Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Animação Facial Sincronizada com a Fala: Visemas Dependentes do Contexto Fonético para o Português do Brasil

Autor: José Mario De Martino

Orientador: Prof. Dr. Léo Pini Magalhães

Co-orientador: Prof. Dr. Fábio Violaro

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica. Área de concentração: **Engenharia de Computação**.

Banca Examinadora

Prof. Dr. Léo Pini Magalhães	DCA/FEEC/UNICAMP
Prof. Dr. Clésio Luis Tozzi	DCA/FEEC/UNICAMP
Prof^a. Dr^a. Wu Shin-Ting	DCA/FEEC/UNICAMP
Prof. Dr. Amauri Lopes	DECOM/FEEC/UNICAMP
Prof Dr. José Luis Encarnação	GRIS/TUD/RFA
Prof. Dr. Plínio Almeida Barbosa	DL/IEL/UNICAMP

Campinas, SP

julho/2005

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

D392a

De Martino, José Mario
Animação facial sincronizada com a fala: visemas dependentes do contexto fonético para o português do Brasil / José Mario De Martino. –Campinas, SP: [s.n.], 2005.

Orientadores: Léo Pini Magalhães e Fábio Violaro.
Tese (doutorado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Computação Gráfica. 2. Animação por Computador. I. Magalhães, Léo Pini. II. Violaro, Fábio. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título

Título em Inglês:	Speech synchronized facial animation: phonetic context dependent visemes for Brazilian portuguese.
Palavras-chave em Inglês:	Computer graphics, Computer animation.
Área de concentração:	Engenharia de Computação.
Titulação:	Doutor em Engenharia Elétrica.
Banca examinadora:	Léo Pini Magalhães, Clésio Luis Tozzi, Wu Shin-Ting, Amauri Lopes, José Luis Encarnação e Plínio Almeida Barbosa.
Data da defesa:	29/07/2005

Resumo

Animação facial por computador diz respeito às técnicas para especificar e controlar a posição, forma e aparência de uma face sintética ao longo do tempo. Animação facial sincronizada com a fala está relacionada ao controle da movimentação da face virtual comandada pelos eventos fonéticos de uma locução. Tal controle implica na manipulação da face virtual de forma coordenada e em sincronismo com o sinal acústico da fala. A coordenação é alcançada pela reprodução, na face virtual, da movimentação articulatória visível necessária à produção dos sons da fala. O objetivo do trabalho é estudar e propor uma metodologia para a definição de representações para os padrões visuais de movimentação articulatória observáveis na face durante a fala, os denominados *visemas*. A metodologia proposta estabelece visemas dependentes do contexto fonético que contemplam o fenômeno da coarticulação perseveratória e antecipatória. Além disso, a partir da descrição geométrica e temporal de visemas estabelecidos pela análise de um corpus lingüístico do português do Brasil, são derivados modelos para a movimentação da articulação temporomandibular e do tecido dos lábios. Apesar do material fonético utilizado no trabalho estar restrito ao português do Brasil, a metodologia proposta é aplicável a outras línguas.

Palavras-chave: Computação Gráfica, Animação Facial, Visemas, Coarticulação, Sincronismo Labial.

Abstract

Computer facial animation refers to the techniques for specifying and controlling the positioning, motion, and appearance of a synthetic face over time. Speech synchronized facial animation addresses the control of a virtual face conducted by the phonetic events of an utterance. Such control implies the manipulation of the virtual face synchronized and coordinated with the speech signal. The coordination is achieved by reproducing on the virtual face the visible articulatory movements necessary for speech production. The objective of the work is to study and propose a methodology to establish representations for the visual articulatory patterns displayed on the face during speech production, the so called *visemes*. The proposed methodology identifies phonetic context dependent visemes that cope with perseverative and anticipatory coarticulation. Additionally, the movements of the temporomandibular joint and the lip tissue are modelled from a set of visemes established by the analysis of a Brazilian Portuguese linguistic corpus. Although the corpus is restricted to Brazilian Portuguese, the methodology is general enough to be applied to other languages.

Keywords: Computer Graphics, Facial Animation, Visual Speech, Talking Heads, Visemes, Coarticulation, Lip Synchronization.

À Eliana, minha amiga, companheira e musa.
Ao Bruno, mais do que um filho, um grande e inspirador amigo.

Agradecimentos

Ao Léo Pini Magalhães e Fábio Violaro, antes de mais nada pela amizade, depois pela atenção, empenho, paciência e compreensão durante esta longa jornada.

Ao Prof. José Luis Encarnação pela generosidade, preocupação e apoio.

A todos os amigos e colegas da Faculdade de Engenharia Elétrica e de Computação pela paciência e tolerância.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq pelo apoio financeiro.

Sumário

Lista de Figuras **xiii**

Lista de Tabelas **xvii**

Glossário **xxi**

1 Introdução **1**

1.1 Introdução 1

2 Animação facial **5**

2.1 Introdução 5

2.2 Manipulação da geometria da face virtual 6

2.2.1 Interpolação de poses-chave 6

2.2.2 Parametrização geométrica 7

2.2.3 Parametrização data-driven 11

2.2.4 Simulação biomecânica 12

2.3 Especificação da animação 14

2.4 Comentários finais 15

3 Produção e representação da fala **17**

3.1 Introdução 17

3.2 Fisiologia da produção da fala 18

3.3 Sons do português do Brasil 20

3.3.1 Segmentos consonantais 21

3.3.2 Segmentos vocálicos 31

3.4 Comentários Finais 38

4 Visualização da Fala **41**

4.1 Introdução 41

4.2 Homofemas de segmentos consonantais 42
4.3 Homofemas para segmentos vocálicos 54
4.4 Efeito da coarticulação nos homofemas 62
4.5 Comentários Finais 66

5 Visemas para o português do Brasil 71
5.1 Introdução 71
5.2 Metodologia e instrumentação 72
5.2.1 Corpus 74
5.2.2 Captura do áudio e vídeo 77
5.2.3 Segmentação do áudio 80
5.2.4 Medida das trajetórias de pontos da face 80
5.2.5 Identificação dos alvos articulatórios 81
5.2.6 Agrupamento dos alvos articulatórios 87
5.2.7 Instante relativo de realização dos visemas 88
5.2.8 Representação paramétrica dos visemas 88
5.3 Resultados 89
5.3.1 Estimativas da precisão das medidas 89
5.3.2 Posição de repouso 93
5.3.3 Visemas consonantais 93
5.3.4 Visemas vocálicos 104
5.4 Comentários Finais 113

6 Modelagem da movimentação facial 115
6.1 Introdução 115
6.2 Comportamento da articulação temporomandibular 116
6.3 Comportamento do lábio inferior 125
6.4 Comportamento do lábio superior 131
6.5 Comportamento do canto da boca 131
6.6 Comentários Finais 132

7 Implementação Piloto 135
7.1 Introdução 135
7.2 Arquitetura do sistema 135
7.3 Conversor fone-visema 137
7.4 Animação facial 142

7.4.1 Comportamento da articulação temporomandibular 144
7.4.2 Comportamento dos lábios inferior e superior 144
7.5 Comentários Finais 149

8 Conclusão **151**

Referências bibliográficas **155**

A Técnica Fotogramétrica **165**
A.1 Introdução 165
A.2 Calibração da Câmera 166
A.3 Visão Estéreo 169

Lista de Figuras

2.1 Estratégias de manipulação da face virtual. 6
2.2 *Feature Points* (FPs) do padrão MPEG-4 (MPEG4 VISUAL, 2001) 9
2.3 Estratégias de controle e definição da dinâmica da animação. 14

3.1 Aparelho Fonador. 19
3.2 Lugares de Articulação. 22
3.3 Posição da língua na produção das vogais /i/, /a/ e /u/. 32
3.4 Diagrama dos segmentos vocálicos orais do português do Brasil. 33
3.5 Diagrama dos segmentos vocálicos nasais do português do Brasil. 36

4.1 Diagrama das vogais e os homofemas vocálicos de Jeffers e Barley (1971) (somente monotongos). 57
4.2 Diagrama das vogais e os homofemas vocálicos de Jackson (1988) com dados de Wozniak e Jackson (1979) - ângulo de observação de 0° (somente monotongos). 58
4.3 Diagrama das vogais e os homofemas vocálicos de Jackson (1988) com dados de Wozniak e Jackson (1979) - ângulo de observação de 90° (somente monotongos). 58
4.4 Diagrama das vogais e os homofemas vocálicos de Montgomery e Jackson (1983) (somente monotongos). 60

5.1 Diagrama de fluxo de dados da metodologia desenvolvida. 73
5.2 Instalações da gravação. 78
5.3 Capacete de referência e pontos de interesse. 79
5.4 Inspeção visual do sinal de áudio: esquerda) forma de onda; direita) espectograma. . 80
5.5 Pontos de interesse cujas trajetórias foram medidas. 81
5.6 Pontos de referência (L_i e R_i , $i = 1, 2, ..., 15$) utilizados para a calibração das câmeras e os pontos de interesse (P_1 a P_4). 82
5.7 Pontos utilizados para o controle e avaliação da precisão das medidas. 82
5.8 Deslocamento na direção X de P_4 no plano sagital durante a produção de [ˈpapɐ]. . . 84

5.9 Deslocamento na direção Y de P_4 no plano sagital durante a produção de [ˈpapɐ]. . . 84
5.10 Deslocamento na direção Z de P_4 no plano sagital durante a produção de [ˈpapɐ]. . . 85
5.11 Deslocamento na direção X de P_4 no plano sagital durante a produção de [ˈaɐ]. 85
5.12 Deslocamento na direção Y de P_4 no plano sagital durante a produção de [ˈaɐ]. 86
5.13 Deslocamento na direção Z de P_4 no plano sagital durante a produção de [ˈaɐ]. 86
5.14 Deslocamento na direção Y de ponto P_4: esquerda) logatoma [ˈpapɐ]; direita) logatoma [ˈaɐ]. 89

6.1 Visão lateral da mandíbula - figura adaptada de Gray (2000). 116
6.2 Articulação Temporomandibular - figura adaptada de Gray (2000). 117
6.3 Seção sagital da Articulação Temporomandibular - figura adaptada de Gray (2000). . 118
6.4 Principais músculos responsáveis pela movimentação da ATM - figura adaptada de Zemlin (2000). 119
6.5 Movimentação da ATM no plano sagital médio. 120
6.6 Rotação da ATM durante [ˈaɐ]: esquerda) medidas; direita) modelo paramétrico dos visemas. 122
6.7 Translação da ATM durante [ˈaɐ]: esquerda) medidas; direita) modelo paramétrico dos visemas. 123
6.8 Rotação da ATM durante [ˈpapɐ]: esquerda) medidas; direita) modelo paramétrico. . 124
6.9 Translação da ATM durante [ˈpapɐ]: esquerda) medidas; direita) modelo paramétrico. 124
6.10 Diagrama das vogais - ângulo de abertura. 125
6.11 Movimentação do lábio inferior no plano sagital médio. 126
6.12 Protrusão do lábio inferior durante a produção do logatoma [ˈiɪ]. 128
6.13 Protrusão do lábio inferior durante a produção do logatoma [ˈiɐ]. 128
6.14 Protrusão do lábio inferior durante a produção do logatoma [ˈiʊ]. 128
6.15 Protrusão do lábio inferior durante a produção do logatoma [ˈaɪ]. 129
6.16 Protrusão do lábio inferior durante a produção do logatoma [ˈaɐ]. 129
6.17 Protrusão do lábio inferior durante a produção do logatoma [ˈaʊ]. 129
6.18 Protrusão do lábio inferior durante a produção do logatoma [ˈuɪ]. 130
6.19 Protrusão do lábio inferior durante a produção do logatoma [ˈuɑ]. 130
6.20 Protrusão do lábio inferior durante a produção do logatoma [ˈuʊ]. 130
6.21 Diagrama das vogais - protrusão lábio inferior. 131
6.22 Diagrama das vogais - protrusão lábio superior. 132
6.23 Diagrama das vogais - extensão da boca. 133

7.1 Diagrama de blocos do sistema de animação facial. 136

7.2 Modelo da cabeça virtual - visualização tonalizada. 142
7.3 Modelo da cabeça virtual - visualização aramada. 143
7.4 Pontos de interesse estendido. 144
7.5 Vértices transformados devido à movimentação da articulação temporomandibular. . 145
7.6 Músculos da face - ilustração adaptada de Sobotta (1990)). 146
7.7 Região de influência do comportamento do lábio superior. 146
7.8 Região de influência do comportamento do lábio inferior. 147
7.9 Efeito da estratégia de deformação. 148

Lista de Tabelas

3.1 Segmentos consonantais do português do Brasil (BARBOSA; ALBANO, 2004). 23
3.2 Classificação dos segmentos consonantais do português do Brasil. 24
3.3 Arquifonemas consonantais do português do Brasil. 25
3.4 Realizações do fone [l]. 25
3.5 Padrões silábicos com um segmento consonantal no ataque e até um segmento consonantal em coda. 27
3.6 Padrões silábicos com dois segmentos consonantais no ataque e até um segmento consonantal em coda. 28
3.7 Padrões silábicos: a) sem ataque e um segmento consonantal em coda; b) um segmento consonantal no ataque e dois em coda. 29
3.8 Pseudo encontros consonantais silábicos e não silábicos. 30
3.9 Segmentos vocálicos orais em posição tônica e pretônica. 34
3.10 Segmentos vocálicos orais em posição postônica. 34
3.11 Contraste entre os fones [ɪ] [i], [ɐ] [a] e [ʊ] [u]. 35
3.12 Segmentos vocálicos nasalizados do português do Brasil. 35
3.13 Contraste entre os fones vocálicos nasalizados e orais. 35
3.14 Segmentos vocálicos nasais do português do Brasil em posições pretônicas, tônicas e postônicas. 36
3.15 Ditongos orais: a) decrescentes; b) crescentes. 37
3.16 Ditongos nasais: a) decrescentes; b) crescentes. 38
3.17 Tritongos: a) orais; b) nasais. 38

4.1 Homofemas consonantais de Nitchie (1950). 43
4.2 Homofemas consonantais de Jeffers e Barley (1971). 43
4.3 Homofemas consonantais de Erber (1974). 45
4.4 Homofemas consonantais de Binnie, Jackson e Montgomery (1976). 45
4.5 Homofemas consonantais de Walden et al. (1977). 46

4.6 Homofemas consonantais de Walden et al. (1981). 46
4.7 Homofemas consonantais de Kricos e Lesner (1982). 47
4.8 Homofemas consonantais de Owens e Blazek (1985). 48
4.9 Identificação dos homofemas /p, b, m/, /f, v/ e /ʃ, ʒ/. 49
4.10 Reconhecimento dos conjuntos /s, z/ e /k, g/ como pertencentes a um mesmo grupo de homofemas. 51
4.11 Reconhecimento dos conjunto /t, d/, /n, l/, /t, d, n/ e /t, d, l/ como pertencentes a um mesmo grupo de homofemas. 52
4.12 Reconhecimento dos conjunto /l/, /n/ e /t, d, n, l/ como pertencentes a um mesmo grupo de homofemas. 53
4.13 Homofemas consonantais do Padrão MPEG-4 (MPEG4 VISUAL, 2001). 54
4.14 Fonemas consonantais do padrão MPEG4 - Tabela C-5 (parcial) Anexo C (MPEG4 VISUAL, 2001) 55
4.15 Mapeamento entre as simbologias MPEG-4 e IPA para segmentos consonantais. . . . 55
4.16 Homofemas vocálicos de Jeffers e Barley (1971) 56
4.17 Homofemas vocálicos de Jackson (1988) com dados de Wozniak e Jackson (1979). . 57
4.18 Matriz de confusão de reconhecimento de vogais e ditongos de Montgomery e Jackson (1983). 59
4.19 Homofemas vocálicos de Montgomery e Jackson (1983). 60
4.20 Homofemas vocálicos do Padrão MPEG-4 (MPEG4 VISUAL, 2001). 61
4.21 Fonemas vocálicos do padrão MPEG4 - Tabela C-5 (parcial) Anexo C (MPEG4 VISUAL, 2001). 61
4.22 Mapeamento entre as simbologias MPEG-4 e IPA para segmentos vocálicos. 62
4.23 Homofemas para a língua francesa de considerando efeitos da coarticulação (BENOÎT et al., 1992). 63
4.24 Agrupamento dos segmentos consonantais em homofemas considerado no trabalho. . 68
4.25 Agrupamento dos segmentos vocálicos em homofemas considerado neste trabalho. . 68

5.1 Homofemas consonantais e fones representantes adotados. 76
5.2 Homofemas vocálicos e fones representantes adotados. 76
5.3 Contextos fonéticos analisados. 77
5.4 Estatística do erro de coordenada dos pontos de controle P_5 a P_{31}. 91
5.5 Estatística do erro de distância dos pontos P_5 a P_{31}. 92
5.6 Posição de repouso dos pontos de interesse. 93
5.7 Visemas consonantais e respectivos contextos fonéticos. 94
5.8 Agrupamento em visemas das diferentes produções do fone [p]. 95

5.9 Alvos Articulatórios dos visemas bilabiais. 95
5.10 Agrupamento em visemas das diferentes produções do fone [f]. 96
5.11 Alvos Articulatórios dos visemas labiodentais. 96
5.12 Agrupamento em visemas das diferentes produções do fone [t]. 97
5.13 Alvos articulatórios dos visemas alveolares plosivos/nasais. 97
5.14 Agrupamento em visemas das diferentes produções do fone [s]. 98
5.15 Alvos articulatórios dos visemas alveolares fricativos. 98
5.16 Agrupamento em visemas das diferentes produções do fone [l]. 99
5.17 Alvos articulatórios dos visemas alveolares laterais. 99
5.18 Agrupamento em visemas das diferentes produções do fone [ʃ]. 100
5.19 Alvos articulatórios dos visemas pós-alveolares. 100
5.20 Agrupamento em visemas das diferentes produções do fone [ʎ]. 101
5.21 Alvos articulatórios dos visemas palatais. 101
5.22 Agrupamento em visemas das diferentes produções do fone [k]. 102
5.23 Alvos articulatórios dos visemas velares plosivos. 102
5.24 Agrupamento em visemas das diferentes produções dos fones [ɣ] e [ɾ]. 103
5.25 Alvos articulatórios dos visemas velares fricativos/tepes. 103
5.26 Visemas vocálicos e respectivos contextos fonéticos. 104
5.27 Agrupamento em visemas das diferentes produções do fone [i]. 105
5.28 Alvos articulatórios dos visemas altos anteriores. 105
5.29 Agrupamento em visemas das diferentes produções do fone [a]. 106
5.30 Alvo articulatório do visema baixo central. 106
5.31 Agrupamento em visemas das diferentes produções do fone [u]. 107
5.32 Alvo articulatório do visema alto posterior. 107
5.33 Agrupamento em visemas das diferentes produções do fone [ɪ]. 108
5.34 Alvo articulatório do visema postônico alto anterior. 108
5.35 Agrupamento em visemas das diferentes produções do fone [ɐ]. 109
5.36 Alvo articulatório do visema postônico baixo central. 109
5.37 Agrupamento em visemas das diferentes produções do fone [ʊ]. 110
5.38 Alvo articulatório do visema postônico alto posterior. 110
5.39 Alvo articulatório do fone [e]. 111
5.40 Distâncias em milímetros do alvo articulatório do fone [e] aos visemas vocálicos. . . 111
5.41 Alvo articulatório do fone [ɛ]. 111
5.42 Distâncias em milímetros do alvo articulatório do fone [ɛ] aos visemas vocálicos. . . 111
5.43 Alvo articulatório do fone [ɔ]. 112

5.44 Distâncias em milímetros do alvo articulatório do fone [ɔ] aos visemas vocálicos. . . 112
5.45 Alvo articulatório do fone [o]. 112
5.46 Distâncias em milímetros do alvo articulatório do fone [o] aos visemas vocálicos. . . 112

6.1 Posição no plano sagital médio do centro da articulação temporomandibular em repouso. 122

7.1 Segmentos vocálicos, contextos fonéticos e respectivos visemas. 138
7.2 Segmentos consonantais e respectivos representantes. 139
7.3 Segmentos consonantais, contextos fonéticos e respectivos visemas. 140

Glossário

Alofone: cada uma das produções sonoras de um mesmo *fonema*.

Alvo articulatório: postura articulatória que caracteriza a conformação do trato vocal associada a um *segmento*.

Arquifonema: neutralização, em determinados contextos fonéticos, do contraste entre *fonemas* .

Articulação: seqüência das etapas da emissão de um som da fala, que são: a) movimento de aproximação dos articuladores ativos; b) sustentação dessa posição; c) afastamento dos articuladores.

Ataque: na estrutura da sílaba (constituída por *ataque*, *núcleo* e *coda*), ataque é posição inicial, não obrigatória, ocupada por até dois *segmentos consonantais* no português do Brasil.

ATM: articulação temporomandibular. A articulação da mandíbula com o osso temporal.

Coda: na estrutura da sílaba (constituída por *ataque*, *núcleo* e *coda*), coda é a posição final, não obrigatória, ocupada por até dois *segmentos consonantais* no português do Brasil.

Ditongo: grupo de dois *segmentos vocálicos* proferidos em uma só sílaba.

Fonema: unidade lingüística abstrata, representativa das produções sonoras que, em conjunto, possuem a mesma função distintiva na estrutura da língua.

Fone: realização acústica de um *segmento*.

Glote: abertura entre as bordas livres das *pregas vocais*.

Homofemas: conjunto de fonemas que não são visualmente distinguíveis.

Logatoma: palavra sem sentido.

Lugar de articulação: ou ponto de articulação, identifica a região do trato vocal onde ocorre a obstrução significativa que caracteriza um *segmento consonantal*.

Modo de articulação: parâmetro utilizado para classificar os *segmentos consonantais*, caracterizando a maneira pela qual estes *segmentos* são articulados.

Monotongo: *segmento vocálico* produzido de maneira a apresentar uma mesma característica acústica, percebida como estável e constante durante a sua produção.

Núcleo: na estrutura da sílaba (constituída por *ataque*, *núcleo* e *coda*), núcleo é a posição central obrigatória ocupada por *segmento vocálico*.

Pregas vocais: duas faixas de tecido muscular na forma de lábios, que se estendem horizontalmente no interior da laringe e que, por movimentação voluntária, podem obstruir a passagem do ar.

Segmento consonantal: *segmento* produzido com a obstrução significativa do trato vocal.

Segmento desvozeado: ou segmento surdo, segmento produzido sem a vibração das *pregas vocais*.

Segmento vocálico: *segmento* produzido sem a obstrução significativa do trato vocal.

Segmento vozeado: ou segmento sonoro, segmento produzido com a vibração das *pregas vocais*.

Segmento: unidade discreta, cuja seqüência compõe o contínuo da fala. Exemplo: a palavra "pá" é constituída pelos segmentos [p] e [a].

Tritongo: grupo de três *segmentos vogais* proferidos em uma só sílaba.

Visema: padrão visual de movimentação articulatória dos *segmentos* de um grupo de *homofemas*.

Capítulo 1

Introdução

1.1 Introdução

A face desempenha um papel significativo na comunicação interpessoal. Já nos primeiros momentos da infância, a criança aprende a reconhecer e identificar indivíduos por suas faces. Ao longo da vida, balizada por experiências sociais, esta habilidade é aperfeiçoada, permitindo a interpretação e identificação de estados emocionais e de suas nuances a partir da inspeção facial. Várias vezes por dia, os indivíduos se engajam em interações face-a-face. Nestas interações, a informação visual apresentada na face muitas vezes complementa e auxilia a compreensão da mensagem transportada pelo sinal acústico da fala. Tais habilidades, desenvolvidas e aperfeiçoadas através de exaustivo treinamento diário, permitem que os indivíduos sejam capazes de perceber variações sutis na expressão facial. O resultado deste exaustivo treinamento é uma acurada sensibilidade e um rigoroso senso crítico na avaliação de animações faciais, tornando a produção deste tipo de animação um grande e envolvente desafio.

O termo animação facial denota os métodos e técnicas para especificar e controlar, ao longo do tempo, a transformação da forma e de outras características visíveis da representação de uma face. O resultado final do processo da animação facial, também denominado animação facial, é um conjunto de imagens, que, ao serem apresentadas em seqüência, a uma taxa apropriada, transmitem a sensação de movimento.

O presente trabalho está restrito à *animação facial por computador*, não abordando as técnicas tradicionais de animação (desenho animado) e a denominada *animação assistida por computador*. Entende-se que o marco divisório entre a animação facial por computador e as outras técnicas reside no mecanismo de produção da movimentação da animação. Movimentação sendo aqui entendida em um sentido amplo, envolvendo qualquer modificação visível ao longo do tempo. Na animação facial por computador, esta movimentação é descrita por um modelo implementado por um algoritmo e co-

dificado em uma linguagem de programação. Já na animação tradicional, assim como na assistida por computador, um animador, baseado em sua experiência, habilidade e sensibilidade, define e controla a movimentação. Excluindo-se os exageros extremos admitidos em uma animação que procurem reproduzir o estilo típico dos desenhos animados, uma da metas da animação facial por computador é o realismo, ou seja, espelhar na animação a movimentação observada em faces reais.

