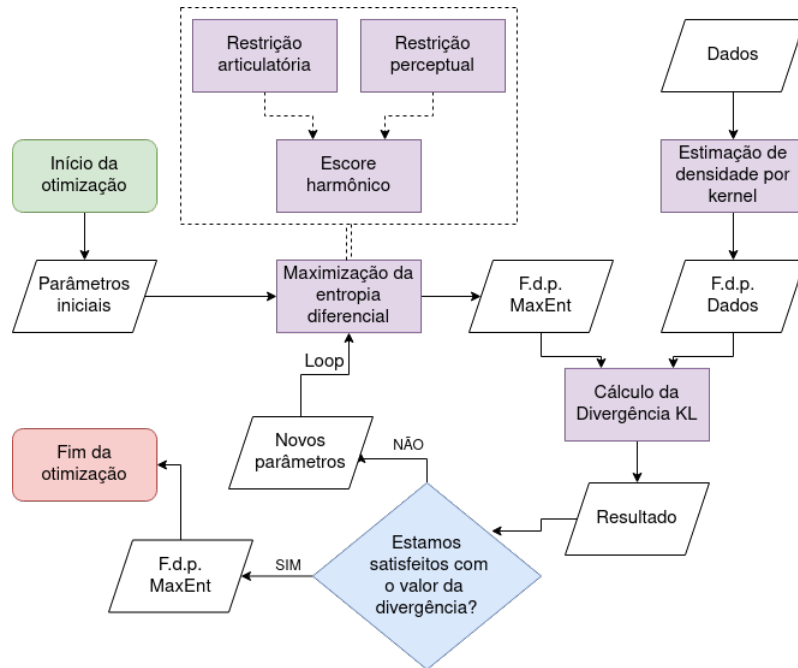


Figura 25: Fluxograma da otimização.

Seguindo o fluxograma acima, nossa otimização começa pela maximização da entropia diferencial para gerar uma f.d.p. que devemos comparar com a f.d.p. estimada dos dados (vide primeira seção deste capítulo). Essa comparação é feita pelo cálculo da divergência entre elas, e serão então ajustados os pesos na tentativa de minimizar essa divergência. Após ajuste, começa outra iteração de maximização de entropia agora com os novos pesos, onde realizaremos uma nova comparação, obteremos uma nova divergência, reajustaremos os pesos e assim em diante quantas vezes forem necessárias. Sendo assim, para viabilizar uma explicação pouco confusa e o menos repetitiva possível, discutiremos o fluxograma em partes menores que serão apresentadas por subseção, conforme indicarei no início de cada uma.

3.7.1 Maximização da entropia diferencial

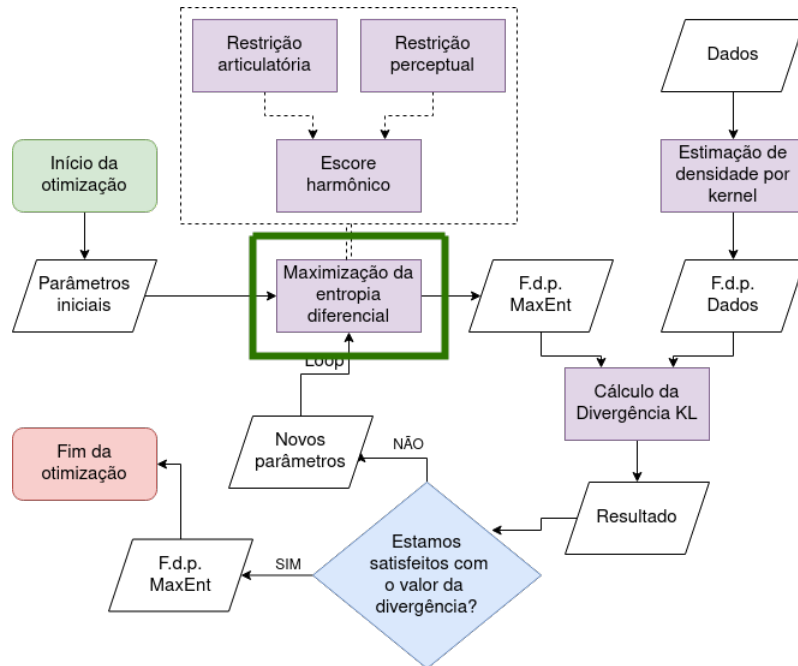


Figura 26: Nesta seção discutiremos a parte destacada. (1)

Vamos começar falando de “entropia”. A entropia é uma medida de incerteza em um sistema.⁶⁰ Isso significa, em termos de probabilidade, que quanto mais homogêneo é um sistema, maior é sua entropia. Quanto mais enviesado, menor. Isso é mais fácil de compreender visualmente:

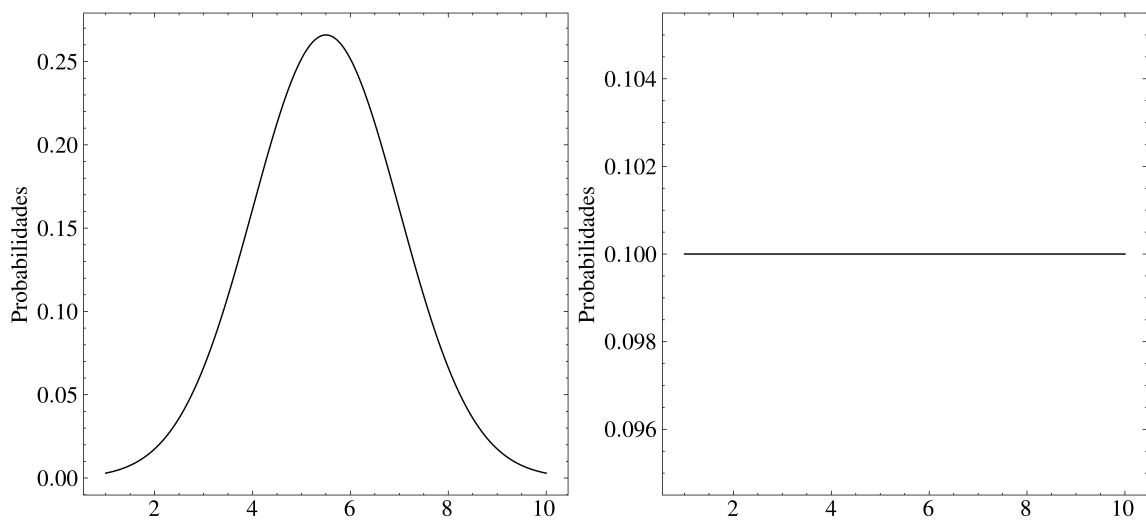


Figura 27: Sistema com menos e mais entropia.

⁶⁰Isso é uma definição simplificada, mas instrumental. O conceito de “entropia” não é simples e adquire diferentes sentidos em diferentes áreas de pesquisa, o que promove bastante confusão entre leigos e até pesquisadores.

Acima, pense que estamos modelando dois sistemas. No gráfico da esquerda, temos uma distribuição de probabilidades normal relativa a 10 números (1 até 10). Isso significa que, se pedíssemos a esse sistema para escolher um número para nós, podemos falar com confiança que há mais chance do número ser 4, 5 ou 6 do que ser 1, 2, 9 ou 10. Por outro lado, o gráfico da direita não parece saber nada sobre a distribuição dos números. Há 10 números e ele divide igualmente as probabilidades entre eles: não há maior ou menor probabilidade para cada número. Se pedíssemos para ele selecionar um número, haveria muito mais *incerteza* sobre qual número seria selecionado. Não poderíamos assumir nada de antemão. É um sistema homogêneo em termos probabilísticos. Sendo assim, podemos afirmar que é um sistema com *maior entropia*. No gráfico do sistema da esquerda, por sua vez, há um sistema com menor grau de incerteza, então pode-se dizer que há *menos entropia*.

Na Fonologia, o conceito de entropia foi introduzido por Sharon Goldwater e Mark Johnson em *Learning OT Constraint Rankings Using a Maximum Entropy Model* (2003). Conforme a metodologia exposta, o princípio de máxima entropia é utilizado para gerar probabilidades com base nos escores harmônicos dos candidatos, *i.e.*, gerar uma distribuição que nada sabe (máxima entropia) além do escore harmônico. A distribuição dessas probabilidades tem como parâmetros os pesos das restrições, pois estes ajustam o escore harmônico. Os autores, após gerar probabilidades, realizam então a maximização da log-verossimilhança entre os dados e as probabilidades resultantes da modelagem. Observa-se a diferença entre probabilidades da modelagem e dados. Ajustam-se os pesos das restrições. E assim em diante repetidas vezes, iterativamente utilizando o método dos gradientes conjugados. E é assim que os pesos, enfim, são otimizados para que eventualmente alcancem uma configuração onde as probabilidades via princípio da máxima entropia coincidam com os dados coletados. Ou seja, até que cheguemos em uma situação em que os parâmetros (pesos das restrições) podem explicar adequadamente os dados.

O procedimento descrito acima por Goldwater e Johnson não é um problema de natureza unicamente estatística. É um problema de aprendizagem fonológica (GOLDWATER; JOHNSON, 2003, p. 1-2). Isso é de extrema importância, pois nos indica que modelos fonológicos não devem ser modelos apenas analíticos, matemáticos ou estatísticos. Eles necessitam ser modelos *mentais* da gramática também. No que diz respeito a esta proposta de modelo, houve pelo menos um esforço em mantermo-nos cognitivamente plausíveis. Nossas restrições são fonologicamente fundamentadas, e todas as operações que descrevemos são análogos contínuos da HG. Em outras palavras, não houve nenhum grande desvio de base teórica até aqui, pois seguimos o “mapa” deixado pelas abordagens discretas. Macroscopicamente, a maneira como candidato, restrições,

violações e harmonia interagem permanece análoga. E, similarmente a Goldwater e Johnson, também iremos trabalhar com o princípio de máxima entropia. No entanto, como de praxe até aqui, devemos trazê-lo para o âmbito das variáveis contínuas e isso significa adotar a entropia diferencial⁶¹. Além disso, agora que nossas restrições são funções matemáticas, há uma maneira conveniente de olhar a relação entre entropia, maximização e restrições fonológicas. Podemos pensar na maximização da entropia como um problema de “otimização com restrições” (*constrained optimization*). E um problema de otimização com restrições, como podemos inferir de seu nome, é um tipo de otimização em que devemos encontrar a melhor solução possível para um problema enquanto respeitamos algumas condições (restrições). Formalmente, podemos pensá-la da seguinte forma:

Para uma interpretação geométrica do processo de otimização de uma função de duas variáveis sujeitas a uma restrição, pense na própria função como uma superfície no espaço tridimensional e na restrição (que é uma equação envolvendo x e y) como uma curva no plano xy . Quando você encontra o máximo ou mínimo da função sujeita à restrição dada, está restringindo sua atenção à parte da superfície que se encontra diretamente acima da curva de restrição. O ponto mais alto nesta porção da superfície é o máximo restrito, e o ponto mais baixo é o mínimo restrito. (HOFFMAN; BRADLEY, 2007, p. 613, tradução nossa)

Visualmente, podemos compreender a passagem acima da seguinte maneira:

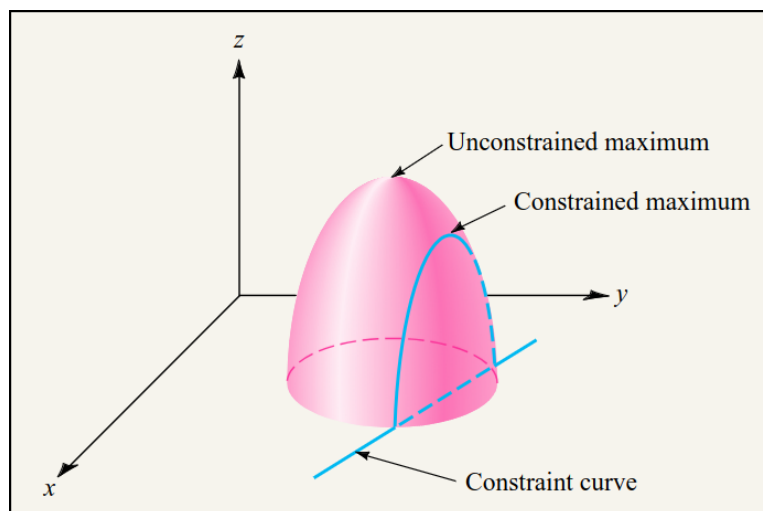


Figura 28: Funcionamento de restrições matemáticas (retirado de HOFFMAN; BRADLEY, 2007, p. 613)

O que faremos, portanto, é tratar a maximização da entropia diferencial como um problema de otimização com restrições. Em nosso caso, analogamente à análise discreta de

⁶¹A entropia diferencial não é exatamente uma “entropia contínua”, pois o conceito de entropia aparentemente não pode ser trazido para variáveis contínuas sem complicações. Assim, podemos pensar a entropia diferencial como *equivalente* ou *análoga*, mas não como *idêntica* à entropia (discreta) de Shannon.

Goldwater e Johnson, a restrição será o escore harmônico. Isso significa que, ao decompô-lo, nossas restrições fonológicas serão as restrições da otimização. Em outras palavras, iremos criar uma f.d.p. que, com exceção de nossas restrições, não assume mais nada. Tudo que ela “sabe” é dado pelas nossas restrições. Em suma, uma f.d.p. dada pela maximização da entropia diferencial por meio de um problema de otimização com restrições, fazendo uso de nossas restrições fonológicas. E o primeiro passo para começar a maximização da entropia é realizar a definição da entropia diferencial (COVER; THOMAS, 2006, p. 243):

$$h(f) = - \int_S f(x) \log(f(x)) dx \quad (22)$$

Sendo $f(x)$ uma f.d.p. para uma variável contínua X e S seu conjunto suporte⁶². Nossa tarefa agora é gerar uma distribuição de probabilidades⁶³ de máxima entropia, por meio de uma otimização com restrições. Efetivamente, isso fará com que tenhamos em mãos uma distribuição de probabilidades que “nada sabe” além das informações de nossas restrições, e cujos parâmetros serão seus pesos. Para esse fim, podemos trabalhar com o método dos multiplicadores de Lagrange. Nesse método, a restrição deve ter o formato $g(x, y, z \dots) = c$. Ou seja devemos especificar tanto a restrição quanto uma constante c que deverá ser a condição dessa restrição, em outras palavras, devemos maximizar a entropia na condição que nossa restrição $g(x, y, z \dots)$ resulte em c . Para nós, é suficiente pensar de maneira fonologicamente intuitiva. Nossa intenção, ao maximizar a entropia modelada pelas restrições, é achar uma distribuição prototípica⁶⁴ de nosso modelo de análise. Ao realizarmos uma modelagem, almejamos a obtenção de uma distribuição originada unicamente pelos pressupostos das restrições.⁶⁵ Assim, como falei, é fonologicamente intuitivo que $c = 0$, pois, se queremos observar uma modelagem prototípica,

⁶²Conjunto suporte é o conjunto em que $f(x) > 0$, ou seja, onde a função $f(x)$ tem valores não-nulos.

⁶³Para quem é de fora da área de estatística, como eu, e nomenclatura é um tanto confusa. Estritamente, me parece, distribuições de probabilidades têm relação com variáveis discretas, mas o termo é usado coloquialmente quando fala-se de elementos matemáticos descritivos de comportamento probabilístico de uma variável em geral, referindo-se também a uma f.d.p. em alguns contextos. No caso, a distribuição de probabilidades de máxima entropia pode ser realizada partindo da entropia diferencial, o que torna-a relativa a variáveis contínuas.

⁶⁴Estou usando o termo de maneira coloquial, pois é uma boa maneira de compreender essa f.d.p. resultante. Não estou fazendo referência à prototipicidade (cf. BOERSMA, 2005) ou algum outro conceito fonológico específico.

⁶⁵Que não assume nada, *i.e.*, a distribuição de probabilidade via maximização da entropia levando em conta as restrições.

estamos falando da distribuição não deve incorrer violações.⁶⁶ Enfim, podemos agora começar expressando o método dos multiplicadores de Lagrange da seguinte forma:

$$\begin{aligned} \text{maximize: } h(f) &= - \int_S f(x) \log(f(x)) dx \\ \text{satisfazendo: } \mathcal{H}(x) &= \sum_{i=1}^n p_i v_i(x) \rightarrow 0 \end{aligned}$$

Utilizei $\mathcal{H}(x) \rightarrow 0$ ao invés de $\mathcal{H}(x) = 0$ para denotar o fato de que, na maioria das análises, o escore harmônico não chegará a exatamente 0.⁶⁷ Mas isso não muda o fato de, para que direcionemos a otimização em $\mathcal{H}(x) \rightarrow 0$, devemos fazer com que $c = 0$. Dito isso tudo, e com nosso problema de otimização definido, podemos proceder com a formulação da *função Lagrangiana* (HILLIER; LIEBERMAN, 2010, p. 999):

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i=1}^n \lambda_i [g_i(x) - c_i] \quad (23)$$

Acima, temos a forma genérica da função Lagrangiana. Observe que o somatório de restrições dessa função apresenta semelhanças com nosso escore harmônico $\sum_{i=1}^n p_i v_i(x)$. Vamos então considerar, como discutimos, que $c = 0$ e, por conveniência de notação, tratar λ como peso de nossas restrições.⁶⁸ A formulação para nosso projeto, então, seria:

$$\mathcal{L}(f, \lambda_i) = h(f) - \sum_{i=1}^n \lambda_i v_i(x) \quad (24)$$

Consequentemente, estaríamos trabalhando com a seguinte equação:

⁶⁶Há uma nota importante aqui. A depender de quais restrições estamos trabalhando e suas expressões matemáticas, é possível que nunca alcancemos exatamente 0 violações. Assim, se formos mais específicos e rigorosos em nossa definição, c deve ser o menor valor possível da restrição para todos os *inputs* no domínio analisado. Em nosso caso, ao realizar $c = 0$, estamos elaborando uma otimização que irá buscar, ao longo de todos os valores possíveis, soluções cujo valor de c aproxima-se de 0. Assim, de maneira geral em uma abordagem de otimização, podemos fazer uso de $c = 0$ violações mesmo que, teoricamente, nossas violações não sejam 0 sob nenhum *input*. Nós, em um ambiente computacional dinâmico que naturalmente tolera certo grau de imprecisão, estamos apenas direcionando as metas da otimização durante suas iterações.

⁶⁷Só pelo fato de fidelidade e marcação oporem-se, já devemos estar cientes de que restrições preferem diferentes intervalos nos candidatos contínuos. Portanto, ao considerar mais de uma restrição em oposição, nunca teremos exatamente 0 violações no escore harmônico: pois isso representaria uma hipotética interseção perfeita entre todas as restrições. A OT e a HG são quadros teóricos de *minimização* de violações que não necessariamente estão buscando candidatos *sem violação alguma*.

