Capítulo 2

Teorias e modelos prosódicos

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

2.1 Quanto à Separação entre Segmentos e Prosódia

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Quanto à Melodia

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

2.2.1 O modelo de Pierrehumbert

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

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

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

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

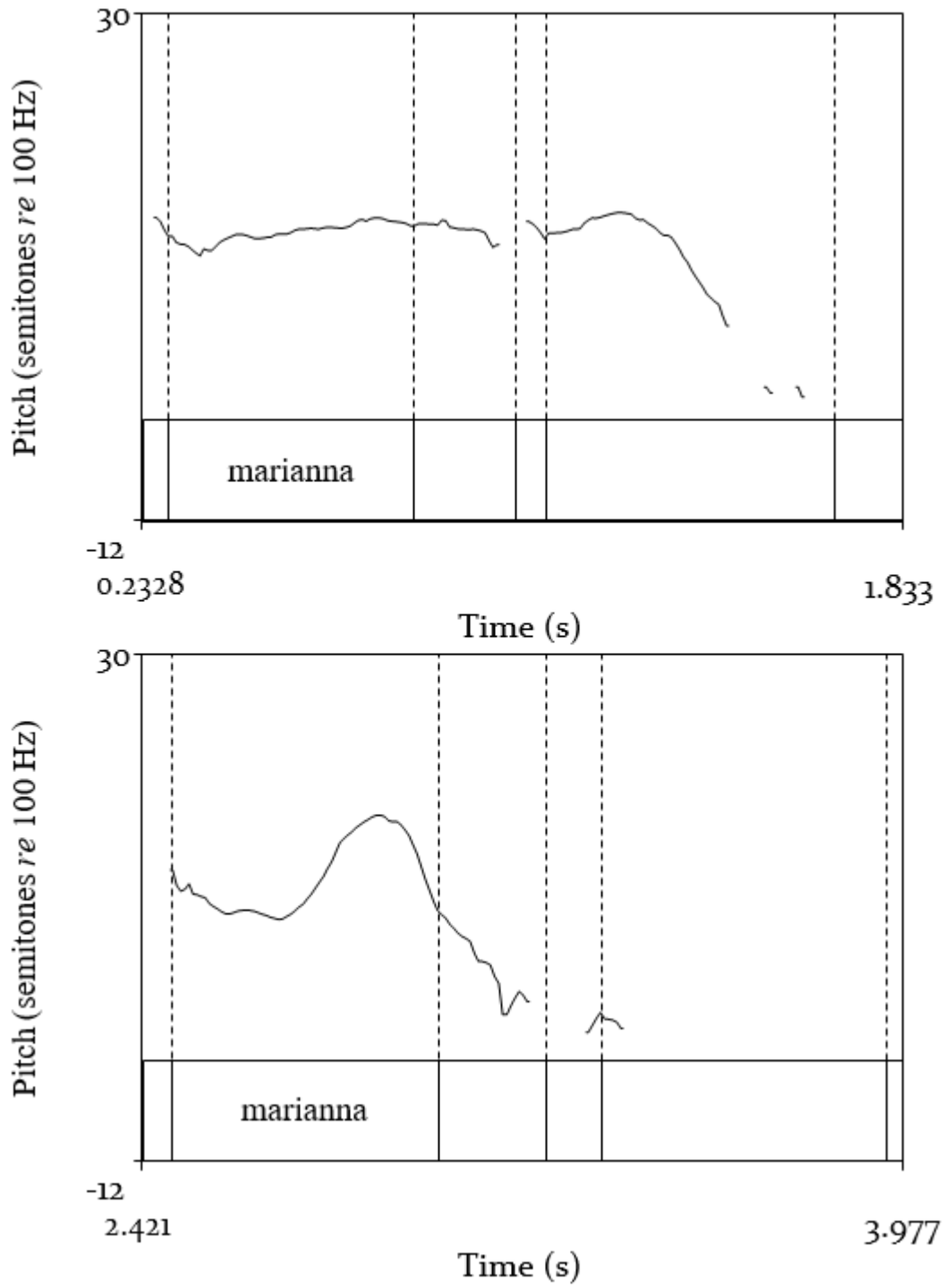

Figura 2.1 – Contraste entre dois enunciados correspondentes à sentença “Marianna made the marmalade” da oficina de aprendizado do ToBI, apenas com a marcação da palavra “Marianna”. A de cima é o enunciado neutro (tom H* na tônica de “Marianna”), a de baixo o enunciado com foco em “Marianna” (tom L+H*).

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