Os movimentos observados em faces reais podem ser agrupados em cinco categorias principais: articulatórios, emocionais, conversacionais, fisiológicos e contorcidos. Os movimentos articulatórios são os sinais visuais produzidos pela movimentação dos órgãos do aparelho fonador durante a produção da fala. Os movimentos emocionais são aqueles associados às expressões de emoções, tais como tristeza, alegria, medo e raiva. Os sinais conversacionais compreendem movimentações faciais que podem acompanhar a fala, sendo utilizados para enfatizar palavras, pontuar visualmente o discurso e gerenciar e controlar o papel receptor-emissor assumidos dinamicamente pelos interlocutores (PELACHAUD; BADLER; STEEDMAN, 1996). Os movimentos fisiológicos, nesta classificação, correspondem a necessidades biológicas como, por exemplo, o piscar dos olhos para a limpeza e manutenção da umidade da córnea. Finalmente, a contorção facial abrange as conformações não usuais da face, resultando nas chamadas caretas.

Dentre estas categorias, a movimentação articulatória tem papel destacado na animação facial sincronizada com a fala. O principal objetivo perseguido neste tipo de animação é a movimentação da face em estreita consonância com o sinal acústico da fala. O conceito chave é o estabelecimento de uma representação apropriada para a movimentação articulatória visível, os chamados visemas. Apesar das quatro primeiras categorias mencionadas acima ocorrerem na comunicação face-a-face, o requisito obrigatório da animação facial sincronizada com a fala é a reprodução da movimentação articulatória. A movimentação associada às emoções, aos sinais conversacionais, às condicionantes fisiológicas, e a eventuais contorções, podem ainda ser incorporadas à animação para modular a movimentação articulatória e emprestar maior expressividade à face virtual.

Para a geração de uma animação facial em sincronia e em harmonia com a locução, é imperativa a reprodução dos movimentos articulatórios associados à realização dos vários fonemas da língua. Para tanto, além da identificação das posturas características dos gestos articulatórios associados aos fonemas, faz-se necessária a representação das transições entre estas posturas considerando os efeitos da coarticulação. Os efeitos da coarticulação se manifestam pela alteração do padrão articulatório de um determinado fone pela influência da articulação de outro adjacente ou, e em menor grau, próximo, mas mais distante, na cadeia da produção sonora. Os efeitos da coarticulação fazem com que, por exemplo, o "p" da palavra "paro" seja visualmente diferente do "p" da palavra "puro". Neste último caso, o movimento articulatório necessário à produção do "u" influencia de maneira significativa os aspectos visíveis da articulação do "p". A coarticulação pode ser classificada como perseveratória ou

antecipatória. Na coarticulação perseveratória, a articulação de um segmento de fala é influenciada pela movimentação articulatória de um segmento que o antecede na cadeia fonética que compõe a locução. Na coarticulação antecipatória, a movimentação articulatória do segmento é influenciada pela articulação de um segmento que o sucede.

O presente trabalho representa um esforço na busca de soluções para a representação da movimentação articulatória visível apropriadas para animação facial, contemplando efeitos da coarticulação. A falta de trabalhos publicados envolvendo a identificação e o estabelecimento de visemas para o português do Brasil empresta a este trabalho característica pioneira. O trabalho também tem carácter multidisciplinar, por trabalhar com conceitos e conhecimentos oriundos da lingüística, produção e percepção visual da fala e computação gráfica.

Através de medidas fotogramétricas das trajetórias tridimensionais de pontos marcados na face de um informante, foi efetuada a análise da movimentação articulatória visível de um corpus cuidadosamente especificado para representar os segmentos fonéticos do português do Brasil. Esta análise revelou um conjunto de 29 visemas, contemplando os efeitos da coarticulação adjacente antecipatória e perseveratória. A partir da descrição geométrica do conjunto de visemas estabelecido, foi desenvolvido um mapeamento geométrico para a manipulação de uma face virtual. Para demonstrar a viabilidade e o potencial da abordagem adotada no estabelecimento dos visemas, foi desenvolvido um sistema piloto de animação facial baseado nos conceitos e resultados alcançados.

As principais contribuições do trabalho podem ser resumidas em:

- definição de uma metodologia para a identificação de visemas;
- identificação de um conjunto de visemas para o português do Brasil, contemplando os efeitos da coarticulação adjacente perseveratória e antecipatória;
- estabelecimento de um conjunto de modelos geométricos para a manipulação da geometria da face virtual a partir dos visemas identificados; e
- implementação piloto de um sistema de animação facial.

O restante do trabalho está organizado na seguinte seqüência:

- Capítulo 2 - Animação facial: caracterização do conceito e recorte adotado no trabalho e revisão da literatura associada;
- Capítulo 3 - Produção e representação da fala: revisão dos mecanismos da produção da fala e caracterização do português do Brasil no contexto deste trabalho;
- Capítulo 4 - Visualização da fala: revisão e discussão dos principais aspectos associados à identificação de visemas;

- Capítulo 5 - Visemas para o português do Brasil: apresentação da metodologia desenvolvida para identificação de visemas, assim como do conjunto de visemas identificados pela metodologia (DE MARTINO; MAGALHÃES, 2004);

- Capítulo 6 - Modelagem da movimentação facial: apresentação dos modelos para a manipulação da geometria da face virtual acionados pela representação dos visemas (DE MARTINO; VIOLARO, 2003);

- Capítulo 7 - Implementação piloto: apresentação de detalhes de sistema de animação facial implementada para demonstrar a viabilidade das soluções propostas.

- Capítulo 8 - Conclusão: capítulo final onde são repassadas as principais contribuições e os desenvolvimentos futuros estimulados pelo presente trabalho.

Capítulo 2

Animação facial

2.1 Introdução

Atualmente é possível identificar duas correntes principais de desenvolvimento envolvendo a animação facial por computador. A linha divisória destas duas vertentes está diretamente relacionada à característica da representação da face virtual. Distinguem-se, neste contexto, a animação baseada em modelo (*model based*) e a animação baseada em imagem (*image based*). Na animação baseada em modelo, a face virtual é descrita por um modelo geométrico, via de regra, tridimensional. O trabalho pioneiro de Parke (1972), citado em Parke e Waters (1996), é referenciado como o marco inicial desta corrente. A animação facial baseada em imagem utiliza imagens fotográficas de poses-chave, as quais são combinadas através de técnicas de *morphing* para a geração da animação (BREGLER; COVELL; SLANEY, 1997) (EZZAT; POGGIO, 1998) (COSATTO; GRAF, 1998) (EZZAT; POGGIO, 2000) (EZZAT; GEIGER; POGGIO, 2002) (BUTTFIELD, 2003). Considerando a utilização de texturas em modelos tridimensionais, as fronteiras entre estas duas vertentes são menos nítidas em algumas abordagens como, por exemplo, em Pighin et al. (1998).

O presente trabalho restringe-se à animação facial baseada em modelos tridimensionais, sendo que neste capítulo são apresentadas as principais linhas de trabalho, restritas a este contexto, encontradas na literatura . O trabalho não aborda os problemas associados à utilização de texturas, assim como os associados à construção e conformação do modelo geométrico para representar uma determinada face. Também não será tratada a animação intraoral que pode incluir a língua, véu palatino e outras aspectos do interior da cavidade oral (ENGWALL, 2002) (BADIN et al., 2002) (MASSARO; LIGHT, 2003) (MASSARO, 2003).

Considerando estes recortes iniciais, é possível identificar duas problemáticas fundamentais associadas à animação facial baseada em modelo. A primeira está associada à estratégia de manipulação da geometria da face virtual. A segunda diz respeito aos mecanismos de definição e controle da di-

nâmica da animação. Este dois aspectos estão inter-relacionados de forma hierárquica, uma vez que o segundo aspecto utiliza os recursos do primeiro para realizar a deformação da face e produzir a animação. As Seções 2.2 e 2.3 abordam trabalhos que permitem a compreensão do estado da arte associado a estes dois aspectos, com ênfase nos aspectos relacionados à animação sincronizada com a fala.

2.2 Manipulação da geometria da face virtual

Na animação baseada em modelo, a superfície da face é tipicamente descrita por uma malha poligonal tridimensional. Durante a animação, a superfície é deformada através do deslocamento apropriado dos vértices dos polígonos que a compõem. As estratégias empregadas para a manipulação da geometria da face e, conseqüentemente, para a realização destes deslocamentos podem ser classificadas em quatro grandes categorias (Figura 2.1): interpolação de poses-chave, parametrização geométrica, parametrização *data-driven* e simulação biomecânica. As Seções 2.2.1 a 2.2.4 apresentam e discutem cada uma destas abordagens.

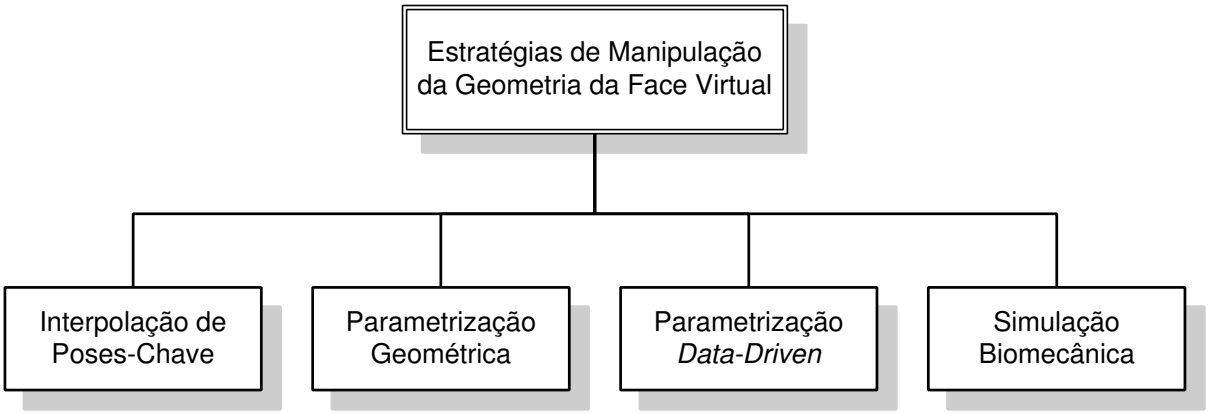

Fig. 2.1: Estratégias de manipulação da face virtual.

2.2.1 Interpolação de poses-chave

O princípio básico desta estratégia consiste na especificação de um conjunto de poses-chave e a interpolação entre estas poses extremas para o cálculo de poses intermediárias. As poses intermediárias descrevem as modificações da face entre as poses-chave e definem a movimentação durante a animação. As poses-chave podem representar expressões ou visemas. Expressões estão associadas, em

geral, a emoções. Visemas procuram descrever poses-chave da movimentação articulatória associada à produção dos sons da fala. Cada pose-chave é definida pelo conjunto de dados que constitui a geometria tridimensional da face ou cabeça virtual. Portanto, cada pose-chave exige o armazenamento das coordenadas de todos os vértices de todos os polígonos que compõem a geometria. Interpolação de poses-chave foi o método utilizado no trabalho pioneiro de Parke (1972).

Em sua versão mais simples, a interpolação é efetuada entre duas poses-chave extremas. Uma variação possível é utilizar quatro poses-chave e efetuar uma interpolação bilinear. Como generalização admitem-se 2^n poses-chave e um processo de interpolação n-dimensional (PARKE; WATERS, 1996). Ressalte-se, entretanto, que o aumento do espaço da interpolação torna o processo pouco intuitivo, diminuindo a sua utilidade prática.

A título de exemplo de aplicação recente deste método é possível mencionar a proposta de Alexa, Behr e Müller (2000) envolvendo a linguagem X3D. X3D (*Extensible 3D*) é uma linguagem de modelagem tridimensional, baseada na sintaxe XML (*Extensible Markup Language*) que objetiva estender e substituir o padrão VRML97 (*Virtual Reality Modelling Language*). Como a linguagem VRML, a linguagem X3D permite a descrição de mundos virtuais para ambientes interativos educacionais ou de entretenimento.

O grande atrativo do método de interpolação de poses-chave é a sua simplicidade de implementação. Este método é suportado pela maioria dos sistemas de modelagem e animação disponíveis no mercado.

Em contraponto à simplicidade de implementação, a estratégia de interpolação de poses-chave apresenta limitações. O espaço de expressões realizáveis é definido pelo número e disparidade das poses-chave utilizadas. Uma expressão fora deste espaço não é realizável. Criar, manter e gerenciar um conjunto significativo de poses-chave, que cubram um leque abrangente de expressões de diferentes indivíduos, não é uma tarefa trivial, tornando o processo extremamente laborioso. Adaptar a estratégia básica da interpolação, dividindo-se a face em regiões independentes e aplicando-se a estratégia separadamente a cada região pode aliviar as dificuldades do processo, sem, entretanto, resolvê-las completamente (KLEISER, 1989).

2.2.2 Parametrização geométrica

A parametrização geométrica engloba os métodos de animação baseados em um conjunto de procedimentos dedicados, cada qual voltado a realizar determinada movimentação ou deformação da face virtual. Os modelos de parametrização geométrica caracterizam-se por oferecer mecanismos geométricos de manipulação da geometria da face virtual sem modelar os mecanismos biomecânicos e fisiológicos associados à movimentação facial.

A "parametrização direta" apresentada em Parke (1982) é um método estritamente geométrico,

cuja principal motivação foi contornar as limitações da interpolação de poses-chave. Neste método, em vez da interpolação entre descrições geométricas distintas, foram adotados procedimentos distintos para a implementação de aspectos diversos da movimentação facial, utilizando uma mesma geometria facial base. Na proposta original, identificam-se cinco procedimentos básicos de manipulação da geometria: construção procedural, interpolação, rotação, escalamento e translação. A composição e sequenciamento da aplicação destes procedimentos resultam na animação do modelo geométrico. Cada procedimento aplicado à face é controlado por um ou mais parâmetros, justificando o nome "parametrização direta" adotado por seu autor. Os parâmetros utilizados por este método são bastantes díspares, abrangendo coordenadas 3D de um ponto, movimentos articulatórios complexos como o posicionamento do lábios nas labiodentais (postura "f"), ou ainda expressões de emoção tais como sorriso e surpresa. Em particular, no que tange à animação facial sincronizada com a fala, nove parâmetros controlam a mandíbula e os lábios: rotação da mandíbula, largura da boca, expressão da boca, posição do lábio superior, protrusão dos lábios, postura "f"do lábio inferior, deslocamentos x, y e z do canto da boca. Os modelos e procedimentos adotados foram desenvolvidos sob inspiração do estilo de animação dos desenhos animados, sem a preocupação de uma análise mais elaborada da anatomia facial e dos mecanismos articulatórios da fala.

O modelo de Parke foi utilizado em Hill, Pearce e Wyvill (1988) para gerar a animação facial simultaneamente a um processo de síntese de fala a partir de texto. Na solução adotada, os visemas foram estimados de imagens, limitadas à visão frontal da face, extraídas de livro sobre leitura orofacial. Na falta de uma descrição de visemas, os autores optaram por uma aproximação pouco precisa pela falta da informação temporal e de profundidade.

Importantes extensões do modelo de Parke foram propostas por Cohen e Massaro, principalmente no que se refere à animação sincronizada com a fala (COHEN; MASSARO, 1990) (COHEN; MASSARO, 1993). As extensões incluem parâmetros adicionais para um melhor controle dos lábios e da língua. Adicionalmente, foi proposta a utilização de funções de dominância derivadas da teoria articulatória de Löfqvist (LöFQVIST, 1990) para a modelagem dos efeitos da coarticulação. Em essência, esta teoria considera que um segmento de fala possui dominância sobre os articuladores, dominância esta que aumenta e diminui no tempo. As funções de dominância de segmentos adjacentes ou próximos na cadeia de produção se sobrepõem resultando na composição dos comandos articulatórios associados aos segmentos. Para a aplicação do modelo de Cohen e Massaro, vários parâmetros que descrevem as várias funções de dominância precisam ser determinados. As funções de dominância permitem uma grande flexibilidade no controle articulatório, entretanto a determinação dos parâmetros das funções é uma questão importante, que não pode ser realizada sem a coleta e análise detalhada de material lingüístico (LE GOFF, 1997) (BENOÎT; Le Goff, 1998) (COSI et al., 2002).

A "parametrização direta" é uma escolha popular entre os pesquisadores da síntese visual da

fala (BESKOW, 2003). Vários sistemas baseados diretamente em descendentes do modelo de Parke têm sido utilizados por diferentes grupos de pesquisa, destacando-se *Baldi* do PSL-UCSC Califórnia (COHEN; MASSARO, 1993), ICP-Grenoble (BENOÎT; Le Goff, 1998), KTH-Stockholm (BESKOW, 1995) (BESKOW, 1997) e LCE-Helsinki (OLIVÈS et al., 1999).

O modelo pseudo-muscular proposto em Kalra et al. (1992) também pode ser classificado como um modelo de parametrização geométrica. Para simular a interação da ação muscular com a pele da face, o modelo utiliza a técnica FFD (Free Form Deformation) (SEDERBERG; PARRY, 1986) para controlar a deformação de regiões da superfície do modelo. As regiões definidas procuram cobrir as regiões de influência dos principais músculos faciais. A especificação das regiões e extensão da deformação de cada músculo devem ser efetuadas de maneira interativa por um animador.

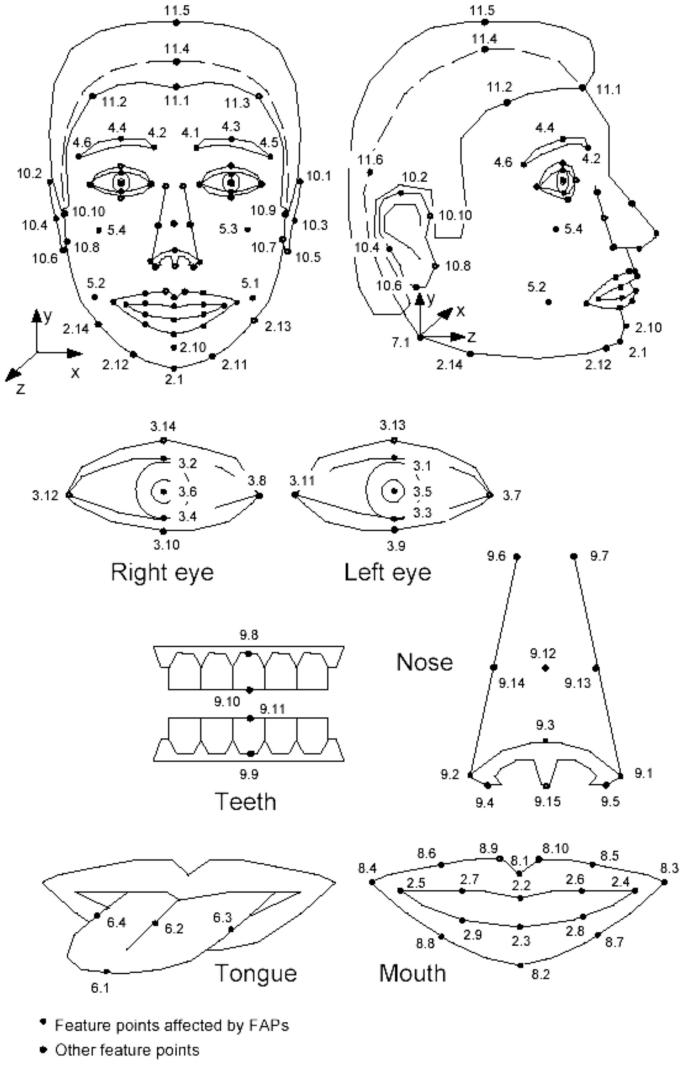

Fig. 2.2: *Feature Points* (FPs) do padrão MPEG-4 (MPEG4 VISUAL, 2001)

.

O modelo adotado pelo padrão ISO/IEC MPEG-4 (MPEG4 SYSTEM, 2001) (MPEG4 VISUAL, 2001) (MPEG4 AUDIO, 2001) também pode ser classificado como de parametrização geométrica. O padrão MPEG-4 trata da codificação e transmissão de cenas áudio-visuais interativas. Além de áudio e vídeo naturais, capturados do mundo real por microfone e câmera, o padrão comporta também objetos e animações gerados sinteticamente. Entre os objetos sintéticos suportados, o padrão inclui faces virtuais. Uma face MPEG-4 é definida por 84 FPs (*Feature Points*), os quais são controlados por 68 FAPs (*Facial Action Parameters*). Dentre os 68 FAPs, 66 são considerados de baixo nível. Cada FAP de baixo nível é utilizado para definir o deslocamento de um FP em uma determinada direção. A Figura 2.2 apresenta os FPs definidos pelo padrão. Os FPs (e os FAPs a eles associados) estão organizados em dez grupos distintos. Na figura, os FPs são identificados pelo número do grupo seguido de um ponto e de um identificador numérico individual. Em particular, os grupos 2, 6 e 8 são de interesse para a movimentação articulatória. O grupo 2 está associado à mandíbula, queixo e pontos internos dos lábios. O grupo 6 abriga pontos da língua e o grupo 8 os pontos externos dos lábios. Existem 31 FAPs associados aos FPs destes grupos. Este conjunto de FAPs está diretamente associado à movimentação articulatória.

A animação facial realista da movimentação articulatória no contexto do MPEG-4 apresenta grandes dificuldades. Os FAPs de baixo nível, apesar dos nomes sugerirem uma interpretação articulatória, são parâmetros puramente geométricos. Por exemplo, o FAP 3 (*Open-Jaw*) atua apenas no FP 2.1 (*Bottom of the chin*) não influenciando, por exemplo, os FPs 9.9 (*Bottom of the lower teeth*) e 9.11 (*Top of the lower teeth*), os quais deveriam se mover com a mandíbula em uma interpretação articulatória. No padrão esta contradição é tratada com a indicação, lacônica como ressaltado em Bailly (2001), que o FAP 3 não afeta a abertura da boca. Adicionalmente, os FAPs de baixo nível não levam em consideração gestos articulatórios característicos como, entre outros, abertura da boca, altura da boca e protrusão labial. O mapeamento destes gestos para um conjunto de FAPs não é imediato, e não é definido no padrão. Para contornar esta dificuldade, o padrão estabelece um conjunto de FAPs de alto nível para representar visemas. Entretanto, o conjunto de 14 visemas definidos possui a grave limitação de apenas contemplar, se tanto, a língua inglesa, sendo deixado para o implementador a decisão de como estes visemas deformam a face. Adicionalmente, é possível, apesar de não recomendável, modificar a face utilizando simultaneamente os parâmetros de alto e os de baixo nível, o que pode levar a resultados imprevisíveis.

Um exemplo do uso deste modelo é oferecido na abordagem descrita em Pelachaud et al. (2001) e Pelachaud (2002), em que as trajetórias de 8 pontos marcados na face de um falante foram medidas durante a produção de um corpus italiano. As medidas foram realizadas com o sistema comercial Elite da empresa BTS Bioengineering (http://www.bts.it). O corpus estudado foi formado de logatomas, ou seja palavras sem sentido, do tipo “aCa”; com C = {/p, f, t, s, k, l, λ, ʃ/}. A partir das medidas,

foram calculados quatro parâmetros articulatórios: LH (*Lip Height*), LW (*Lip Width*), UP (*Upper Lip protrusion*) e LP (*Lower Lip Protrusion*). Através de processo de minimização realizado com uma rede neural, as trajetórias dos parâmetros articulatórios foram aproximadas por funções RBFs (*Radial Basis Function*). A face virtual, compatível com o padrão MPEG-4, foi manipulada através de mapeamento geométrico das trajetórias dos parâmetros articulatórios. Para o mapeamento geométrico, foram definidas regiões de influência dos parâmetros articulatórios.

2.2.3 Parametrização data-driven

Assim como a parametrização geométrica, a parametrização *data-driven* objetiva a reprodução da deformação da superfície da face sem simular os mecanismos biomecânicos e fisiológicos associados. Entretanto, em vez de se basear em observações qualitativas, como na parametrização geométrica, a medida do comportamento real é fundamental. Na parametrização *data-driven*, os parâmetros são estabelecidos através da medida e análise estatística de dados obtidos durante a produção da fala. A análise é, em geral, baseada em variantes da análise multidimensional PCA (*Principal Component Analysis*) (JOBSON, 1992). Na parametrização *data-driven*, o método de captura e de processamento de dados é um aspecto importante.

Em Kuratate, Yehia e Vatikiotis-Bateson (1998) é utilizado um conjunto de 8 modelos geométricos tridimensionais de alta resolução, representando as configurações estáticas de 5 vogais /a, i, u, e, o/ do japonês e de 3 poses extremas: boca escancarada, boca fechada, porém relaxada, e boca fechada com os dentes cerrados. A aquisição dos dados foi realizada utilizando um informante da língua japonesa. Os modelos geométricos foram capturados com o auxílio de um digitalizador 3D da empresa Cyberware (http://www.cyberware.com). A análise PCA realizada, envolvendo os 8 modelos, resultou em 7 componentes principais. Adicionalmente, através da análise de um conjunto de 18 pontos de referência, definidos e identificados em cada um dos 8 modelos geométricos, foram calculados estimadores para os pesos das componentes principais para a realização de posturas faciais a partir da posição dos 18 pontos de referência. A geração da animação propriamente dita foi baseada em procedimento de captura de movimento, em que as trajetórias dos 18 pontos de referência foram medidas na face de um informante durante a fala e reproduzidas no modelo virtual. As medidas foram realizadas com o auxílio de equipamento Optotrak da empresa NDI (http://www.ndigital.com).