⁶⁸Isto é, para nos adequarmos à notação da Lagrangiana. Na prática, o multiplicador lagrangiano λ é um peso de restrição de funcionamento idêntico aos nossos pesos de restrições fonológicas. Não faz diferença prática, mas faz diferença para evitar confusão na notação tradicional.

$$\mathcal{L}(f, \lambda_0, \lambda_1, \lambda_2) = h(f) - \lambda_0 N - \lambda_1 v_P(F1) - \lambda_2 v_A(F1, F2) \quad (25)$$

Para clareza, antes de continuarmos devemos ter em mente que:

$h(f)$ é a entropia diferencial de uma f.d.p. f .

$\lambda_{0,1,2}$ são multiplicadores lagrangianos.

$\lambda_{1,2}$ podem ser considerados os pesos de nossas restrições fonológicas.

$v_P(F1), v_A(F1, F2)$ são restrições fonológicas.

N é uma restrição de normalização, que diz que a integral da máxima entropia deve ser 1, ou seja, $\int_{-\infty}^{\infty} f(x) dx - 1 = 0$, conforme Bishop (2006, p. 54).

Vamos rearranjar os elementos de nossa Lagrangiana agora no funcional⁶⁹ $\mathbf{F}$:

$$\mathbf{F} = - \int_S f(x) \log(f(x)) dx - \lambda_0 \left(\int_{-\infty}^{\infty} f(x) dx - 1 \right) \quad (26)$$

Perceba que desconsideramos nossas restrições fonológicas na elaboração do funcional $\mathbf{F}$. A razão disso é que são independentes de $f(x)$, ou seja, não variam conforme a f.d.p. $f(x)$ varia. Podemos pensar, simplificada e de forma específica a nosso caso, que é porque o cômputo de violações $e^{R_i(x) - \hat{f}_{\text{dados}}(x)}$ não contém, em termos matemáticos, a f.d.p. sendo otimizada – apenas a f.d.p. dos dados. Vamos agora remover as integrações para podermos focar exclusivamente no comportamento desse funcional:

$$\mathbf{F} = -f(x) \log(f(x)) - \lambda_0 (f(x) - 1) \quad (27)$$

Considerando que $\frac{df(x)}{dx} = 0$ é uma condição necessária para que $x = x_{\text{extremum}}^*$ (HILLIER; LIEBERMAN, 2010, p. 998), devemos diferenciar o funcional com relação à f.d.p. $f(x)$ e igualá-lo a 0. O diferencial de um funcional pode ser calculado com a equação de Euler (conhecida também como equação de Euler-Lagrange):

⁶⁹Um funcional é um operador que toma uma função $f(x)$ qualquer e retorna um *output* $\mathbf{F}$. Um exemplo de funcional é a própria entropia diferencial de uma variável contínua x , pois seu *input* é uma função de densidade de probabilidades e seu *output* é um valor escalar representando a entropia de x sobre tal densidade (BISHOP, 2006, p. 703). Em muitos casos de otimização, o funcional pode ser entendido com a função objetivo que está sendo otimizada. Conforme Gelfand e Fomin em *Calculus of Variations* (1963, p. 1), “one might say that a functional is a kind of function, where the independent variable is itself a function (or curve)”.

Seja $J[y]$ um funcional da forma $\int_a^b \mathbf{F}(x, y, y') dx$ definido no conjunto de funções $y(x)$ e satisfaça as condições de contorno $y(a) = A, y(b) = B$. **Então, uma condição necessária para $J[y]$ ter um extremo para uma dada função $y(x)$ é que $y(x)$ satisfaça a equação de Euler $\mathbf{F}_y - \frac{d}{dx} \mathbf{F}_{y'} = 0$.** (GELFAND; FOMIN, 1963, p. 15, tradução nossa, grifo nosso)

Sabendo que nosso funcional não depende da derivada da f.d.p. em relação a x (*i.e.*, f'), e assim consequentemente o segundo termo $-\frac{d}{dx} \mathbf{F}_{f'}$ para nosso caso é zero, podemos simplificar a equação de Euler-Lagrange para apenas seu primeiro termo $\mathbf{F}_f$. O que resultaria na expressão:

$$\frac{\partial \mathbf{F}}{\partial f} = \frac{\partial}{\partial f} (-f(x) \log(f(x)) - \lambda_0 (f(x) - 1)) = 0$$

Podemos diferenciar $-f(x) \log(f(x))$ pela regra dos produtos, $\frac{d}{dx}(u \cdot v) = u' \cdot v + u \cdot v'$:

$$\begin{aligned} \frac{\partial}{\partial f} (-f(x) \log(f(x))) &= -(1 \cdot \log(f(x)) + f(x) \frac{1}{f(x)}) \\ &= -(\log(f(x)) + 1) \\ &= -\log(f(x)) - 1 \end{aligned}$$

E a diferenciação da restrição de normalização $-\lambda_0 (f(x) - 1)$ é igual a $-\lambda_0$, pois a derivada de constantes é 0 (então a derivada de -1 é 0) e a derivada de $f(x)$ relativa a $f(x)$ é 1, fazendo com que $\frac{\partial}{\partial f} (-\lambda_0 (f(x) - 1)) = -\lambda_0(1 - 0) = -\lambda_0$. Agora temos o seguinte resultado:

$$\mathbf{F}_f = -\log(f(x)) - 1 - \lambda_0 = 0$$

Podemos considerar que essa é nossa função objetivo⁷⁰, pois é o diferencial do funcional da entropia de uma f.d.p. que pretendemos maximizar. Ao trabalharmos com seu diferencial, buscamos compreender como mudanças no *input* causam mudanças no *output*, similarmente a como trabalhei o conceito de estabilidade do escore harmônico. Logo, se a entropia (*i.e.*, o funcional) é uma métrica, precisamos trabalhar com seu diferencial para entender *como* essa

⁷⁰A função que estamos tentando otimizar.

métrica muda, para ajustá-la e alcançar o maior *output* possível no funcional. Com essa equação em mãos, podemos entender como mudanças irão impactar a entropia e sistematicamente ajustar nossas restrições fonológicas para maximizá-la. Agora que concluímos a aplicação da equação de Euler-Lagrange sob nosso funcional, chegamos em:

$$-\log(f(x)) - 1 - \lambda_0 - \lambda_1 v_P(F1) - \lambda_2 v_A(F1, F2) = 0 \quad (28)$$

Nossas restrições fonológicas mantiveram-se as mesmas, visto que são independentes de $f(x)$. Assim, não fizeram parte do funcional e, consequentemente, não foram “computadas” pela equação de Euler-Lagrange. Mas, enfim, acima temos a versão final de nossa máxima entropia. Talvez fique mais óbvio ao realizar uma simplificação:

$$\begin{aligned} -\log(f(x)) - 1 - \lambda_0 - \lambda_1 v_P(F1) - \lambda_2 v_A(F1, F2) &= 0 \\ -1 - \lambda_0 - \lambda_1 v_P(F1) - \lambda_2 v_A(F1, F2) &= \log(f(x)) \\ e^{-1-\lambda_0-\lambda_1 v_P(F1)-\lambda_2 v_A(F1, F2)} &= e^{\log(f(x))} \\ e^{-1-\lambda_0-\lambda_1 v_P(F1)-\lambda_2 v_A(F1, F2)} &= f(x) \end{aligned}$$

Chegamos agora na definição de distribuição de probabilidade via maximização da entropia diferencial⁷¹. Devemos aplicar essa definição, agora, ao nosso problema:

$$f_{\text{MaxEnt}}(F1, F2) = e^{-1-\lambda_0-\lambda_1 v_P(F1)-\lambda_2 v_A(F1, F2)}$$

Onde $\lambda_{1,2}$ são equivalentes a nossos pesos de restrições $p_{P,A}$. Essa será nossa fórmula para a função de densidade de probabilidades de máxima entropia. De maneira genérica, poderíamos escrevê-la da seguinte forma:

$$f_{\text{MaxEnt}}(x) = e^{-1-\lambda_0-\sum_{i=1}^n \lambda_i v_i(x)} \quad (29)$$

Tendo essa fórmula em mente, agora por via das dúvidas, podemos recorrer à literatura para saber se estamos no caminho certo. Cover e Thomas (2006, p. 410) definem uma fórmula geral de densidade que maximiza a entropia:

$$f(x) = e^{\lambda_0 - 1 + \sum_{i=1}^m \lambda_i r_i(x)} \quad (30)$$

⁷¹Calculada com log de base e e, consequentemente, em *nats*. Discutiremos isso brevemente em outra nota de rodapé, de maneira mais contextualizada.

Podemos perceber que a única diferença entre as duas fórmulas é o sinal dos multiplicadores lagrangianos⁷². Em outras palavras, nossa f.d.p. de máxima entropia pode ser considerada adequada. Talvez tivesse sido mais fácil que simplesmente pegássemos a definição geral de Cover e Thomas, porém seria consideravelmente menos intuitivo. No longo prazo, conforme manipulamos as expressões matemáticas em nosso quadro teórico da HG, parece necessário desenvolver uma intuição e compreensão sobre todos os aspectos com os quais trabalhamos, para não cair em possíveis desalinhamentos entre *aspectos fonológicos sendo descritos* e *as operações matemáticas que os subjazem*.

Munidos da equação de f.d.p. que maximiza entropia, podemos passar agora para a próxima etapa de nossa otimização. Devemos, finalmente, calcular os pesos de nossas restrições fonológicas. Além disso, conforme vimos, devemos também calcular λ_0 , que surgiu em virtude da restrição de normalização de nossa f.d.p.

3.7.2 Divergência de Kullback-Leibler

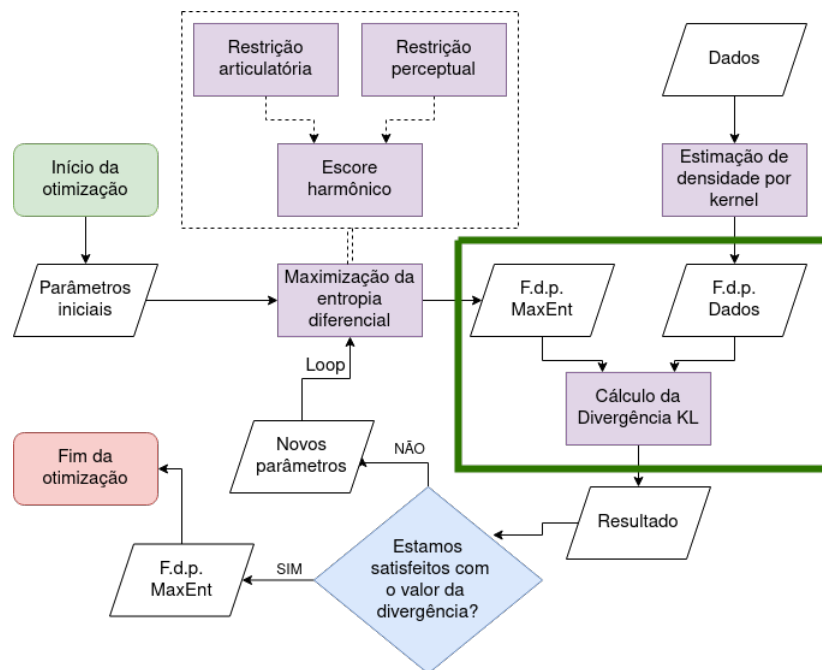


Figura 29: Nesta seção discutiremos a parte destacada. (2)

Definiremos da seguinte maneira os objetivos desta etapa: devemos achar pesos para nossas restrições (e, além disso, um valor para λ_0) que garantam que nossa f.d.p. de máxima

⁷²Hillier e Lieberman (2010, p. 999) descrevem a fórmula genérica da Lagrangiana com multiplicadores negativos, enquanto Bishop (2006, p. 708), Cover e Thomas (2006, p. 410) com multiplicadores positivos. Ao tratarmos de restrições de igualdade, como as nossas, o sinal dos multiplicadores é uma questão de convenção e não afeta o comportamento da função.

entropia assemelhe-se aos dados coletados. Para tanto, precisamos definir o que entendemos por “assemelhar-se” aos dados. Tradicionalmente, isso é feito por meio da maximização da log-verossimilhança (cf. GOLDWATER; JOHNSON, 2003; HAYES; WILSON, 2007), que é um tipo de estimação de parâmetros. Neste protótipo, faremos uso de um método equivalente. Conforme vimos no fluxograma e discutimos durante a seção passada, realizaremos comparações entre f.d.p. de máxima entropia *versus* dos dados, e isso significa que a redução do grau de *diferença* entre f.d.p. seja uma medida relativamente intuitiva para nós. Assim, ao invés de *maximizar a verossimilhança*, iremos *minimizar sua divergência*. Essas duas operações são equivalentes (cf. EGUCHI; COPAS, 2006). Em nosso caso, em que a preocupação da otimização é a adequação do comportamento de uma distribuição à outra, é metodologicamente coerente que trabalhem com a redução de divergência entre elas. Mais especificamente, trabalharemos com a minimização da divergência de Kullback-Leibler (cf. KULLBACK; LEIBLER, 1951), também conhecida como divergência KL ou entropia relativa. Em suma:

Um estimador é dito consistente se eventualmente recupera os verdadeiros parâmetros que geraram os dados à medida que o tamanho da amostra tende ao infinito, i.e., $\hat{\theta}(D) \rightarrow \theta^*$ à medida que $|D| \rightarrow \infty$ (onde a seta denota convergência em probabilidade). É claro que esse conceito só faz sentido se os dados vêm do modelo especificado com parâmetros θ^* , o que normalmente não é o caso com dados reais. Apesar disso, pode ser uma propriedade teórica útil. Pode-se mostrar que *MLE* [maximum likelihood estimation, estimação de máxima verossimilhança] é um estimador consistente. **A razão intuitiva é que maximizar a verossimilhança é equivalente a minimizar KL** $(p(\cdot|\theta^*)||p(\cdot|\hat{\theta}))$, onde $p(\cdot|\theta^*)$ é a distribuição verdadeira e $p(\cdot|\hat{\theta})$ é nossa estimativa. Podemos alcançar 0 de divergência KL se e somente se $\hat{\theta} = \theta^*$. (MURPHY, 2012, p. 200, tradução nossa, grifo nosso)

Vamos agora formalmente definir a divergência KL. Ela é dada pela seguinte expressão (adaptado de MURPHY, 2012, p. 57):

$$\text{KL}(p||q) = \int_{-\infty}^{\infty} p \log \left(\frac{p}{q} \right) dx \quad (31)$$

Que pode ser reformulada de maneira mais intuitiva⁷³ como uma diferença entre integrais pela propriedade da fração dos logaritmos (adaptado de BISHOP, 2006, p. 55):

$$\text{KL}(p||q) = \int_{-\infty}^{\infty} p \log(p) dx - \int_{-\infty}^{\infty} p \log(q) dx \quad (32)$$

No âmbito da modelagem proposta neste trabalho, p e q correspondem respectivamente a $\hat{f}_{\text{dados}}(F1, F2)$ e $f_{\text{MaxEnt}}(F1, F2)$. A divergência KL é o número de *bits* ou *nats* adicionais necessários para codificar (*encode*) informação utilizando uma distribuição q ao invés

⁷³Intuitiva no sentido de que “diferenças” ou “divergências” são associadas a subtração.

de uma distribuição original ou ideal p (MURPHY, 2012, p. 58). *Bits* e *nats* são unidades de entropia.⁷⁴ Isso significa que, se estamos falando de uma quantidade extra de *bits*, a divergência KL nunca será negativa e só alcançará 0 quando $p = q$ (*Ibid.*, p. 58). Similarmente a como o funcional da Lagrangiana foi nossa função objetivo da maximização da entropia, aqui a divergência KL será também nossa função objetivo. Nossa meta agora é *minimizar* a divergência KL. Sendo assim, pode-se dizer que estamos realizando uma operação análoga à que realizamos na seção anterior, mas “no sentido contrário”: minimização ao invés de maximização. Mas diferentemente da maximização da entropia, em que buscávamos uma formulação genérica para a f.d.p., aqui iremos ajustar parâmetros da distribuição (pesos e λ_0) para que a diferença de *nats* entre $\hat{f}_{\text{dados}}(F1, F2)$ e $f_{\text{MaxEnt}}(F1, F2)$ seja a menor possível. Em outras palavras, estamos realizando uma aproximação não-paramétrica da f.d.p.⁷⁵

Agora que definimos nossa distância entre f.d.p., gostaria de notar uma de suas propriedades de particular interesse para nós. A divergência KL não é uma métrica em sentido estrito, pois não satisfaz a condição de simetria entre dois pontos ou dois lados. Isso significa que a divergência KL entre p e q não resulta no mesmo valor da divergência KL entre q e p , ou seja, $\text{KL}(p||q) \neq \text{KL}(q||p)$ (BISHOP, 2006, p. 55). Essa direcionalidade é importante para nossa metodologia, pois permite o controle sobre qual distribuição deve ser considerada *originária* e qual seria *modelada*, o que alinha-se com a estrutura de nossa modelagem ($\hat{f}_{\text{dados}}(F1, F2)$ e $f_{\text{MaxEnt}}(F1, F2)$).

Nesta subseção, não pretendo realizar uma demonstração prática da divergência KL, pois sinto que será mais esclarecedor se realizar essa operação na subseção seguinte ao tratar da *minimização* da divergência KL. Assim, realizamos uma demonstração mais bem contextualizada nos conformes do nosso problema. Desse modo, podemos agora prosseguir para a discussão do algoritmo de otimização.

⁷⁴*Bits* é uma abreviação de “dígitos binários”, a unidade básica de informação binária, e é a unidade resultante de análises que fazem uso de $\log_2$. Por outro lado, *nats* são a unidade básica da informação natural, ou seja, da informação resultante de $\log_e$ (MURPHY 2012, p. 56). Em termos práticos, ao buscarmos redução a divergência, não é uma grande preocupação para nós se estamos reduzindo *bits* ou *nats* – assim como não seria se estivéssemos reduzindo centímetros ou polegadas se nossa intenção fosse aproximar a unidade de 0 – mas, para manter uma consistência entre unidades, usaremos $\log_e$ para calcular a divergência KL, pois calculamos a entropia com $\log_e$ e viemos praticando a notação $\log = \ln$.

⁷⁵Uma metodologia semelhante foi publicada por Mazzoni e Reucher como *Quasi-Continuous Maximum Entropy Distribution Approximation with Kernel Density* (2010). Além disso, a lógica de “maximização da entropia e minimização da entropia cruzada” também já estava disposta em *Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy* de Shore e Johnson (1980). Desse modo, apesar da metodologia específica desse projeto ser de minha autoria, dificilmente pode-se dizer que é uma ideia original, visto que “maximizar entropia e minimizar diferença” é um conceito relativamente bem documentado.