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

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

2.2.2 O Sistema DaTo de Notação Entoacional

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

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

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

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

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

tulo 5.

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

2.2.3 O Modelo de Fujisaki

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

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

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

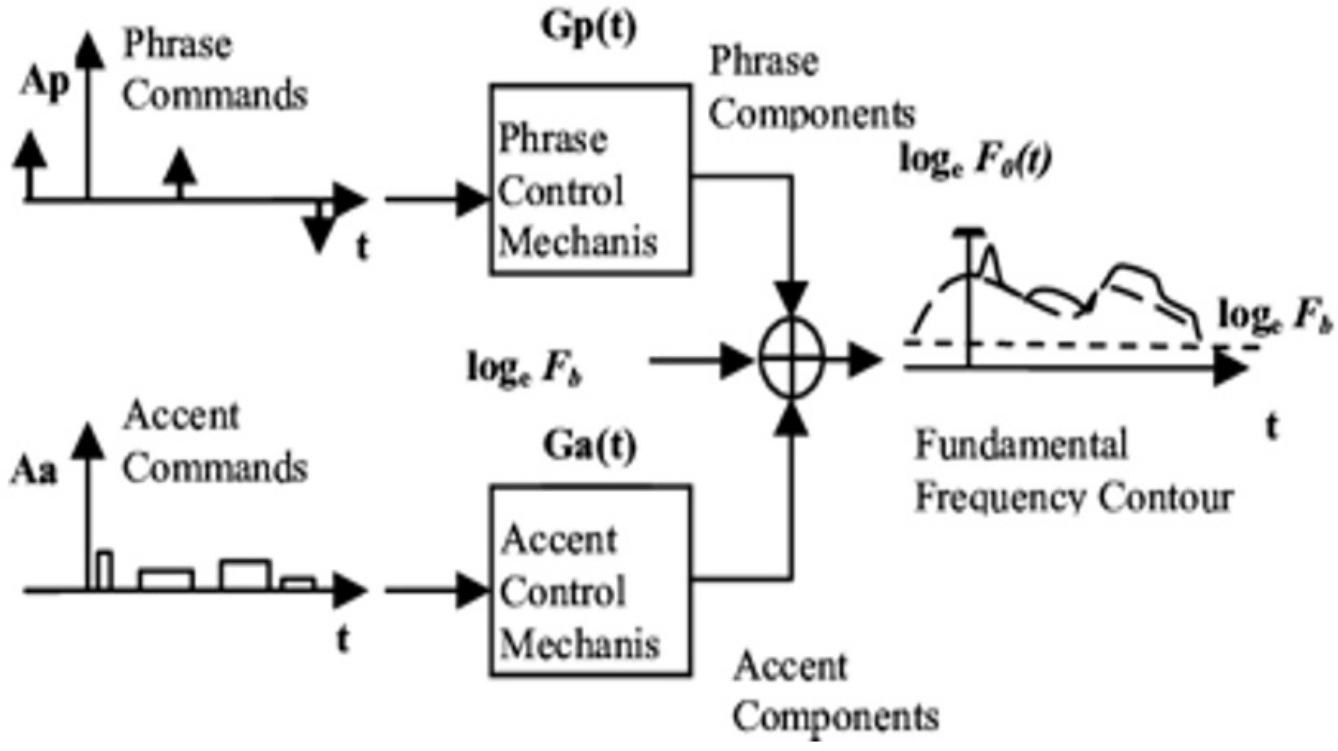

Figura 2.2 – Componentes do modelo de Fujisaki explicado no texto, reproduzida com autorização do autor, Keikichi Hirose. Fonte: Hirose e Fujisaki (1982).