Em Elisei et al. (2001) foram efetuadas medidas fotogramétricas da posição de 168 pontos especialmente marcados na face de um falante. Foram processadas 10 vogais do francês, em posição sustentada hiperarticuladas e 8 posições de fechamento consonantais em contextos vocálico simétrico VCV (em que V representa uma vogal e C uma consoante), abrangendo as vogais /i, a, u/. Através da aplicação sucessiva da análise PCA em sub-conjuntos dos dados capturados, foram estabelecidos 6 parâmetros de controle descrevendo aproximadamente 90% da variância das poses-chave captura-

das. A abordagem foi utilizada para reproduzir em uma face sintética a movimentação da face de um falante, através de um processo de minimização do erro entre a imagem produzida sinteticamente e a capturada.

A abordagem *data driven* permite potencialmente reproduzir poses-chave articulatórias com um pequeno conjunto de parâmetros, os pesos das componentes principais. Observa-se que, em geral, a animação é produzida através da estimativa dos valores dos parâmetros baseada na captura dos movimentos observados em uma situação real. A abordagem é, portanto, especialmente vantajosa para aplicações como vídeo conferência, que envolvem a transmissão e reprodução da informação visual produzida durante a fala por uma pessoa. Entretanto, como as componentes principais nem sempre admitem uma interpretação articulatória clara, não é intuitivo qual é a composição linear destas componentes necessária para a reprodução de uma determinada postura facial. Para aplicações que não comportam a captura de movimento, a abordagem pode apresentar dificuldades e levar à produção de expressões pouco realistas.

2.2.4 Simulação biomecânica

Na parametrização geométrica discutida anteriormente, não há, em princípio, restrições quanto ao tipo e extensão das deformações que os parâmetros exercem sobre o modelo. Apesar desta permitir uma grande flexibilidade no controle das deformações, é necessário cuidado em sua utilização, uma vez que facilmente pode-se chegar a poses fisiologicamente impossíveis. Em animação onde se procura o realismo, tais distorções não são toleradas. Da mesma forma, a combinação das componentes principais da abordagem *data-driven* nem sempre garante que a pose da face virtual seja compatível com as realizáveis na prática. A abordagem biomecânica procura reduzir o espaço de configurações possíveis àquelas fisicamente realizáveis. Para tanto, a abordagem procura considerar as limitações e restrições biomecânicas, buscando simular o comportamento estrutural e dinâmico da face. A animação biomecânica possui o potencial de permitir a geração de animação com alto grau de realismo.

Na abordagem proposta em Platt e Badler (1981), um músculo é aproximado por um elemento linear que pode contrair e relaxar. Este elemento tem uma extremidade fixa em um osso e a outra, que se move, entremeada com o tecido facial. O músculo é um elemento elástico que une estas duas extremidades. Os deslocamentos dos vértices do modelo geométrico devido à contração muscular são modelados com translações na direção da atuação do músculo. O efeito na pele é representado por um conjunto de pontos 3D na superfície da face, definindo a região afetada pela ação muscular. Para codificar as ação musculares primitivas foi utilizado o sistema FACS (*Facial Action Coding System*) (EKMAN; FRIESEN, 1978), citado em (PARKE; WATERS, 1996). O sistema FACS é uma notação usualmente utilizada em animação facial para descrever e codificar expressões faciais. FACS, entre-

tanto, foi idealizada como ferramenta para a descrição de estados emocionais visíveis na face, não sendo apropriado para a descrição da movimentação articulatória.

No modelo muscular proposto em Waters (1987), tem-se uma simulação mais elaborada da ação muscular. Três tipos de músculos são tratados no modelo: músculo linear, que se contrai na direção da sua origem (ponto de ligação do músculo ao osso), músculo-laminar (*sheet-muscle*), que cobre uma região; músculo esfíncter, músculo anelar que se contrai em torno de um centro. No cálculo do deslocamento a ser aplicado aos vértices do modelo geométrico, é considerado que ocorre a dissipação da força muscular do ponto de inserção ao ponto de origem, assim como nas regiões adjacentes. Neste modelo, a elasticidade da pele é modelada implicitamente pela redução gradual dos efeitos da atuação muscular em função da distância do ponto de origem do músculo. O modelo de Waters foi adotado na abordagem de Bondy et al. (2001)

Um modelo mais elaborado para representar as características do tecido facial é apresentado em Terzopoulos e Waters (1990) e aperfeiçoado em Lee, Terzopoulos e Walters (1995). Neste modelo, as várias camadas do tecido facial que recobrem os ossos são representadas por três camadas massa-mola. Adicionalmente às forças elásticas entre os nós das várias camadas, o modelo proposto impõe restrições à movimentação para garantir a preservação do volume do tecido facial e evitar a penetração do tecido no esqueleto. Uma característica interessante deste modelo é a sua potencial capacidade para tratar rugas e dobras da pele, resultando em expressões mais naturais. A maior desvantagem é a complexidade computacional. Além disso, os parâmetros associados à características físicas do tecido, tais como espessura das camadas, características viscoelásticas do tecido facial, precisam ser estimados. Em Waters Terzopoulos e Waters (1990) os autores utilizam contornos deformáveis, chamados "*snakes*" (KASS; WITKIN; TERZOPOULOS, 1987), para detectar e rastrear, em uma seqüência de vídeo, os lábios, contorno do queixo, sulco nasolabial e sobrancelhas, especialmente destacados com o uso de maquiagem. Através da análise do vídeo foram estimadas as contrações da musculatura e assim os parâmetros do modelo, os quais foram utilizados pra acionar uma face virtual. Em Lucero e Munhall (1999) são apresentados valores para os parâmetros baseados em uma compilação de dados anatômicos disponíveis na literatura e em medidas eletromiográficas - EMG. Para evitar o procedimento invasivo associados às medidas EMG, a abordagem proposta em Pitermann e Munhall (2001) utiliza o equipamento Optotrak para a captura de movimentos de pontos da face. Através destas medidas são estimadas as ações musculares. Em Kähler, Haber e Seidel (2001) é apresentado um modelo muscular de três camadas similar ao proposto em Terzopoulos e Waters (1990).

Envolvendo um refinamento ainda maior na abordagem, é possível identificar a simulação de cirurgias faciais utilizando modelos de elementos finitos (KOCH et al., 1996) (KEEVE et al., 1996) (CHABANAS; PAYAN, 2000) (PAYAN et al., 2002). Tais soluções envolvem, em geral, a utilização de so-

fisticados dispositivos de imageamento médico como tomógrafos computadorizados, equipamentos de raio-x e de ressonância magnética. Esta abordagem objetiva a simulação acurada da estrutura e dinâmica facial e a sua aplicação típica voltada ao planejamento cirúrgico comporta os altos custos computacionais, além dos operacionais.

2.3 Especificação da animação

Os mecanismos de especificação da animação podem ser reduzidos a grandes abordagens (Figura 2.3): sistemas interativos, sistemas baseados em script e sistemas baseados em captura de movimento.

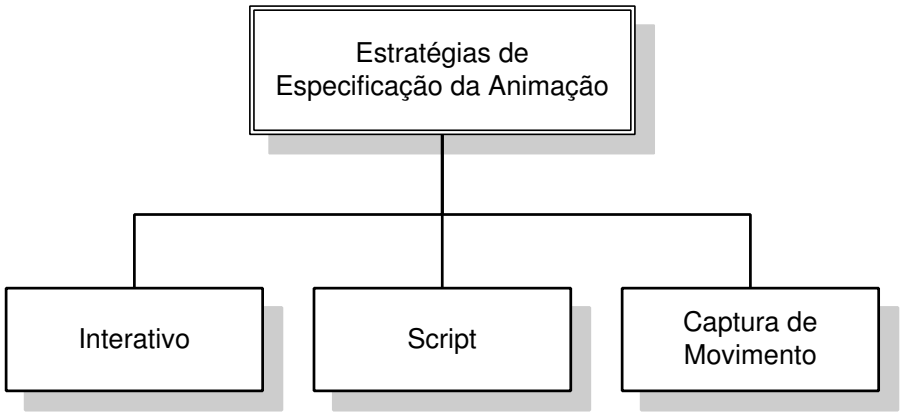

Fig. 2.3: Estratégias de controle e definição da dinâmica da animação.

Os sistemas interativos se apóiam fortemente na habilidade e talento dos animadores, estando associados, em geral, a um intenso e árduo trabalho manual para a definição dos movimentos desejados. A grande maioria dos sistemas comerciais de animação, tais como, entre outros, os pacotes Maya da empresa Alias|Wavefront, 3DStudioMax da empresa Autodesk e Softimage da empresa SoftImage, suporta esta abordagem.

Os sistemas baseados em script são caracterizados pela existência de um formalismo de descrição da dinâmica da animação, o denominado script. O script, em geral expresso em formato textual, especifica comandos que, ao serem interpretados, acionam os algoritmos disponíveis para a manipulação da geometria e produção da seqüência de deformações especificada (KALRA et al., 1991) (PELACHAUD; BADLER; STEEDMAN, 1996). O mecanismo de controle da animação sincronizada com a fala está restrito ao controle e predição da movimentação facial a partir de eventos fonéticos e, portanto, enquadra-se naturalmente em um controle baseado em script. Neste tipo de animação, a descrição

dos eventos fonéticos é realizada através da transcrição fonética da locução complementada com a informação de sua temporização, expressa, tipicamente, pela duração e instante de início.

Já os sistemas de captura de movimento baseiam-se na captura e respectiva reprodução dos movimentos executados por um operador ou ator. A captura pode ser efetuada utilizando-se equipamentos tão diversos como, *joysticks*, *data gloves*, fantoches e estruturas mecânicas dotadas de sensores, conhecidas como *Waldos* (MENACHE, 2000), ou por conjuntos de câmeras de vídeo gravando simultaneamente, em diferentes ângulos, um ator (STURMAN, 1998). Animações com alto grau de realismo são geradas utilizando captura de movimento. As duas principais desvantagens desta abordagem são o custo e necessidade da repetição da captura dos dados para cada nova animação. O custo está associado à soma do valor do equipamento, da necessidade do uso de atores e do armazenamento e processamento de grandes volumes de dados.

2.4 Comentários finais

A face é uma estrutura complexa formada por epiderme, tecido conjuntivo, músculos, cartilagens e ossos. Grande parte de sua superfície visível é flexível. A movimentação da face produz detalhes sutis como dobras e rugas que estamos acostumados a reconhecer. A familiaridade com a face advinda de anos de observação implica em um natural, acurado e exigente senso crítico na apreciação de animações faciais por computador.

As atividades de pesquisa voltadas para a animação facial por computador estendem-se já por aproximadamente 35 anos. Neste período, passou-se de representações hoje consideradas grosseiras, que mal conseguiam piscar os olhos e abrir a boca de forma realista, para modelos sofisticados que procuram modelar a realidade biodinâmica e fisiológica da face.

Os desenvolvimentos da animação facial foram aqui acomodados em duas grandes categorias, aqueles baseados na manipulação de uma representação geométrica, via de regra tridimensional, e aqueles baseados na manipulação de imagens.

O reconhecimento da importância das pistas visuais presentes na face durante a fala tem motivado um grande número de pesquisas voltadas à animação realista sincronizada com a fala. A questão fundamental que permeia estas pesquisas é a reprodução da movimentação articulatória visível na face virtual.

O presente trabalho aborda a animação facial sincronizada com a fala enfatizando a reprodução realista da movimentação articulatória visível, e se insere no contexto da animação facial de modelo geométrico tridimensional. Nos capítulos que se seguem, a principal preocupação é o estabelecimento de representações visuais para os fones do português do Brasil, os denominados visemas. Os visemas estabelecem a referência a ser seguida durante a animação para a reprodução da movimentação

articulatória. A estratégia básica de manipulação da geometria da face virtual adotada no presente trabalho pode ser classificada como parametrização geométrica.

No que se refere ao controle da animação, a estratégia adotada é baseada em script, o qual descreve a seqüência dos eventos fonéticos e as respectivas temporizações do sinal da fala. Tal script pode ser derivado como subproduto de um processo de conversão texto-fala, ou através da análise, manual ou automática, do sinal acústico da fala.

Capítulo 3

Produção e representação da fala

3.1 Introdução

A animação facial engloba as técnicas para a especificação e controle da movimentação de uma face virtual. Excetuando animações caricaturais, em que se aceitam exageros e distorções da realidade, uma das forças que impulsiona o desenvolvimento das técnicas de animação facial é a busca do realismo. Dentro deste contexto, destaca-se a simulação dos movimentos visíveis na face ocasionados durante a produção da fala pela movimentação dos órgãos que compõem o trato vocal. A animação facial realista sincronizada com fala baseia-se no inter-relacionamento entre os eventos sonoros e os articulatórios visíveis e necessários à produção acústica. O formalismo fonético utilizado para a descrição da fala é o tema principal deste capítulo. O objetivo principal do capítulo é estabelecer uma base de conceitos e nomenclaturas associados, principalmente, à descrição dos aspectos articulatórios da produção da fala. Tais conceitos são utilizados no desenvolvimento dos capítulos subseqüentes. O conteúdo do capítulo é um recorte de conceitos bem estabelecidos no campo do conhecimento lingüístico denominado *Fonética Articulatória*.

Na Seção 3.2 são apresentados os aspectos articulatórios envolvidos na produção da fala, com ênfase naqueles que contribuem para a formação de pistas visuais observáveis na face do falante.

Na Seção 3.3 são abordados conceitos básicos da fonética como segmento, fonema e alvo articulatório. Um recorte importante efetuado nesta seção é a identificação de um conjunto de fonemas, que serão considerados como os representantes do português falado no Brasil. Ainda na seção, estes fonemas são analisados e classificados segundo aspectos articulatórios em consoantes e vogais. Além disso, são apresentadas as possíveis seqüências de ocorrência de vogais e consoantes no português do Brasil.

3.2 Fisiologia da produção da fala

O conjunto de órgãos envolvidos na produção da fala é denominado Aparelho Fonador. A Figura 3.1 apresenta de forma esquemática os principais órgãos que compõem este aparelho. Levando-se em conta a função desempenhada por seus elementos, o aparelho fonador pode ser decomposto em três grandes seções:

1. *Seção Respiratória*;

2. *Seção Fonatória*;

3. *Seção Articulatória*.

A *Seção Respiratória* é composta pelos pulmões, músculos pulmonares, brônquios e traquéia. A principal função desta seção é fornecer a corrente de ar necessária para a produção da fala. Em particular, os sons do português são produzidos exclusivamente pela corrente expiratória advinda dos pulmões. O caminho da corrente de ar pulmonar, na ausência de qualquer obstrução, é representado na Figura 3.1 pela linha pontilhada que sai do pulmão, passa pela traquéia e laringe, se divide na faringe, indo parte para a cavidade oral e parte para as cavidades nasais, terminando por sair pela boca e nariz.

A *Seção Fonatória* é constituída essencialmente pela laringe. A laringe é uma caixa de cartilagem e tecido mole, localizada acima da traquéia, no interior da qual localizam-se as pregas vocais. As pregas vocais são duas faixas de tecido muscular na forma de lábios, que se estendem horizontalmente no interior da laringe e que, por movimentação voluntária, podem obstruir a passagem do ar. À abertura entre as pregas vocais é dada a denominação de glote. A glote encontra-se aberta durante a respiração normal, sendo fechada para a fonação. Durante a fonação, a pressão do fluxo de ar expiratório força a abertura da glote, permitindo a passagem de um pulso de ar em alta velocidade. Este pulso de ar, ao escapar, produz uma compensação momentânea de pressão entre as regiões supra e infraglótica, promovendo o fechamento da glote. Mantida a pressão do fluxo de ar expiratório e a tensão das pregas vocais, o ciclo se repete, resultando na produção de uma série de pulsos de curta duração que excitam a coluna de ar supralaríngea, produzindo um sinal sonoro que caracteriza a fonação. Este sinal sonoro, denominado de tom laríngeo, tem característica harmônica, tendo espectro constituído de freqüência fundamental e de múltiplos desta freqüência. A característica espectral do tom laríngeo pode, em certo grau, ser controlada durante a fala através de uma maior ou menor contração dos músculos das pregas vocais, permitindo a produção de tons, respectivamente, mais agudos ou mais graves, ou, equivalentemente, com freqüência fundamental maior ou menor.

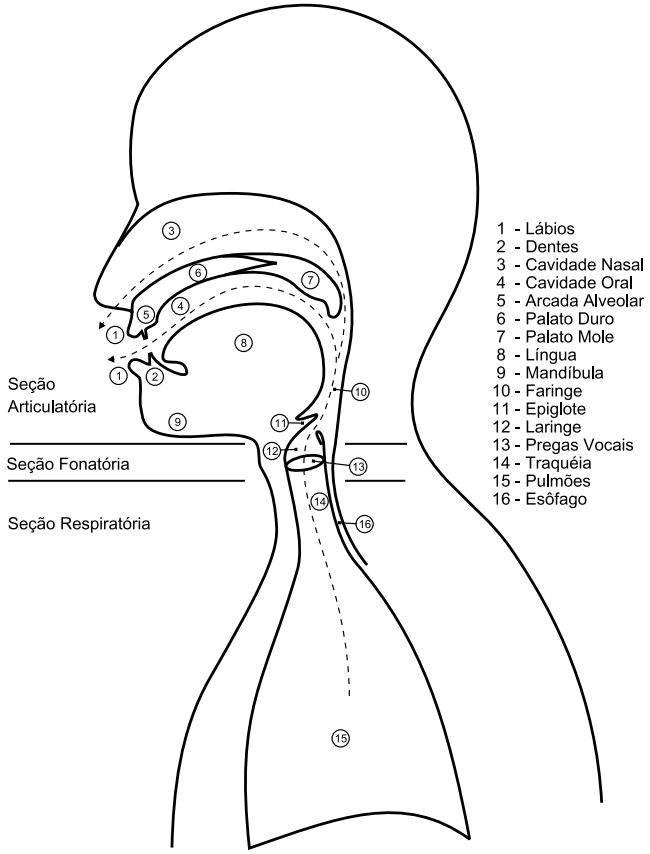

Fig. 3.1: Aparelho Fonador.

A *Seção Articulatória* é constituída pela faringe, cavidade oral, vestíbulo oral, cavidades nasais, língua, dentes e lábios. O trato vocal, formado pela faringe, cavidade oral, vestíbulo oral e cavidades nasais, atua como filtro ressonador do tom laríngeo, atenuando determinadas freqüências e reforçando outras. A resposta em freqüência deste filtro é definida pela forma do trato vocal, em especial, das cavidades orais e nasais e do vestíbulo oral. A forma da cavidade oral é essencialmente modificada pela movimentação da língua e da mandíbula. Já as cavidades nasais, que possuem forma e volume fixos, são inseridas ou retiradas do processo de produção da fala com o abaixamento ou levantamento do palato mole ou véu palatino. O abaixamento, ao conectar as cavidades nasais ao caminho do fluxo de ar expiratório, aumenta o comprimento e a complexidade do trato vocal, resultando em uma modificação das características do som emitido. O vestíbulo oral ou cavidade bucal é o pequeno espaço limitado pelos lábios, bochechas e, internamente, por gengivas e dentes. Este espaço também participa na definição das características do filtro, atuando como cavidade ressonadora, sendo altamente variável quanto à forma e ao volume, dependendo principalmente da postura dos lábios. Os órgãos ou partes do aparelho fonador que se movem para modificar a configuração das cavidades supralaríngeas durante a fala são denominados de articuladores ativos. Exemplos de articuladores ativos são:

língua, lábio inferiore e palato mole. Considera-se articulador passivo, o órgão, ou parte do aparelho fonador, do qual o articulador ativo se aproxima para a configuração do trato vocal. Lábio superior, dentes incisivos superiores, arcada alveolar e palato duro são exemplos de articuladores passivos.

3.3 Sons do português do Brasil

A fonética é o ramo da lingüística voltado ao estudo da produção e percepção dos sons da fala em seus aspectos articulatórios e auditivos. Uma premissa básica da análise fonética consiste em considerar o contínuo da fala como composto por uma seqüência de unidades discretas, os denominados segmentos. A segmentação fonética apóia-se no conhecimento lingüístico dos instantes onde mudanças no sinal acústico resultam em alterações do significado das palavras. Assim, por exemplo, a palavra "mata" é composta por quatro segmentos distintos: "m", "a", "t" e "a". Para esta palavra é possível identificar as seguintes alterações que resultam em mudanças lingüísticas significativas, dando lugar a palavras diferentes: "pata", "meta", "mapa", "mato". A realização acústica de um segmento é denominada fone.

A fonética classifica ainda os segmentos em duas categorias fundamentais, a saber, consonantal e vocálico. A caracterização fonética dos segmentos é tradicionalmente efetuada através da descrição de alvos articulatórios ("articulatory targets") (IPA, 1999). Um *Alvo Articulatório* pode ser entendido como a representação estática da conformação do trato vocal característica do segmento. Em essência, os segmentos consonantais são caracterizados por posturas articulatórias associadas à obstrução significativa na linha sagital média do trato vocal na postura representativa. Já os segmentos vocálicos são produzidos sem uma obstrução significativa, mas com um posicionamento característico da língua e na maioria das vezes com a vibração das pregas vocais.

No português do Brasil, assim como em outras línguas, é possível identificar palavras com significados diferentes, cujas cadeias sonoras se distinguem apenas por um segmento. Os segmentos que distinguem estas palavras constituem os fonemas da língua. Um fonema pode ser entendido como uma unidade lingüística abstrata, representativa das produções sonoras que, em conjunto, possuem a mesma função distintiva na estrutura da língua. Produções sonoras diferentes, porém com a mesma função distintiva, ou seja, associadas a um mesmo fonema, são denominadas alofones. Os pares de palavras, como por exemplo "mata" e "pata", que permitem a identificação de fonemas "m" e "p" constituem os denominados pares mínimos. Um par mínimo permite a identificação de fonemas por contraste. Mantendo-se a convenção usualmente empregada, na representação dos fones (realização acústica), serão utilizados símbolos entre colchetes, por exemplo [p]. Para a representação dos fonemas (unidade distintiva abstrata), os símbolos serão apresentados entre barras oblíquas, por exemplo /p/.

Na produção da fala, os segmentos se agregam para formar unidades maiores, as sílabas. Toda locução é constituída por pelo menos uma sílaba. Do ponto de vista articulatório, os sucessivos estreitamentos do trato vocal, para a produção de segmentos consonantais, e aberturas, para a produção de segmentos vocálicos, estabelecem uma referência para a identificação das sílabas (IPA, 1999). A estrutura interna da sílaba pode ser considerada como tendo três partes: ataque, núcleo e coda. Das três partes, apenas o núcleo é obrigatório, sendo constituído, no que se refere a este trabalho, de um a três segmentos vocálicos. Já o ataque e a coda são opcionais, e podem ser formados por até dois segmentos consonantais cada um no português do Brasil.

A seguir são descritos os segmentos consonantais e vocálicos encontrados tipicamente na variante da região urbana do Estado de São Paulo. Neste trabalho, esta variante será considerada genericamente como representante do português falado no Brasil. Para a representação textual dos segmentos da fala, será utilizado o Alfabeto Fonético Internacional (IPA - International Phonetic Alphabet) (IPA, 1999).

3.3.1 Segmentos consonantais

É possível descrever e classificar os segmentos consonantais com base em três parâmetros articulatórios:

1. *Estado da Glote*;
2. *Lugar de Articulação* (ou *Ponto de Articulação*) ;
3. *Modo de Articulação*

O *Estado da Glote*, se aberta ou fechada, permite identificar duas categorias de segmentos consonantais: segmentos vozeados ou sonoros, e segmentos desvozeados ou surdos. Nos segmentos vozeados, a glote encontra-se fechada, ocorrendo a vibração das pregas vocais. Já nos segmentos desvozeados, a glote encontra-se aberta, não ocorrendo a vibração das pregas vocais.

O *Lugar de Articulação* identifica a região do trato vocal onde ocorre a obstrução significativa que caracteriza o segmento consonantal. Os seguintes lugares de articulação são relevantes no português do Brasil (veja Figura 3.2):

- Bilabial: o articulador ativo é o lábio inferior e o articulador passivo é o lábio superior. A aproximação dos lábios é conduzida pela movimentação da mandíbula e pela ação de músculos faciais que atuam comprimindo os lábios;
- Labiodental: o articulador ativo é o lábio inferior e como articulador passivo têm-se os dentes incisivos superiores;

- Alveolar: o articulador ativo é o ápice da língua e como articulador passivo tem-se a arcada alveolar;

- Pós-alveolar: o articulador ativo é a parte anterior da língua e o articulador passivo é a parte central do palato duro;

- Palatal: o articulador ativo é a parte média da língua e o articulador passivo é a parte final do palato duro;

- Velar: o articulador ativo é a parte posterior da língua e o articulador passivo é o véu palatino.