3.7.3 Algoritmo de Broyden-Fletcher-Goldfarb-Shanno

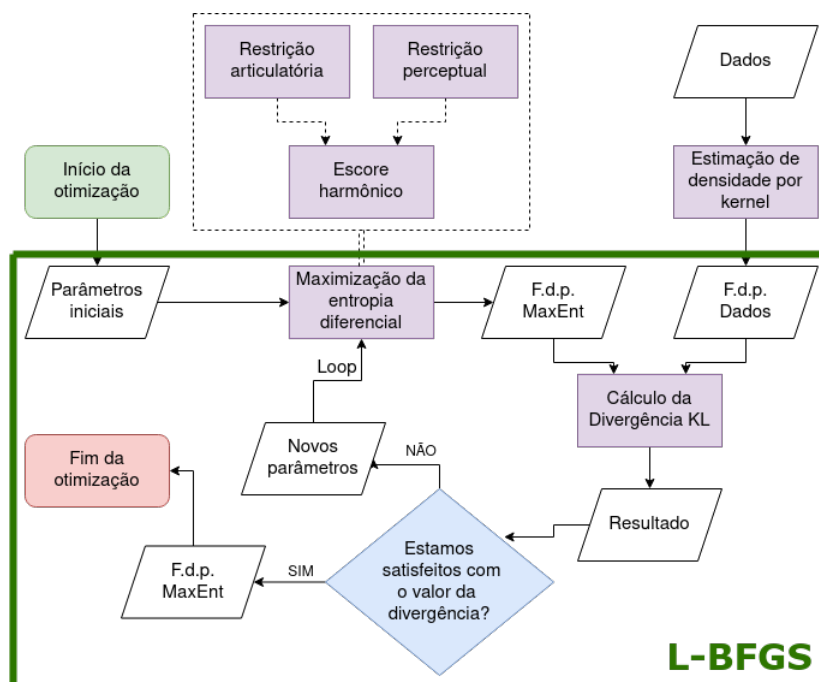


Figura 30: Nesta seção discutiremos a parte destacada. (3)

Como podemos observar no fluxograma acima, faremos agora uso de tudo que discutimos ao longo desta seção. Realizaremos a minimização da divergência KL entre f.d.p. modelada vs. f.d.p. estimada dos dados e, para esse fim, utilizaremos o algoritmo de Broyden-Fletcher-Goldfarb-Shanno, conhecido como BFGS. Mais especificamente, aplicarei sua versão de memória limitada (cf. BYRD; LU; NOCEDAL, 1995), intitulada L-BFGS. Esse é um algoritmo não-linear, pois nossas restrições fonológicas não são lineares e a minimização da divergência KL não é uma operação linear, visto que envolve logaritmos⁷⁶.

Tradicionalmente, a otimização em gramáticas MaxEnt é realizada pelo método de gradientes conjugados (cf. GOLDWATER; JOHNSON, 2003; HAYES; WILSON, 2007). Inicialmente, tentei realizar a otimização por esse mesmo método, porém me deparei com dois problemas. O primeiro é que o método dos gradientes conjugados não apresentou boa performance durante testagem, sendo relativamente mais lento quando comparado com o algoritmo L-BFGS. Além disso, o segundo problema que encontrei foi a constante negatização dos pesos das restrições. Não há maneira nativa ao método dos gradientes conjugados para evitar valores

⁷⁶Sistemas lineares devem atender aos princípios de Aditividade ($f(x + y) = f(x) + f(y)$) e Homogeneidade ($f(cx) = cf(x)$). A presença de um logaritmo na função impossibilita a satisfação dessas condições, graças às suas propriedades. Perceba: $\log(x \cdot y) = \log(x) + \log(y)$, que não satisfaz a condição de Aditividade e, assim, introduz não-linearidade no sistema.

negativos, diferentemente do algoritmo L-BFGS que suporta a imposição de limites inferiores e superiores à solução.⁷⁷ Sendo assim, optei pelo uso do algoritmo L-BFGS ao invés do método dos gradientes conjugados já existente na literatura fonológica. Como a escolha de um algoritmo de otimização é extremamente sensível ao problema trabalhado, faz sentido que nossa otimização – ao trabalhar um quadro teórico matemático distinto – vá requerer um algoritmo distinto. Vale a pena notar, porém, que algoritmos de memória limitada, como L-BFGS, evoluíram gradualmente como tentativas de melhorar métodos de gradiente conjugados não-lineares (NO-CEDAL; WRIGHT, 1999, p. 227). Ou seja, apesar de serem algoritmos distintos, são métodos historicamente e matematicamente relacionados.

Para compreender o algoritmo L-BFGS, é mais fácil apresentá-lo de maneira prática, expondo suas fórmulas passo a passo (*Ibid.*, p. 224). Vamos começar retomando nossa definição de modelo de f.d.p.:

$$f_{\text{MaxEnt}}(F1, F2) = e^{-1-\lambda_0-\lambda_1 v_P(F1)-\lambda_2 v_A(F1, F2)}$$

Que será utilizada conjuntamente com a f.d.p. estimada dos dados $\hat{f}_{\text{dados}}(F1, F2)$ na equação da divergência de Kullback-Leibler:

$$\mathbf{KL}(\hat{f}_{\text{dados}}(F1, F2) || f_{\text{MaxEnt}}(F1, F2)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p \log \left(\frac{\hat{f}_{\text{dados}}(F1, F2)}{f_{\text{MaxEnt}}(F1, F2)} \right) dF1 dF2$$

Nosso problema agora é a minimização dessa divergência, que pode ser escrito da seguinte maneira:

$$\min_{\lambda} \mathbf{KL}(\hat{f}_{\text{dados}}(F1, F2) || f_{\text{MaxEnt}}(F1, F2)), \lambda = (\lambda_{0,1,2})$$

Em outras palavras, o objetivo da otimização é minimizar a divergência KL entre a f.d.p. dos dados e a f.d.p. modelada com relação aos parâmetros λ que são λ_0 (relativo à normalização da f.d.p.), λ_1 (peso da restrição perceptual) e λ_2 (peso da restrição articulatória). Devemos também especificar que queremos apenas valores positivos para nossos pesos:

$$\lambda_i > 0, \text{ para } i \in \{0, 1, 2\}$$

⁷⁷Não quis lidar com valores negativos pois “pesos de restrições negativos” não é algo intuitivo de ser interpretado fonologicamente. É possível que valores negativos nos levassem, em alguns casos, a uma menor divergência entre f.d.p. MaxEnt e dos dados, mas isso seria uma solução unicamente *matemática* e não *fonológica*. Assim, optei que os pesos das restrições sempre fossem positivos, em linha com a literatura fonológica. Talvez pudéssemos pensar que essa, em si, seja uma outra restrição em nosso quadro teórico, instalada diretamente no algoritmo de otimização.

Com o problema de otimização formalizado, podemos aplicar o algoritmo L-BFGS. Nós começamos nossa otimização com um valor pré-determinado para todos os parâmetros (em nosso caso, usarei 1) e uma aproximação para a matriz hessiana⁷⁸ inversa $H^{(0)}$, gerada automaticamente em nossa aplicação⁷⁹. Essa é a etapa de inicialização. Podemos definir esse primeiro passo 0 do seguinte modo:

$$\lambda^{(0)} = (\lambda_0^{(0)}, \lambda_1^{(0)}, \lambda_2^{(0)})$$

Depois, a cada iteração k , computaremos o gradiente da divergência KL relativo aos parâmetros λ_i :

$$\nabla \mathbf{KL}^{(k)} = \left(\frac{\partial \mathbf{KL}}{\partial \lambda_0}, \frac{\partial \mathbf{KL}}{\partial \lambda_1}, \frac{\partial \mathbf{KL}}{\partial \lambda_2} \right)^{(k)}$$

Com o gradiente em mãos, computaremos a direção da busca $p^{(k)}$ ao aplicar a aproximação da matriz hessiana inversa $H^{(k)}$ no gradiente. Essa “direção da busca” é um vetor que indica numericamente para onde a otimização deve seguir para minimizar nossa função objetivo. Vejamos:

$$p^{(k)} = -H^{(k)} \nabla \mathbf{KL}^{(k)}$$

Agora que temos uma direção para ajustar nossos parâmetros, obedeceremos a seguinte regra de atualização:

$$\lambda^{(k+1)} = \lambda^{(k)} + \alpha^{(k)} p^{(k)}$$

Acima, $\alpha^{(k)}$ é uma grandeza escalar que define o tamanho do “salto” que as coor-

⁷⁸A matriz hessiana é uma matriz quadrada que contém as derivadas parciais de segunda ordem da função objetivo, armazenando, portanto, informações sobre a curvatura da função objetivo em todas as direções. Ela é frequentemente utilizada em problemas de otimização, pois nos auxilia a buscar a direção da otimização de maneira eficiente. No entanto, o uso direto do hessiano não é sempre útil para atualização dos passos entre iterações na otimização, pois apenas nos diz sobre a curvatura. Por outro lado, o seu inverso, em razão das propriedades de inversão em álgebra linear, é capaz de ajustar o vetor gradiente, indicando tamanho do passo e direção da descida por iteração. O uso da matriz hessiana inversa é originário dos algoritmos newtonianos e encontra amplo uso nos métodos quase-newtonianos também, como o L-BFGS. Geralmente, é utilizada uma *aproximação* da matriz hessiana inversa, pois cálculo de matrizes e suas inversões é computacionalmente intensivo e pouco eficiente. E, em nosso caso usando L-BFGS, utilizaremos uma aproximação *simplificada* da matriz hessiana inversa por meio de informações *recentes* entre iterações, como veremos. É aí que L-BFGS (versão de memória limitada) difere do algoritmo original BFGS, que calcula uma aproximação completa da matriz hessiana inversa.

⁷⁹Com o uso da biblioteca *scipy* e módulo *minimize* na linguagem de programação Python3.

denadas $(\lambda_0^{(k)}, \lambda_1^{(k)}, \lambda_2^{(k)})$ darão na direção em que o vetor $p^{(k)}$ aponta em cada iteração. Essa grandeza escalar é selecionada de acordo com critérios específicos que garantem que os passos não sejam excessivamente grandes ou demasiadamente pequenos. Não abordarei aqui estes critérios, pois não são aspectos computacionalmente relevantes para nós.⁸⁰ Com isso, geraremos novos valores para nossos parâmetros em cada iteração, movendo-os na direção da minimização da divergência. Agora, após chegarmos nesse novo valor, atualizaremos a matriz hessiana inversa para a próxima iteração. Devemos computar a diferença entre parâmetros e a diferença entre gradientes:

$$s^{(k)} = \lambda^{(k+1)} - \lambda^{(k)}$$

$$y^{(k)} = \nabla \text{KL}^{(k+1)} - \nabla \text{KL}^{(k)}$$

E usá-los para calcular um fator de escala $\rho^{(k)}$ que, sendo o produto interno da diferença entre dois gradientes consecutivos e dois pontos consecutivos no espaço de parâmetros, garante que a matriz hessiana inversa aproximada seja atualizada de modo que reflète as curvaturas locais da função objetivo:

$$\rho^{(k)} = \frac{1}{y_{(k)}^T s^{(k)}}$$

Isso nos permite atualizar as iterações sem depender da matriz hessiana inversa aproximada completa⁸¹, que é computacionalmente mais complexa e menos eficiente em problemas de larga escala. As únicas informações que cada iteração depende para atualização são $s^{(k)}$ e $y^{(k)}$, diferenças recentes nos gradientes e vetores de posição. Daí o nome dessa versão do algoritmo: L-BFGS, algoritmo de Broyden-Fletcher-Goldfarb-Shanno *de memória limitada*. Assim, finalmente atualizamos a matriz hessiana inversa para iniciar uma nova iteração:

$$H^{(k+1)} = (I - \rho^{(k)} s^{(k)} y_{(k)}^T) H^{(k)} (I - \rho^{(k)} y_{(k)} s_{(k)}^T) + \rho^{(k)} s^{(k)} s_{(k)}^T$$

Sendo I a matriz identidade. Em nosso caso, com 3 parâmetros λ , a matriz identidade

⁸⁰Em termos computacionais, a seleção de α é dada automaticamente, então não é preciso que nos ocupemos com isso.

⁸¹Como explicado em nota de rodapé anterior, ao trabalharmos com a versão de memória limitada deste algoritmo, não fazemos uso da matriz hessiana inversa aproximada completa e sim de uma versão simplificada que apenas armazena informações relativas à última iteração. No caso, isso é viabilizado por $s^{(k)}$, $y^{(k)}$ e $\rho^{(k)}$.

I é uma matriz 3×3 com valores 1 na sua diagonal e 0 no resto:

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Podemos, enfim, resumir esse processo da seguinte maneira:

0. Gerar f.d.p. inicial $f_{\text{MaxEnt}}^{(0)}(F1, F2)$ com $\lambda^{(0)} = (1, 1, 1)$, atualizando cada iteração k .
- (k) : 1. Calcular $\mathbf{KL}^{(k)}(\hat{f}_{\text{dados}}(F1, F2) || f_{\text{MaxEnt}}^{(k)}(F1, F2))$ e $\nabla \mathbf{KL}^{(k)}$.
- (k) : 2. Buscar a direção $p^{(k)} = -H^{(k)} \nabla \mathbf{KL}^{(k)}$.
- (k) : 3. Selecionar $\alpha^{(k)}$ e encontrar novos parâmetros $\lambda^{(k+1)} = \lambda^{(k)} + \alpha^{(k)} p^{(k)}$.
- (k) : 4. Atualizar a matriz hessiana inversa $H^{(k+1)} = V_{(k)}^T H^{(k)} V_{(k)} + \rho^{(k)} s^{(k)} s_{(k)}^T$.
- (k) : 5. Gerar nova f.d.p. $f_{\text{MaxEnt}}^{(k+1)}(F1, F2)$ com novos parâmetros $\lambda^{(k+1)}$.
- $(k+1)$: 1. Calcular $\mathbf{KL}^{(k)}(\hat{f}_{\text{dados}}(F1, F2) || f_{\text{MaxEnt}}^{(k+1)}(F1, F2))$ e $\nabla \mathbf{KL}^{(k+1)}$.
- $(k+1)$: 2. Buscar a direção $p^{(k+1)} = -H^{(k+1)} \nabla \mathbf{KL}^{(k+1)}$.
- $(k+1)$: 3. Vide passo 3 da iteração k anterior e assim sucessivamente para os passos seguintes.

Esse procedimento ocorre recursivamente até completarmos nossa minimização e/ou alcançarmos convergência. Normalmente critérios de convergência podem ser mudanças muito pequenas na divergência KL ou nos parâmetros, ou podemos impôr um limite na quantidade de iterações. Abaixo demonstro visualmente essas etapas de minimização ao aplicar o algoritmo L-BFGS na equação $f(x, y) = (x - 1)^4 + (y - 2)^4$, função que selecionei por sua simplicidade.

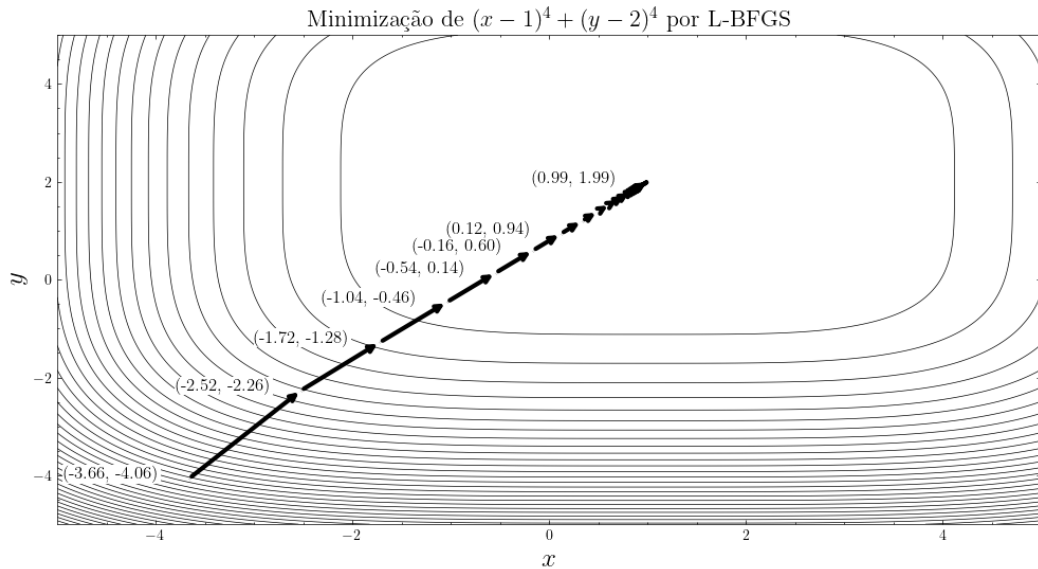


Figura 31: Demonstração do algoritmo L-BFGS.

O gráfico acima mostra como são os “passos” ou “saltos” por iteração e apresenta de maneira intuitiva o que realizaremos na minimização da divergência KL ao otimizar nossos parâmetros. Dito isso, ao realizarmos a minimização da divergência KL da maneira como discutimos acima com as restrições e f.d.p. que abordamos ao longo deste capítulo, chegaremos nos seguintes valores:

Restrição	Parâmetro	Valor otimizado
Normalização	λ_0	1.027
Perceptual	λ_1 ou p_P	0.417
Articulatória	λ_2 ou p_A	0.018

Tabela 19: Valores otimizados dos pesos das restrições e λ_0 .

Esses valores foram confirmados ao longo de mais de uma rodada de otimização, partindo de diferentes palpites iniciais. Como os valores dos pesos foram consistentemente os mesmos (ou, pelo menos, extremamente próximos) em diversas testagens independentes com palpites iniciais distintos, podemos falar com confiança que o processo de otimização teve êxito e converge nos mesmos valores. Além disso, como podemos observar abaixo, houve boa diminuição na divergência KL em relação aos valores iniciais:

	Parâmetros	Divergência
Pré-otimização	$\lambda = \{1, 1, 1\}$	4.233
Pós-otimização	$\lambda = \{1.027, 0.417, 0.018\}$	0.351

Tabela 20: Divergência KL pré- e pós-otimização (decimais dos parâmetros abreviados).