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

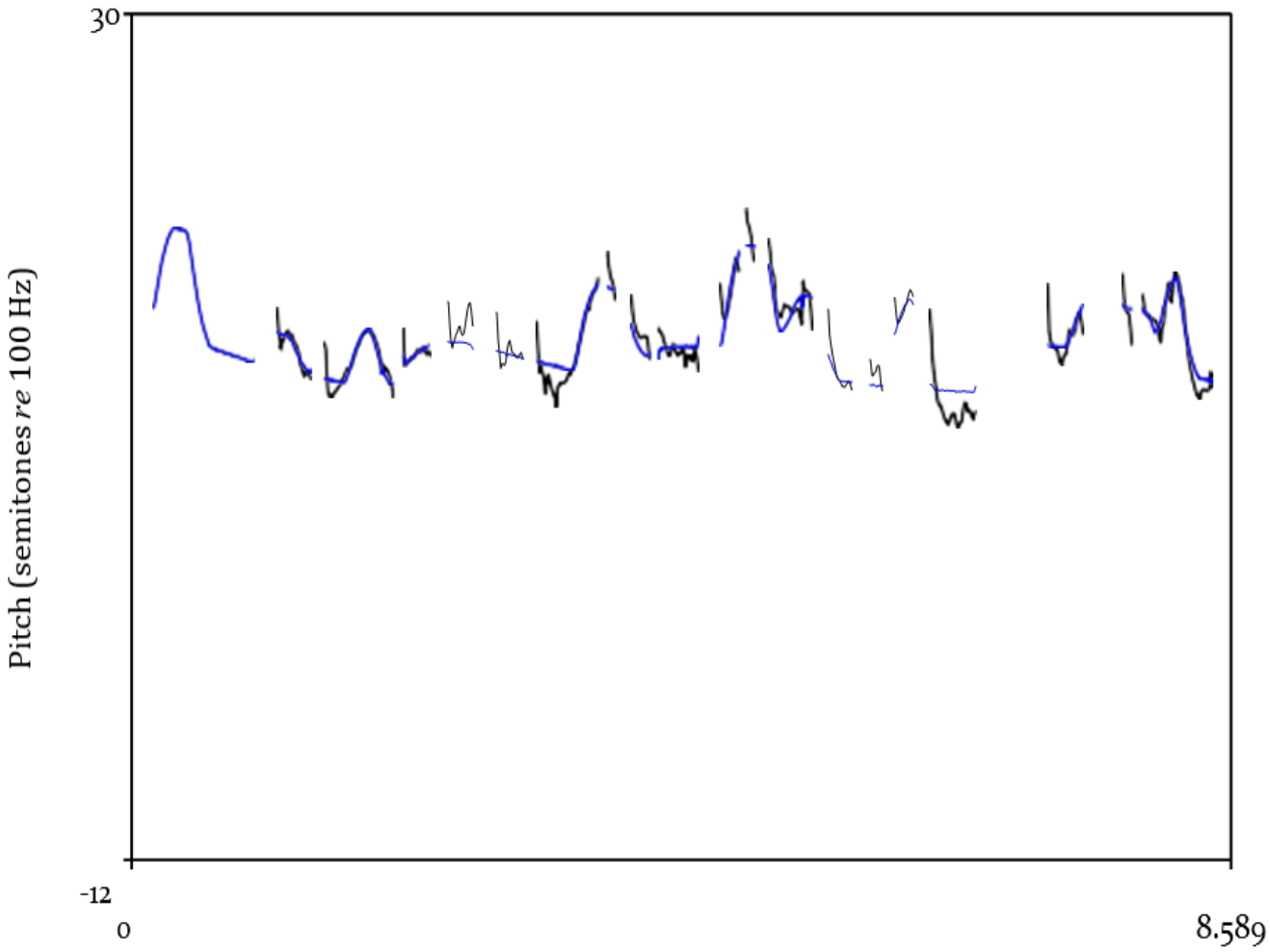

Figura 2.3 – Curvas melódicas (Fo) do enunciado “Quando ouvia os sinos a chamá-los, enroscava-se debaixo da manta com os joelhos quase chegando à testa e pensava: ‘talvez se esqueçam de mim’ ” lido por locutora paulista. Em preto a curva original e, mais clara, a curva gerada pelo modelo de Fujisaki.