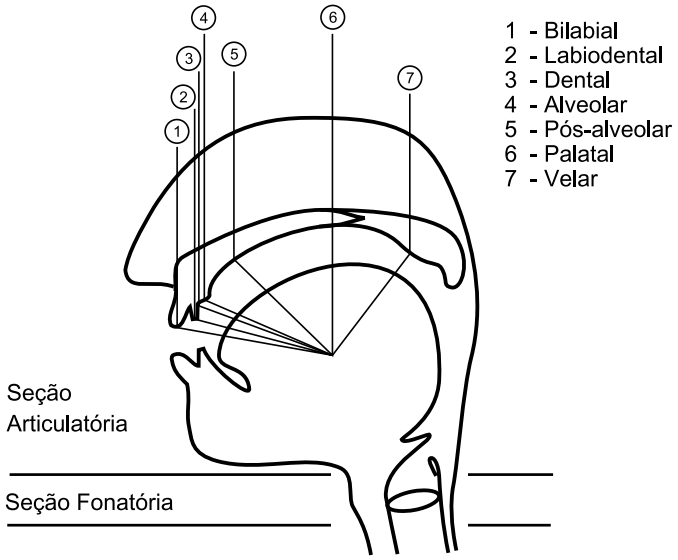

Fig. 3.2: Lugares de Articulação.

O *Modo de Articulação* refere-se ao grau e à natureza da obstrução/constrição do trato vocal característica do segmento, estando intimamente associado à posição relativa dos articuladores, indicando como e em que grau o fluxo de ar pulmonar é obstruído. Os seguintes modos de articulação são de interesse para o português do Brasil:

- Plosivo ou Oclusivo: os articuladores produzem obstrução total do fluxo de ar pulmonar. O véu palatino está levantado e o ar que vem dos pulmões é forçado para a cavidade oral. A abertura rápida da oclusão, forçada pelo fluxo de ar, transitando para segmento vocálico produz o som do segmento.

- Nasal: os articuladores produzem a obstrução total da passagem da corrente de ar pela boca. O véu palatino encontra-se abaixado e o ar que vem dos pulmões é desviado para a cavidade nasal.
- Fricativo: os articuladores não obstruem totalmente o trato vocal, mas ocorre aproximação suficiente para produzir fricção e turbulência no fluxo de ar pulmonar.
- Vibrante: o articulador ativo toca algumas vezes o articulador passivo causando vibração.
- Tepe (ou "Tap"): o articulador ativo toca uma única vez o articulador passivo ocasionando uma rápida obstrução do fluxo de ar.
- Lateral: o articulador ativo toca o articulador passivo de forma a causar a obstrução do fluxo de ar na linha sagital média do trato vocal. O ar é expelido pelos lados da obstrução, caracterizando o escape lateral e o nome do modo de articulação.

		Lugar de Articulação					
		Bilabial	**Labio-dental**	**Alveolar**	**Pós-alveolar**	**Palatal**	**Velar**
Modo de Articulação	**Plosivo**	p b		t d			k g
	Nasal	m		n		ɲ	
	Vibrante						
	Tepe			ɾ			
	Fricativo		f v	s z	ʃ ʒ		ɣ
	Lateral			l		ʎ	

Tab. 3.1: Segmentos consonantais do português do Brasil (BARBOSA; ALBANO, 2004).

A Tabela 3.1 apresenta os segmentos consonantais do português do Brasil. Estes segmentos são representativos da variante regional urbana do Estado de São Paulo (BARBOSA; ALBANO, 2004). Nas células da tabela, onde símbolos aparecem aos pares, o segmento à esquerda é desvozeado e o à direita vozeado. As linhas da tabela expressam a classificação dos segmentos consonantais segundo o modo de articulação, como indicado pelos rótulos à esquerda. As colunas da tabela expressam a classificação dos segmentos segundo o lugar de articulação, como indicado pelos rótulos no topo da

tabela. Os lugares de articulação são apresentados, da esquerda para direita, a partir dos lugares mais externos aos mais internos da cavidade oral.

Símbolo	**Classificação**	**Exemplos Ortográficos**	
		Posição de Ataque	
		Início Palavra	**Intervocálica**
p	Plosivo bilabial desvozeado	**p**ata	ri**p**a
b	Plosivo bilabial vozeado	**b**ata	ri**b**a
t	Plosivo alveolar desvozeado	**t**ata	la**t**o
d	Plosivo alveolar vozeado	**d**ata	la**d**o
k	Plosivo velar desvozeado	**c**ata	va**c**a
g	Plosivo velar vozeado	**g**ata	va**g**a
f	Fricativo labiodental desvozeado	**f**aca	lu**f**a
v	Fricativo labiodental vozeado	**v**aca	lu**v**a
s	Fricativo alveolar desvozeado	**s**aca	ca**ç**a
z	Fricativo alveolar vozeado	**z**aca	ca**s**a
ʃ	Fricativo pós-alveolar desvozeado	**ch**aga	quei**x**o
ʒ	Fricativo pós-alveolar vozeado	**j**aca	quei**j**o
m	Nasal bilabial vozeado	**m**ata	ca**m**a
n	Nasal alveolar vozeado	**n**ata	ca**n**a
ɲ	Nasal palatal vozeado	**nh**ambi	ga**nh**o
l	Lateral alveolar vozeado	**l**ata	ga**l**o
ʎ	Lateral palatal vozeado	**lh**ama	ga**lh**o
ɾ	Tepe alveolar vozeado		ca**r**o
ɣ	Fricativo velar vozeado	**r**ata	ca**rr**o

Tab. 3.2: Classificação dos segmentos consonantais do português do Brasil.

Arqui-fonema	Fone	Exemplos Ortográficos		
		Posição de Coda		
		Final palavra [a]	Antes de C desvozeada	Antes de C vozeada
/S/	[s]	mê**s**	ca**s**ca	
	[z]			ra**s**ga, me**s**mo
/R/	[ɾ]	ma**r**	ma**r**ca	ama**r**ga [b]
	[ɣ]			

Tab. 3.3: Arquifonemas consonantais do português do Brasil.

[a] Mais precisamente antes de pausa silenciosa.

[b] A realização de /R/ é bastante variável no português do Brasil, sendo que o tepe é a realização mais comum da variante considerada neste trabalho (BARBOSA; ALBANO, 2004)

Arqui-fonema	Fone	Exemplos Ortográficos						
		Posição de Ataque				Posição de Coda		
		Início Pala-vra	Após C mesma silaba	Entre vogais	Após C de outra sílaba	Final de Palavra	Antes de C Desvoze-ada	Antes de C Voze-ada
/L/	[l]	**l**ata	p**l**aca	ga**l**o	bur**l**a			
	[ʊ̯]					ma**l**	go**l**pe	mo**l**de

Tab. 3.4: Realizações do fone [l].

Na Tabela 3.2 são apresentados exemplos de palavras que ilustram os segmentos consonantais considerados neste trabalho. A primeira coluna da tabela apresenta os símbolos IPA dos segmentos

consonantais. A segunda explicita a classificação do segmento conforme os parâmetros articulatórios apresentados anteriormente. Nos exemplos ortográficos, a letra, ou letras, em negrito constituem a representação ortográfica do segmento em questão. Os exemplos ortográficos estão organizados em duas colunas, ambas associadas às sílabas nas quais o segmento aparece em posição de ataque (início de sílaba). A primeira coluna exemplifica a ocorrência do segmento no início de uma palavra. Na segunda, tem-se exemplos da ocorrência do segmento consonantal entre segmentos vocálicos (veja Seção 3.3.2 que trata dos segmentos vocálicos).

Os exemplos ortográficos da Tabela 3.2 apresentam pares mínimos que permitem a identificação por contraste dos 19 fonemas consonantais do português do Brasil: /p/, /b/, /t/, /d/, /k/, /g/, /f/, /v/, /s/, /z/, /ʃ/, /ʒ/, /m/, /n/, /ɲ/, /l/, ʎ, /ɾ/ e /ɣ/. Três pares destes fonemas, entretanto, perdem o contraste fonêmico em contextos bem determinados e em distribuição complementar, ou seja, quando um ocorre o outro não ocorre, e vice-versa. O contraste fonêmico de /s/ e /z/ é neutralizado na posição de coda (final de sílaba), ocorrendo sistematicamente como [z] quando seguido de segmento consonantal vozeado e como [s] nas situações complementares - posição de coda, seguido de segmento consonantal desvozeado ou diante de pausa silenciosa (final de palavra). O termo arquifonema é utilizado para designar a perda do contraste entre fonemas. O arquifonema associado a /s/ e /z/ é usualmente representado por /S/ (MONARETTO; QUEDNAU; HORA, 2001) (SILVA, 2002) (BARBOSA; ALBANO, 2004). Nos mesmos contextos lingüísticos, os fonemas /ɾ/ e /ɣ/ também perdem o caráter distintivo, dando lugar ao arquifonema /R/. A Tabela 3.3 apresenta exemplos ilustrativos de situações de neutralização de contraste associadas aos arquifonemas /S/ e /R/.

Já o arquifonema /L/ manifesta-se em todas as posições, excetuando a de coda, como um segmento consonantal lateral alveolar vozeado - [l]. Já na posição de coda, na maioria dos dialetos do português brasileiro, ocorre a vocalização deste fonema (SILVA, 2002). Neste caso, o mesmo pode ser foneticamente expresso por um segmento consonantal lábio velar vozeado [w] ou pela vogal assilábica [ʊ̯] (veja Seção 3.3.2 que trata dos segmentos vocálicos). As duas representações fonéticas, [w] e [ʊ̯], são equivalentes no que tange às características articulatórias e acústicas do som produzido. A diferença entre as duas representações reside na interpretação fonológica (se consoante ou se vogal) que se deseja emprestar ao segmento. A Tabela 3.4 apresenta exemplos de palavras envolvendo o fonema /l/, com a indicação dos fones associados a cada uma das realizações articulatórias. Neste trabalho, será adotada a vogal assilábica para representar a realização do /l/ em posição de coda.

Como mencionado na Seção 3.3, o padrão silábico do português do Brasil admite sílabas com zero, um ou dois segmentos consonantais, tanto no ataque como em posição de coda. As Tabelas 3.5 a 3.7 apresentam padrões silábicos envolvendo segmentos consonantais encontrados no português do Brasil considerado neste trabalho. Nos rótulos apresentados no topo das tabelas, a letra "C" indica a ocorrência de um segmento consonantal e a letra "V" a ocorrência de segmento vocálico.

Seg.	Exemplo					
	CV	**CVV**	**CVVV**	**CVC**	**CVVC**	**CVVVC**
/p/	**pa**ta	cha**péu**		**pes**ca, **pes**ga, **per**da	cha**péus**	
/b/	**ba**ta	**boi**na		**bis**po, **bis**bilho, **bor**da	**bois**	
/t/	**ta**ta	a**teu**		**tos**ta, sa**tis**dar, **tor**ta	a**teus**	
/d/	**da**ta	ju**deu**		**des**te, **des**de, **dar**do	a**deus**	
/k/	**ca**ta	pe**cou**	**qual**	**cas**pa, **cas**bá, **car**pa	man**cais**	**quais**
/g/	**ga**ta	**gau**lês	**guai**aca	**gas**ta, **gos**ma, **gor**da	le**gais**	i**guais**
/f/	**fa**ca	mo**fou**		**fis**co, **fis**go, **for**ca	tro**féus**	
/v/	**va**ca	ca**vei**		**vis**co, **vis**go, **ver**te	casca**véis**	
/s/	**sa**ca	ca**çou**		**cis**co, **ses**go, **cer**co	bo**cais**	
/z/	**za**ca	ca**sou**		na**zis**ta, na**zis**mo, ca**ser**na	an**zóis**	
/ʃ/	**cha**ga	**chei**ro		**xis**to, **chus**ma, **char**co	mare**chais**	
/ʒ/	**ja**ca	**jei**to		**jus**to, silo**gis**mo, **jar**da	bre**jais**	
/m/	**ma**ta	**moi**ta		**mús**culo, **mus**go, **mar**ca	nor**mais**	
/n/	**na**ta	**noi**te		**nes**te, **nes**ga, **ner**vo	tú**neis**	
/ɲ/	ga**nho**	ba**nhei**ro		pe**nhas**co, monta**nhis**mo, ba**nhar**	pu**nhais**	
/l/	ga**lo**	**lei**te		**les**te, **les**ma, **ler**do	fa**lais**	
/ʎ/	ga**lho**	o**lhei**		traba**lhis**ta, traba**lhis**mo, o**lhar**	i**lhéus**	
/ɾ/	ca**ro**	sa**rei**		ca**ras**, ma**ras**mo, a**rar**	he**róis**	
/ɣ/	ca**rro**	**rei**no		**res**friar, **res**gate, ba**rrar**	cu**rrais**	

Tab. 3.5: Padrões silábicos com um segmento consonantal no ataque e até um segmento consonantal em coda.

Encontro	**Exemplo**			
	CCV	**CCVV**	**CCVC**	**CCVVC**
/pɾ/	**pra**ta	**prai**a	**pres**crito, **pres**biopia, com**prar**	cú**preos**
/pl/	**pla**ca	**plai**na	**plás**tico, **plas**ma, exem**plar**	pim**pleus**
/bɾ/	**bra**ço	he**breu**	ca**bres**to, maca**bris**mo, co**brar**	he**breus**
/bl/	**ble**fe	nu**blou**	**blas**to, bi**blis**mo, nu**blar**	dá**blios**
/tɾ/	**tri**go	**trei**no	**tras**te, **tras**go, fil**trar**	vi**trais**
/tl/	a**tle**ta	ca**tléi**a	a**tlas**	
/dɾ/	**dra**ga	la**drou**	**drás**tico, enxa**dris**mo, la**drar**	cate**drais**
/kɾ/	**cri**me	**cria**	**cris**ta, **cris**ma, lu**crar**	sepul**crais**
/kl/	**cli**ma	mes**clou**	ci**clis**ta, ci**clis**mo, te**clar**	**claus**tro
/gɾ/	**gra**ma	lo**grei**	a**gres**te, **gras**nar, lo**grar**	jo**grais**
/gl/	**glo**bo	**glau**coma	si**glas**, si**glar**	si**glais**
/fɾ/	**fro**ta	**frau**de	**fras**co, **fres**nel, so**frer**	alcacho**frais**
/fl/	**fle**cha	**fleu**ma	**flós**culo, in**flar**	in**flais**
/vɾ/	la**vra**	li**vrei**	li**vres**co, li**vrar**	li**vrais**
/vl/	**Vla**dmir			

Tab. 3.6: Padrões silábicos com dois segmentos consonantais no ataque e até um segmento consonantal em coda.

a)

Segmento	Exemplo	
	VC	VVC
/s/	**es**pada	**aus**pício
/z/	**es**boço	
/ɣ/	**ar**te	**iur**ta

b)

Segmento	Exemplo
	CVCC
/p/	**pers**pectiva
/t/	in**ters**tício
/v/	**vers**ta

Tab. 3.7: Padrões silábicos: a) sem ataque e um segmento consonantal em coda; b) um segmento consonantal no ataque e dois em coda.

A Tabela 3.8 apresenta outros casos possíveis de encontros consonantais, que, neste trabalho, serão considerados como pseudo encontros consonantais, pois será assumida a existência da vogal epentética /ɪ/ entre os segmentos consonantais. Assim , por exemplo, a palavra "pneu" seria transcrita foneticamente como [pɪ.neʊ], composta de duas sílabas. Na transcrição fonética apresentada, o ponto indica a separação silábica.

As Tabelas 3.5 a 3.8 permitem as seguintes observações:

- Na posição de ataque, o português do Brasil admite zero, uma ou duas consoantes;
- Em sílabas, com ataque constituído por apenas uma consoante, quando o núcleo for composto por até duas vogais, pode ocorrer qualquer um dos fonemas consonantais, sendo que no início de palavra o tepe e o segmento nasal palatal não ocorrem;
- Na posição de ataque, quando o núcleo for composto por três segmentos vocálicos, ocorrem apenas os fonemas /k/ e /g/;
- Em sílabas com dois segmentos consonantais na posição de ataque, o segundo fonema consonantal é /ɾ/ ou /l/, sendo que os fonemas que podem ocorrer na primeira posição são restritos aos elementos do conjunto: /p/, /b/, /k/, /g/, /t/, /d/, /f/ e /v/.
- Em sílabas com coda constituída por apenas um segmento consonantal, ocorrem apenas os arquifonemas /S/ e /R/ nesta posição. Na posição de coda, o fonema /l/ se manifesta como o segmento vocálico [ʊ̯] (ou [w]);
- Em sílabas com coda constituída por dois segmentos consonantais, ocorre apenas o encontro /ɾs/. É importante observar que no contexto deste trabalho, a ocorrência do encontro /ns/ não será tratada como um encontro consonantal, mas sim como uma vogal nasal seguida de /S/;

- Os seguintes padrões silábicos envolvendo segmentos consonantais são relevantes para o português do Brasil no contexto deste trabalho: CV, CVV, CVVV, CVC, CVVC, CVVC, CCV, CCVV, CCVC, CCVVC, VC, VVC e CVCC;

- Considerando os padrões silábicos e as possibilidades de combinação em uma palavra ou entre palavras, identifica-se que a maior seqüência de segmentos consonantais possível de ser realizada em palavras e entre palavras no português do Brasil é VCCCCV, como, por exemplo, na palavra "perstrição".

Encontro	Exemplo
/pt/	**pte**rodátilo, ade**pt**o
/pn/	**pn**eu, a**pn**éia
/ps/	**ps**oríase, autó**ps**ia, fórce**ps**
/bp/	su**b-p**adrão
/bb/	su**b-b**ase
/bt/	o**bt**uso
/bd/	**bd**élio
/bc/	su**bc**onjunto
/bg/	su**bg**rupo
/gm/	su**bm**isso
/bn/	a**bn**egado
/bf/	su**bf**aturamento
/bv/	ó**bv**io
/bs/	a**bs**oluto
/bz/	su**bz**ona
/bʃ/	su**bch**efe
/bʒ/	o**bj**eto
/bɣ/	su**b-r**otina
/tb/	poli**tb**uro

Encontro	Exemplo
/tk/	vie**tc**ongue
/tm/	**tm**ese, ri**tm**o
/tn/	e**tn**ia
/ts/	**ts**é-tsé
/tz/	**tz**ar, quar**tz**o, her**tz**
/tʃ/	**tx**ucarramãe , pa**tch**uli
/tʒ/	lu**tj**anídeo
/dp/	a**dp**resso
/dk/	vo**dc**a
/dm/	a**dm**itir
/dn/	a**dn**ato
/dv/	a**dv**erso
/ds/	a**ds**tringente
/dz/	**dz**ubucuá, a**dz**âneni
/dʒ/	a**dj**etivo
/kp/	e**cp**iesma
/kb/	e**cb**ólico
/kt/	la**ct**ose
/kd/	e**cd**êmico

Encontro	Exemplo
/km/	a**cm**e
/kn/	a**cn**e
/kf/	e**cf**onema
/ks/	**cs**i, proli**xo**
/kz/	**cz**ar, e**cz**ema
/gb/	**gb**ari
/gd/	ba**gd**ali
/gm/	do**gm**a
/gn/	**gn**omo, i**gn**ição
/gf/	esta**gf**lação
/gs/	tun**gs**tênio
/mn/	**mn**emônico
/ms/	a**ms**terdamês
/ft/	o**ft**álmico, afta
/fn/	hó**fn**io
/vn/	czare**vn**a

Tab. 3.8: Pseudo encontros consonantais silábicos e não silábicos.

3.3.2 Segmentos vocálicos

Os segmentos vocálicos são sempre realizados com a vibração das pregas vocais. As características acústicas destes segmentos são influenciadas essencialmente pela língua, mandíbula e lábios, que moldam a forma do trato vocal. Um segmento vocálico sempre é pronunciado com a ponta da língua abaixada, formando no interior da cavidade oral uma superfície convexa, cujo ponto mais alto define o local de maior constrição do trato vocal que, em grande parte, caracteriza o segmento vocálico. Esta constrição, entretanto, não causa a obstrução total do trato ou ruído de fricção pela passagem do fluxo de ar pulmonar. A Figura 3.3 apresenta de forma esquemática a vista em corte no plano sagital médio do trato vocal. Na figura, a língua é representada em três posturas diferentes sempre formando uma superfície convexa no interior da cavidade oral. A título de ilustração são indicados na figura os pontos de maior constrição que caracterizam os segmentos vocálicos /i/, /a/ e /u/.

Três características articulatórias são usualmente utilizadas para classificar os segmentos vocálicos. Duas destas características caracterizam a posição do ponto mais alto da língua no interior da cavidade oral; a outra descreve a postura dos lábios. Estas três características são:

1. Posicionamento Vertical da Língua (altura da língua);

2. Posicionamento Anterior/Posterior da Língua; e

3. Postura dos Lábios.

Para a caracterização dos segmentos vocálicos segundo o *Posicionamento Vertical da Língua*, são considerados quatro níveis de referência para a altura da língua, mais especificamente, do ponto mais alto da língua: alta, média-alta, média-baixa, e baixa. A língua na posição alta encontra-se na posição mais elevada e mais próxima do articulador passivo; a língua na posição mais baixa encontra-se na posição mais distante do articulador passivo, em posição de repouso no chão da boca. Dois pontos adicionais eqüidistantes definem as posições intermediárias média-alta e média-baixa. Observa-se que no posicionamento vertical da língua, a mandíbula também atua, estando mais fechada na posição alta e mais aberta na posição baixa. As denominações fechada, meio-fechada, meio-aberta e aberta também são utilizadas por alguns autores para classificar os segmentos vocálicos, com a seguinte equivalência: alta/fechada; média-alta/meio-fechada; média-baixa/meio-aberta; baixa/aberta (IPA, 1999) (SILVA, 2002).

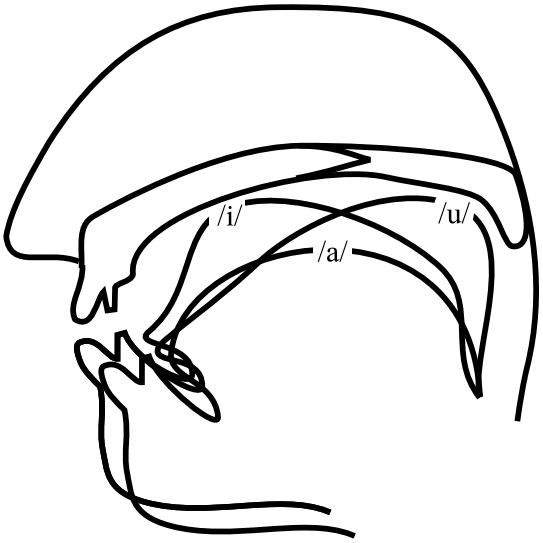

Fig. 3.3: Posição da língua na produção das vogais /i/, /a/ e /u/.

Para a classificação segundo o *Posicionamento Anterior/Posterior da Língua*, é considerado que a cavidade oral é constituída por três regiões contíguas e distintas: uma localizada na frente (anterior); outra localizada atrás (posterior); e a terceira localizada entre estas (central). Durante a articulação de um segmento vocálico, a língua é posicionada de modo a produzir o ponto de maior constrição do trato vocal em uma das três regiões. Considerando este posicionamento, os segmentos vocálicos podem ser classificados em anterior, central ou posterior.

O critério da *Postura dos Lábios* permite classificar os segmentos vocálicos em arredondados e não-arredondados. Na postura arredondada, os lábios se encontram protraídos e arredondados, sendo que na postura não-arredondada isto não ocorre. No português do Brasil, arredondamento e protrusão ocorrem apenas nos segmentos vocálicos posteriores. Nestes segmentos, o arredondamento e a conseqüente protrusão é maior no segmento alto (fechado), diminuindo com a altura do segmento vocálico.

Os segmentos vocálicos do português do Brasil admitem ainda a nasalização, ou seja, a produção com o véu palatino abaixado. Assim, é possível ainda classificar os segmentos vocálicos nas categorias *Oral* e *Nasalizado* (ou *Nasal*).

O português é uma língua que possui um ritmo da fala marcado por acentos silábicos, existindo sílabas com proeminência acentual. As sílabas com proeminência são denominadas sílabas tônicas, em contraposição às outras que são denominadas átonas. A vogal que compõe a sílaba tônica é denominada vogal tônica. Na realização de uma palavra, as sílabas tônicas carregam o acento mais forte, denominado acento primário. Os segmentos vocálicos átonos podem ser ainda pretônicos ou postônicos, caso antecedam ou sucedam o segmento tônico. A simbologia IPA utiliza o apóstrofo

para marcar a sílaba tônica, por exemplo [ˈpatɐ].

A Figura 3.4 apresenta os segmentos vocálicos orais do português do Brasil em posições idealizadas. Os segmentos vocálicos estão representados no diagrama com símbolos do Alfabeto Fonético Internacional - IPA. O contorno exterior do trapézio apresentado na figura delimita a área vocálica. A área vocálica define os limites anterior/posterior e alto/baixo idealizados dos possíveis posicionamentos da língua na produção dos segmentos. Os pontos intermediários são definidos como os pontos articulatoriamente eqüidistantes tanto na horizontal como na vertical. O português do Brasil apresenta sete segmentos vocálicos orais que ocorrem em posição tônica ou pretônica [i, e, ɛ, a, ɔ, o, u]. Os segmentos [ɪ, ɐ, ʊ] ocorrem em posição postônica.