Sendo assim, dada a redução da divergência, sabemos que a f.d.p. modelada aproximou-se em comportamento dos dados coletados. Agora, ao aplicar nossos resultados na equação de nossa distribuição, alcançaremos a seguinte expressão:

$$f_{\text{MaxEnt}}(F1, F2) = e^{-2.027 - 0.417v_P(F1) - 0.018v_A(F1, F2)}$$

E a expressão acima resulta visualmente na seguinte na distribuição:

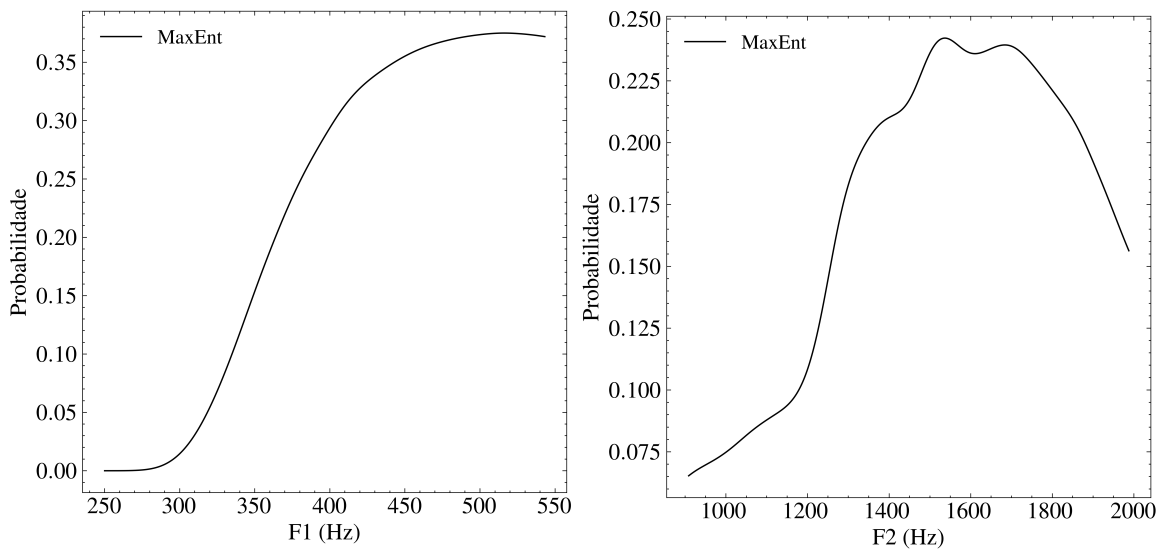


Figura 32: F.d.p. de máxima entropia.

Para trabalhar com candidatos contínuos, precisamos lidar com uma distribuição de probabilidades contínua. A imagem acima é um gráfico da função de densidade de probabilidades modelada pelo escore harmônico dos dados coletados. Ela é uma distribuição de probabilidades em que maximizamos a incerteza, fazendo com que apenas “conhecesse” nossas violações das restrições fonológicas. É uma distribuição, então, que *emula* os dados empíricos, pois é conhecida pelo escore harmônico e assemelha-se a uma versão abstrata dos dados. Em outras palavras, se assumirmos que criamos restrições fonológicas adequadas e temos quantidade suficiente de dados, poderíamos pensar que essas são distribuição prototípicas dos fenômenos

analisados, *i.e.*, distribuições de categorias fonológicas sem ruído. Mas essa é uma afirmação ainda especulativa, especialmente nessa etapa de desenvolvimento do método, pois depende do refinamento, testagem e validação dessa metodologia como um todo.

Finalmente, assumindo uma boa prototipicidade das distribuições, a ideia por trás dessa metodologia é a realização de previsões assim como faríamos em um quadro teórico MaxEnt discreto. Iremos derivar probabilidades dessa f.d.p. de máxima entropia que representa os dados transformados pelas violações incorridas sobre eles. Mas é evidente que, como consequência de sua modelagem, a qualidade dessas previsões é inteiramente dependente da qualidade das restrições (e da seleção dos parâmetros específicos de cada modelagem). As restrições de exemplo que discutimos resultaram na curvatura acima. Podemos compará-la com a f.d.p. dos dados:

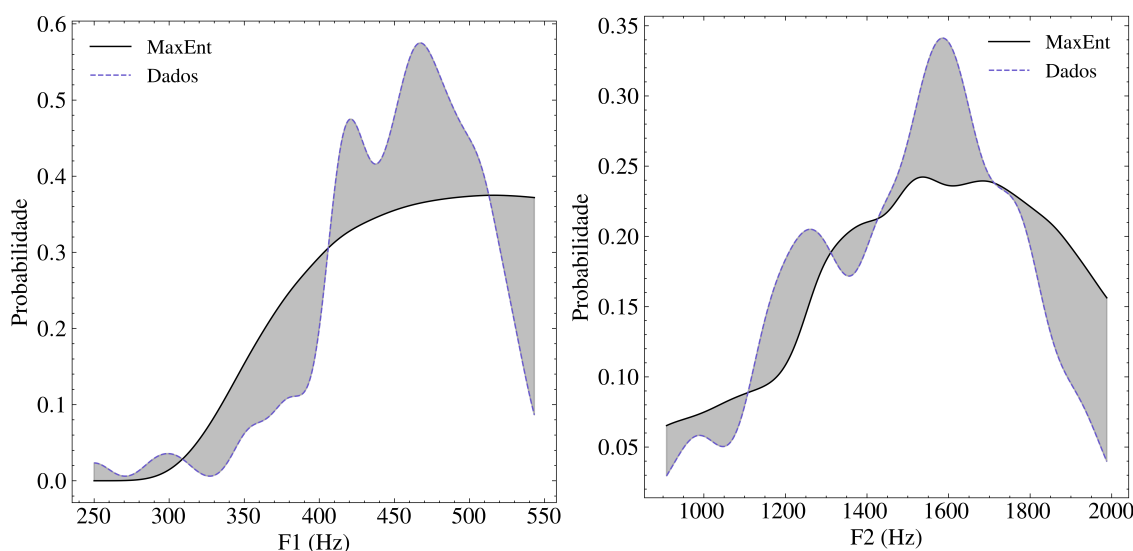


Figura 33: F.d.p. de máxima entropia *versus* f.d.p. dos dados.

Ambas restrições (perceptual e articulatória) requerem um cuidado maior em sua arquitetura. Mas, na condição de demonstração, pode-se dizer com certa confiança que obtivemos algum grau de êxito em nossa proposta (adaptar a HG para lidar com candidatos contínuos). Digo isso pois, mesmo que a distribuição MaxEnt acima não pareça ter capturado muito bem o comportamento dos dados, o quadro teórico matemático/computacional em si funcionou adequadamente e foi plenamente capaz de gerar uma distribuição com base nas violações das restrições. Resta agora ajustar parâmetros, ajustar restrições, refinar aspectos computacionais e realizar experimentos. Bastante coisa. Além disso, podemos deixar aberta a possibilidade de realizar modelagens em que as restrições *opõem-se* aos dados coletados, ao invés de alinhar-se a eles. Ao invés de modelar concordância entre dados e restrições (como neste capítulo), poderíamos

modelar *discordância* entre eles. Poderíamos, por exemplo, utilizar restrições para gerar uma distribuição de referência e analisar como contato dialetal, ao longo dos anos em um grupo de falantes, causa mudanças na produção de vogais específicas – e, relativo à distribuição de referência, como a distribuição das vogais vêm movimentando-se ao longo do tempo. Ao invés de comparar as vogais *entre grupos de falantes*, agora poderíamos compará-las a uma distribuição prototípica gerada pelas restrições. Desse modo, percebemos que há várias maneiras em que essas distribuições podem ser úteis para análise.

Podemos passar, finalmente, para a última etapa de uma análise de HG: o *output*, onde demonstro como podemos tirar as probabilidades dessa f.d.p. do score harmônico, analogamente à gramática MaxEnt tradicional.

3.8 Output

Em uma análise discreta de HG ou OT probabilísticas, o *output* é dado em termos de diferentes probabilidades assinaladas a candidatos. Conforme o princípio que vem guiando nosso quadro teórico, devemos realizar operações em espaço contínuo que são análogas às operações discretas da HG. Logo, devemos agora obter probabilidades com base em nossa modelagem. Com nossa f.d.p. modelada em mãos, basta integrar faixas de formantes (candidatos contínuos) de nosso interesse. Formalmente, as probabilidades serão dadas por:

$$P_{\text{variável}} = \int_a^b f_{\text{MaxEnt}}(x) dx \quad (33)$$

Onde $[a, b]$ é o intervalo analisado, *i.e.*, o candidato. Vou realizar algumas demonstrações com base na f.d.p. que obtivemos na última seção, extraindo probabilidades de F1, de F2 e de F1,F2 ocorrerem conjuntamente. Para obter as probabilidades separadas de F1 e F2, devemos marginalizar a f.d.p. de máxima entropia, como foi feito com a f.d.p. estimada dos dados durante o cômputo de violações. Em relação ao nosso caso específico, podemos definir as seguintes 3 fórmulas de análise por variável nos candidatos:

$$\begin{aligned} P_{F1,F2} &= \int_{a_{F2}}^{b_{F2}} \int_{a_{F1}}^{b_{F1}} f_{\text{MaxEnt}}(F1, F2) dF1 dF2 \\ P_{F1} &= \int_{a_{F1}}^{b_{F1}} \left(\int_{-\infty}^{\infty} f_{\text{MaxEnt}}(F1, F2) dF2 \right) dF1 \\ P_{F2} &= \int_{a_{F2}}^{b_{F2}} \left(\int_{-\infty}^{\infty} f_{\text{MaxEnt}}(F1, F2) dF1 \right) dF2 \end{aligned}$$

Podemos agora elaborar em uma tabela os resultados encontrados pela aplicação

das equações acima. Vamos colocar a faixa de formantes relevante (*i.e.*, o candidato), o escore harmônico e a estabilidade calculados agora com os devidos pesos das restrições e as probabilidades encontradas pelas equações acima. Vejamos:

F1 (Hz)	F2 (Hz)	$\mathcal{H}$	$\mathcal{E}$	$P_{F1,F2}$	P_{F1}	P_{F2}
250 até 543	908 até 1987	22.365	N/A	0.998	0.998	0.999
250 até 350	908 até 1300	11.859	-8.831	0.008	0.057	0.205
350 até 450	1300 até 1700	0.983	-0.612	0.204	0.421	0.476
450 até 543	1700 até 1987	0.574	+0.138	0.163	0.517	0.314

Tabela 21: Probabilidades de formantes, *i.e.*, o *output*.

Na tabela acima, podemos observar a riqueza de informações que candidatos contínuos podem oferecer. Obtivemos o escore harmônico (soma de violações), uma medida de estabilidade (taxa de mudança e direção), predições conjuntas (F1 e F2) e separadas (F1 ou F2). Uma coisa a se notar é que o escore harmônico não corresponde mais *precisamente* com maiores taxas de produção prevista. Apesar disso, ele ainda conserva a tendência geral de que um escore harmônico mais próximo de 0 é atrelado a uma maior chance de produção. Isso nos leva a pensar no papel das derivadas parciais do escore harmônico e se, ao invés de um somatório do gradiente, devemos elaborar uma nova definição matemática para harmonia que intrinsecamente inclua sua taxa de mudança e direcionalidade. Mas isso teria consequências para o processo de otimização, que agora seria mais computacionalmente intensivo. Além disso, precisaríamos pensar nas adaptações matemáticas necessárias em outros elementos que dependem do escore harmônico, *etc.* Por enquanto, a definição análoga à discreta (trabalhada aqui) parece ser adequada, salvo essa única peculiaridade.

Como os candidatos não são discretizados – apenas observado em partes, como discutimos ao longo deste capítulo – podemos também analisar regiões de interseção entre os candidatos da tabela acima:

F1 (Hz)	F2 (Hz)	$\mathcal{H}$	$\mathcal{E}$	$P_{F1,F2}$	P_{F1}	P_{F2}
300 até 400	1200 até 1500	2.184	-1.219	0.071	0.229	0.300
400 até 500	1500 até 1900	0.578	-0.129	0.251	0.522	0.487

Tabela 22: Probabilidades de formantes, *i.e.*, o *output*.

Mas há limitações em quão microscopicamente somos capazes de analisar nossas

variáveis. A tabela abaixo é um exemplo disso:

F1 (Hz)	F2 (Hz)	$\mathcal{H}$	$\mathcal{E}$	$P_{F1,F2}$	P_{F1}	P_{F2}
500.00 até 500.99	1800.00 até 1800.99	0.004	-0.026	0.000	0.005	0.000
500.00 até 500.99	1000 até 1300	0.013	-1.241	0.000	0.005	0.171
500.00 até 500.99	1300 até 1600	0.006	-0.509	0.001	0.005	0.350
500.00 até 500.99	1600 até 1900	0.005	-0.021	0.001	0.005	0.359
300 até 400	1800.00 até 1800.99	1.029	-0.268	0.000	0.229	0.000
400 até 500	1800.00 até 1800.99	0.437	-0.069	0.000	0.522	0.000

Tabela 23: Probabilidades de formantes, *i.e.*, o *output*.

As previsões conjuntas encontram-se entre 0% e 0.1%, em virtude de F1 ou F2 serem tão estreitos que suas previsões individuais tendem a 0%. Como consequência, se um dos formantes teoricamente não há chance de ser realizado, é matematicamente compreensível que conjuntamente isso resulte no fato de que não haveria chance de ocorrer aquela produção. Mas isso é empiricamente incorreto. A tabela acima não pode ser dita verdadeira, com exceção dos valores individuais de P_{F1} ou P_{F2} não-estreitos. Esse problema decorre do funcionamento normal de uma f.d.p., o que nos leva a enquadrar esse fenômeno mais como sendo uma *limitação da metodologia* do que como um *erro* de fato. Como já havia dito anteriormente, ao adotarmos uma abordagem contínua, nós nos sujeitamos a novas limitações metodológicas que requerem as devidas considerações. O que acontece na tabela acima pode ser entendido, em sua versão mais extrema e estreita, da seguinte maneira:

$$\lim_{\epsilon \rightarrow 0} P_{\text{variável}} = \lim_{\epsilon \rightarrow 0} \int_a^{a+\epsilon} f_{\text{MaxEnt}}(x) dx$$

Conforme diminuísimos nosso candidato contínuo e o aproximamos de um candidato *discreto*, a área integrada aproxima-se de 0: como pode ser visto acima, $\epsilon \rightarrow 0$. Isso significa que, por definição⁸², o resultado dessa integral também é 0, pois a área sob a curva é 0. Sendo assim, não é que a probabilidade seja de fato 0 empiricamente, mas sim que não somos capazes de extrair probabilidades de elementos discretos de uma f.d.p. que é contínua. Dessa maneira, nossa modelagem deve exclusivamente tratar de candidatos contínuos e não tentar realizar discretizações. Isso é, admitidamente, uma limitação, mas nossa intenção não era tratar de candidatos discretos em primeiro lugar. A elaboração desse quadro teórico, muito pelo

⁸²Formalmente, podemos entender que a integral é a *área sob uma curva*.

contrário, se justificou por ser uma análise não-discretizante. Sendo assim, essa limitação é algo que devemos ter em mente, porém que não irá afetar nossa análise.

Gostaria de concluir essa seção com algumas afirmações. Primeiramente, essa é uma metodologia experimental. Antes de escrever este protótipo, minha maior preocupação era que essa modelagem não fosse uma abordagem viável. É por isso que me senti compelido a realizar um primeiro modelo – disposto ao longo deste capítulo – que demonstrasse a viabilidade dessa modelagem de candidatos contínuos. Há ainda muito o que ajustar e calibrar, mas espero que eu tenha sido capaz de convencer o leitor de que é *possível* trabalhar com candidatos contínuos ou infinitos. Esse é o objetivo desta dissertação. Em pesquisas posteriores, esta modelagem deve ser refinada a fim de solucionar as muitas questões ainda em aberto dispostas até aqui.

3.9 A hipótese e a plausibilidade cognitiva

Primeiramente, gostaria de explicitar o princípio que vem nos acompanhando ao longo deste capítulo. Isto é, a hipótese que subjaz os princípios matemáticos apresentados neste capítulo: **fenômenos fonéticos representados como *distribuições* oferecem *insights* valiosos para a Fonologia**. Essas distribuições, por sua vez, podem ser enquadradas como candidatos contínuos em um quadro teórico matemático inspirado pela HG. Ao trabalharmos com esses candidatos, somos capazes de criar distribuições que almejam prototipicidade fonológica. Tendo deixado evidente minha hipótese de pesquisa, gostaria de falar um pouco sobre plausibilidade cognitiva, pois é um assunto que cerceia qualquer hipótese fonológica.

Todo modelo de análise fonológica deve estar atento aos pressupostos cognitivos que o sustentam. Caso contrário, nossa metodologia fonológica torna-se uma análise estatística, indistinguível de uma modelagem matemática abstrata e desconectada de elementos biológicos, gramaticais, cognitivos, articulatórios, e assim em diante. Há uma grande diferença entre tomar nota de inferências de análises estatísticas, como são realizadas algumas abordagens quantitativas na Linguística, e realizar modelagens biologicamente plausíveis, como é o caso da Fonologia. Sendo assim, a estrutura do quadro teórico apresentado seguiu elemento por elemento da estrutura da HG. A HG, e outros quadros teóricos da Otimalidade, são metodologias cognitivamente plausíveis, inspiradas por modelos conexionistas da mente (LEGENDRE; MIYATA; SMOLENSKY, 1990a,b). Além disso, ao inaugurar a abordagem MaxEnt na Fonologia, Goldwater e Johnson escrevem o seguinte:

Gradiente Ascendente é um algoritmo de otimização popular, mas não muito eficiente, que pode produzir curvas de aprendizado semelhantes às humanas, embora não tenhamos investigado isso aqui: mais uma vez, os pesos de restrições aos quais ele converge

devem ser os mesmos que os obtidos usando o Gradiente Conjugado. Deixamos para trabalhos futuros a questão de qual algoritmo de otimização melhor modela o caminho do aprendizado humano. (GOLDWATER; JOHNSON, 2003, p. 7, tradução nossa)

Assim, ao estabelecermos uma metodologia, é preciso um olhar crítico não apenas relativo aos resultados encontrados, mas à plausibilidade fonológica da modelagem – como podemos inferir partindo da preocupação dos autores na citação acima. Não basta chegar em pesos das restrições, é preciso também considerar se o caminho da otimização é análogo à cognição humana. Sendo assim, gostaria de dizer que a elaboração desse quadro teórico matemático é apenas uma primeira parte nas considerações metodológicas. Agora, com a testagem e validação desse quadro teórico, é preciso refiná-lo ainda mais para que adeque-se melhor a modelos gramaticais e cognitivos. Mas esse é um esforço que já demos início na própria concepção do método, pois foi pensado em torno de elementos perceptuais e articulatórios de maneira a pavimentar o caminho para diálogos com outras vertentes não-otimalistas da Fonologia e com a Fonética. Por fim, digo que a metodologia deste trabalho é baseada na HG em quesito arquitetura, mas suas restrições foram inspiradas por vertentes gestuais da Fonologia e pelas publicações de Paul Boersma e Edward Flemming, que venho repetidamente fazendo referência ao longo deste trabalho.