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

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

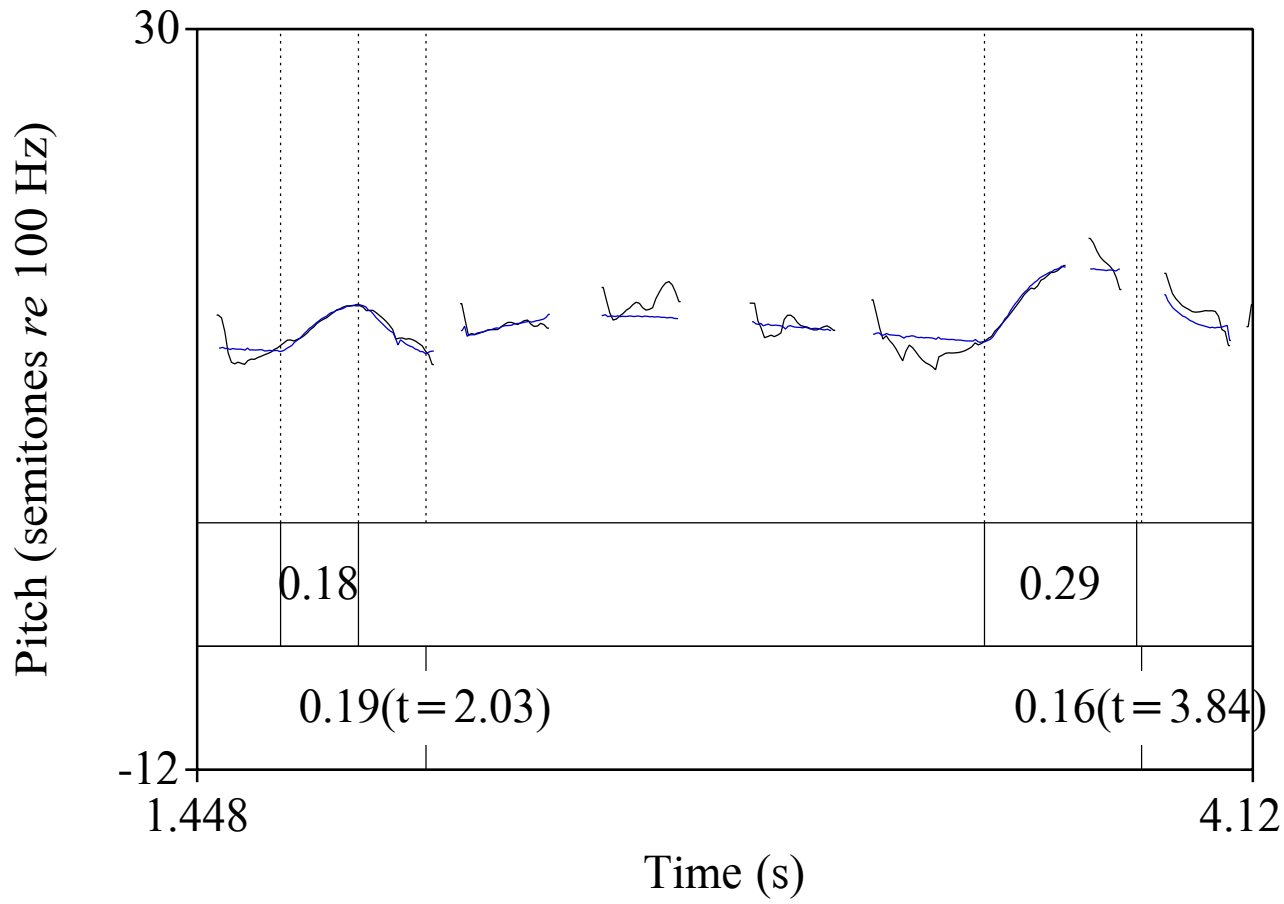

Figura 2.4 – Trecho **a chamá-los, enroscava-se debaixo da manta**do enunciado da Figura 2.3 comparando a curva melódica original (preto) com a gerada pelo modelo de Fujisaki (mais clara) com os comandos de sintagma (camada de baixo) e de acento (camada de cima) assinalados.