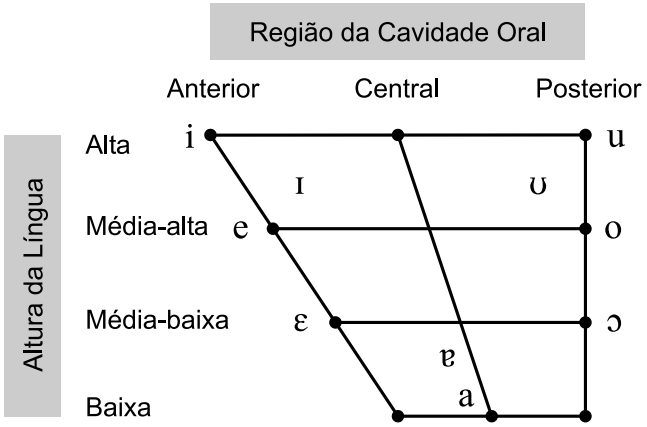

Fig. 3.4: Diagrama dos segmentos vocálicos orais do português do Brasil.

A Tabela 3.9 apresenta exemplos que ilustram a ocorrência dos segmentos vocálicos orais em posição tônica e pretônica. A primeira coluna apresenta o símbolo IPA do segmento vocálico. Na segunda tem-se a classificação do segmento de acordo com as características Posicionamento Vertical e Posicionamento Anterior/Posterior da Língua e Postura dos Lábios. Exemplos ortográficos envolvendo os segmentos são apresentados na terceira e quarta colunas, sendo que as colunas quatro e seis apresentam as respectivas transcrições fonéticas. Nos exemplos da coluna três, o segmento vocálico ocorre em posição tônica. A ocorrência em posição pretônica é ilustrada nos exemplos da coluna cinco.

Já a Tabela 3.10 apresenta exemplos de ocorrência dos segmentos vocálicos em posição postônica. A organização da Tabela 3.10 é análoga à da Tabela 3.9.

A Tabela 3.11 apresenta pares mínimos que permitem a caracterização dos fones [ɪ], [ɐ] e [ʊ] em contraste com [i], [a] e [u].

Símbolo	Classificação	Posição Tônica		Posição Pretônica	
		Exemplo	Transcrição	Exemplo	Transcrição
i	Alta anterior não-arredondada oral	sico	[ˈsikʊ]	final	[fiˈnaʊ̯]
e	Média-alta anterior não-arredondada oral	seco	[ˈsekʊ]	elevador	[elevaˈdoɾ]
ɛ	Média-baixa anterior não-arredondada oral	(eu) seco	[ˈsɛkʊ]	belíssima	[bɛˈlisɪmɐ]
a	Baixa central não-arredondada oral	saco	[ˈsɑkʊ]	café	[kaˈfɛ]
ɔ	Média-baixa posterior arredondada oral	(eu) soco	[ˈsɔkʊ]	sozinho	[sɔˈziɲʊ]
o	Média-alta posterior arredondada oral	soco	[ˈsokʊ]	moral	[moˈɾɑʊ̯]
u	Alta posterior arredondada oral	suco	[ˈsukʊ]	duvidar	[duviˈdɑɾ]

Tab. 3.9: Segmentos vocálicos orais em posição tônica e pretônica.

Símbolo	Classificação	Posição Tônica	
		Exemplo	Transcrição
ɪ	Alta anterior não-arredondada oral postônica	saque	[ˈsakɪ]
ɐ	Baixa central não-arredondada oral postônica	saca	[ˈsakɐ]
ʊ	Alta posterior arredondada oral postônica	saco	[ˈsakʊ]

Tab. 3.10: Segmentos vocálicos orais em posição postônica.

Fone	Exemplo	Transcrição	Fone	Exemplo	Transcrição
[ɪ]	vive	[ˈvivɪ]	[i]	vivi	[viˈvi]
[ɐ]	cara	[ˈkaɾɐ]	[a]	cará	[kaˈɾa]
[ʊ]	tato	[ˈtatʊ]	[u]	tatu	[taˈtu]

Tab. 3.11: Contraste entre os fones [ɪ] [i], [ɐ] [a] e [ʊ] [u].

Símbolo	Classificação	Posição Tônica	
		Exemplo	Transcrição
ĩ	Alta anterior não arredondada nasal	cinco	[ˈsĩkʊ]
ẽ	Média-alta anterior não-arredondada nasal	senda	[ˈsẽdɐ]
ɐ̃	Baixa central não-arredondada nasal	manto	[ˈmɐ̃tʊ]
õ	Média-alta posterior arredondada nasal	conto	[ˈkõtʊ]
ũ	Alta posterior arredondada nasal	sunga	[ˈsũgɐ]

Tab. 3.12: Segmentos vocálicos nasalizados do português do Brasil.

Fone	Exemplo	Transcrição	Fone	Exemplo	Transcrição
[ĩ]	cinco	[ˈsĩkʊ]	[i]	sico	[ˈsikʊ]
[ẽ]	senda	[ˈsẽdɐ]	[e]	seda	[ˈsedɐ]
			[ɛ]	(ele) seda	[ˈsɛdɐ]
[ɐ̃]	manto	[ˈmɐ̃tʊ]	[a]	mato	[ˈmatʊ]
[õ]	conto	[ˈkõtʊ]	[ɔ]	(eu) coto	[ˈkɔtʊ]
			[o]	coto	[ˈkotʊ]
[ũ]	sunga	[ˈsũgɐ]	[u]	suga	[ˈsugɐ]

Tab. 3.13: Contraste entre os fones vocálicos nasalizados e orais.

O português do Brasil admite ainda cinco segmentos vocálicos nasais [ĩ, ẽ, ɐ̃, õ, ũ] (veja Tabela 3.12). Estes segmentos se justificam através dos pares mínimos apresentados na Tabela 3.13. A posição idealizada destes segmentos na área vocálica é mostrada na Figura 3.5. Na Tabela 3.14 tem-se exemplos de palavras que ilustram a ocorrência dos segmentos vocálicos nasais em posição pretônica, tônica e postônica.

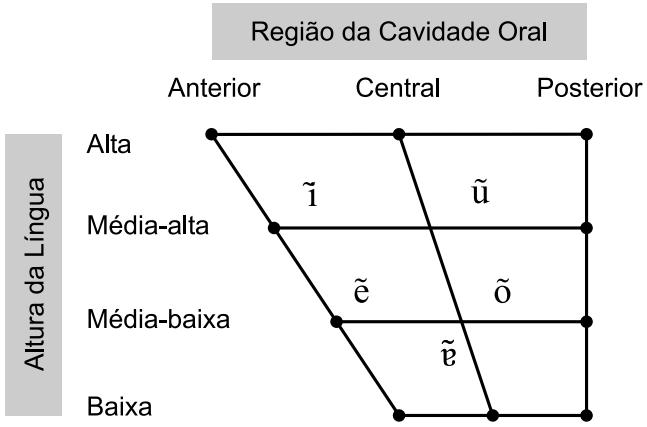

Fig. 3.5: Diagrama dos segmentos vocálicos nasais do português do Brasil.

Símbolo	**Posição pretônica**		**Posição tônica**		**Posição postônica**	
	Exemplo	**Transcrição**	**Exemplo**	**Transcrição**	**Exemplo**	**Transcrição**
ĩ	infeliz	[ĩfeˈlis]	cinco	[ˈsĩkʊ]	ínterim	[ˈĩteɾĩ]
ẽ	entrada	[ẽˈtrada]	senda	[ˈsẽdɐ]	hífen	[ˈifẽ]
ɐ̃	cantina	[kɐ̃ˈtinɐ]	manto	[ˈmɐ̃tʊ]	ímã	[ˈimɐ̃]
õ	construir	[kõstɾuˈɪɾ]	conto	[ˈkõtʊ]	cólon	[ˈkɔlõ]
ũ	umbigo	[ũˈbigʊ]	sunga	[sũgɐ]	álbum	[ˈaʊ̯bũ]

Tab. 3.14: Segmentos vocálicos nasais do português do Brasil em posições pretônicas, tônicas e postônicas.

Um segmento vocálico articulado de maneira a apresentar uma mesma característica acústica, percebida como estável e constante durante a sua produção, é denominado de monotongo. Se em uma sílaba, durante a produção de um segmento vocálico, ocorrer a variação dos articuladores para a

produção de outro segmento vocálico, tem-se um ditongo. Um tritongo é realizado em uma mesma sílaba, através da variação dos articuladores para a formação em seqüência de três segmentos vocálicos, sendo um deles central com maior intensidade. A ocorrência de um hiato é realizada pela produção de dois segmentos vocálicos em dois pulsos silábicos distintos que se sucedem na cadeia da produção sonora. As Tabelas 3.15 a 3.17 apresentam listas de ditongos e tritongos do português do Brasil.

Para acomodar o conceito de que em uma sílaba só ocorre um segmento vocálico, apenas um dos segmentos vocálicos dos ditongos e tritongos é considerado para a análise fonológica como vogal. Os segmentos vocálicos que acompanham esta vogal nos ditongos e tritongos são considerados vogais assilábicas, semivogais, semiconsoantes ou glides, dependendo do autor (CAGLIARI, 1981) (SILVA, 2002). Na simbologia IPA, uma vogal assilábica é expressa pelo diacrítico "◌̯" colocado em baixo do símbolo do segmento vocálico, como apresentado nas Tabelas 3.15 a 3.17.

a)

Representação Fonética	Exemplo
[eɪ̯]	lei
[ɛɪ̯]	anéis
[aɪ̯]	pai
[ɔɪ̯]	dói
[oɪ̯]	boi
[uɪ̯]	fui
[iʊ̯]	viu
[eʊ̯]	meu
[ɛʊ̯]	céu
[aʊ̯]	mau
[ɔʊ̯]	sol
[oʊ̯]	sou
[uʊ̯]	sul

b)

Representação Fonética	Exemplo
[ɪ̯ɪ]	série
[ɪ̯ɐ]	séria
[ɪ̯ʊ]	sério
[ʊ̯ɪ]	tênue
[ʊ̯ɐ]	árdua
[ʊ̯ɔ]	quota
[ʊ̯o]	quotizar
[ʊ̯ʊ]	vácuo

Tab. 3.15: Ditongos orais: a) decrescentes; b) crescentes.

a)

Representação Fonética	Exemplo
[ɐ̃ɪ̯]	mãe
[õɪ̯]	põe
[ũɪ̯]	muito
[ɐ̃ʊ̯]	pão

b)

Representação Fonética	Exemplo
[ʊ̯ĩ]	qüinqüenal
[ʊ̯ẽ]	freqüente
[ʊ̯ɐ̃]	quando

Tab. 3.16: Ditongos nasais: a) decrescentes; b) crescentes.

a)

Representação Fonética	Exemplo
[ʊ̯iʊ̯]	delinqüiu
[ʊ̯eɪ̯]	averigüei
[ʊ̯aɪ̯]	quais
[ʊ̯aʊ̯]	qual
[ʊ̯ɔɪ̯]	sequóia
[ʊ̯oʊ̯]	averiguou

b)

Representação Fonética	Exemplo
[ʊ̯ɐ̃ʊ̯]	quão
[ʊ̯õɪ̯]	saguões

Tab. 3.17: Tritongos: a) orais; b) nasais.

3.4 Comentários Finais

A animação facial realista sincronizada com a fala procura reproduzir em uma face virtual a movimentação articulatória visível associada à produção da fala. Faz parte da estratégia para se atingir este objetivo, estabelecer o conjunto de sons possíveis da língua. A partir dos sons pode-se, caracterizar a movimentação articulatória necessária à sua produção para, então, estabelecer uma representação visual da movimentação.

O repertório de fonemas de uma língua estabelece o conjunto de sons com características distintivas, com os quais é possível formar qualquer locução. Distintivo no sentido de que a mudança do som em uma palavra pode alterar o seu significado (por exemplo, [patɐ], [matɐ]).

O conjunto de fonemas adotado para representar o português do Brasil expresso nas Tabelas 3.1, 3.9 e 3.12 é utilizado, neste trabalho, como referência para o desenvolvimento de representações visuais dos sons da fala. Observa-se que o conjunto de fonemas adotado, embora baseado na variante urbana do Estado de São Paulo, pode ser estendido para contemplar outras variantes encontradas no país. Os métodos e conceitos utilizados ao longo do trabalho são genéricos o suficiente para contemplar estas extensões.

No próximo capítulo são discutidas questões associadas à visibilidade da movimentação articulatória necessárias a produção dos sons do português do Brasil.

Capítulo 4

Visualização da Fala

4.1 Introdução

A comunicação pela fala é efetuada principalmente, quando não exclusivamente, através da produção, percepção e interpretação da informação acústica emitida por um falante. Não obstante, as pistas visuais induzidas na face do locutor pela movimentação articulatória necessária à produção acústica também podem contribuir para a percepção da fala. Em situação de degradação do sinal acústico, a inteligibilidade pode ser melhorada através da observação da face do locutor (ERBER, 1975). Mesmo em situações para as quais o sinal acústico é claro e sem distorções, a informação visual contribui para a percepção (REISBERG; MCLENAY; GOLDFIELD, 1987). É aceito que a percepção da fala envolve a integração das informações auditivas e visuais (KOZLOWSKI, 1997) (MASSARO, 2004).

Extrair das pistas visuais da face informações lingüísticas é denominado leitura orofacial, ou ainda, leitura labial. A leitura orofacial é praticada, em maior ou menor grau, por todos, mesmo que de forma inconsciente (JEFFERS; BARLEY, 1971). Entretanto, a ruptura da coerência visual-auditiva da fala pode levar à percepção distorcida e incorreta da mensagem transmitida, como revelado pelo efeito McGurk (MCGURK; MACDONALD, 1976). A descoberta do efeito McGurk teve lugar quando imagens de vídeo de um locutor articulando "ga" foram dubladas com o áudio de "ba". A produção resultante foi percebida como "da". O efeito McGurk revela a importância de se contemplar com fidedignidade os aspectos articulatórios da produção da fala quando da implementação de uma cabeça falante virtual, que inerentemente tem a intenção de explorar o canal visual para a comunicação.

Os sons da fala são produzidos pela modificação controlada do fluxo de ar pulmonar em sua passagem pelo trato vocal. Estas modificações são efetuadas essencialmente pelo posicionamento das pregas vocais, do véu palatino, da língua, da mandíbula e dos lábios. Entretanto, grande parte destes movimentos articulatórios ocorre no interior da cavidade oral sem que seja possível a fácil visualização. Conseqüentemente, o contraste visual dos segmentos é reduzido a um conjunto mais restrito

de parâmetros do que o conjunto total das possibilidades articulatórias, tornando a percepção visual menos eficaz para a discriminação entre segmentos do que a percepção auditiva. Assim, na percepção visual da fala, os padrões de movimentação articulatória visualmente contrastáveis acabam por ser associados a mais de um segmento sonoro. Notadamente, o vozeamento e a nasalidade apresentam efeito acústico marcante, porém não permitem o contraste visual (ERBER, 1972). Segmentos sonoros que não são possíveis de ser diferenciados visualmente são denominados, neste trabalho, de homofemas (homo + (morf)ema).

A identificação de homofemas está associada à identificação e ao contraste de diferentes padrões visíveis de movimentação articulatória. O grau de contraste destes padrões, entretanto, depende de um conjunto de fatores, tais como, a composição fonética da locução, a habilidade articulatória do falante, as características físicas da face do locutor, a taxa de elocução, a capacidade do ouvinte/observador de perceber as pistas visuais, a iluminação, a distância e o ângulo de visão do ouvinte/observador (JEFFERS; BARLEY, 1971). Este fatores colocam um grande entrave para o estabelecimento de conjuntos de homofemas gerais e universais. Nas Seções 4.2 e 4.3 é apresentada uma compilação de trabalhos envolvendo a identificação de homofemas consonantais e vocálicos, respectivamente. É importante observar que os trabalhos discutidos nestas seções foram desenvolvidos tendo por contexto a língua inglesa. A Seção 4.4 aborda os efeitos da coarticulação na visualização dos homofemas.

Este capítulo tem como foco principal identificar os aspectos da produção da fala que moldam as pistas visuais apresentadas na face do falante. A Seção 4.5 sintetiza os principais aspectos discutidos no capítulo, estabelecendo as premissas teóricas que nortearam a identificação de um conjunto de visemas para o português do Brasil, cujo detalhamento é apresentado no Capítulo 5. É importante destacar que a principal contribuição do presente capítulo é o conjunto de homofemas apresentados nas Tabelas 4.24 e 4.25 derivadas da análise realizada ao longo do capítulo.

4.2 Homofemas de segmentos consonantais

A proposta mais tradicional de se agrupar os segmentos consonantais em homofemas é apresentada na Tabela 4.1 (NITCHIE, 1950), citado em (KRICOS; LESNER, 1982). Esta classificação espelha fortemente a grade lugar/modo de articulação da tabela consonantal, com as nasais sendo confundidas com as plosivas de mesmo lugar de articulação. Os homofemas da Tabela 4.1 foram estabelecidos para o ensino de técnicas de leitura orofacial a portadores de deficiência auditiva, tendo sido desenvolvidos a partir de observações clínicas. Experimentos de percepção realizados para a identificação visual dos segmentos da fala através da leitura orofacial, entretanto, indicam que os homofemas da Tabela 4.1, apesar de plausíveis teoricamente, não parecem ser totalmente discriminados por observadores.

Homofemas Consonantais		
/p, b, m/	/θ, ð/	/ʃ, ʒ, tʃ, tʒ/
/f, v/	/t, d, n/	/j/
/w/	/l/	/k, g, ŋ/
/r/	/s, z/	/h/

Tab. 4.1: Homofemas consonantais de Nitchie (1950).

Também baseado na experiência advinda do trato de deficientes auditivos, o trabalho de Jeffers e Barley (1971) propõe classificar os homofemas de acordo com a visibilidade dos movimentos articulatórios. As seguintes categorias são definidas: visível, moderadamente visível, ocasionalmente visível e raramente visível. Tal classificação reflete a dependência da identificação dos homofemas de fatores como iluminação, posição relativa entre locutor e ouvinte, precisão articulatória do falante, e reafirma a convicção de que uma definição universal e geral para qualquer contexto não é possível.

A Tabela 4.2 apresenta a classificação proposta em Jeffers e Barley (1971). Nesta proposta, o conceito de visibilidade está associado não só à movimentação característica de um determinado grupo de homofemas, como também à robustez com que este movimento se reproduz independentemente da taxa de elocução e do contexto fonético. Na classificação da Tabela 4.2 é considerada uma condição de visualização altamente favorável com as palavras sendo articuladas de forma convencional e com movimentos amplos, porém sem exageros. A taxa de elocução é lenta, porém percebida como normal.

Homofemas Visíveis		**Homofemas Obscuros**	
Visíveis	**Moderamente Visíveis**	**Ocasionalmente Visíveis**	**Raramente Visíveis**
/f, v/ /w, hw, r/ /p, b, m/	/θ, ð/ /ʃ, ʒ, tʃ, dʒ/ /s, z/	/j/ /t, d, n, l/	/k, g, ŋ/

Tab. 4.2: Homofemas consonantais de Jeffers e Barley (1971).

É interessante observar que a classificação de Jeffers e Barley (1971) também reflete fortemente a grade modo/lugar de articulação das consoantes. Em particular, considerando apenas os segmentos consonantais do português do Brasil (ver Tabela 3.1), e as classificações de Nitchie (1950) e Jeffers e Barley (1971), têm-se os seguintes grupos de homofemas:

- /f, v/ fricativas labiodentais (desvozeada e vozeada);
- /p, b, m/ plosivas bilabiais (desvozeada e vozeada) e nasal bilabial,
- /ʃ, ʒ/ fricativas pós-alveolares (desvozeada e vozeada),
- /s, z/ fricativas alveolares (desvozeada e vozeada),
- /t, d, n/ plosivas (desvozeada e vozeada) e nasal alveolares (NITCHIE, 1950), ou /t, d, n, l/ plosivas (desvozeada e vozeada), nasal e lateral alveoalares (JEFFERS; BARLEY, 1971); e
- /k, g/ plosivas velares (desvozeada e vozeada).

As Tabelas 4.3 a 4.8 apresentam os resultados de estudos voltados ao reconhecimento visual de consoantes. Em tais estudos foram realizados testes de reconhecimento das consoantes a partir da observação por indivíduos normais e com deficiência auditiva de imagens de um falante gravadas em vídeo ou película. Aqui se fará a análise destes trabalhos sem a distinção do grupo de estudo empregado, uma vez que os resultados de Erber (1972), Benguerel e Pichora-Fuller (1982) e Owens e Blazek (1985) indicam que não há diferença significativa na capacidade de reconhecimento entre indivíduos normais e portadores de deficiência auditiva. Os trabalhos analisados foram selecionados por permitirem a normalização do critério estatístico de identificação dos grupos de homofemas. O critério utilizado, proposto em (BINNIE; JACKSON; MONTGOMERY, 1976), estabelece que, considerando a matriz de confusão estímulo/resposta, a taxa de reconhecimento de grupo é dada pelo número total de respostas atribuídas ao grupo dividido pelo número total de estímulos pertencentes ao grupo. Os resultados apresentados nas tabelas refletem uma taxa de reconhecimento de grupo maior do que 75%. Nas tabelas, a primeira coluna indica o(s) contexto(s) fonético(s) utilizado(s) no reconhecimento das consoantes. Na especificação do contexto fonético, as vogais são expressas na notação IPA e a consoante pelo símbolo "C", sendo indicado também o conjunto de segmentos consonantais utilizado. A três outras colunas apresentam os homofemas identificados. Fonemas, cujas identificações não atenderam o critério de taxa de reconhecimento de grupo maior do que 75%, são apresentados nas tabelas com o rótulo "Indefinido".

Considerando-se apenas as consoantes comuns ao português do Brasil, a análise das Tabelas 4.3 a 4.8 revela uma coerência da maioria dos resultados na identificação dos grupos de homofemas /p, b, m/, /f, v/ e /ʃ, ʒ/. Por sua vez, os resultados da classificação das outras consoantes relevantes para o português do Brasil, /t, d, n, s, z, k, g, l/, apresentam-se de maneira menos nítida. A Tabela 4.9 apresenta a compilação dos resultados associados ao reconhecimento dos grupos de homofemas /p, b, m/, /f, v/ e /ʃ, ʒ/. Na tabela, a correta identificação do grupo de homofemas em um dado trabalho é identificada por "sim", caso contrário por "não". A última linha da tabela apresenta a porcentagem,

Contexto	**Homofemas**		
	/i/C/i/	**/ɑ/C/ɑ/**	**/u/C/u/**
/i/C/i/ /ɑ/C/ɑ/ /u/C/u/ C = {/p, b, t, d, k, g, f, v, θ, ð, s, z, ʃ, ʒ, m, n, w, r, l, h/}	/p, b, m/ /f, v/ /w, r/ /θ, ð/ /ʃ, ʒ/ /s, z/ /l/ /k, g, h/ /t, d, n/	/p, b, m/ /f, v/ /w, r/ /θ, ð/ /ʃ, ʒ/ /t, d, n, s, z/ /l/ /k, g/ /h/	/p, b, m/ /f, v/ /θ, ð/ /s, z, ʃ, ʒ/ /k, g, h/ /l/ Indefinido: /t, d, n, w, r/

Tab. 4.3: Homofemas consonantais de Erber (1974).

Contexto	**Homofemas**
C/ɑ/ C = {/p, b, t, d, k, g, f, v, θ, ð, s, z, ʃ, ʒ, m, n, w, r, l, j/}	/p, b, m/ /f, v/ /w/ /θ, ð/ /ʃ, ʒ/ /r/ /t, d, s, z / /n, l/ Indefinido: /k, g, j/

Tab. 4.4: Homofemas consonantais de Binnie, Jackson e Montgomery (1976).

Contexto	**Homofemas**	
	Sem Treinamento	**Com Treinamento**
C/ɑ/C C = {/p, b, t, d, k, g, f, v, θ, ð, s, z, ʃ, ʒ, m, n, w, r, l, j/}	/p, b, m/ /f, v/ /w/ /θ, ð/ /s, z, ʃ, ʒ/ Indefinido: /t, d, n, k, g, j, r, l/	/p, b, m/ /f, v/ /w/ /r/ /θ, ð/ /ʃ, ʒ/ /s, z/ /t, d, n, k, g, j/ /l/

Tab. 4.5: Homofemas consonantais de Walden et al. (1977).

Contexto	**Homofemas**	
	Sem Treinamento	**Com Treinamento**
/ɑ/C/ɑ/ C = {/p, b, t, d, k, g, f, v, θ, ð, s, z, ʃ, ʒ, tʃ, dʒ, m, n, w, r, l, j/}	/p, b, m/ /f, v/ /w, r/ /θ, ð/ /ʃ, ʒ, tʃ, dʒ/ Indefinido: /t, d, n, s, z, j, k, g, l/	/p, b, m/ /f, v/ /w, r/ /θ, ð/ /ʃ, ʒ, tʃ, dʒ/ /t, d, n, s, z, k, g, j, l/

Tab. 4.6: Homofemas consonantais de Walden et al. (1981).