4 Implementação computacional

Para viabilizar o tipo de modelagem proposto neste trabalho, acessibilidade é a palavra-chave. Não basta apenas que a metodologia seja matematicamente sólida, pois aspectos de pesquisa também são decididos por limitações práticas. Desse modo, um dos objetivos finais de minha pesquisa foi a confecção de um *software* simples, uma interface visual semelhante ao OTSoft ou MaxEnt de Bruce Hayes, capaz de modelar automaticamente os dados para o usuário. Isso permitiria a modelagem para leigos em matemática ou computação. Esse *software*, apelidado homonimamente de GradiEnt, já encontra-se funcional e seu código pode ser acessado na íntegra em meu GitHub pelo *link*: <https://github.com/alexandre-barroso/GradiEnt/>. Vejamos uma captura de tela do GradiEnt:

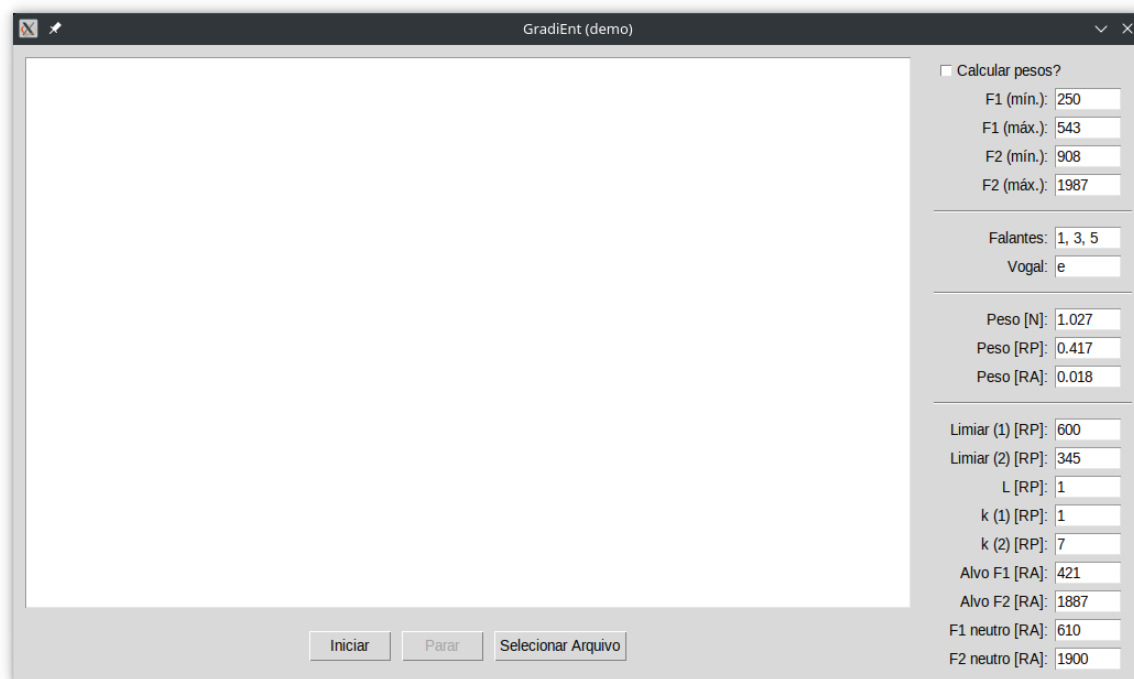


Figura 34: Captura de tela do GradiEnt (1).

Em sua interface visual, todos os parâmetros que discuti no capítulo **Fundamentos matemáticos** podem ser ajustados pelo usuário. Na mesma pasta dos *scripts* do GradiEnt, o usuário deve colocar um arquivo de texto com seus dados e poderá ajustar os parâmetros e candidato conforme necessário. Perceba que essa versão de demonstração⁸³ lida especificamente

⁸³É uma versão considerada “de demonstração” pois sua interface visual faz a modelagem seguindo as restrições apresentadas aqui, que admitidamente existem em condição de *placeholders*. Não há, impedimento nenhum, no entanto de usuários avançados modificarem o *script* com as suas próprias funções matemáticas para elaboração de outras restrições.

com formantes – mas isso não é uma limitação dos fundamentos matemáticos. Futuramente, em uma versão posterior já em desenvolvimento⁸⁴, a interface visual GradiEnt deve ser demonstrada com variáveis contínuas além de formantes. Dito isso, podemos nos focar em algumas partes do GradiEnt para compreender seu uso:

Interface do GradiEnt (2) com os seguintes valores:

F1 (mín.):	250
F1 (máx.):	543
F2 (mín.):	908
F2 (máx.):	1987

Figura 35: Captura de tela do GradiEnt (2).

Esse é o candidato contínuo. Aqui o usuário pode definir a combinação de intervalos analisados. Como dito, essa demonstração ocupa-se com formantes, então as amostras do usuários – nesta versão da interface visual – devem ser formantes F1 e F2.

Interface do GradiEnt (3) com os seguintes valores:

Falantes:	1, 3, 5
Vogal:	e

Figura 36: Captura de tela do GradiEnt (3).

A imagem acima mostra como podemos selecionar falantes e vogais específicas em nosso banco de dados. Isso está relacionado com a maneira como os arquivos de dados devem ser organizados para o GradiEnt. Vejamos um exemplo de arquivo de dados (já pré-processados) contendo informações de 5 falantes e 2 vogais:⁸⁵

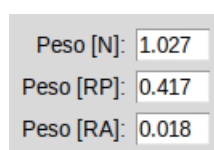
Falante	Vogal	F1	F2
1	o	488.237	1174.454
1	e	424.911	1523.522
2	e	377.647	1643.607
2	o	373.667	911.659
3	o	511.052	1112.049
3	e	473.596	1861.898
4	o	498.117	1388.768
4	e	479.78	1606.562

⁸⁴Em meus estudos de doutoramento.

⁸⁵Arquivo completo disponível no *link* do Github.

5	o	418.017	1282.108
5	e	440.263	1723.166

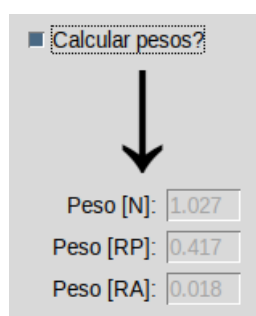
Assim, podemos seleccionar quais falantes (e combinações de falantes) e vogais (e combinações de vogais) queremos analisar de cada vez. Nas análises realizadas no capítulo **Fundamentos matemáticos**, além de seleccionar a vogal “e”, selecionei também falantes 1, 3 e 5, pois são todos do sexo masculino – em comparação com 2 e 4 que são do sexo feminino.



Peso [N]: 1.027
 Peso [RP]: 0.417
 Peso [RA]: 0.018

Figura 37: Captura de tela do GradiEnt (4).

Estes são os pesos, etiquetados [RP] para restrição perceptual e [RA] para restrição articulatória. Se nós já soubermos os pesos das restrições, ou se quisermos manipulá-los, podemos manualmente inseri-los. Por outro lado, podemos optar por rodar o algoritmo de otimização e descobrir os pesos, seleccionando a opção “Calcular pesos?”:



■ Calcular pesos?
 ↓
 Peso [N]: 1.027
 Peso [RP]: 0.417
 Peso [RA]: 0.018

Figura 38: Captura de tela do GradiEnt (5).

O procedimento de otimização pode demorar desde alguns minutos até algumas horas, dependendo da quantidade de dados e dos parâmetros seleccionados. Então, depois de rodá-lo algumas vezes e tendo confirmado o peso das restrições para sua análise, é mais prático inseri-los manualmente e pular a otimização. Por último, temos os parâmetros das restrições:

Limiar (1) [RP]:	600
Limiar (2) [RP]:	345
L [RP]:	1
k (1) [RP]:	1
k (2) [RP]:	7
Alvo F1 [RA]:	421
Alvo F2 [RA]:	1887
F1 neutro [RA]:	610
F2 neutro [RA]:	1900

Figura 39: Captura de tela do GradiEnt (6).

Novamente, [RP] e [RA] indicam a qual restrição o parâmetro pertence. Aqui, vale a pena relembrar as expressões matemáticas de nossas restrições para melhor entender de onde se origina cada parâmetro acima:

Restrição perceptual:

$$R_P(F1) = L - \frac{L^2}{(1+e^{k_1(F1-F1_{\text{limiar1}})})(1+e^{-k_2(F1-F1_{\text{limiar2}})})}$$

Restrição articulatória:

$$R_A(F1, F2) = \sqrt{(C_{F1} - A_{F1})^2 + (C_{F2} - A_{F2})^2} \cdot \frac{\sqrt{(C_{F1} - N_{F1})^2 + (C_{F2} - N_{F2})^2}}{\sqrt{(A_{F1} - N_{F1})^2 + (A_{F2} - N_{F2})^2}}$$

Enfim, o usuário pode utilizar seus próprios dados, filtrá-los e manipular os parâmetros das restrições de maneira conveniente. Para iniciar a análise, basta selecionar o arquivo com os dados (que já deve estar na formatação adequada) e clicar “Iniciar”. Em funcionamento, o usuário acompanha em tempo real os cálculos e modelagem, vide:

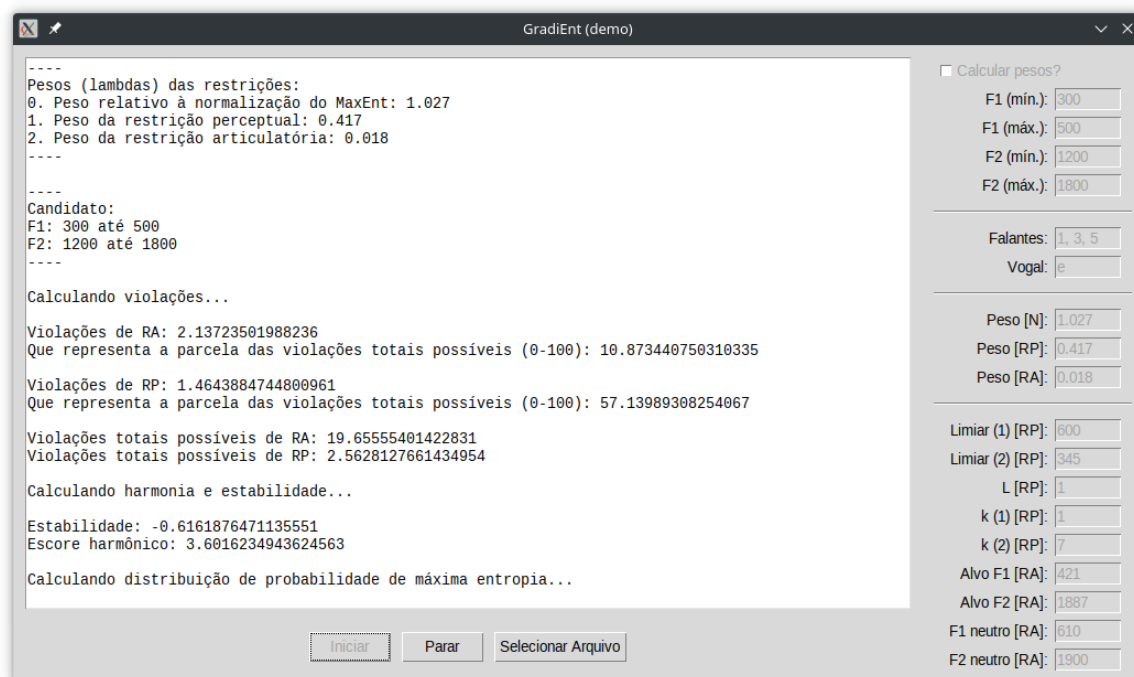


Figura 40: Captura de tela do GradiEnt (7).

Depois de concluir a análise, o programa também gera um arquivo de texto com as informações importantes da análise, além dos gráficos relevantes (vide os gráficos que venho utilizando no último capítulo). Abaixo, coloco um exemplo desse arquivo de relatório gerado pelos parâmetros da imagem acima:

Relatório

Arquivo utilizado: dados.txt

Vogal(is): ['e']

Falante(s): [1, 3, 5]

F1: 300 até 500

F2: 1200 até 1800

Peso da restrição perceptual (RP): 0.417

Peso da restrição articulatória (RA): 0.018

Violações da RP: 1.4643884744800961

Violações da RA: 2.13723501988236

Escore harmônico: 3.6016234943624563

Estabilidade: -0.6161876471135551

Probabilidade F1: 0.7532445074701037

Probabilidade F2: 0.6784355058865299

Probabilidade conjunta F1 e F2: 0.5153641190921973

Divergência KL (modelo vs. dados): 0.3515552133490507

Os *scripts* necessários para rodar o GradiEnt estão em meu GitHub, conforme o *link* citado, e lá é possível também ler maiores instruções sobre como utilizá-lo. Além disso, um recorte do código do *script* dedicado aos cálculos do capítulo **Fundamentos matemáticos** pode ser encontrado no **anexo A** deste trabalho.

Antes de encerrar este capítulo, gostaria de frisar algumas informações importantes. Apesar de funcional, a edição atual do GradiEnt é demonstrativa, pois ela é apenas capaz de analisar F1 e F2 e apenas com as restrições apresentadas no capítulo anterior. Dessa maneira, é uma versão limitada desse tipo de modelagem que, teoricamente, é capaz de analisar variáveis contínuas em geral (*e.g.*, formantes, duração, intensidade, *onset*, e assim em diante) – na condição que as restrições sejam devidamente adaptadas às necessidades da pesquisa. Para uma implementação mais abrangente, seria possível desenvolver uma interface visual que permitisse a manipulação livre das expressões matemáticas das restrições. No entanto, considerando que esta pesquisa apresenta os elementos basilares dessa modelagem/metodologia e constitui sua primeira demonstração sistemática, optou-se por focar nos aspectos fundamentais. A implementação de recursos mais avançados poderá ser desenvolvida em uma pesquisa posterior, após a consolidação dos princípios teóricos aqui apresentados. Vale ressaltar que o formato atual já permite realizar as demonstrações e testagens necessárias para validar os conceitos fundamentais propostos.

5 Considerações finais

Ao longo deste trabalho, tive como intenção apresentar os elementos nucleares de uma possível modelagem fonológica de variáveis fonéticas contínuas. Aproximei-me dessa tarefa por dois ângulos: (i) a elaboração de um quadro teórico matemático condizente com a literatura fonológica relevante; (ii) a confecção de uma interface visual capaz de viabilizar essa modelagem para testagem. Minha motivação foi a escassez de literatura fonológica que lida com variáveis contínuas *sem discretizá-las*. Posso dizer que, em relação ao que pretendi realizar em meu ponto de partida, alcancei marcos importantes, mas que requerem um maior refinamento.

A modelagem apresentada inicia-se em duas etapas de pré-processamento, normalização pelo escore Z e a estimação de densidade por kernel (EDK). A EDK cria a função de densidade de probabilidades (f.d.p.) que representa (estima) a distribuição dos dados coletados. Essa f.d.p. é um conceito chave para a modelagem, pois ela redefine o que compreendemos como candidatos na abordagem otimalista, que agora podem ser analisados em intervalos *enquanto* mantém-se como uma única variável contínua através de toda a análise. Inspirado na Fonologia Funcional e em sua gramática bidirecional (articulação *versus* percepção), foram introduzidas dois tipos de restrições: perceptual (pertencente à família das restrições de fidelidade) e articulatória (pertencente à família das restrições de marcação). A restrição perceptual foi inspirada por conceitos computacionais, como o *perceptron*, enquanto a restrição articulatória é inspirada por abordagens fonéticas motoras. Os conceitos de escore harmônico e *output* permaneceram praticamente idênticos aos da abordagem otimalista discreta, com alterações matemáticas que permitiram o tratamento de variáveis contínuas dentro da modelagem desenvolvida. Como contribuição original, foi introduzido o conceito de *estabilidade* para demonstrar novas aplicações possíveis da Otimalidade em ambiente contínuo. Foi também repensado o cômputo de violações como *interação* entre candidatos e restrições de maneira mais explícita. E, além disso, o quadro de otimização dos pesos foi repensado – inspirado nas otimizações de máxima entropia de Goldwater e Johnson – fazendo uso da entropia diferencial e da divergência de Kullback-Leibler para criar um *output* contínuo de onde podem ser extraídas probabilidades.

Conforme a própria OT, a modelagem apresentada aqui é altamente customizável – as restrições podem ser alteradas tanto com a introdução de diferentes valores quanto com a própria alteração de sua forma e funcionamento. Essa modularidade, que é típica da OT, permite que pesquisas futuras façam alterações em um ou outro elemento sem necessariamente ter que repensar a modelagem matemática como um todo. Os cálculos da estimação da f.d.p. dos dados, das violações, do escore harmônico, da estabilidade, da otimização e da extração de

probabilidades compõem os elementos basilares desse modelo – dessa abordagem otimalista contínua – e já encontram-se prontos para uso. As restrições, por sua vez, devem adequar-se às investigações que as empregam, mas já encontram-se aqui em forma de prova de conceito, apesar de ainda um tanto rudimentares. Por fim, foram elaborados e disponibilizados alguns *scripts* que conjuntamente operam uma interface gráfica para testagem desse modelo, demonstrando (i) reprodutibilidade, (ii) acessibilidade e (iii) viabilidade deste tipo de modelagem.

Essa pesquisa também, como pano de fundo, levanta uma hipótese sobre a constituição das categorias fonológicas como distribuições de probabilidades. Isso é uma consequência lógica da própria modelagem, que procede naturalmente ao inseri-la em um contexto mais amplo sobre suas implicações. No entanto, considerando o recorte e proposta delimitados, discussões sobre aspectos da cognição da gramática fonológica permaneceram em aberto. Ainda assim, é fundamental destacar que esta modelagem, como já discutido anteriormente, não é puramente estatística, pois sua construção parte de fundamentos fonológicos explícitos. A adoção de restrições perceptuais e articulatórias, inspiradas em abordagens da Fonologia, assegura que a modelagem permaneça vinculada a pressupostos gramaticais e não apenas a cálculos probabilísticos. Assim, ainda que discussões mais aprofundadas sobre a cognição fonológica não tenham sido o foco central desta dissertação, o modelo aqui proposto mantém-se ancorado na teoria fonológica e abre caminho para investigações futuras sobre sua plausibilidade cognitiva. E, com certeza, estes são elementos que devem e serão explorados futuramente.