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

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

2.2.4 O Modelo PENTA

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

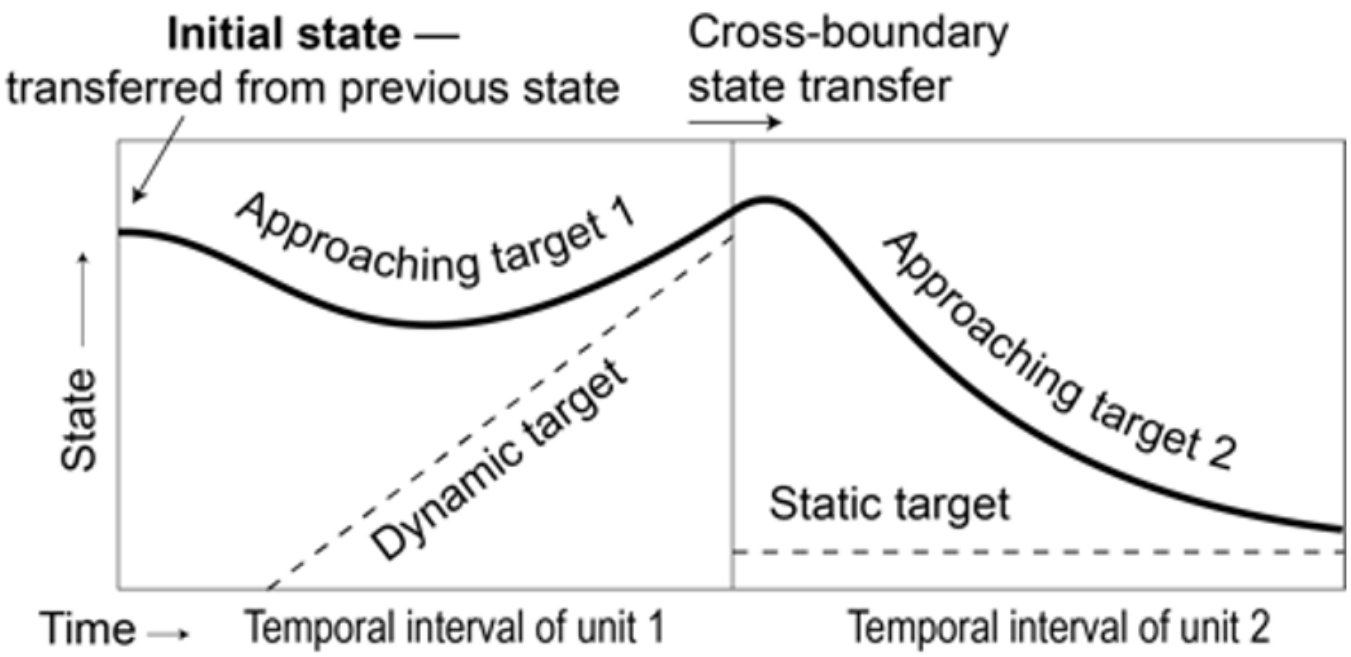

Figura 2.5 – Curva melódica básica gerada pelo modelo PENTA, conforme explicado no texto, reproduzida com autorização do autor, Yi Xu. Fonte: Xu e Wang (2001).