<table>
<tr><th rowspan="2">Contexto</th><th colspan="2">Homofemas</th></tr>
<tr><th>Falante 1</th><th>Falante 2</th></tr>
<tr><td rowspan="6">/ɑ/C/ɑ/
C = {/p, b, t, d, k, g, f, v, θ, ð, s, z, ʃ, ʒ, tʃ, dʒ, m, n, ŋ, w, r, l, h, j/}</td><td>/p, b, m/
/f, v/
/r, w/
/θ, ð/
/ʃ, ʒ, tʃ, dʒ/
/t, d, s, z/
/l/
/k, g, ŋ, j, h/
Indefinido:
/n/</td><td>/p, b, m/
/f, v/
/r, w/
/θ, ð/
/ʃ, ʒ, tʃ, dʒ/
/t, d, s, z/
/l/
/n, k, g, ŋ, j, h/</td></tr>
<tr><td>Falante 3</td><td>Falante 4</td></tr>
<tr><td>/p, b, m/
/f, v, r/
/w/
/θ, ð/
/ʃ, ʒ, tʃ, dʒ/
/k, g/
Indefinido:
/t, d, n, s, z, ŋ, l, j, h/</td><td>/p, b, m/
/f, v/
/r, w/
/θ, ð/
/ʃ, ʒ, tʃ, dʒ/
/t, d, s, z/
Indefinido:
/k, g, n, ŋ, l, j, h/</td></tr>
<tr><td>Falante 5</td><td>Falante 6</td></tr>
<tr><td>/p, b, m/
/f, v, s, z/
/r, w/
/ʃ, ʒ, tʃ, dʒ/
Indefinido:
/t, d, n, θ, ð, k, g, ŋ, l, j, h/</td><td>/p, b, m/
/r, w, θ, ð/
/ʃ, ʒ, tʃ, dʒ/
/t, d, n, s, z, l, j, h/
Indefinido:
/f, v, k, g, ŋ/</td></tr>
</table>

Tab. 4.7: Homofemas consonantais de Kricos e Lesner (1982).

Contexto	Homofemas	
	/i/C/i/	**/ɑ/C/ɑ/**
/i/C/i/ /ɑ/C/ɑ/ /ʌ/C/ʌ/ /u/C/u/ C = {/p, b, t, d, k, g, f, v, θ, ð, s, z, ʃ, ʒ, tʃ, dʒ, n, m, w, r, l, h, j}	/p, b, m/ /f, v/ /θ, ð/ /w, r/ /ʃ, ʒ, tʃ, dʒ/ /t, d, s, z/ Indefinido: /n, k, g, h, l, j/	/p, b, m/ /f, v/ /θ, ð/ /w, r/ /ʃ, ʒ, tʃ, dʒ/ /n, k, g, l/ /h/ Indefinido: /t, d, s, z, j/
	/ʌ/C/ʌ/	**/u/C/u/**
	/p, b, m/ /f, v/ /θ, ð/ /w, r/ /ʃ, ʒ, tʃ, dʒ/ /t, d, s, z/ Indefinido: /n, k, g, h, l, j/	/p, b, m/ /f, v/ Indefinido: /θ, ð, w, r, ʃ, ʒ, tʃ, dʒ, t, d, n, s, z, k, g, l, h, j/

Tab. 4.8: Homofemas consonantais de Owens e Blazek (1985).

considerando todos os contextos de todos os trabalhos, de identificação destes homofemas. Observa-se que o percentual de reconhecimento do grupo /p, b, m/ é maior do que o dos grupos /f, v/, e /ʃ, ʒ/, sugerindo um decréscimo da visibilidade dos homofemas.

	/p, b, m/	**/f, v/**	**/ʃ, ʒ/** [a]
Erber (1974) - /i/C/i/	sim	sim	sim
Erber (1974) - /ɑ/C/ɑ/	sim	sim	sim
Erber (1974) - /u/C/u/	sim	sim	não
Binnie, Jackson e Montgomery (1976) - C/ɑ/	sim	sim	sim
Walden et al. (1977) - Sem treinamento	sim	sim	não
Walden et al. (1977) - Com treinamento	sim	sim	sim
Walden et al. (1981) - Sem treinamento	sim	sim	sim
Walden et al. (1981) - Com treinamento	sim	sim	sim
Kricos e Lesner (1982) - Falante 1	sim	sim	sim
Kricos e Lesner (1982) - Falante 2	sim	sim	sim
Kricos e Lesner (1982) - Falante 3	sim	não	sim
Kricos e Lesner (1982) - Falante 4	sim	sim	sim
Kricos e Lesner (1982) - Falante 5	sim	não	sim
Kricos e Lesner (1982) - Falante 6	sim	não	sim
Owens e Blazek (1985) -/i/C/i/	sim	sim	sim
Owens e Blazek (1985) - /ɑ/C/ɑ/	sim	sim	sim
Owens e Blazek (1985) - /ʌ/C/ʌ/	sim	sim	sim
Owens e Blazek (1985) - /u/C/u/	sim	sim	não
Porcentagem Identificação	100%	83,3%	83,3%

Tab. 4.9: Identificação dos homofemas /p, b, m/, /f, v/ e /ʃ, ʒ/.

[a] Foi considerado que /ʃ, ʒ/ e /ʃ, ʒ, tʃ, tʒ/ representam a mesma classe de homofemas.

Estes resultados não são discordantes com a classificação proposta em Jeffers e Barley (1971) apresentada na Tabela 4.2. A relativa boa visibilidade destes homofemas é justificada pela movimentação característica, envolvendo os lábios e a mandíbula, normalmente utilizada para a produção

dos segmentos. Observe-se que /p, b, m/ e /f, v/ são articulados com o envolvimento dos lábios e em lugares relativamente exteriores da cavidade vocal (confira Figura 3.2). Já as consoantes /ʃ, ʒ/ estão associadas ao movimento de protrusão labial característico destes segmentos, o que facilita a percepção visual.

Por seu lado, a análise do agrupamento das consoantes /s, z, t, d, n, l, k, g/ apresenta-se como mais complexa. Tal fato pode ser justificado pela pouca visibilidade destas consoantes, como apontado em Jeffers e Barley (1971), e confirmado pela dispersão dos resultados dos trabalhos listados nas Tabelas 4.3 a 4.8.

Não obstante, considerando os resultados apresentados e os conjuntos de fonemas /s, z/, /k, g/, é possível construir a Tabela 4.10. Nesta tabela é indicado se o grupo de fonemas está (indicado pelo rótulo 'sim'), ou não (indicado pelo rótulo 'não'), contido em algum dos conjuntos de homofemas identificados no trabalho. Utiliza-se a notação '?' para indicar os casos em que todos os segmentos consonantais do conjunto em análise não tiveram classificação definida, ou seja, não atenderam ao critério de taxa de reconhecimento de grupo maior do que 75%. A penúltima linha da tabela apresenta a porcentagem em que o conjunto de fonemas indicado no topo da tabela, está contido em algum conjunto de homofemas identificado nos trabalhos analisados. A porcentagem à esquerda, considera os casos indefinidos como um conjunto extra de homofemas - o conjunto dos homofemas que não permitem uma identificação robusta. A porcentagem à direita, considera os casos indefinidos como a não classificação dos homofemas. A porcentagem da esquerda permite uma avaliação menos rigorosa da pertinência da classificação dos conjuntos /s, z/ e /k, g/ como homofemas. A porcentagem à direita reflete uma abordagem mais restritiva. A última linha da tabela apresenta a média aritmética destas duas porcentagens. Estes resultados são concordantes com a classificação de visibilidade proposta em Jeffers e Barley (1971), em que o grupo de homofemas /s, z/ possui visibilidade moderada e o grupo /k, g/ é de difícil visualização (confira Tabela 4.2).

Já a análise dos fonemas /t, d, n, l/ indica uma maior concordância com a classificação de Nitchie (1950) do que a de Jeffers e Barley (1971). As Tabelas 4.11 e 4.12 indicam as porcentagens de reconhecimento dos grupos de homofemas indicados na primeira linha da tabela. Estas tabelas estão organizadas da mesma forma que a Tabela 4.10. Das Tabelas 4.11 e 4.12 é possível identificar que há uma certa tendência de se reconhecer o grupo /t, d/ como homofemas. Os resultados indicam também que o fonema /n/ pode ser confundido como participante do grupo de homofemas /t, d, n/, assim como do grupo /n, l/, sem mostrar tendência significativa de associação com qualquer um dos dois grupos. Já o fonema /l/ apresenta a tendência um pouco maior de ser reconhecido de forma isolada do que formando o grupo /n, l/. Estes resultados confirmam a classificação da Tabela 4.1, sem discordar, entretanto, da classificação de Jeffers e Barley (1971), a qual indica que estes fonemas são de difícil visualização.

	/s, z/		/k, g/	
Erber (1974) - /i/C/i/	sim		sim	
Erber (1974) - /ɑ/C/ɑ/	sim		sim	
Erber (1974) - /u/C/u/	sim		sim	
Binnie, Jackson e Montgomery (1976)	sim		?	
Walden et al. (1977) - com treinamento	sim		?	
Walden et al. (1977) - sem treinamento	sim		sim	
Walden et al. (1981) - com treinamento	?		?	
Walden et al. (1981) - sem treinamento	sim		sim	
Kricos e Lesner (1982) - Falante 1	sim		sim	
Kricos e Lesner (1982) - Falante 2	sim		sim	
Kricos e Lesner (1982) - Falante 3	?		sim	
Kricos e Lesner (1982) - Falante 4	sim		?	
Kricos e Lesner (1982) - Falante 5	sim		?	
Kricos e Lesner (1982) - Falante 6	sim		?	
Owens e Blazek (1985) - /i/C/i/	sim		?	
Owens e Blazek (1985) - /ɑ/C/ɑ/	?		sim	
Owens e Blazek (1985) - /ʌ/C/ʌ/	sim		?	
Owens e Blazek (1985) - /u/C/u/	?		?	
Porcentagem Agrupamento	100%	78%	100%	50%
Porcentagem Média	89%		50%	

Tab. 4.10: Reconhecimento dos conjuntos /s, z/ e /k, g/ como pertencentes a um mesmo grupo de homofemas.

	/t, d/		/n, l/		/t, d, n/		/t, d, l/	
Erber (1974) - /i/C/i/	sim		não		sim		não	
Erber (1974) - /ɑ/C/ɑ/	sim		não		sim		não	
Erber (1974) - /u/C/u/	?		não		?		não	
Binnie, Jackson e Montgomery (1976)	sim		sim		não		não	
Walden et al. (1977) - sem treinamento	?		?		?		?	
Walden et al. (1977) - com treinamento	sim		não		sim		não	
Walden et al. (1981) - - sem treinamento	?		?		?		?	
Walden et al. (1981) - com treinamento	sim		sim		sim		sim	
Kricos e Lesner (1982) - Falante 1	sim		não		não		não	
Kricos e Lesner (1982) - Falante 2	sim		não		não		não	
Kricos e Lesner (1982) - Falante 3	sim		?		?		?	
Kricos e Lesner (1982) - Falante 4	sim		?		não		não	
Kricos e Lesner (1982) - Falante 5	?		?		?		?	
Kricos e Lesner (1982) - Falante 6	sim		sim		sim		sim	
Owens e Blazek (1985) - /i/C/i/	sim		?		não		não	
Owens e Blazek (1985) - /ɑ/C/ɑ/	?		sim		não		não	
Owens e Blazek (1985) - /ʌ/C/ʌ/	sim		?		não		não	
Owens e Blazek (1985) - /u/C/u/	?		?		?		?	
Porcentagem Agrupamento	100%	67%	67%	22%	61%	28%	39%	11%
Porcentagem Média	83,5%		44,5%		44,5%		25%	

Tab. 4.11: Reconhecimento dos conjunto /t, d/, /n, l/, /t, d, n/ e /t, d, l/ como pertencentes a um mesmo grupo de homofemas.

	/l/		/n/		/t, d, n, l/	
Erber (1974) - /i/C/i/	sim		não		não	
Erber (1974) - /ɑ/C/ɑ/	sim		não		não	
Erber (1974) - /u/C/u/	sim		?		não	
Binnie, Jackson e Montgomery (1976)	não		não		não	
Walden et al. (1977) - sem treinamento	?		?		?	
Walden et al. (1977) - com treinamento	sim		não		não	
Walden et al. (1981) - - sem treinamento	?		?		?	
Walden et al. (1981) - com treinamento	não		não		sim	
Kricos e Lesner (1982) - Falante 1	sim		?		não	
Kricos e Lesner (1982) - Falante 2	sim		não		não	
Kricos e Lesner (1982) - Falante 3	?		?		?	
Kricos e Lesner (1982) - Falante 4	?		?		não	
Kricos e Lesner (1982) - Falante 5	?		?		?	
Kricos e Lesner (1982) - Falante 6	não		não		sim	
Owens e Blazek (1985) - /i/C/i/	?		?		não	
Owens e Blazek (1985) - /ɑ/C/ɑ/	não		não		não	
Owens e Blazek (1985) - /ʌ/C/ʌ/	?		?		não	
Owens e Blazek (1985) - /u/C/u/	?		?		?	
Porcentagem Agrupamento	78%	33%	56%	0%	39%	11%
Porcentagem Média	55,5%		28%		25%	

Tab. 4.12: Reconhecimento dos conjunto /l/, /n/ e /t, d, n, l/ como pertencentes a um mesmo grupo de homofemas.

Uma possível explicação para a dificuldade de classificação dos segmentos alveolares, /s, z, t, d, n, l/, e velares /k, g/, pode ser atribuída à combinação de um conjunto de fatores. Diferentemente das consoantes /f, v/, /p, b, m/ e /ʃ, ʒ/, as consoantes /t, d, n, s, z, k, g, l/ não apresentam articulação com característica visual marcante que permita fácil contraste. Adicionalmente, há uma quantidade relativamente grande de consoantes alveolares /t, d, n, s, z, l/ comparada ao conjunto de consoantes associadas a outros lugares de articulação, tornando a discriminação visual ainda mais difícil. De seu lado, as velares possuem lugar de articulação relativamente profundo no interior da cavidade oral, não permitindo uma fácil visualização.

O padrão MPEG-4 (MPEG4 VISUAL, 2001) adota os grupos de homofemas consonantais apresentados na Tabela 4.13. É necessário salientar que o padrão é omisso quanto a segmentos consonantais existentes em outras línguas que não a inglesa. Mesmo nesta, observa-se que o fonema /ʒ/ não é apresentado no documento MPEG4 Visual (2001). Na Tabela 4.13, o segmento foi incluído no grupo indicado pela maioria dos trabalhos analisados (consulte Tabela 4.9).

Homofemas Consonantais		
/p, b, m/	/t, d/	/s, z/
/f, v/	/k, g/	/n, l/
/θ, ð/	/ʃ, ʒ, tʃ, tʒ/	/r/

Tab. 4.13: Homofemas consonantais do Padrão MPEG-4 (MPEG4 VISUAL, 2001).

É importante observar que a simbologia adotada na documentação do padrão MPEG-4 (MPEG4 VISUAL, 2001) não utiliza o alfabeto fonético internacional para a especificação dos fonemas. A tabela como apresentada na documentação MPEG-4 é reproduzida na Tabela 4.14. A Tabela 4.15 apresenta o mapeamento do fonemas especificados na Tabela 4.14 para a simbologia IPA considerando a pronúncia britânica padrão (PAERSALL; TRUMBLE, 1996).

4.3 Homofemas para segmentos vocálicos

Os diferentes segmentos vocálicos são produzidos pela conformação da cavidade oral realizada principalmente pelo posicionamento da língua. Observa-se que a movimentação da língua ocorre no interior da cavidade oral e não é, em geral, visível, dificultando a precisa distinção visual dos segmentos vocálicos. Adicionalmente, a distinção visual entre segmentos vocálicos orais e nasais não é possível, já que a movimentação do véu palatino ocorre no interior da cavidade oral. Em

Phonemes	Example
p, b, m	put, bed, mill
f, v	far, voice
T, D	think, that
t, d	tip, doll
k, g	call, gas
tS, dZ, S	chair, join, she
s, z	sir, zeal
n, l	lot, not
r	red

Tab. 4.14: Fonemas consonantais do padrão MPEG4 - Tabela C-5 (parcial) Anexo C (MPEG4 VISUAL, 2001) .

MPEG-4	IPA	MPEG-4	IPA	MPEG-4	IPA
p	p	n	n	S	ʃ
b	b	f	f	tS	tʃ
t	t	v	v	dZ	dʒ
d	d	T	θ	r	r
k	k	D	ð	l	l
g	g	s	s		
m	m	z	z		

Tab. 4.15: Mapeamento entre as simbologias MPEG-4 e IPA para segmentos consonantais.

contrapartida, a protrusão/arredondamento dos lábios, presente na produção das vogais posteriores do português do Brasil, é uma característica de fácil observação.

Diferentemente dos segmentos consonantais que permitem a identificação visual do lugar de articulação, em que pese o aumento da dificuldade de visualização dos lugares de articulação mais exteriores para os menos exteriores, não há dois segmentos vocálicos orais que sejam produzidos exatamente com um mesmo padrão de movimentação articulatória (JACKSON, 1988). Entretanto, muitos destes padrões são suficientemente parecidos, para permitir na prática a identificação de homofemas (JEFFERS; BARLEY, 1971).

Jeffers e Barley (1971) adotam dois parâmetros para a identificação dos homofemas de segmentos vocálicos: *Postura dos Lábios* e *Abertura da Boca*. Para a *Postura dos Lábios* admite-se quatro situações: *Protraídos*, *Arredondados*, *Relaxados* e *Retraídos*. Para a *Abertura da Boca* são admitidas as configurações *Estreita* e *Moderada*. A Tabela 4.16 apresenta a classificação dos segmentos vocálicos segundo Jeffers e Barley (1971). Para esta classificação é considerada uma condição de visualização altamente favorável com as palavras sendo articuladas de forma convencional e com movimentos amplos, porém sem exageros. A taxa de elocução é lenta, porém ainda considerada normal. Três grupos de homofemas (/ɑʊ/, /ɔɪ/ e /aɪ/) são descritos através da indicação da configuração inicial e final da movimentação articulatória associada à produção dos fonemas. A Figura 4.1 apresenta a distribuição dos homofemas no diagrama das vogais, considerando apenas os monotongos.

Visíveis	**Protruídos/Estreita**	/u, ʊ, o, oʊ, ɝ/
	Retraídos/Estreita	/i, ɪ, eɪ, ʌ/
	Arredondados/Moderada	/ɔ/
	Relaxados/Moderada para Protruídos/Estreita	/ɑʊ/
Obscuros	**Relaxados/Moderada**	/ɛ, æ, ɑ/
	Arredondados/Moderada para Retraídos/Estreita	/ɔɪ/
	Relaxados/Moderada para Retraídos/Estreita	/aɪ/

Tab. 4.16: Homofemas vocálicos de Jeffers e Barley (1971)

Em Wozniak e Jackson (1979) é apresentado estudo da influência do ângulo de visão na percepção de vogais e ditongos. Neste estudo, um falante foi gravado em vídeo de frente (ângulo de observação de 0°) e de perfil (ângulo de observação de 90°). Foram utilizados 16 logatomas do tipo /h/V/g/, com V = /i, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ, ɝ, eɪ, oʊ, ɑɪ, ɑʊ, ɔɪ, ju/. As matrizes de confusão estímulo/resposta

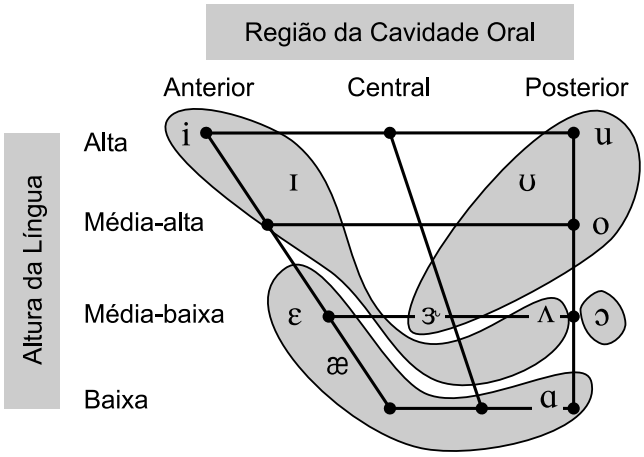

Fig. 4.1: Diagrama das vogais e os homofemas vocálicos de Jeffers e Barley (1971) (somente monotongos).

do estudo foram processadas em Jackson (1988) para o levantamento da taxa de reconhecimento de grupo como proposta em Binnie, Jackson e Montgomery (1976). Os resultados da classificação são apresentados na Tabela 4.17. A primeira coluna da tabela apresenta o contexto fonético utilizado no experimento. Os valores entre parênteses nas colunas 2 e 3 indicam as taxas de reconhecimento de grupo dos homofemas. A Figura 4.2 e Figura 4.3 apresentam as distribuições dos grupos de homofemas no diagrama das vogais para as duas condições analisadas.

Contexto	**Homofemas**	
	Ângulo de Observação O°	**Ângulo de Observação 9O°**
/h/V/g/ V = {/i, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ, ɜˠ, eɪ, oʊ, ɑɪ, ɑʊ, ɔɪ, ju/}	/ɑʊ/ (99%) /ɛ, æ, ɑ, ɑɪ, eɪ/ (96,2%) /oʊ/ (96%) /ɔ, ɔɪ/ (94%) /i, ɪ/ (85,5%) /ʌ/ (81%) /u, ju/ (96,5%) /ʊ, ɜˠ/ (83,5%)	/ɑʊ/ (100%) /ɛ, æ, ɑ, ɑɪ, eɪ/ (94,8%) /oʊ/ (91,0%) /ɔ, ɔɪ/ (99,0%) /i, ɪ/ (89%) /ʌ/ (91%) /u, ʊ, ɜˠ, ju/ (95,8%)

Tab. 4.17: Homofemas vocálicos de Jackson (1988) com dados de Wozniak e Jackson (1979).

A Tabela 4.18 apresenta a matriz de confusão estímulo/resposta do estudo realizado em Montgomery e Jackson (1983). A Tabela 4.19 apresenta a identificação de homofemas vocálicos a partir

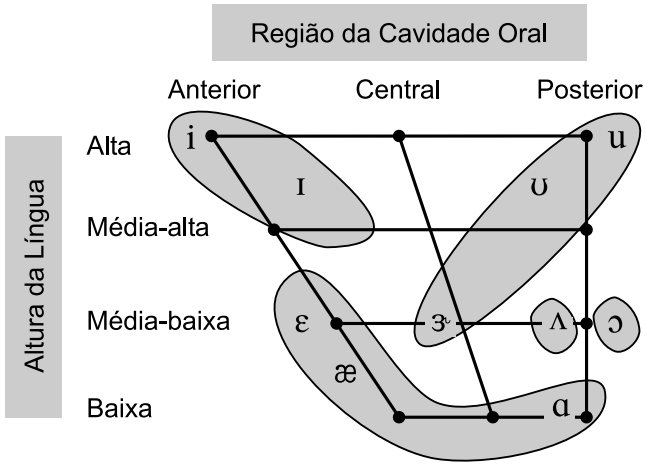

Fig. 4.2: Diagrama das vogais e os homofemas vocálicos de Jackson (1988) com dados de Wozniak e Jackson (1979) - ângulo de observação de 0° (somente monotongos).

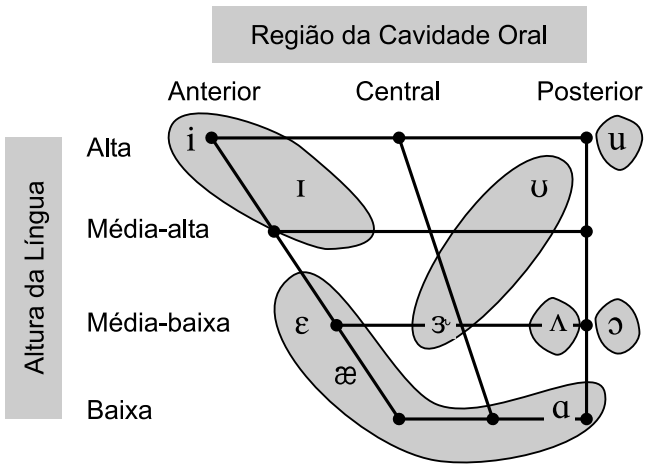

Fig. 4.3: Diagrama das vogais e os homofemas vocálicos de Jackson (1988) com dados de Wozniak e Jackson (1979) - ângulo de observação de 90° (somente monotongos).

dos dados da Tabela 4.18. A Figura 4.4 apresenta a distribuição destes homofemas no diagrama das vogais.