Gostaria de finalizar esta dissertação reiterando que a pesquisa apresentada representa um debate inicial sobre os tópicos discutidos até aqui, estabelecendo bases para investigações futuras. É verdadeiro que diversos aspectos ainda necessitam de reflexões mais profundas, correções e atualizações, mas os resultados obtidos e apresentados indicam um caminho promissor para pesquisas posteriores. A proposta da dissertação – assertar a viabilidade da Otimalidade ao lidar com candidatos contínuos – pode ser dita como tendo sido realizada suficientemente. É, sim, possível trabalhar com candidatos não-discretos no quadro teórico da Teoria da Otimalidade. Pretendo, portanto, dar continuidade a esta modelagem em minha tese de doutorado, expandindo seu quadro teórico, realizando as devidas correções apontadas por mim ao longo da dissertação e abordando desafios empíricos com um maior número de dados.

6 Referências bibliográficas

- BALL, M. J.; RAHILLY, J. **Phonetics: The Science of Speech**. New York: Routledge, 2013.
- BEGUŠ, G. **Unnatural Phonology: A Synchrony-Diachrony Interface Approach**. 2018. Tese (Doutorado em Linguística) – Harvard University, Cambridge, 2018.
- BERTHOZ, A. **The Brain's Sense of Movement**. Cambridge, MA: Harvard University Press, 1997.
- BISHOP, C. M. **Neural Networks for Pattern Recognition**. Oxford: Clarendon Press, 1995.
- BOERSMA, P. **How we learn variation, optionality, and probability**. Institute of Phonetic Sciences, University of Amsterdam, Proceedings 21 (1997), p. 43-58.
- BOERSMA, P. **Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives**. The Hague: Holland Academic Graphics, 1998
- BOERSMA, P. Prototypicality judgments as inverted perception. **Rutgers Optimality Archive**, vol. 742, 2005.
- BOERSMA, P.; CHLÁDKOVÁ, K.; BENDERS, T. Phonological features emerge substance-freely from the phonetics and the morphology. **Canadian Journal of Linguistics/Revue canadienne de linguistique**, vol. 67, n. 4, p. 611-669, 2022.
- BOERSMA, P.; ESCUDERO, P.; HAYES, R. Learning Abstract Phonological from Auditory Phonetic Categories: An Integrated Model for the Acquisition of Language-Specific Sound Categories. In: **Proceedings of the 15th International Congress of Phonetic Sciences**, Barcelona, 2003.
- BYRD, R. H.; LU, P.; NOCEDAL, J. A Limited Memory Algorithm for Bound Constrained Optimization. **SIAM Journal on Scientific Computing**, vol. 16, n. 5, p. 1190-1208, 1995.
- CATFORD, J. C. **A Practical Introduction to Phonetics**. Oxford: Oxford University Press,

2001.

CHLÁDKOVÁ, K.; BOERSMA, P.; BENDERS, T. **The perceptual basis of the feature vowel height**. In: **18th International Congress of Phonetic Sciences**, University of Glasgow, p. 1-5, 2015.

CHOMSKY, N.; HALLE, M. **The Sound Pattern of English**. NY: Harper & Row, 1968.

COVER, T.; THOMAS, J. **Elements of Information Theory**. New Jersey, US: John Wiley & Sons, 2006.

DIEHL, R. L.; KLUENDER, K. R. On the objects of speech perception. **Ecological psychology**, v. 1, n. 2, p. 121-144, 1989.

EGUCHI, S.; COPAS, J. Interpreting Kullback–Leibler divergence with the Neyman–Pearson lemma. **Journal of Multivariate Analysis**, v. 97, p. 2034-2040, 2006.

FLEMMING, E. **Auditory Representations in Phonology**. Tese (Doutorado em Linguística) – University of California, Los Angeles, 1995.

FLEMMING, E. Phonetic Optimization: Compromise in Speech Production. **University of Maryland Working Papers in Linguistics 5: Selected phonology papers from H-OT-97**, 1997.

FLEMMING, E. Scalar and categorical phenomena in a unified model of phonetics and phonology. **Phonology**, vol. 18, p. 7-44, 2001.

FLEMMING, E. Contrast and Perceptual Distinctiveness. In: HAYES, B.; KIRCHNER, R.; STERIADE, D. **Phonetically-Based Phonology**. Cambridge: Cambridge University Press, 2004.

FLEMMING, E. The Grammar of Coarticulation. In: EMBARKI, M.; DODANE, C. **La coarticulation: Des indices à la représentation**. Editions L'Harmattan, 2011.

FLEMMING, E.; CHO, H. The phonetic specification of contour tones: evidence from the Mandarin rising tone. **Phonology**, vol. 34, p. 1-40, 2017.

FLEMMING, E. Comparing MaxEnt and Noisy Harmonic Grammar. **Glossa: a journal of general linguistics**, vol. 6(1), p. 1-42, 2021.

FOWLER, C. A. An event approach to the study of speech perception from a direct-realist perspective. **Journal of Phonetics**, v. 14, n. 1, p. 3-28, 1986.

FOWLER, C. A. Listeners do hear sounds, not tongues. **The Journal of the Acoustical Society of America**, v. 99, n. 3, p. 1730-1741, 1996.

GELFAND, I. M.; FOMIN, S. V. **Calculus of Variations**. Englewood Cliffs, NJ: Prentice-Hall, 1963.

GOLDWATER, S.; JOHNSON, M. **Learning OT Constraint Rankings Using a Maximum Entropy Model**. In: **Proceedings of the Workshop on Variation within Optimality Theory**, Brown University, p. 112-120, 2003.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2a ed. New York: Springer, 2008.

HAYES, B.; WILSON, C. **A Maximum Entropy Model of Phonotactics and Phonotactic Learning**. p. 1-67, ago. 2007. Disponível em <https://linguistics.ucla.edu/people/hayes/Phonotactics/HayesWilsonMaximumEntropyPhonotacticsAugust2007.pdf>. Acesso em: 11 de abril de 2024.

HILLIER, F. S.; LIEBERMAN, G. J. **Introduction to Operations Research**. 9a ed. New York: McGraw-Hill, 2010.

HOFFMAN, L. D.; BRADLEY, G. L. **Calculus: For Business, Economics, and the Social and Life Sciences (Brief Edition)**. 10a ed. New York: McGraw-Hill, 2007.

JOHNSON, K.; FLEMMING, E.; WRIGHT, R. The Hyperspace Effect: Phonetic Targets Are Hyperarticulated. **Language**, vol. 64, no. 3, p. 505-528, 1993.

KAGER, R. **Optimality Theory**. Cambridge, UK: Cambridge University Press, 2004.

KENSTOWICZ, M.; SANDALO, F. Pretonic Vowel Reduction in Brazilian Portuguese: Harmony and Dispersion. **Journal of Portuguese Linguistics**, vol. 15, n. 6, p. 1-19, 2016.

KULLBACK, S.; LEIBLER, R. A. On Information and Sufficiency. **Annals of Mathematical Statistics**, vol. 22, n. 1, p. 79-86, 1951.

LEE, S. H. Contraste das vogais no PB e OT. **ESTUDOS LINGÜÍSTICOS**, vol. 39, n. 1, p. 35-44, 2010.

LEFKOWITZ, L. M. **Maxent Harmonic Grammars and Phonetic Duration**. 2017. Tese (Doutorado em Linguística) – University of California, Los Angeles, 2017.

LEGENDRE, G.; MIYATA, Y.; SMOLENSKY, P. Harmonic Grammar: A Formal Multi-level Connectionist Theory of Linguistic Well-formedness: Theoretical Foundations. **Report: Institute of Cognitive Science**, Boulder, vol. 90-94, 1990.

LEGENDRE, G.; MIYATA, Y.; SMOLENSKY, P. Can Connectionism Contribute to Syntax? Harmonic Grammar, with an Application. **Report: Institute of Cognitive Science**, Boulder, vol. 90-12, 1990.

LIBERMAN, A. M.; MATTINGLY, I. G. The motor theory of speech perception revised. **Cognition**, vol. 21, n. 1, p. 1-36, 1985.

LIEBERMAN, P.; BLUMSTEIN, S. E. **Speech Physiology, Speech Perception, and Acoustic Phonetics**. Cambridge: Cambridge University Press, 1988.

LOBANOV, B. M. Classification of Russian Vowels Spoken by Different Speakers. **The Journal of the Acoustical Society of America**, vol. 49, n. 2B, p. 606-608, 1971.

MADRUGA, M. R. **A fonética e fonologia da harmonia de [ATR] no Português Brasileiro**. 2017. Tese (Doutorado em Linguística) – Universidade Estadual de Campinas, São Paulo, 2017.

MAGRI, G. **Implicational Universals in Stochastic Constraint-Based Phonology**. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, 2018, Bruxelas. Conferência. 2018. p. 1-10.

MAZZONI, D.; REUCHER, E. **Quasi-continuous maximum entropy distribution approximation with kernel density**. In: **Diskussionsbeitrag Nr. 447**, 2010, FernUniversität in Hagen. Conferência. p. 1-17.

MCCULLOCH, W. S.; PITTS, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. **Bulletin of Mathematical Biology**, vol. 52, n. 1/2, p. 99-115, 1990.

MICONI, T. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. **Elife** v. 6, e20899, 23 fev. 2017. ISSN: 2050-084X. DOI: 10.7554/eLife.20899.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient Estimation of Word Representations in Vector Space. In: **Proceedings of the International Conference on Learning Representations**, 2013.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. Cambridge: The MIT Press, 2012.

NEAREY, T. M. Speech perception as pattern recognition. **The Journal of the Acoustical Society of America**, v. 101, n. 6, p. 3241-3254, 1997.

NOCEDAL, J.; WRIGHT, S. J. **Numerical Optimization**. New York: Springer, 1999.

ODDEN, D. **Introducing Phonology**. Cambridge: Cambridge University Press, 2013.

OHALA, J. J. Speech perception is hearing sounds, not tongues. **The Journal of the Acoustical Society of America**, v. 99, n. 3, p. 1718-1725, 1996.

PRINCE, A.; SMOLENSKY, P. Optimality Theory: Constraint Interaction in Generative Grammar. **Technical reports of the Rutgers Center for Cognitive Science**, NJ, vol. 2, 1993.

ROSENBLATT, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. **Psychological Review**, vol. 65, n. 6, p. 386-408, 1957.

ROSENBLATT, F. **Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms**. Washington, D.C.: Spartan Books, 1962.

SCHMIDT, R. A. A schema theory of discrete motor skill learning. **Psychological Review**, vol. 82, n. 4, p. 225-260, 1975.

SHORE, J.; JOHNSON, R. Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. **IEEE Transactions on Information Theory**, vol. 26, n. 1, p. 26-37, 1980.

STEVENS, K. N. The quantal nature of speech: evidence from articulatory-acoustic data. In: DAVID, E. E.; DENES, P. B. (Ed.). **Human communication: a unified view**. New York: McGraw-Hill, 1972. p. 51-66.

A Script para modelagem

```

1 import os
2 import sys
3 import random
4 import logging
5 from datetime import datetime
6 import pandas as pd
7 import numpy as np
8 from scipy.stats import gaussian_kde, ttest_1samp, pearsonr, spearmanr
9 from scipy.integrate import trapz, nquad, quad, dblquad, simpson
10 from scipy.optimize import fsolve, minimize, Bounds
11 from scipy.misc import derivative
12 from scipy.interpolate import interp1d
13 from sklearn.feature_selection import mutual_info_regression
14 from sklearn.gaussian_process import GaussianProcessRegressor
15 from sklearn.gaussian_process.kernels import RBF, ConstantKernel as C, WhiteKernel
16 from sklearn.preprocessing import StandardScaler, MinMaxScaler
17 from sklearn.neighbors import KernelDensity
18 from sklearn.model_selection import KFold
19 import matplotlib.pyplot as plt
20 from mpl_toolkits.mplot3d import Axes3D
21 from sympy import Symbol, integrate
22 from tqdm import tqdm
23 from PIL import Image
24 import scienceplots
25 plt.style.use(['science', 'ieee'])
26 import cProfile
27 import pstats
28 import functools
29 import argparse
30 profiler = cProfile.Profile()
31 profiler.enable()
32
33 if len(sys.argv) == 1:
34     sys.argv.extend([
35         '--alvo_F1=421',
36         '--a_F1=250',
37         '--b_F1=543',
38         '--a_F2=908',
39         '--b_F2=1987',
40         '--limiar_1=600',
41         '--limiar_2=345',
42         '--L=1',
43         '--k_1=1',
44         '--k_2=7',
45         '--alvo_F1=421',
46         '--alvo_F2=1887',
47         '--neutro_F1=610',
48         '--neutro_F2=1900',

```

```

49     '--lambda_zero=1.027',
50     '--lambda_RP=0.417',
51     '--lambda_RA=0.018',
52     '--vogais=e',
53     '--caminho_do_arquivo=dados.txt'
54 ])
55
56
57 def parse_parametros():
58     parser = argparse.ArgumentParser()
59     parser.add_argument('--otimizar', action='store_true', help="Enable optimization if this
60         flag is set")
61     parser.add_argument('--alvo_F1', type=int, default=421)
62     parser.add_argument('--alvo_F2', type=int, default=1887)
63     parser.add_argument('--limiar_1', type=int, default=600)
64     parser.add_argument('--limiar_2', type=int, default=345)
65     parser.add_argument('--neutro_F1', type=int, default=610)
66     parser.add_argument('--neutro_F2', type=int, default=1900)
67     parser.add_argument('--L', type=int, default=1)
68     parser.add_argument('--k_1', type=int, default=1)
69     parser.add_argument('--k_2', type=int, default=7)
70     parser.add_argument('--a_F1', type=int, default=250)
71     parser.add_argument('--b_F1', type=int, default=543)
72     parser.add_argument('--a_F2', type=int, default=908)
73     parser.add_argument('--b_F2', type=int, default=1987)
74     parser.add_argument('--lambda_zero', type=float, default=1.027)
75     parser.add_argument('--lambda_RA', type=float, default=0.018)
76     parser.add_argument('--lambda_RP', type=float, default=0.417)
77     parser.add_argument('--vogais', type=str, default="e")
78     parser.add_argument('--entrevistados', type=str, default="1,3,5")
79     parser.add_argument('--caminho_do_arquivo', type=str, default="dados.txt")
80     args = parser.parse_args()
81     params_todos = vars(args)
82
83     args.vogais = args.vogais.split(',')
84     args.entrevistados = list(map(int, args.entrevistados.split(',')))
85
86     dados_keys = ['alvo_F1', 'alvo_F2', 'limiar_1', 'limiar_2', 'neutro_F1', 'neutro_F2',
87         'L', 'k_1', 'k_2', 'a_F1', 'b_F1', 'a_F2', 'b_F2']
88
89     pesos_keys = ['lambda_zero', 'lambda_RA', 'lambda_RP']
90
91     otimizar = params_todos['otimizar']
92     params_arq = {'entrevistados': args.entrevistados, 'vogais': args.vogais, '
93         caminho_do_arquivo': args.caminho_do_arquivo}
94     params_dados = {k: params_todos[k] for k in dados_keys}
95     params_pesos = {k: params_todos[k] for k in pesos_keys}
96
97     return params_dados, params_pesos, otimizar, params_arq
98
99 def definir_diretorio():

```

```

98     if getattr(sys, 'frozen', False):
99         os.chdir(sys._MEIPASS)
100     else:
101         os.chdir(os.path.dirname(os.path.abspath(__file__)))
102
103 definir_diretorio()
104
105 def ler_dados(caminho_do_arquivo, vogais, entrevistados):
106     dados = pd.read_csv(caminho_do_arquivo, sep=" ", quotechar='"', header=0)
107     dados = dados[(dados['F1'] != 'NA') & (dados['F2'] != 'NA')]
108     dados['F1'] = pd.to_numeric(dados['F1'])
109     dados['F2'] = pd.to_numeric(dados['F2'])
110     candidatos = dados[dados["Vogal"].isin(vogais) & dados["Falante"].isin(entrevistados)].copy()
111     return candidatos[['F1', 'F2']]
112
113 def calcular_kde(dados, largura=0.2):
114     print('\nCalculando valores para a f.d.p. dos dados...')
115     scaler = StandardScaler()
116     valores = np.vstack([dados['F1'], dados['F2']]).T
117     valores_normalizados = scaler.fit_transform(valores)
118     kde = gaussian_kde(valores_normalizados.T, bw_method=largura)
119     limits = [(-np.inf, np.inf), (-np.inf, np.inf)]
120     integral, error = nquad(lambda x, y: kde(np.vstack([x, y])), limits)
121     print(f"\nChecagem do integral da f.d.p.: {integral}")
122     if not np.isclose(integral, 1, atol=1e-3):
123         print('    |')
124         print("    |--> Valores EDK não estão normalizados.")
125     else:
126         print('    |')
127         print("    |--> Valores EDK estão normalizados.")
128     return kde, scaler
129
130 def criar_fdp_marginal(kde, params):
131     min_F1 = params['min_F1']
132     max_F1 = params['max_F1']
133     min_F2 = params['min_F2']
134     max_F2 = params['max_F2']
135     resolucao = params['resolucao']
136     grade_F1 = np.linspace(min_F1, max_F1, resolucao)
137     grade_F2 = np.linspace(min_F2, max_F2, resolucao)
138     F1_mesh, F2_mesh = np.meshgrid(grade_F1, grade_F2)
139     pontos_grade = np.vstack([F1_mesh.ravel(), F2_mesh.ravel()])
140     valores_kde = kde(pontos_grade).reshape(F1_mesh.shape)
141     dx = (max_F1 - min_F1) / (resolucao - 1)
142     dy = (max_F2 - min_F2) / (resolucao - 1)
143     valores_marginalizados_F1 = np.sum(valores_kde, axis=1) * dy
144     integral_marginal_F1 = np.sum(valores_marginalizados_F1) * dx
145     if integral_marginal_F1 != 1:
146         valores_marginalizados_F1 /= integral_marginal_F1
147     funcao_interpolada = interp1d(grade_F1, valores_marginalizados_F1, kind='cubic',