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

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

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

2.3 Quanto à Organização Temporal

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

2.3.1 Modelos Segmentais

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

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

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

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

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

onze regras.

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

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

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

2.3.2 Modelos Acima do Segmento

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

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

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

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

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

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

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

2.3.3 Modelos Dinâmicos do Ritmo da Fala

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

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

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

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

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

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

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

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

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

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

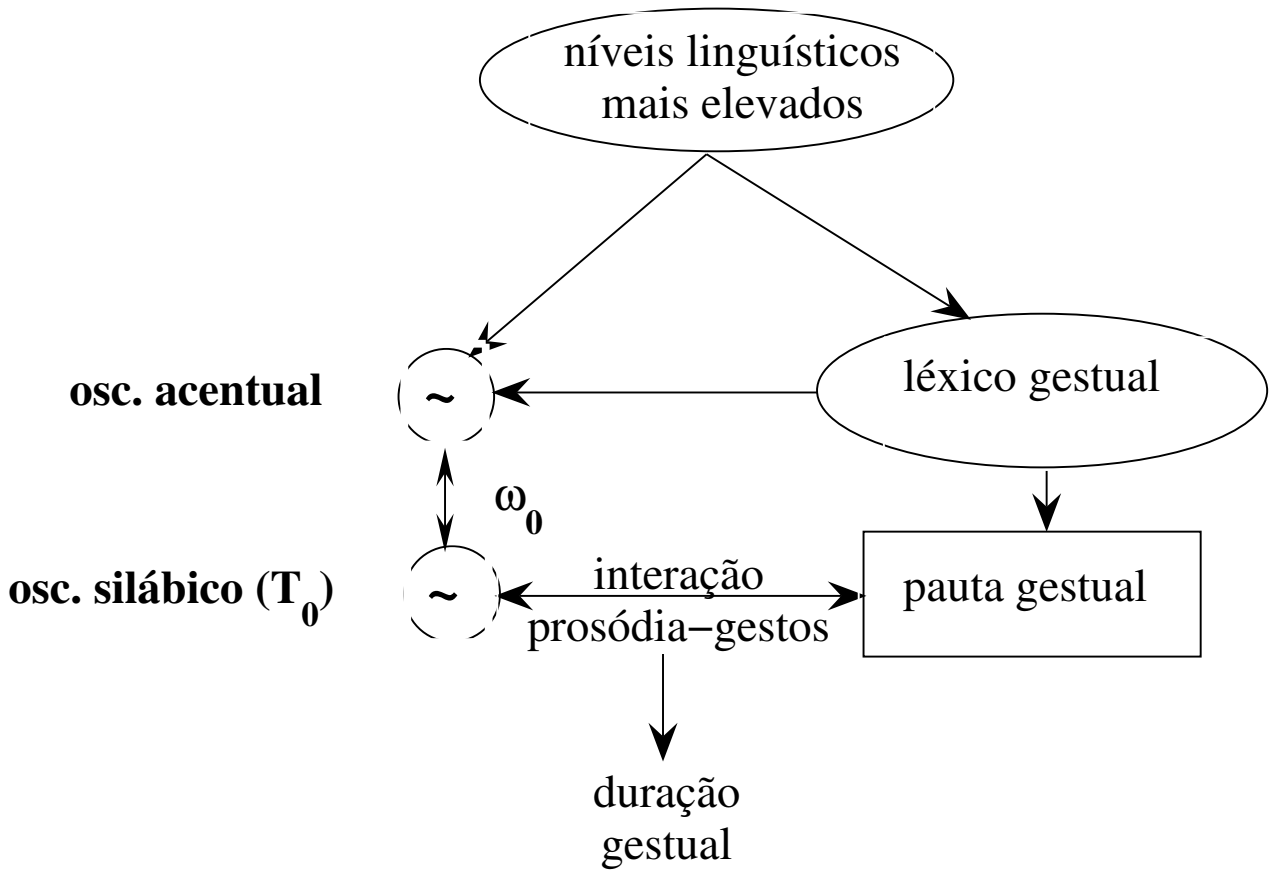

Figura 2.6 – Diagrama do modelo dinâmico do ritmo da fala de Barbosa.

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

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

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

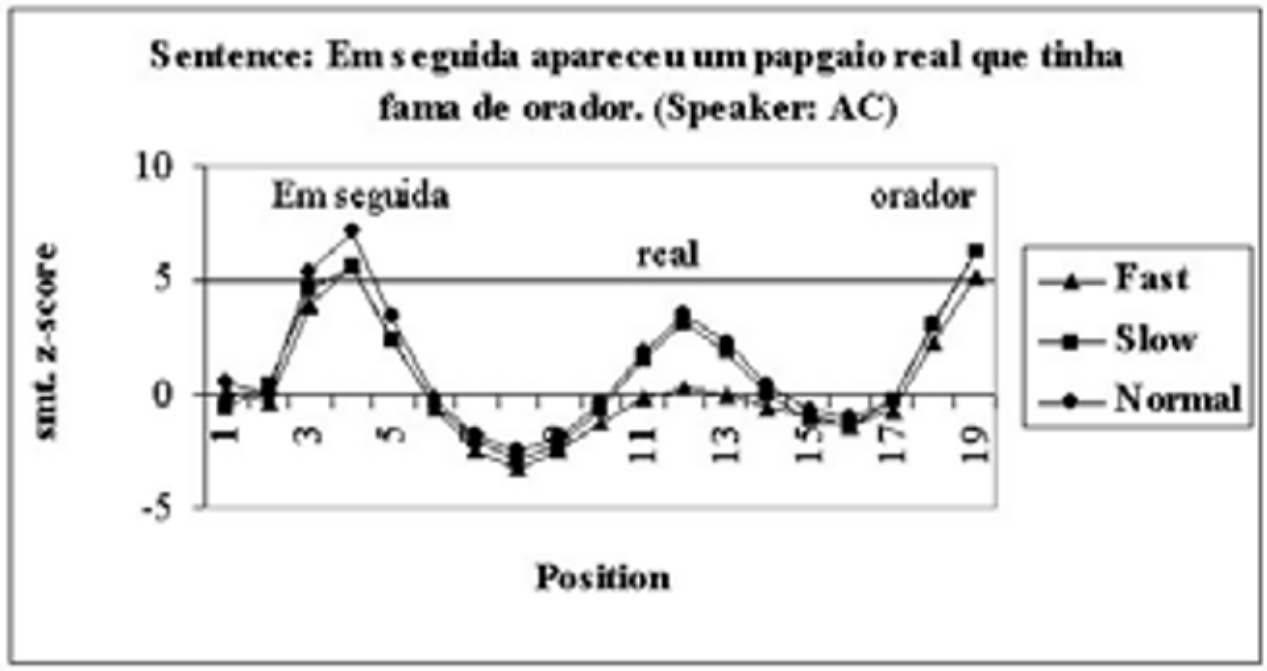

Figura 2.7 – Comparação de duração natural normalizada (acima) e duração gerada pelo modelo dinâmico (abaixo) para as unidades VV da sentença "Em seguida apareceu uma papagaio real que tinha fama de orador." em três taxas de elocução.

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

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

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

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

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