		Resposta														
		i	**ɪ**	**ɛ**	**æ**	**eɪ**	**ɑ**	**ɔ**	**ʌ**	**ɑɪ**	**ɔɪ**	**ʊ**	**u**	**ɜ**	**ɑʊ**	**oʊ**
	i	77	14	10	6	9							1			
	ɪ	46	31	16	1	19			1	1		1		2		
	ɛ		9	25	26	32	5	4	2	3		2	1	8	1	
E	**æ**	1	7	18	50	20	13	3	2				2		1	1
s	**eɪ**	6	8	33	14	39	5	2	3	6		2				
t	**ɑ**	1	2	1	15	1	53	11	5	28					1	
í	**ɔ**						9	77	1	1	18	1	1	1	5	4
m	**ʌ**	2	2	4	3		35	14	42	12		1	3			
u	**ɑɪ**	3	2	6	13	10	7	7	4	65						1
l	**ɔɪ**	1			1		1				97	2	4		2	10
o	**ʊ**		2				1	2	4	1	1	48	25	32	1	1
	u							1			2	9	96		5	4
	ɜ	2	2	1					4		1	19	22	65	1	1
	ɑʊ					1		3			1		1		102	10
	oʊ						1		3		3	3	10	1	6	90

Tab. 4.18: Matriz de confusão de reconhecimento de vogais e ditongos de Montgomery e Jackson (1983).

Em Jackson, Montgomery e Binnie (1976) é efetuada uma avaliação das principais características visuais responsáveis pela identificação dos grupos de homofemas vocálicos. Como resultado, os autores identificaram duas características mais proeminentes: o arredondamento dos lábios e a abertura vertical dos lábios. A característica de arredondamento dos lábios está fortemente correlacionada com a distância entre os cantos da boca, comportando, em um extremo, a postura estendida associada à produção do segmento vocálico /i/ do inglês e, no outro, a postura fortemente arredondada do /u/. O conceito de arredondamento, como entendido pelos autores, é detalhado em Montgomery e Jackson (1983) e exposto a seguir.

Para os autores a característica de arredondamento não se manifesta somente como a forma circular dos lábios, mas é expressa por um conjunto de eventos, incluindo a protrusão, a tensão, a abertura

Contexto	Homofemas
/h/V/g/ V = {/i, ɪ, ɛ, æ, eɪ, ɑ, ɔ , ʌ, ɑɪ, ɔɪ, ʊ, u, ɜ, ɑʊ, oʊ/}	/i, ɪ, ɛ, æ, eɪ/ (87,9%) /ɑ, ɔ, ʌ, ɑɪ, ɔɪ/ (82,5%) /u, ʊ, ɜ/ (89,5%) /ɑʊ/ (86,4%) /oʊ/ (76,9%)

Tab. 4.19: Homofemas vocálicos de Montgomery e Jackson (1983).

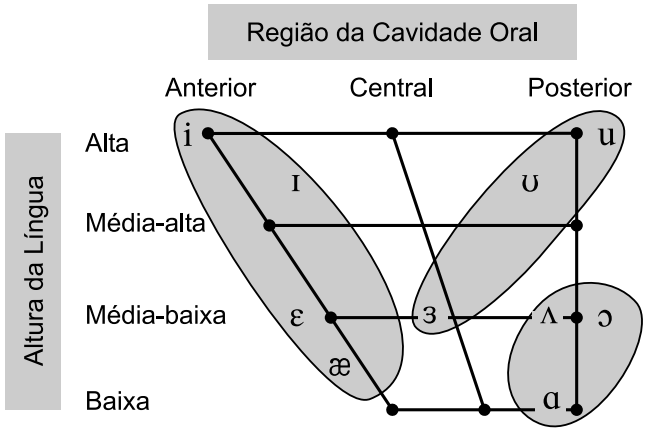

Fig. 4.4: Diagrama das vogais e os homofemas vocálicos de Montgomery e Jackson (1983) (somente monotongos).

reduzida dos lábios e a postura característica destes para a produção de /ɔ/ e /ɜ/. Este conjunto de eventos possui em geral grande visibilidade e não está apenas e necessariamente associado a um formato circular do contorno interior dos lábios. Segundo Jackson (1988), a característica do arredondamento influencia o segundo formante dos segmentos vocálicos; a abertura vertical dos lábios, o primeiro formante.

O padrão MPEG-4 (MPEG4 VISUAL, 2001) adota os grupos de homofemas vocálicos apresentados na Tabela 4.20. Esta tabela representa a conversão para o alfabeto fonético internacional da representação adotada na documentação ISO/IEC e reproduzida na Tabela 4.21. Para a conversão entre representações foi utilizada como referência a pronúncia padrão britânica (PAERSALL; TRUMBLE, 1996). É interessante observar que o padrão não especifica como os outros fonemas vocálicos existentes na língua inglesa, ou ainda existentes em outras línguas, devem ser associados a estes grupos de homofemas.

Homofemas Vocálicos
/ɑ/
/e/
/ɪ/
/ɒ/
/ʊ/

Tab. 4.20: Homofemas vocálicos do Padrão MPEG-4 (MPEG4 VISUAL, 2001).

Phonemes	**Example**
A:	c<u>a</u>r
e	b<u>e</u>d
I	t<u>i</u>p
Q	t<u>o</u>p
U	b<u>oo</u>k

Tab. 4.21: Fonemas vocálicos do padrão MPEG4 - Tabela C-5 (parcial) Anexo C (MPEG4 VISUAL, 2001).

MPEG-4	IPA
A:	ɑ
e	e
I	ɪ
Q	ɒ
U	ʊ

Tab. 4.22: Mapeamento entre as simbologias MPEG-4 e IPA para segmentos vocálicos.

4.4 Efeito da coarticulação nos homofemas

A coarticulação se manifesta pela alteração do padrão articulatório de um dado segmento pela influência da articulação de outro adjacente ou, e em menor grau, próximo na cadeia da produção sonora. É possível distinguir a coarticulação perseveratória da coarticulação antecipatória. A coarticulação perseveratória tem lugar quando a articulação de um segmento é influenciada por outro que o antecede na cadeia da produção sonora. Na coarticulação antecipatória, o segmento é influenciado por segmento que o sucede. Se o efeito da coarticulação não impacta significativamente na percepção acústica, permitindo que alofones gerados em contexto fonéticos diferentes sejam associados a um mesmo fonema, o mesmo não se aplica tão diretamente à percepção visual. A característica visual pode, principalmente para as consoantes, sofrer variações significativas a ponto de alterar a percepção dos grupos de homofemas.

Em estudo envolvendo a língua francesa, foram identificados os 21 homofemas apresentados na Tabela 4.23 (BENOÎT et al., 1992). Como indicado na segunda linha da tabela, os contextos fonéticos utilizados para a identificação foram V**V**V C**V**C e V**C**V, onde V representa um segmento vocálico e C um segmento consonantal. A classificação refere-se ao segmento em negrito. Para os contextos CVC e VCV, tem-se C = {/b, v, z, ʒ, r, l/} e V = {/i, y, a/}. Para o contexto VVV, tem-se V = {/i, y, e, ø, ɛ, œ, a, ɔ, o, u, ɛ̃, œ̃, ã, õ/}. A classificação foi efetuada a partir da análise de imagens de vídeo de um falante. Os seguintes parâmetros, tomados no centro da realização acústica do segmento, foram medidos: distância horizontal entre os cantos da boca (interna e externa); distância vertical entre os lábios no centro da boca (interna e externa), largura e altura do arco de cupido, posição vertical da ponta do queixo, protrusão do lábio superior, protrusão do lábio inferior, protrusão de um dos cantos da boca, área da zona vermelha do lábio superior, área da zona vermelha do lábio inferior, área entre

Homofemas			
ID	**VVV**	**CVC**	**VCV**
1	/aaa/	/rar/, /lal/	/ira/, /ara/, /ari/, /ala/
2	/iii/	/zaz/	
3	/yyy/, /øøø/, /ooo/, /uuu/, /õõõ/	/byb/, /zyz/, /ʒyʒ/, /ryr/, /lyl/	/yzy/, /yʒy/, /yry/, /yly/
4	/eee/, /ɛɛɛ/, /ɛ̃ɛ̃ɛ̃/	/rir/, /lil/	/iri/, /ili/, /ila/, /ali/
5	/ɔɔɔ/		
6	/œœœ/, /œ̃œ̃œ̃/		
7	/ããã/		
8		/bab/, /vav/	
9		/ʒaʒ/	
10		/bib/, /viv/	
11		/ziz/	/izi/, /iza/, /azi/, /aza/
12		/ʒiʒ/	/iʒi/, /iʒa/, /aʒi/, /aʒa/
13		/vyv/	/yvy/
14			/ibi/, /iba/, /abi/, /aba/
15			/iby/, /aby/, /ybi/, /yba/
16			/ivi/, /iva/, /avi/, /ava/
17			/ivy/, /avy/, /yvi/, /yva/, /izy/, /azy/, /yzi/, /yza/
18			/iry/, /ary/, /yri/, /yra/
19			/ily/, /aly/, /yli/, /yla/
20			/iʒy/, /aʒy/, /yʒi/, /yʒa/
21			/yby/

Tab. 4.23: Homofemas para a língua francesa de considerando efeitos da coarticulação (BENOÎT et al., 1992).

lábios da abertura da boca. Através de procedimentos de análise multidimensional, envolvendo os parâmetros medidos, os 21 homofemas foram revelados. Os parâmetros considerados como mais significativos no processo de agrupamento foram: a distância horizontal e vertical entre lábios e protrusão do canto da boca. Desta classificação é possível observar que as 14 vogais nos contexto VVV formaram 7 grupos diferentes de homofemas. Com a exceção de /ã/, os outros segmentos vocálicos nasais, /ɛ̃, œ̃, õ/, são homofemas dos correspondentes segmentos orais.

A Tabela 4.23 permite observar o efeito da coarticulação na percepção dos homofemas. Por exemplo, o homofema 14 está associado à produção de /b/ em contexto VCV, onde as vogais não são arredondadas. O homofema 15 está associado ao /b/ em contextos VCV onde uma das vogais é vogal arredondada /y/. Finalmente, o homofema 21 representa a produção de /b/ entre as vogais arredondadas.

Os resultados de Erber (1974) (Tabela 4.3), Benguerel e Pichora-Fuller (1982) e Owens e Blazek (1985) (Tabela 4.8) também ilustram aspectos do efeito da coarticulação na percepção visual. Os resultados dos três trabalhos apontam para uma redução da visibilidade dos segmentos consonantais no contexto VCV, quando V é a vogal arredondada /u/. Ainda nos resultados de Erber (1974) é possível verificar que o arredondamento aparentemente afetou /s, z/ em tal grau, que acabaram sendo associadas ao mesmo padrão visual que /ʃ, ʒ/, sendo que estes últimos segmentos possuem um padrão arredondado.

É possível identificar na literatura outros estudos, comentados a seguir, que procuram estabelecer um modelo descritivo da dinâmica do fenômeno da coarticulação, não havendo entretanto consenso sobre a validade e aplicabilidade dos mesmos. Há várias razões para esta indefinição. A fala é o resultado de um ato motor complexo, controlado por processos neurais, cujos detalhes são pouco conhecidos, que atuam de forma coordenada em vários sistemas neuromusculares, cada qual com complexas propriedades biomecânicas. São grandes as dificuldades inerentes ao estudo deste complexo sistema, incluindo o fato de que o objeto de estudo é de difícil acesso físico. A captura e a análise dos parâmetros relevantes é um grande desafio, e potencialmente os parâmetros mais interessantes são aqueles do processo neural subjacente da produção da fala, que, em geral, não estão disponíveis (PERKELL, 1990).

Para a coarticulação antecipatória tem-se três modelos principais: modelo *look-ahead*, modelo *time-locked*, e modelo híbrido.

O modelo *look-ahead* (ÖHMAN, 1966) (ÖHMAN, 1967) (BENGUEREL; COWAN, 1974) advoga que a movimentação de um determinado articulador, característico de um fonema, tem início, na cadeia da produção sonora, tão logo não exista mais conflito com o posicionamento articulatório necessário para a realização dos fonemas atual e intermediários. Para este modelo é possível que os efeitos da coarticulação antecipatória se estendam por até seis consoantes na cadeia da produção (LUBKER,

1982).

Por outro lado, o modelo *time-locked* (BELL-BERTI; HARRIS, 1979) (BELL-BERTI; HARRIS, 1982) estabelece que o efeito da coarticulação não é tão extenso e que a movimentação do articulador sempre tem início em um instante pré-definido antes do início da produção sonora do fonema.

O modelo híbrido (PERKELL, 1990) procura compatibilizar os outros dois modelos e define duas fases para a movimentação do articulador. A primeira fase tem início gradual tão logo seja possível, como no modelo *look-ahead*. Durante a segunda fase, a movimentação é mais acentuada, tendo início como no modelo 'time-locked'.

Uma justificativa para a existência de diferentes modelos estaria nas características dos diferentes procedimentos experimentais, que produziram dados experimentais aparentemente conflitantes (LUBKER, 1982) (GELFER; BELL-BERTI; HARRIS, 1989).

O modelo de produção gestual proposto em Löfqvist (1990) foi adaptado por Cohen e Massaro (1993) para modelar a coarticulação. Este modelo gestual estabelece que os diferentes segmentos da fala têm função de dominância, sobre os articuladores, que aumenta, atinge um máximo e depois decresce no tempo. Um segmento possui diferentes dominâncias sobre os diferentes articuladores. Segmentos adjacentes podem apresentar funções de dominância que se sobrepõem temporalmente, sofrendo um processo de combinação. Em Cohen e Massaro (1993) a dominância sobre os articuladores é modelada por função exponencial crescente, para a fase inicial, e decrescente, para fase a final da articulação. A combinação das funções de dominância é efetuada por uma soma ponderada entre as funções envolvidas. A definição apropriada dos parâmetros das funções de dominância permitem a modelagem dos vários modelos de coarticulação. No trabalho original, os parâmetros das funções de dominância e os pesos das misturas são definidos interativamente e a partir de observações informais.

Os parâmetros das funções de dominância do modelo de coarticulação de Cohen e Massaro foram estimados em LE GOFF (1997) e Benoît e Le Goff (1998) através da análise das imagens de vídeo associado à produção de um corpus de 722 palavras. O estudo buscou estimar os efeitos da coarticulação para o conjunto de visemas independentes do contexto fonético para a língua francesa proposto em BENOÎT et al. (1992). Na análise foram considerados trifones do tipo VCV e CVC (onde "V" indica vogal e "C" consoante), sendo que a análise das vogais foi limitada aos segmentos extremos /a, i, y/.

Recentemente, Beskow (2004) estimou os parâmetros dos modelos de coarticulação de Cohen e Massaro (1993) e de Öhman (1967), através de medidas fotogramétricas com quatro câmeras e processamento baseado em redes neurais.

Em Albrecht, Haber e Seidel (2002) a influência de funções de dominância do modelo de Cohen e Massaro foi limitada a sete segmentos para frente e sete para trás, com o intuito de reduzir o custo computacional dos cálculos dos efeitos da coarticulação perseveratória e antecipatória.

Em Revéret, Bailly e Badin (2000) e Elisei et al. (2001) são apontadas deficiências do modelo de Cohen e Massaro principalmente no tratamento das consoantes bilabiais oclusivas e labiodentais fricativas, sendo proposto um modelo de coarticulação baseado nas observações de Öhman (1967).

No sistema descrito em Pelachaud (1991) e Pelachaud, Badler e Steedman (1996) a coarticulação é modelada utilizando a estratégia "look-ahead". Na solução adotada, os fonemas são classificados conforme o seu grau de plasticidade, ou seja, como estão sujeitos a influências de outros segmentos. Esta classificação é análoga ao conceito de visibilidade proposto em Jeffers e Barley (1971) apresentado anteriormente. Dado um segmento consonantal com alto grau de plasticidade (baixa visibilidade), o algoritmo procura, para frente e para trás, a primeira vogal com o menor grau de plasticidade. Todos os segmentos consonantais entre a vogal e a consoante em questão são deformados, assumindo as características visuais da vogal com um fator de atenuação pré-definido. Dependendo se a vogal menos deformável (mais visível) esteja antes ou depois do fonema em questão, tem-se, respectivamente, o tratamento da coarticulação perseveratória ou antecipatória. Após esta primeira estimativa, o algoritmo verifica se a solução encontrada satisfaz restrições temporais e espaciais. Se o tempo entre dois fones não satisfaz restrições de contração e relaxamento muscular, é efetuada a ponderação utilizando uma aproximação polinomial de terceira ordem que interpola entre as posições limites totalmente relaxada e totalmente contraída. A intensidade da ação que resulta na conformação geométrica da face também é ponderada levando em conta restrições espaciais pré-definidas. Uma matriz pré-definida estabelece fatores de atenuação entre as várias posições possíveis.

O padrão MPEG-4 suporta apenas a coarticulação entre segmentos adjacentes através da soma ponderada da contribuição de cada um deles (OSTERMANN, 2002).

4.5 Comentários Finais

O conceito de fonema pode ser entendido como a menor unidade de percepção auditiva da fala. Um fonema representa um conjunto de realizações sonoras que são percebidas como similares, possuindo a mesma função lingüística. Os fonemas permitem a distinção e contraste lingüístico entre os segmentos da fala. O mesmo princípio pode ser atribuído ao conceito de homofema. Muitos segmentos ao serem articulados apresentam um mesmo padrão visual que permite a distinção entre os vários grupos de homofemas de uma língua. Da mesma forma que os fonemas admitem alofones, ou seja, variações na produção acústica, que, entretanto, são reconhecidos como uma mesma unidade contrastável, um grupo de homofemas também está associado a variações do padrão visual da realização articulatória. No caso de um homofema, tais variações podem estar associadas à produção de um ou mais fonemas distintos. Um padrão visual de movimentação articulatória representante de um grupo de homofemas é denominado neste trabalho de *Visema*. Uma das questões principais deste trabalho

é, além da identificação de grupos de homofemas para o português do Brasil, o estabelecimento de visemas representativos destes grupos de homofemas.

A identificação de homofemas está associada à identificação e ao contraste entre diferentes padrões visíveis de movimentação articulatória, ou seja, entre diferentes visemas. A percepção de um visema, e a respectiva identificação dos fonemas associados, é uma tarefa que depende de capacidades inerentes ao ouvinte e ao falante, assim como de fatores ambientais e lingüísticos. A coarticulação, definida principalmente pelo contexto fonético e taxa de elocução, é um importante fator lingüístico que influencia significativamente a movimentação articulatória e, conseqüentemente, o padrão visual apresentado na face do locutor e deve, portanto, ser considerada na implementação de uma cabeça virtual falante.

Na revisão da literatura não é possível identificar uma classificação universal de homofemas que contemple todos os fonemas em todos os contextos visuais e lingüísticos. Entretanto, apesar das discordâncias na literatura, é possível identificar alguns pontos em comum. Considerando apenas os fonemas do português do Brasil existentes nos estudos da língua inglesa, observa-se que os grupos /f, v/, /p, b, m/ e /ʃ, ʒ/ são identificados como grupos de homofemas na maioria dos estudos. Os outros fonemas consonantais, /t, d, n, l, s, z, k, g/, são menos visíveis. Não obstante, o critério do lugar de articulação proposto em Nitchie (1950) parece ser uma boa aproximação para a classificação. Visando a redução do número de casos a serem estudados, esta também será a estratégia a ser adotada para os fonemas /ʎ, ɲ/.

A Tabela 4.24 apresenta os agrupamentos de fonemas consonantais que serão considerados como homofemas neste trabalho. A primeira coluna da tabela apresenta os agrupamentos dos fonemas em homofemas, sendo que a segunda coluna apresenta a designação que será utilizada para se referir aos homofemas da primeira coluna.

Os poucos estudos publicados sobre homofemas vocálicos são um indicativo da dificuldade da classificação destes segmentos. Diferentemente dos fonemas consonantais, onde é possível identificar grupos com características visuais comuns, como por exemplo o fechamento dos lábios das bilabiais, cada um dos segmentos vocálicos é articulado de maneira diferente. A característica visual mais marcante dos segmentos vocálicos, no que tange à língua portuguesa, é o arredondamento dos lábios das vogais posteriores. A abertura da mandíbula, que está associada ao posicionamento da língua no interior da cavidade oral, é uma característica menos diferençável visualmente, mas que fornece pistas para a identificação do segmento vocálico. Os segmentos vocálicos nasais, considerando que a movimentação do véu palatino não é, em geral, visível, são homofemas dos respectivos segmentos vocálicos orais, como indicado pelos resultados apresentados em (BENOÎT et al., 1992) para a língua francesa.

Assumir que a característica visual mais marcante dos segmentos vocálicos também é, no que

Homofema	**Designação**
[p, b, m]	Bilabial
[f, v]	Labiodental
[t, d, n]	Alveolar plosivo/nasal
[s, z]	Alveolar fricativo
[ɾ]	Alveolar tepe
[l]	Alveolar lateral
[ʃ, ʒ]	Pós-alveolar
[ʎ, ɲ]	Palatal
[k, g]	Velar plosivo
[ɣ]	Velar fricativo

Tab. 4.24: Agrupamento dos segmentos consonantais em homofemas considerado no trabalho.

Homofema	**Designação**
[i, ĩ]	Alto anterior
[e, ẽ]	Médio-alto anterior
[ɛ]	Médio-baixo anterior
[a, ɐ̃]	Baixo central
[ɔ]	Médio-baixo posterior
[o, õ]	Médio-alto posterior
[u, ũ]	Alto posterior
[ɪ]	Postônico alto anterior
[ɐ]	Postônico baixo central
[ʊ]	Postônico alto posterior

Tab. 4.25: Agrupamento dos segmentos vocálicos em homofemas considerado neste trabalho.

tange à língua portuguesa falada no Brasil, o arredondamento dos lábios das vogais posteriores, é compatível com as classificações apresentadas nas Tabelas 4.16, 4.17 e 4.19. Adicionalmente, considerar que a abertura horizontal dos lábios, que está associada à abertura da mandíbula, que por sua vez está associada ao posicionamento da língua no interior da cavidade oral, é uma característica visualmente menos perceptível, mas que fornece algumas pistas para a identificação dos segmentos vocálicos, também é justificável, uma vez que em nenhum dos estudos houve a confusão das vogais altas anteriores com a vogal baixa /ɑ/ (confira Figuras 4.1 a 4.4).

A Tabela 4.25 apresenta o agrupamento dos segmentos vocálicos orais e nasais que serão considerados, neste trabalho, como homofemas. Esta tabela está organizada da mesma maneira que a Tabela 4.24.

No próximo capítulo é apresentada a metodologia desenvolvida para a identificação e descrição de visemas associados aos homofemas consonantais e vocálicos para o português do Brasil identificados no presente capítulo.

Capítulo 5

Visemas para o português do Brasil

5.1 Introdução

Como discutido no Capítulo 4, visemas são padrões de movimentação articulatória que permitem a identificação e diferenciação visual dos homofemas de uma língua. No que tange à animação facial, a caracterização de um visema pode ser reduzida a três aspectos fundamentais: 1) a descrição geométrica de posturas articulatórias representativas da conformação do trato vocal, principalmente na região da boca, que caracterizam a produção dos segmentos da língua; 2) os instantes de realização destas posturas articulatórias em relação ao sinal acústico da fala; 3) a transição entre as posturas alvo dos diversos segmentos que constituem a cadeia da locução, levando em conta os efeitos da coarticulação.

No presente capítulo é apresentada uma metodologia para a identificação, caracterização e modelagem destes três aspectos a partir de medidas das trajetórias de pontos marcados na face de um falante durante a produção de um corpus. O corpus utilizado procura expor de maneira sistemática os efeitos da coarticulação adjacente antecipatória e perseveratória.

É importante observar que não é do conhecimento do autor nenhum estudo anterior que tenha estabelecido um conjunto de visemas para o português do Brasil. A falta dos visemas para o português do Brasil motivou o desenvolvimento da metodologia objeto deste capítulo.

A metodologia proposta estabelece uma descrição geométrica e temporal das posições de pontos ao redor da boca, caracterizando um conjunto de visemas para o português do Brasil. Estes visemas dependem não só do segmento que está sendo produzido, como também do contexto fonético de sua realização. A interpolação dos parâmetros geométricos, atendendo a restrições temporais dos instantes de realização das posturas articulatórias alvo e duração da locução, resulta na simulação da dinâmica articulatória visível englobando os efeitos da coarticulação adjacente.

A Seção 5.2 detalha a metodologia e instrumentação utilizada para a identificação de um conjunto

de visemas para o português do Brasil. Uma avaliação da precisão das medidas e os resultados da aplicação desta metodologia são apresentados na Seção 5.3.

5.2 Metodologia e instrumentação

A metodologia adotada para o estabelecimento de visemas baseia-se na medida da trajetória de pontos de interesse marcados na face de um falante durante a produção de logatomas (palavras sem sentido). A partir destas medidas procura-se identificar posturas articulatórias alvo, ou simplesmente alvos articulatórios, característicos dos vários segmentos da língua. A movimentação articulatória visível durante a fala é resultado da transição entre os alvos articulatórios de cada um dos segmentos da seqüência que compõe um locução. As principais premissas da metodologia são:

- Cada segmento de uma língua pode ser caracterizado por um alvo articulatório, que estabelece a conformação característica do trato vocal necessária à sua produção;
- A movimentação articulatória durante uma locução pode ser expressa pela transição entre os alvos articulatórios da seqüência de segmentos que compõem a fala;
- O alvo articulatório de um segmento, devido aos efeitos da coarticulação, depende do contexto fonético de sua produção.