```

```

        fill_value="extrapolate")
148     return funcao_interpolada
149
150 def arquivo_EDK(kde, params, degrau=0.5):
151     limite_F1 = (params['min_F1'], params['max_F1'])
152     limite_F2 = (params['min_F2'], params['max_F2'])
153     with open('amostra_valores.txt', 'w') as file:
154         file.write("F1, F2, Valor EDK\n")
155         for F1 in np.arange(limite_F1[0], limite_F1[1], degrau):
156             for F2 in np.arange(limite_F2[0], limite_F2[1], degrau):
157                 valores = [F1, F2]
158                 valores_kde = kde(valores)[0]
159                 file.write(f"{F1}, {F2}, {valores_kde}\n")
160
161 def integral_fdp(kde, params):
162     fator_superior = 1.5
163     fator_inferior = 1
164     min_F1, max_F1 = params['min_F1'] * fator_inferior, params['max_F1'] * fator_superior
165     min_F2, max_F2 = params['min_F2'] * fator_inferior, params['max_F2'] * fator_superior
166     def integrando(F1, F2):
167         return kde(np.array([F1, F2]))
168     integral, _ = nquad(integrando, [[min_F1, max_F1], [min_F2, max_F2]])
169     return integral
170
171 def verificacao_kde(kde, params, tolerancia=0.02):
172     integral = integral_fdp(kde, params)
173     return np.isclose(integral, 1.0, atol=tolerancia)
174
175 def restricao_articulatoria(F1, F2, kde, params):
176     alvo_F1 = params['alvo_F1']
177     alvo_F2 = params['alvo_F2']
178     neutro_F1 = params['neutro_F1']
179     neutro_F2 = params['neutro_F2']
180     esforco_alvo = np.sqrt((alvo_F1 - neutro_F1) ** 2 + (alvo_F2 - neutro_F2) ** 2)
181     esforco_producao = np.sqrt((F1 - neutro_F1) ** 2 + (F2 - neutro_F2) ** 2)
182     distancia = np.sqrt((F1 - alvo_F1) ** 2 + (F2 - alvo_F2) ** 2)
183     dif_esforco = (esforco_producao + 1e-6) / (esforco_alvo + 1e-6)
184     RA = distancia * dif_esforco
185     return RA
186
187 def integral_ra(kde, params):
188     a_F1 = params['a_F1']
189     b_F1 = params['b_F1']
190     a_F2 = params['a_F2']
191     b_F2 = params['b_F2']
192     funcao_ra = lambda F2, F1: np.exp(restricao_articulatoria(F1, F2, kde, params) - kde([F1,
193     F2]))[0])
194     integral, erro = dblquad(funcao_ra, a_F2, b_F2, lambda F2: a_F1, lambda F2: b_F1)
195     return integral
196 def restricao_perceptual(F1, kde, params):

```

```

197     limiar_1 = params['limiar_1']
198     limiar_2 = params['limiar_2']
199     L = params['L']
200     k_1 = params['k_1']
201     k_2 = params['k_2']
202     produto = (L**2 / ( (1 + np.exp(k_1 * (F1 - limiar_1))) * (1 + np.exp(-k_2 * (F1 -
203         limiar_2))) ) )
204     RP = (L - produto)
205     return RP
206
207 def integral_rp(kde, params, fdp_marginalizada):
208     a_F1 = params['a_F1']
209     b_F1 = params['b_F1']
210     min_F2 = params['a_F2']
211     max_F2 = params['b_F2']
212     integral, erro = quad(lambda F1: np.exp(restricao_perceptual(F1, kde, params) -
213         fdp_marginalizada(F1)) , a_F1, b_F1)
214     return integral
215
216 def escore_estabilidade(kde, scaler, fdp_marginalizada, params, lambda_RA, lambda_RP):
217     a_F1 = params['a_F1']
218     b_F1 = params['b_F1']
219     a_F2 = params['a_F2']
220     b_F2 = params['b_F2']
221     volume = (b_F1 - a_F1) * (b_F2 - a_F2)
222     def funcao_harmonia(F1, F2):
223         RA = np.exp(restricao_articulatoria(F1, F2, kde, params) - kde([F1, F2])[0])
224         RP = np.exp(restricao_perceptual(F1, kde, params) - fdp_marginalizada(F1))
225         return lambda_RA * RA + lambda_RP * RP
226     def derivada_funcao_harmonia(F1, F2):
227         def derivada_F1(F1, F2):
228             h = 1e-5
229             return (funcao_harmonia(F1 + h, F2) - funcao_harmonia(F1, F2)) / h
230         def derivada_F2(F1, F2):
231             h = 1e-5
232             return (funcao_harmonia(F1, F2 + h) - funcao_harmonia(F1, F2)) / h
233         return (derivada_F1(F1, F2) + derivada_F2(F1, F2))
234     integral, erro = dblquad(derivada_funcao_harmonia, a_F1, b_F1, lambda F1: a_F2, lambda F1:
235         b_F2, epsabs=1.0e-3, epsrel=1.0e-3)
236     estabilidade = integral/volume
237     return estabilidade
238
239 def entropia_diferencial(kde, params):
240     a_F1 = params['min_F1']
241     b_F1 = params['max_F1']
242     a_F2 = params['min_F2']
243     b_F2 = params['max_F2']
244     def integrando(F1, F2):
245         fdp = kde([F1, F2])[0]
246         epsilon = 1e-10
247         return -fdp * np.log(fdp + epsilon)

```

```

245     entropia, erro = nquad(integrando, [[a_F1, b_F1], [a_F2, b_F2]])
246     return entropia
247
248 def equacao_maxent(F1, F2, fdp_marginalizada, lambda_zero, lambda_RA, lambda_RP, kde, params):
249     RA = np.exp( restricao_articulatoria(F1, F2, kde, params) - kde([F1, F2])[0] )
250     RP = np.exp( restricao_perceptual(F1, kde, params) - fdp_marginalizada(F1) )
251     formula_maxent = np.exp(-1 - lambda_zero - lambda_RA * RA - lambda_RP * RP)
252     return formula_maxent
253
254 def calculo_maxent(fdp_marginalizada, lambda_zero, lambda_RA, lambda_RP, kde, params):
255     def fdp_maxent(F1, F2):
256         maxent = equacao_maxent(F1, F2, fdp_marginalizada, lambda_zero, lambda_RA, lambda_RP,
257                                 kde, params)
258         return maxent
259     return fdp_maxent
260
261 def arquivo_maxent(fdp_marginalizada, lambda_zero, lambda_RA, lambda_RP, kde, params):
262     min_F1 = params['min_F1']
263     max_F1 = params['max_F1']
264     min_F2 = params['min_F2']
265     max_F2 = params['max_F2']
266     resolucao = params['resolucao']
267     fdp_maxent = calculo_maxent(fdp_marginalizada, lambda_zero, lambda_RA, lambda_RP, kde,
268                                params)
269     valores_F1 = np.linspace(min_F1, max_F1, resolucao)
270     valores_F2 = np.linspace(min_F2, max_F2, resolucao)
271     grade_F1, grade_F2 = np.meshgrid(valores_F1, valores_F2)
272     with open('valores_MaxEnt.txt', 'w') as file:
273         file.write("F1, F2, Valor MaxEnt\n")
274         for i in range(resolucao):
275             for j in range(resolucao):
276                 valor_F1 = grade_F1[i, j]
277                 valor_F2 = grade_F2[i, j]
278                 valor_maxent = fdp_maxent(valor_F1, valor_F2)
279                 file.write(f"{valor_F1}, {valor_F2}, {valor_maxent}\n")
280
281 def fdps_otimizacao(fdp_marginalizada, lambda_zero, lambda_RA, lambda_RP, kde, params):
282     print('\n---> LAMBDA zero: ', lambda_zero)
283     print('\n---> Lambda RA: ', lambda_RA)
284     print('\n---> Lambda RP: ', lambda_RP)
285     print('\n--- INÍCIO DE ITERAÇÃO DA OTIMIZAÇÃO ---')
286     min_F1 = params['min_F1']
287     max_F1 = params['max_F1']
288     min_F2 = params['min_F2']
289     max_F2 = params['max_F2']
290     resolucao = params['resolucao']
291     fdp_maxent = calculo_maxent(fdp_marginalizada, lambda_zero, lambda_RA, lambda_RP, kde,
292                                params)
293     print('OTIMIZAÇÃO: CÁLCULO DA MAXENT CONCLUÍDO')
294     grade_F1 = np.linspace(min_F1, max_F1, resolucao)

```

```

293 grade_F2 = np.linspace(min_F2, max_F2, resolucao)
294 grade_F1, grade_F2 = np.meshgrid(grade_F1, grade_F2)
295 print('OTIMIZAÇÃO: CRIAÇÃO DO GRID')
296 pontos_grade = np.vstack([grade_F1.ravel(), grade_F2.ravel()])
297 print('OTIMIZAÇÃO: STACKING OS PONTOS DAS GRADES EM ARRAY 2D')
298 valores_kde = kde(pontos_grade).reshape(grade_F1.shape)
299 print('OTIMIZAÇÃO: CRIAÇÃO DOS VALORES DA PDF DOS DADOS CONCLUÍDO')
300 valores_maxent = np.zeros_like(grade_F1)
301 print('OTIMIZAÇÃO: INICIALIZAR ARRAY 2D PARA VALORES MAXENT CONCLUÍDO')
302 print('OTIMIZAÇÃO: INÍCIO DO PREENCHIMENTO DO ARRAY COM VALORES MAXENT')
303 for i in range(grade_F1.shape[0]):
304     for j in range(grade_F1.shape[1]):
305         valor_F1 = grade_F1[i, j]
306         valor_F2 = grade_F2[i, j]
307         valores_maxent[i, j] = fdp_maxent(valor_F1, valor_F2)
308 print('OTIMIZAÇÃO: CRIAÇÃO DOS VALORES MAXENT CONCLUÍDO')
309 print('--- FIM DE ITERAÇÃO DA OTIMIZAÇÃO ---')
310 return valores_maxent, valores_kde
311
312 def criar_fdp(fdp_marginalizada, lambda_zero, lambda_RA, lambda_RP, kde, params):
313     min_F1 = params['min_F1']
314     max_F1 = params['max_F1']
315     min_F2 = params['min_F2']
316     max_F2 = params['max_F2']
317     resolucao = params['resolucao']
318     fdp_maxent = calculo_maxent(fdp_marginalizada, lambda_zero, lambda_RA, lambda_RP, kde,
319                                params)
320     grade_F1 = np.linspace(min_F1, max_F1, resolucao)
321     grade_F2 = np.linspace(min_F2, max_F2, resolucao)
322     grade_F1, grade_F2 = np.meshgrid(grade_F1, grade_F2)
323     pontos_grade = np.vstack([grade_F1.ravel(), grade_F2.ravel()])
324     valores_kde = kde(pontos_grade).reshape(grade_F1.shape)
325     valores_maxent = np.zeros_like(grade_F1)
326     for i in range(grade_F1.shape[0]):
327         for j in range(grade_F1.shape[1]):
328             valor_F1 = grade_F1[i, j]
329             valor_F2 = grade_F2[i, j]
330             valores_maxent[i, j] = fdp_maxent(valor_F1, valor_F2)
331     dx = (max_F1 - min_F1) / resolucao
332     dy = (max_F2 - min_F2) / resolucao
333     integral_maxent = np.sum(valores_maxent) * dx * dy
334     print('Integral FDP MaxEnt pré-normalização:', integral_maxent)
335     integral_kde = np.sum(valores_kde) * dx * dy
336     print('Integral FDP dos dados pré-normalização:', integral_kde)
337     if integral_maxent != 1:
338         valores_maxent /= integral_maxent
339     if integral_kde != 1:
340         valores_kde /= integral_kde
341     integral_maxent_posn = np.sum(valores_maxent) * dx * dy
342     print('Integral FDP MaxEnt pós-normalização:', integral_maxent_posn)
343     integral_kde_posn = np.sum(valores_kde) * dx * dy

```

```

343     print('Integral FDP dos dados pós-normalização:', integral_kde_posn)
344     return valores_maxent, valores_kde
345
346 def verificacao_fdps(valores_maxent, valores_kde_dimensionados, params):
347     fator_superior = 1
348     fator_inferior = 1
349     min_F1 = params['min_F1'] * fator_inferior
350     max_F1 = params['max_F1'] * fator_superior
351     min_F2 = params['min_F2'] * fator_inferior
352     max_F2 = params['max_F2'] * fator_superior
353     resolucao = params['resolucao']
354     grade_F1 = np.linspace(min_F1, max_F1, resolucao)
355     grade_F2 = np.linspace(min_F2, max_F2, resolucao)
356     def calcula_integral_e_verifica_normalizacao(valores):
357         integral_F1 = trapz(valores, grade_F1, axis=0)
358         integral_total = trapz(integral_F1, grade_F2)
359         normalizada = np.isclose(integral_total, 1, atol=0.05)
360         return integral_total, normalizada
361     integral_maxent, normalizada_maxent = calcula_integral_e_verifica_normalizacao(
362         valores_maxent)
363     integral_kde, normalizada_kde = calcula_integral_e_verifica_normalizacao(
364         valores_kde_dimensionados)
365     print("\nResultados da Verificação e Comparação das PDFs:")
366     print(f"Integral da MaxEnt: {integral_maxent}, Normalizada: {'Sim' if normalizada_maxent
367         else 'Não'}")
368     print(f"Integral da EDK: {integral_kde}, Normalizada: {'Sim' if normalizada_kde else 'Não
369         '}")
370     return integral_maxent, integral_kde, normalizada_maxent, normalizada_kde
371
372 def kullback_leibler(valores_maxent, valores_kde, params):
373     epsilon = 1e-10
374     valores_maxent = np.clip(valores_maxent, epsilon, None)
375     valores_kde = np.clip(valores_kde, epsilon, None)
376     min_F1 = params['min_F1']
377     max_F1 = params['max_F1']
378     min_F2 = params['min_F2']
379     max_F2 = params['max_F2']
380     resolucao = params['resolucao']
381     grade_F1 = np.linspace(min_F1, max_F1, resolucao)
382     grade_F2 = np.linspace(min_F2, max_F2, resolucao)
383     dx = np.abs(grade_F1[1] - grade_F1[0])
384     dy = np.abs(grade_F2[1] - grade_F2[0])
385     def normalizar_fdp(valores, dx, dy):
386         if valores.ndim == 2:
387             soma_integral = np.trapz(np.trapz(valores, dx=dx, axis=1), dx=dy)
388         elif valores.ndim == 1:
389             soma_integral = np.trapz(valores, dx=dx)
390         else:
391             raise ValueError("Erro: o input deve ser 1D ou 2D")
392         return valores / soma_integral
393     valores_maxent_normalizados = normalizar_fdp(valores_maxent, dx, dy)

```

```

390 valores_kde_normalizados = normalizar_fdp(valores_kde, dx, dy)
391 def verificar_integral_fdp_normalizada(valores, dx, dy, nome_fdp):
392     if valores.ndim == 2:
393         integral = np.sum(np.trapz(np.trapz(valores, dx=dx, axis=1), dx=dy))
394     elif valores.ndim == 1:
395         integral = np.trapz(valores, dx=dx)
396     print(f"Integral da {nome_fdp} normalizada: {integral}")
397 verificar_integral_fdp_normalizada(valores_maxent_normalizados, dx, dy, "MaxEnt")
398 verificar_integral_fdp_normalizada(valores_kde_normalizados, dx, dy, "EDK")
399 p_log_p = valores_kde_normalizados * np.log(valores_kde_normalizados)
400 p_log_q = valores_kde_normalizados * np.log(valores_maxent_normalizados)
401 def integral_kl(valores, dx, dy):
402     integral_intermediaria = np.trapz(valores, dx=dx, axis=1)
403     return np.trapz(integral_intermediaria, dx=dy)
404 integral_p_log_p = integral_kl(p_log_p, dx, dy)
405 integral_p_log_q = integral_kl(p_log_q, dx, dy)
406 divergencia_kl = integral_p_log_p - integral_p_log_q
407 return divergencia_kl
408
409 class OtimizadorKL:
410     def __init__(self, kde, params_normalizados_ref):
411         self.kde = kde
412         self.params = params_normalizados_ref
413         self.contador_zero = {'Lambda_zero': 0, 'RA': 0, 'RP': 0}
414     def funcao_objetivo(self, lambdas):
415         print('Contador zero:', self.contador_zero)
416         min_F1 = self.params['min_F1']
417         max_F1 = self.params['max_F1']
418         min_F2 = self.params['min_F2']
419         max_F2 = self.params['max_F2']
420         resolucao = self.params['resolucao']
421         lambda_zero, peso_RA, peso_RP = lambdas[0], lambdas[1], lambdas[2]
422         epsilon = 1e-4
423         limiar_zero = 1
424         if lambda_zero <= 0:
425             self.contador_zero['Lambda_zero'] += 1
426             if self.contador_zero['Lambda_zero'] >= limiar_zero:
427                 lambda_zero = epsilon
428         else:
429             self.contador_zero['Lambda_zero'] = 0
430         if peso_RA <= 0:
431             self.contador_zero['RA'] += 1
432             if self.contador_zero['RA'] >= limiar_zero:
433                 peso_RA = epsilon
434         else:
435             self.contador_zero['RA'] = 0
436         if peso_RP <= 0:
437             self.contador_zero['RP'] += 1
438             if self.contador_zero['RP'] >= limiar_zero:
439                 peso_RP = epsilon
440         else:

```

```

441         self.contador_zero['RP'] = 0
442         valores_maxent, valores_kde = fdps_otimizacao(fdp_marginalizada, lambda_zero, peso_RA,
443             peso_RP, kde, self.params)
444         print('\nCalculando divergência KL')
445         divergencia_kl = kullback_leibler(valores_maxent, valores_kde, self.params)
446         print('\n----> Divergência KL:', divergencia_kl)
447         return divergencia_kl
448
449 def extrair_probabilidades(valores_maxent, params, scaler):
450     print('\n----')
451     min_F1, max_F1 = params['min_F1'], params['max_F1']
452     min_F2, max_F2 = params['min_F2'], params['max_F2']
453     a_F1, a_F2 = params['a_F1'], params['a_F2']
454     b_F1, b_F2 = params['b_F1'], params['b_F2']
455     resolucao = params['resolucao']
456     grade_F1 = np.linspace(min_F1, max_F1, resolucao)
457     grade_F2 = np.linspace(min_F2, max_F2, resolucao)
458     dx = grade_F1[1] - grade_F1[0]
459     dy = grade_F2[1] - grade_F2[0]
460     marginal_F1_fdp = np.trapz(valores_maxent, dx=dy, axis=1)
461     marginal_F2_fdp = np.trapz(valores_maxent, dx=dx, axis=0)
462     marginal_F1_fdp_normalized = marginal_F1_fdp / np.trapz(marginal_F1_fdp, dx=dx)
463     marginal_F2_fdp_normalized = marginal_F2_fdp / np.trapz(marginal_F2_fdp, dx=dy)
464     indices_F1 = np.logical_and(grade_F1 >= a_F1, grade_F1 <= b_F1)
465     indices_F2 = np.logical_and(grade_F2 >= a_F2, grade_F2 <= b_F2)
466     relevant_valores_maxent = valores_maxent[np.ix_(indices_F1, indices_F2)]
467     prob_conjunta = np.trapz(np.trapz(relevant_valores_maxent, dx=dx, axis=1), dx=dy)
468     prob_F1 = np.trapz(marginal_F1_fdp_normalized[indices_F1], dx=dx)
469     prob_F2 = np.trapz(marginal_F2_fdp_normalized[indices_F2], dx=dy)
470     print(f"Probabilidade conjunta F1, F2: {prob_conjunta:.9f}")
471     print(f"Probabilidade F1: {prob_F1:.12f}")
472     print(f"Probabilidade F2: {prob_F2:.12f}")
473     print('----')
474     return prob_conjunta, prob_F1, prob_F2
475
476 params_dados, params_pesos, otimizar, params_arq = parse_parametros()
477 entrevistados = params_arq['entrevistados']
478 vogais = params_arq['vogais']
479 caminho_do_arquivo = params_arq['caminho_do_arquivo']
480
481 candidatos = ler_dados(caminho_do_arquivo, vogais, entrevistados)
482
483 print('\nArquivo lido com sucesso:', caminho_do_arquivo)
484
485 print('\nFalante(s) selecionado(s):', entrevistados)
486
487 print('\nVogal(is) selecionada(s):', vogais)
488
489 largura_customizada = 0.15
490 largura_scott = 'scott'

```

```

491 largura_silverman = 'silverman'
492 kde, scaler = calcular_kde(candidatos)
493
494 print('\nOs valores para a f.d.p. foram estimados por EDK com sucesso.')
495
496 params_dados['resolucao'] = 1000
497 params_dados['min_F1'] = min(candidatos['F1'])
498 params_dados['max_F1'] = max(candidatos['F1'])
499 params_dados['min_F2'] = min(candidatos['F2'])
500 params_dados['max_F2'] = max(candidatos['F2'])
501
502 params_normalizados = {
503     'resolucao': params_dados['resolucao'],
504     'alvo_F1': scaler.transform([[params_dados['alvo_F1'], 0]])[0][0],
505     'limiar_1': scaler.transform([[params_dados['limiar_1'], 0]])[0][0],
506     'limiar_2': scaler.transform([[params_dados['limiar_2'], 0]])[0][0],
507     'neutro_F1': scaler.transform([[params_dados['neutro_F1'], 0]])[0][0],
508     'neutro_F2': scaler.transform([[params_dados['neutro_F2'], 0]])[0][0],
509     'a_F1': scaler.transform([[params_dados['a_F1'], 0]])[0][0],
510     'b_F1': scaler.transform([[params_dados['b_F1'], 0]])[0][0],
511     'min_F1': scaler.transform([[params_dados['min_F1'], 0]])[0][0],
512     'max_F1': scaler.transform([[params_dados['max_F1'], 0]])[0][0],
513     'alvo_F2': scaler.transform([[0, params_dados['alvo_F2']]])[0][1],
514     'a_F2': scaler.transform([[0, params_dados['a_F2']]])[0][1],
515     'b_F2': scaler.transform([[0, params_dados['b_F2']]])[0][1],
516     'min_F2': scaler.transform([[0, params_dados['min_F2']]])[0][1],
517     'max_F2': scaler.transform([[0, params_dados['max_F2']]])[0][1],
518     'L': params_dados['L'],
519     'k_1': params_dados['k_1'],
520     'k_2': params_dados['k_2'],
521 }
522 params_ref = {
523     'resolucao': params_dados['resolucao'],
524     'alvo_F1': params_dados['alvo_F1'],
525     'alvo_F2': params_dados['alvo_F2'],
526     'limiar_1': params_dados['limiar_1'],
527     'limiar_2': params_dados['limiar_2'],
528     'neutro_F1': params_dados['neutro_F1'],
529     'neutro_F2': params_dados['neutro_F2'],
530     'L': params_dados['L'],
531     'k_1': params_dados['k_1'],
532     'k_2': params_dados['k_2'],
533     'a_F1': candidatos['F1'].min(),
534     'b_F1': candidatos['F1'].max(),
535     'a_F2': candidatos['F2'].min(),
536     'b_F2': candidatos['F2'].max(),
537     'min_F1': candidatos['F1'].min(),
538     'max_F1': candidatos['F1'].max(),
539     'min_F2': candidatos['F2'].min(),
540     'max_F2': candidatos['F2'].max(),
541 }

```

```

542 params_normalizados_ref = {
543     'resolucao':1000,
544     'alvo_F1': scaler.transform([[params_ref['alvo_F1'], 0]])[0][0],
545     'limiar_1': scaler.transform([[params_ref['limiar_1'], 0]])[0][0],
546     'limiar_2': scaler.transform([[params_ref['limiar_2'], 0]])[0][0],
547     'neutro_F1': scaler.transform([[params_ref['neutro_F1'], 0]])[0][0],
548     'neutro_F2': scaler.transform([[params_ref['neutro_F2'], 0]])[0][0],
549     'a_F1': scaler.transform([[params_ref['a_F1'], 0]])[0][0],
550     'b_F1': scaler.transform([[params_ref['b_F1'], 0]])[0][0],
551     'min_F1': scaler.transform([[params_ref['min_F1'], 0]])[0][0],
552     'max_F1': scaler.transform([[params_ref['max_F1'], 0]])[0][0],
553     'alvo_F2': scaler.transform([[0, params_ref['alvo_F2']]])[0][1],
554     'a_F2': scaler.transform([[0, params_ref['a_F2']]])[0][1],
555     'b_F2': scaler.transform([[0, params_ref['b_F2']]])[0][1],
556     'min_F2': scaler.transform([[0, params_ref['min_F2']]])[0][1],
557     'max_F2': scaler.transform([[0, params_ref['max_F2']]])[0][1],
558     'L': params_ref['L'],
559     'k_1': params_ref['k_1'],
560     'k_2': params_ref['k_2'],
561 }
562
563 arquivo_EDK(kde, params_normalizados_ref)
564
565 print('\nArquivo com amostra de valores normalizados e da f.d.p. foi criado.')
566
567 fdp_marginalizada = criar_fdp_marginal(kde, params_normalizados_ref)
568
569 print('\nF.d.p. marginalizada foi criada.')
570
571 entropia = entropia_diferencial(kde, params_normalizados_ref)
572
573 SLSQP = 'SLSQP'
574 BFGS = 'BFGS'
575 COBYLA = 'COBYLA'
576 L_BFGS_B = 'L-BFGS-B'
577
578 otimizador = OtimizadorKL(kde, params_normalizados_ref)
579 lambdas_iniciais = [1, 1, 1]
580 limites = Bounds([1e-8, 1e-8, 1e-8], [np.inf, np.inf, np.inf])
581
582 if otimizar:
583     print('\nIniciando otimização...')
584     otimizacao = minimize(otimizador.funcao_objetivo, lambdas_iniciais, method='L-BFGS-B',
585                          bounds=limites, options={'maxiter': 1000})
586     lambdas_otimizados = otimizacao.x
587     lambda_zero = lambdas_otimizados[0]
588     lambda_RA = lambdas_otimizados[1]
589     lambda_RP = lambdas_otimizados[2]
590     print('\nOtimização concluída com sucesso.')
591 else:

```

```

592     lambda_zero = params_pesos['lambda_zero']
593     lambda_RA = params_pesos['lambda_RA']
594     lambda_RP = params_pesos['lambda_RP']
595     print('\nValores de lambda foram especificados.')
596
597     print('\n----')
598     print('Pesos (lambdas) das restrições:')
599     print('0. Peso relativo à normalização do MaxEnt:', lambda_zero)
600     print('1. Peso da restrição perceptual:', lambda_RP)
601     print('2. Peso da restrição articulatória:', lambda_RA)
602     print('----')
603
604
605     valor_integral_fdp = integral_fdp(kde, params_normalizados_ref)
606     violacoes_ra = integral_ra(kde, params_normalizados)
607     violacoes_ra_pesadas = violacoes_ra*lambda_RA
608     violacoes_rp = integral_rp(kde, params_normalizados, fdp_marginalizada)
609     violacoes_rp_pesadas = violacoes_rp*lambda_RP
610     violacoes_ra_total = integral_ra(kde, params_normalizados_ref)
611     violacoes_ra_total_pesadas = violacoes_ra_total*lambda_RA
612     violacoes_rp_total = integral_rp(kde, params_normalizados_ref, fdp_marginalizada)
613     violacoes_rp_total_pesadas = violacoes_rp_total*lambda_RP
614     violacoes_ra_normalizado = violacoes_ra_pesadas * 100/violacoes_ra_total_pesadas
615     violacoes_rp_normalizado = violacoes_rp_pesadas * 100/violacoes_rp_total_pesadas
616
617     print('\n----')
618     print('Candidato:')
619     print('F1:', params_dados['a_F1'], 'até', params_dados['b_F1'])
620     print('F2:', params_dados['a_F2'], 'até', params_dados['b_F2'])
621     print('----')
622
623     print('\nCalculando violações...')
624
625     print('\nViolações de RP:', violacoes_rp_pesadas)
626     print('Que representa a parcela das violações totais possíveis (0-100):',
627           violacoes_rp_normalizado)
627     print('\nViolações de RA:', violacoes_ra_pesadas)
628     print('Que representa a parcela das violações totais possíveis (0-100):',
629           violacoes_ra_normalizado)
629
630     print('\nViolações totais possíveis de RA:', violacoes_ra_total_pesadas)
631     print('Violações totais possíveis de RP:', violacoes_rp_total_pesadas)
632
633     estabilidade = escore_estabilidade(kde, scaler, fdp_marginalizada, params_normalizados,
634                                       lambda_RA, lambda_RP)
634     soma_violacoes = violacoes_ra_pesadas + violacoes_rp_pesadas
635     soma_violacoes_norm = (violacoes_ra_pesadas + violacoes_rp_pesadas) * 100 / (
636         violacoes_ra_total_pesadas + violacoes_rp_total_pesadas)
636
637     print('\nCalculando harmonia e estabilidade...')
638     print('\nEstabilidade:', estabilidade)

```

```

639 print('Escore harmônico:', soma_violacoes)
640
641 print('\nCalculando distribuição de probabilidade de máxima entropia...')
642
643 f_maxent_normalizado = calculo_maxent(fdp_marginalizada, lambda_zero, lambda_RA, lambda_RP,
    kde, params_normalizados_ref)
644 fdp_maxent_formatada, fdp_dados_formatada = criar_fdp(fdp_marginalizada, lambda_zero,
    lambda_RA, lambda_RP, kde, params_normalizados_ref)
645 verificacao_fdp(fdp_maxent_formatada, fdp_dados_formatada, params_normalizados_ref)
646
647 print('\nDistribuição de probabilidade de máxima entropia criada e normalizada.')
648
649 divergencia_kl = kullback_leibler(fdp_maxent_formatada, fdp_dados_formatada,
    params_normalizados_ref)
650
651 print("\nValor da divergência de Kullback-Leibler:", divergencia_kl)
652
653 prob_conjunta, prob_F1, prob_F2 = extrair_probabilidades(fdp_maxent_formatada,
    params_normalizados, scaler)
654
655 def gerar_relatorio(caminho_do_arquivo, vogais, entrevistados, largura_customizada,
    valor_integral_fdp,
656                     violacoes_ra, violacoes_rp, violacoes_ra_total, violacoes_rp_total,
657                     violacoes_ra_normalizado, violacoes_rp_normalizado, estabilidade,
    soma_violacoes,
658                     entropia, divergencia_kl, prob_conjunta, prob_F1, prob_F2, a_F1, a_F2,
    b_F1, b_F2,
659                     lambda_zero, lambda_RA, lambda_RP):
660
661     with open('relatorio.txt', 'w', encoding='utf-8') as arquivo:
662         arquivo.write("Relatório\n")
663         arquivo.write("-----\n\n")
664         arquivo.write(f"Arquivo utilizado: {caminho_do_arquivo}\n\n")
665         arquivo.write(f"Vogal(is): {vogais}\n")
666         arquivo.write(f"Falante(s): {entrevistados}\n\n")
667         arquivo.write(f"F1: {a_F1} até {b_F1}\n")
668         arquivo.write(f"F2: {a_F2} até {b_F2}\n\n")
669         arquivo.write(f"Peso da restrição perceptual (RP): {lambda_RP}\n")
670         arquivo.write(f"Peso da restrição articulatória (RA): {lambda_RA}\n\n")
671         arquivo.write(f"Violações da RP: {violacoes_rp}\n")
672         arquivo.write(f"Violações da RA: {violacoes_ra}\n\n")
673         arquivo.write(f"Escore harmônico: {soma_violacoes}\n")
674         arquivo.write(f"Estabilidade: {estabilidade}\n\n")
675         arquivo.write(f"Probabilidade F1: {prob_F1}\n")
676         arquivo.write(f"Probabilidade F2: {prob_F2}\n")
677         arquivo.write(f"Probabilidade conjunta F1 e F2: {prob_conjunta}\n\n")
678         arquivo.write(f"Divergência KL (modelo vs. dados): {divergencia_kl}\n\n")
679
680     gerar_relatorio(caminho_do_arquivo, vogais, entrevistados, largura_customizada,
        valor_integral_fdp,
681                    violacoes_ra_pesadas, violacoes_rp_pesadas, violacoes_ra_total,

```

```
violacoes_rp_total,  
682         violacoes_ra_normalizado, violacoes_rp_normalizado, estabilidade,  
soma_violacoes,  
683         entropia, divergencia_kl, prob_conjunta, prob_F1, prob_F2, params_dados['a_F1'  
], params_dados['a_F2'],  
684         params_dados['b_F1'], params_dados['b_F2'], lambda_zero, lambda_RA,  
lambda_RP)  
685  
686 print('\nArquivo de relatório gerado e análise concluída.')  
687  
688 if __name__ == "__main__":  
689     args = parse_pesos()  
690     print(f"Lambda Zero: {args.lambda_zero}")  
691     print(f"Lambda RA: {args.lambda_RA}")  
692     print(f"Lambda RP: {args.lambda_RP}")
```

B Dados utilizados

Neste anexo, apresento os dados utilizados. Em meu Github, <https://github.com/alexandre-barroso/GradiEnt/>, poderão também ser encontrados dados da vogal [o] pretônica para todos os falantes, além de 2 outros falantes (2 e 4) do sexo feminino não considerados nesta pesquisa. Vide abaixo um resumo dos dados utilizados nesta dissertação:

Falante	Número de Dados	Origem
1	48	Maranhão
3	32	Ceará
5	34	Maranhão

Tabela 24: Sobre os dados.

E abaixo encontram-se os dados listados:

```

1 Falante Vogal F1 F2
2 1 e 424.911 1523.522
3 1 e 440.76 1391.502
4 1 e 472.388 1708.746
5 1 e 516.169 1489.957
6 1 e 438.913 1550.124
7 1 e 310.483 1691.094
8 1 e 530.1 1535.953
9 1 e 390.512 1844.246
10 1 e 268.621 1596.12
11 1 e 270.469 1972.785
12 1 e 399.822 1557.334
13 1 e 487.313 1638.883
14 1 e 407.285 1806.207
15 1 e 444.456 1754.742
16 1 e 365.352 1474.045
17 1 e 313.255 1535.207
18 1 e 428.678 1768.913
19 1 e 419.367 1364.9
20 1 e 426.83 1372.11
21 1 e 386.816 1699.796
22 1 e 391.436 1750.515
23 1 e 334.648 1541.422
24 1 e 471.464 1744.299
25 1 e 350.498 1529.737
26 1 e 497.548 908.675
27 1 e 401.67 1870.849
28 1 e 328.181 1679.409
29 1 e 319.794 1864.633
30 1 e 477.079 1691.84

```

31 1 e 267.626 1579.462
32 1 e 343.959 1504.129
33 1 e 250.0 1606.811
34 1 e 349.574 1642.364
35 1 e 437.065 1929.275
36 1 e 543.177 1638.883
37 1 e 449.147 1665.237
38 1 e 383.049 1642.364
39 1 e 443.532 1924.054
40 1 e 476.155 1607.805
41 1 e 496.624 1582.943
42 1 e 490.085 1317.91
43 1 e 455.686 1535.207
44 1 e 450.071 1485.482
45 1 e 437.989 1739.824
46 1 e 410.057 1504.875
47 1 e 410.981 1281.611
48 1 e 469.616 1766.427
49 1 e 454.762 1664.491
50 3 e 473.596 1861.898
51 3 e 406.006 1712.227
52 3 e 491.365 1886.761
53 3 e 350.711 1282.854
54 3 e 318.799 1061.579
55 3 e 289.73 1785.82
56 3 e 342.253 1775.626
57 3 e 364.783 1801.731
58 3 e 338.486 1822.119
59 3 e 466.063 1804.964
60 3 e 436.994 1711.481
61 3 e 425.764 1705.763
62 3 e 341.329 1571.506
63 3 e 527.967 1699.299
64 3 e 503.589 1748.526
65 3 e 387.313 1673.193
66 3 e 276.581 1882.286
67 3 e 422.921 1838.776
68 3 e 343.177 1423.575
69 3 e 416.382 1855.434
70 3 e 481.059 1874.578
71 3 e 425.764 1630.927
72 3 e 526.119 1918.087
73 3 e 317.875 1744.051
74 3 e 454.833 1833.804
75 3 e 484.826 1558.578
76 3 e 489.517 1506.118
77 3 e 391.009 1635.402
78 3 e 512.971 1628.441
79 3 e 481.983 1651.314
80 3 e 336.638 1615.015
81 3 e 461.372 1806.953

82 5 e 372.246 1516.311
83 5 e 440.263 1723.166
84 5 e 395.7 1670.21
85 5 e 446.162 1574.738
86 5 e 439.126 1548.384
87 5 e 358.173 1932.508
88 5 e 315.956 1117.022
89 5 e 418.017 1405.674
90 5 e 493.07 1779.355
91 5 e 450.853 1753.747
92 5 e 514.179 1695.818
93 5 e 416.809 1811.428
94 5 e 437.918 1702.282
95 5 e 356.965 1847.478
96 5 e 327.683 1860.407
97 5 e 331.166 1783.333
98 5 e 361.656 1853.942
99 5 e 490.725 1567.528
100 5 e 457.889 1611.037
101 5 e 408.635 1737.587
102 5 e 428.536 1724.907
103 5 e 421.5 1926.043
104 5 e 389.872 1497.167
105 5 e 373.383 1987.702
106 5 e 469.616 1501.145
107 5 e 436.78 1658.276
108 5 e 423.845 1691.094
109 5 e 389.872 1705.514
110 5 e 448.507 1937.231
111 5 e 386.318 1512.334
112 5 e 356.965 1783.333
113 5 e 364.001 1810.682
114 5 e 310.057 1890.739
115 5 e 425.053 1521.035