A Figura 5.1 apresenta um diagrama de fluxo de dados da metodologia empregada. No diagrama, os arcos direcionados representam as informações que fluem entre as várias etapas de processamento, representadas pelos círculos. O resultado final, indicado pelo retângulo cinza, é um conjunto de visemas para o português do Brasil. Resumidamente, a metodologia é composta pelas seguintes etapas:

- **Captura em áudio e vídeo**: Gravação em áudio e vídeo da produção de corpus constituído de um conjunto de frases sem sentido. O corpus analisado é apresentado e discutido na Seção 5.2.1. O procedimento e a instrumentação utilizada na gravação são apresentados na Seção 5.2.2;
- **Segmentação do áudio**: Identificação na trilha de áudio das realizações dos vários segmentos do português do Brasil e suas respectivas temporizações. A Seção 5.2.3 detalha o procedimento adotado para a segmentação do áudio;
- **Medida das trajetórias de pontos da face**: Levantamento das trajetórias no espaço tridimensional de pontos de interesse marcados na face do falante durante a produção dos logatomas do corpus. A Seção 5.2.4 aborda a técnica e os procedimentos utilizados para as medidas;

- **Identificação dos alvos articulatórios**: Análise das trajetórias dos pontos de interesse para a identificação dos alvos articulatórios característicos dos segmentos, levando em conta diferentes contextos fonéticos. Este procedimento é tratado na Seção 5.2.5;

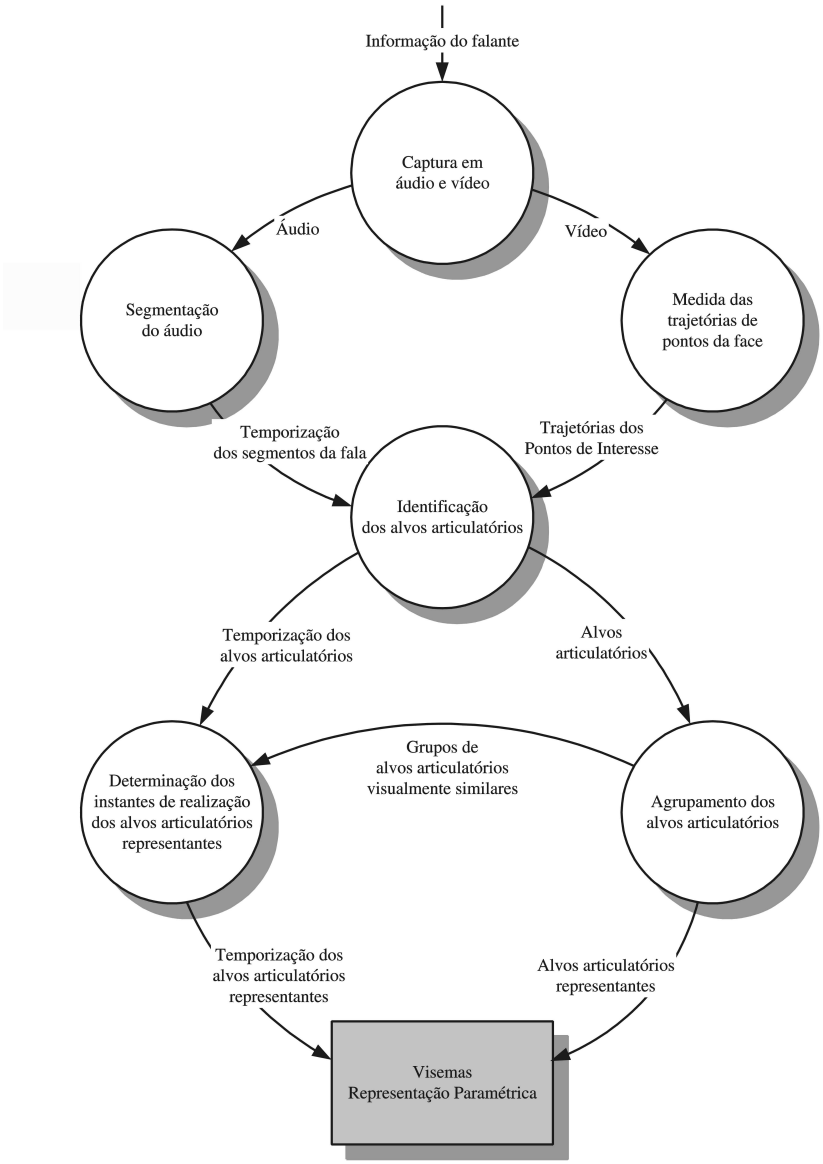

Fig. 5.1: Diagrama de fluxo de dados da metodologia desenvolvida.

- **Agrupamento dos alvos articulatórios**: Processo de agrupamento por similaridade (*clustering*) dos alvos articulatórios de um mesmo segmento em diversos contextos fonéticos. O centróide da nuvem de alvos que forma um grupo é tomado como o alvo articulatório representante

do grupo e estabelece os parâmetros geométricos estáticos da postura articulatória característica de um visema. Na Seção 5.2.6 tem-se o detalhamento do procedimento adotado;

- **Determinação dos instantes de realização dos alvos articulatórios representantes**: Nesta etapa são determinados os instantes de realização, em relação à produção acústica, dos alvos articulatórios representantes de cada um dos grupos identificados. A Seção 5.2.7 apresenta a estratégia adotada.

A aplicação da seqüência de processamentos descrita acima resulta na caracterização da geometria e da temporização de um conjunto de visemas. O padrão de movimentação articulatória expresso pelos visemas é reproduzido durante a animação facial pela interpolação da geometria dos alvos articulatórios representantes, temporizada pelos seus instantes de realização e em sincronia com a produção acústica da fala (Seção 5.2.8).

5.2.1 Corpus

A análise lingüística realizada no escopo deste trabalho foi centrada em logatomas paroxítonos do tipo $'C_1V_1C_2V_2$, onde C_1 e C2 representam segmentos consonantais, e V_1 e V_2 segmentos vocálicos, sendo que os segmentos C_1 e C_2 são opcionais. O símbolo IPA " ' ", antes de C_1, precede e marca a sílaba acentuada. O tipo de logatoma adotado permite o estudo da coarticulação adjacente antecipatória e perseveratória nos contextos:

- #**C**V - segmento consonantal entre silêncio e segmento vocálico (o símbolo # representa o silêncio);
- V**C**V - segmento consonantal entre dois segmentos vocálicos;
- #**V**C - segmento vocálico entre silêncio e segmento consonantal;
- #**V**V - segmento vocálico entre silêncio e segmento vocálico;
- C**V**C - segmento vocálico entre dois segmentos consonantais;
- C**V**V - segmento vocálico entre segmento consonantal e segmento vocálico;
- C**V**# - segmento vocálico entre segmento consonantal e silêncio;
- V**V**# - segmento vocálico entre segmento vocálico e silêncio.

Observa-se que a escolha de logatomas paroxítonos para a análise explora o fato de que no português a grande maioria das palavras possui este tipo de tonicidade.

Considerando os logatomas de tipo $'C_1V_1C_2V_2$ e os segmentos do português do Brasil conforme discutido no Capítulo 3, é possível identificar 36.480 produções diferentes, uma vez que:

- 19 possibilidades diferente podem ocupar C_1 - 18 segmentos consonantais no início de palavra (confira Tabelas 3.1 e 3.2) e a ausência de segmento;
- 12 segmentos vocálicos (7 orais e 5 nasais - Tabelas 3.9 e 3.12) podem ocupar a posição V_1 na sílaba tônica;
- 20 possibilidades diferentes podem ocupar C_2 - 19 segmentos consonantais para a posição intervocálica (Tabela 3.1e 3.2) e a ausência de segmento; e
- 8 segmentos vocálicos (3 orais e 5 nasais - Tabelas 3.10 e Tabela 3.12) na posição postônica V_2.

Para viabilizar este trabalho, dado o elevado número de casos, optou-se por uma análise mais restrita explorando a similaridade visual existente por definição nos grupos de homofemas. Assim, para cada grupo de homofemas, apenas um segmento representante foi selecionado para a análise.

A Tabela 5.1 apresenta os grupos de homofemas consonantais considerados (Seções 4.2 e 4.5). Nesta tabela, a primeira coluna apresenta o agrupamento dos segmentos em homofemas; a segunda apresenta a designação do grupo; e a última apresenta o segmento adotado como representante. A análise dos padrões de movimentação articulatória visível foi efetuada utilizando-se apenas o segmento representante. É assumido tacitamente que os outros segmento de um mesmo grupo apresentam as mesmas pistas visuais da movimentação articulatória na face do falante. De forma análoga, a Tabela 5.2 apresenta homofemas vocálicos e os respectivos segmentos representantes.

Considerando os logatomas do tipo $'C_1V_1C_2V_2$ e a simplificação das possibilidades advindas do agrupamento em homofemas, o número de casos se reduz para $10 \times 7 \times 11 \times 3 = 2.310$. Dentre este conjunto foi efetuada a gravação das produções do tipo $'C_1V_1C_2V_2$, com as consoantes opcionais e C_1 = {[p, t, k, f, s, ʃ, l, ʎ, ɣ]}, C_2 = {[p, t, k, f, s, ʃ, l, ʎ, ɾ]}, V_1 = {[i, e, ɛ, a, ɔ, o, u]} e V_2 = {[ɪ, ɐ, ʊ]}.

Devido ao tempo de processamento, a análise do material gravado associado a todos os logatomas revelou-se inviável para o escopo deste trabalho. Em média, o tempo de processamento de cada logatoma foi de aproximadamente 2 dias. Assim, optou-se pela análise detalhada de um grupo mais restrito de casos, deixando para um trabalho futuro o refinamento dos visemas com a incorporação dos resultados da análise destes outros casos. Para a análise foram selecionados os seguintes 102 casos:

Homofema	**Designação**	**Representante**
[p, b, m]	Bilabial	[p]
[f, v]	Labiodental	[f]
[t, d, n]	Alveolar plosivo/nasal	[t]
[s, z]	Alveolar fricativo	[s]
[ɾ]	Alveolar tepe	[ɾ]
[l]	Alveolar lateral	[l]
[ʃ, ʒ]	Pós-alveolar	[ʃ]
[ʎ, ɲ]	Palatal	[ʎ]
[k, g]	Velar plosivo	[k]
[ɣ]	Velar fricativo	[ɣ]

Tab. 5.1: Homofemas consonantais e fones representantes adotados.

Homofema	**Designação**	**Representante**
[i, ĩ]	Alto anterior	[i]
[e, ẽ]	Médio-alto anterior	[e]
[ɛ]	Médio-baixo anterior	[ɛ]
[a, ɐ̃]	Baixo central	[a]
[ɔ]	Médio-baixo posterior	[ɔ]
[o, õ]	Médio-alto posterior	[o]
[u, ũ]	Alto posterior	[u]
[ɪ]	Postônico alto anterior	[ɪ]
[ɐ]	Postônico baixo central	[ɐ]
[ʊ]	Postônico alto posterior	[ʊ]

Tab. 5.2: Homofemas vocálicos e fones representantes adotados.

- ˈCV_1CV_2 com C = {[p, t, k, f, s, ʃ, l, ʎ, (ɣ)ɾ]}, V_1 = {[i, a, u]} e V_2 = {[ɪ, ɐ, ʊ]}, perfazendo um total de 81 casos. Note-se que aqui foi feito $C_1 = C_2 = C$. Observa-se também que, uma vez que o tepe [ɾ] não ocorre em início de palavra, optou-se pela análise dos logatomas do tipo ˈ[ɣ]V_1[ɾ]V_2 para o tratamento dos segmentos [ɾ] e [ɣ];

- ˈV_1V_2, com V_1 = {[i, e, ɛ, a, ɔ, o, u]} e V_2 = {[ɪ, ɐ, ʊ]}, perfazendo um total de 21 casos. As informações dos logatomas deste tipo foram coletadas para fornecer subsídios à classificação dos segmentos vocálicos em visemas. Como estratégia de trabalho foi estabelecido que os segmentos vocálicos [e, ɛ, ɔ, o] seriam associados, por critério de similaridade, a algum dos visemas dos segmentos [i, a, o, ɪ, ɐ, ʊ].

O corpus efetivamente tratado permite a análise de 18 produções diferentes para cada um dos segmentos consonantais [p, t, k, f, s, ʃ, l, ʎ, (ɣ)ɾ]; de 30 produções para cada um dos segmentos vocálicos [i, a, u]; de 34 casos para as vogais [ɪ, ɐ, ʊ]; e, finalmente, de 3 casos para os segmentos [e, ɛ, ɔ, o]. A Tabela 5.3 lista os contextos fonéticos presentes no corpus.

Segmento	**Contextos**
C = {[p, t, k, f, s, ʃ, l, ʎ, (ɣ)ɾ]}	#CV, onde V = {[i, a, u]} (3 repetições de cada)
	V_1CV_2, onde V_1 = {[i, a, u]} e V_2 = {[ɪ, ɐ, ʊ]}
V = {[i, a, u]}	CVC, onde C = {[p, t, k, f, s, ʃ, l, ʎ, (ɣ)ɾ]} (3 repetições de cada)
	#VV_1, onde V_1= {[ɪ, ɐ, ʊ]}
V = {[ɪ, ɐ, ʊ]}	CV#, onde C = {[p, t, k, f, s, ʃ, l, ʎ, (ɣ)ɾ]} (3 repetições de cada)
	#V_1V, onde V_1 = {[i, e, ɛ, a, ɔ, o, u]}
V = {[e, ɛ, ɔ, o]}	#VV_1, onde V_1= {[ɪ, ɐ, ʊ]}

Tab. 5.3: Contextos fonéticos analisados.

5.2.2 Captura do áudio e vídeo

Para a identificação de visemas para o português do Brasil, foram efetuadas gravações em vídeo de um falante nascido e criado na capital do Estado de São Paulo, ao enunciar logatomas do tipo ˈCV_1CV_2 (Seção 5.2.1). Foram gravados um total de 2.100 logatomas.

O falante foi gravado com auxílio de duas câmeras de vídeo JVC KY27C, devidamente sincronizadas e posicionadas ortogonalmente entre si. Para a captura do áudio foi utilizado um microfone

Shure SM58 posicionado frontalmente e a uma distância de aproximadamente 1 metro do falante. As imagens e o áudio foram gravadas em fitas S-VHS com o auxílio de 2 gravadores de vídeo - um JVC BR-S622 DXU e outro JVC BR-S822 DXU. O material bruto capturado totalizou 3 horas de gravação. A gravação foi realizada nas dependências do Estúdio de Produção de Vídeo do Departamento de Multimeios do Instituto de Artes da Universidade Estadual de Campinas. Os logatomas foram lidos da tela de um computador portátil posicionado à frente do falante. A Figura 5.2, capturada poucos instantes antes do início das gravações, permite uma visão do estúdio e do posicionamento das câmeras, microfone e computador portátil. Na imagem à direita é possível observar o computador. Na imagem à esquerda, em seu centro, aparece o microfone localizado à frente do falante.

Fig. 5.2: Instalações da gravação.

Os logatomas foram apresentados individualmente com letras de corpo tamanho 72 pontos e em uma seqüência aleatória previamente sorteada e desconhecida do falante. O início da apresentação de cada logatoma foi marcado por sinal sonoro de 1 kHz com duração de 0,5 segundo. O instante de apresentação do próximo logatoma foi comandado por um operador com o auxílio de aplicativo especialmente desenvolvido para tal fim em execução no computador portátil. Para facilitar o processo de gravação, este aplicativo foi dotado de capacidades básicas de navegação pela base de dados, tais como, ir para o próximo logatoma, repetir o atual, voltar para a anterior, ou ir para um determinado logatoma, gerando o tom de 1 KHz antes da apresentação de cada logatoma.

A equipe de gravação foi formada, além do falante, por outros três participantes: um para monitorar, dentro da cabine de gravação do estúdio, os sinais de áudio e vídeo; outro para operar o computador portátil e comandar a apresentação dos logatomas e um terceiro para acompanhar e conferir, com a lista dos logatomas, a correta locução destes. Ao se detectar uma falha de locução, optou-se por repetir imediatamente o logatoma, até a sua produção correta. O falante foi instruído a pronunciar cada logatoma apenas após o tom de 1 KHz e para articulá-los a partir de uma posição de

repouso, com boca fechada e os dentes cerrados, devendo retornar para esta posição após a produção acústica.

Na face do locutor foram previamente marcados, com tinta branca, pontos de interesse em torno da boca, no queixo, na bochecha e na ponta do nariz. Um capacete, visto na Figura 5.3, foi especialmente desenvolvido e construído para estabelecer, mesmo com movimentos involuntários da cabeça, um sistema de referência para as medidas da posição dos pontos de interesse através da técnica fotogramétrica apresentada no Apêndice A. A adoção do capacete, fixo à cabeça do locutor, evita a necessidade de dispositivos, nem sempre confortáveis, para fixar e garantir a imobilidade de cabeça do locutor, como, por exemplo, o adotado em Fujimura (FUJIMURA, 1980).

O capacete é composto de um suporte ajustável para fixação de cabeça e uma estrutura de alumínio, fixa a este suporte, pintada de branco, onde foram marcados pontos pretos de 20 em 20 milímetros. Os pontos marcados da estrutura de alumínio foram utilizados como pontos de referência para a calibração da câmera na técnica fotogramétrica (veja Apêndice A) e como pontos de controle para avaliação da precisão das medidas realizadas (Seção 5.3.1).

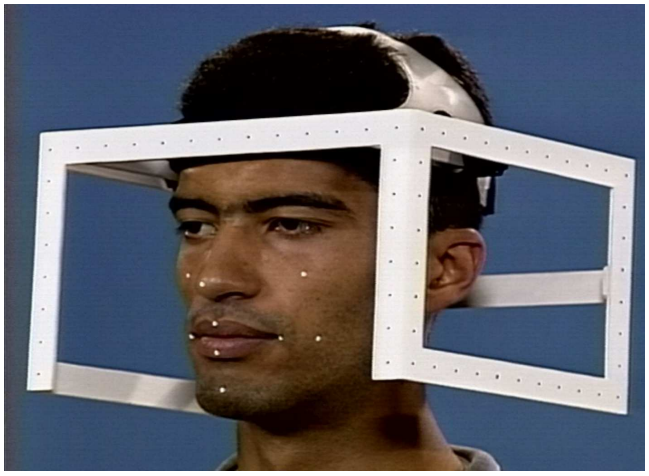

Fig. 5.3: Capacete de referência e pontos de interesse.

As fitas de vídeo contendo as locuções realizadas, foram digitalizadas e armazenadas no formato Mini-DV utilizando um gravador e reprodutor de vídeo dual digital/analógico Mini-DV/S-VHS JVC SR-VS10U. Após a digitalização o material foi transferido, através de interface digital IEEE 1394 (firewire), para uma ilha de edição não-linear iFinish V60 versão 3.2, Media 100. O material foi segmentado e rotulado manualmente nas fronteiras de produção dos logatomas, tomando-se como referência o tom de 1 kHz. Dos 2.100 logatomas gravados, 102 foram selecionados para a análise (Seção 5.2.1).

5.2.3 Segmentação do áudio

Através da análise do sinal de áudio de cada logatoma, foram identificados os instantes de início e fim da produção acústica de cada segmento fonético do logatoma. Esta segmentação foi efetuada de forma visual e auditiva com o auxílio do pacote Cool Edit versão 96, da empresa Syntrillium Software Corporation. A análise visual baseou-se na inspeção visual da onda do sinal de áudio (gráfico da amplitude em função do tempo) e de seu espectrograma (amplitude das componentes espectrais em função do tempo). A título de exemplo, a Figura 5.4 apresenta a forma de onda e o espectograma do sinal de áudio associados à locução ['kakɐ]. Nas imagens da figura, as linhas verticais em cor clara indicam a segmentação do áudio.

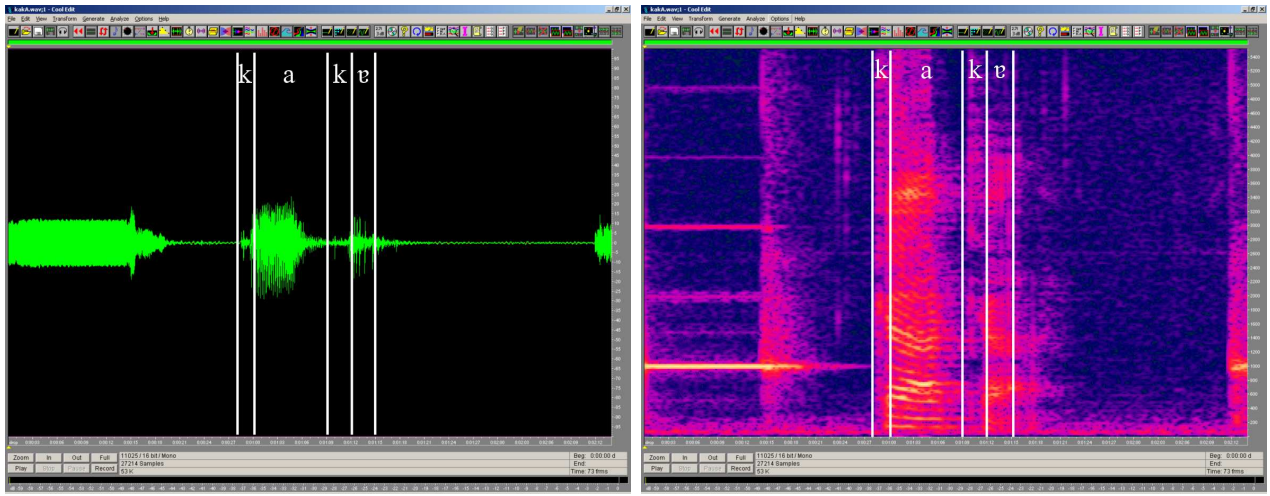

Fig. 5.4: Inspeção visual do sinal de áudio: esquerda) forma de onda; direita) espectograma.

5.2.4 Medida das trajetórias de pontos da face

A Figura 5.5 identifica os pontos da face cujas trajetórias foram analisadas durante a articulação dos logatomas do corpus. Estes pontos, denominados de pontos de interesse, foram marcados com tinta branca na face do falante. A figura indica também a orientação do sistema de coordenadas cartesianas convencionado para as medidas. As trajetórias dos pontos de interesse foram calculadas utilizando processo fotogramétrico, cuja formulação é apresentada no Apêndice A.

Para automatizar o processo de medida da trajetória espacial dos pontos de interesse durante a articulação dos segmentos fonéticos, foi desenvolvida uma ferramenta de *software*. Esta utiliza técnicas de processamento de imagem para identificar, na imagem inicial da produção de cada logatoma, pontos de referência (pontos no capacete) e pontos de interesse marcados na face do falante. Em seguida, calcula as coordenadas dos pontos de interesse em todas as outras imagens da seqüência de vídeo.

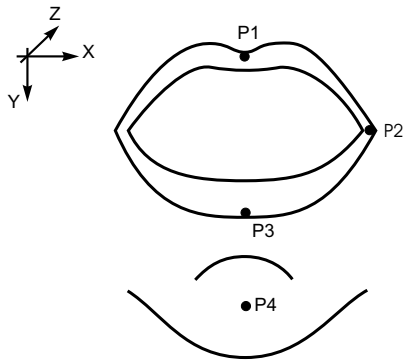

Fig. 5.5: Pontos de interesse cujas trajetórias foram medidas.

Para a imagem inicial de cada locução foram utilizados 15 pontos de referência para o processo de calibração de cada câmera. Os pontos de referência para a imagem esquerda (esquerda do locutor) da Figura 5.6 foram identificados com os rótulos L_i , $i = 1, 2..., 15$. Os pontos de referência da imagem direita foram identificados com os rótulos R_i , $i = 1, 2, ..., 15$.

Após a calibração, o sistema foi configurado para calcular a trajetória de 31 pontos: 4 pontos de interesse localizados na face do locutor e 27 pontos de controle localizados na estrutura do capacete. Os 4 pontos da face são (Figuras 5.5 e 5.6): lábio superior (P_1); canto dos lábios (P_2); lábio inferior (P_3); e queixo (P_4). Os pontos de controle P_5 a P_{31} são mostrados na Figura 5.7.

Por convenção adotou-se o ponto L_3 (ou de forma equivalente R_3) como a origem do sistema cartesiano de medidas. Na convenção adotada, o eixo X aponta para a esquerda do falante, o eixo Y para baixo e o eixo Z para trás. Os valores das coordenadas (x, y, z) da posição dos pontos em cada quadro foram exportados e armazenados em arquivo em formato compatível com o pacote de processamento matemático MatLab versão 6.1.0.450 release 12.1, da empresa The MathWorks.

Com o auxílio do pacote MatLab, os dados brutos das trajetórias foram filtrados, para redução do ruído de captura, por um filtro passa-baixa Butterworth de sexta ordem com freqüência de corte de aproximadamente 4,5 Hz , equivalendo a $(1/6)$ da freqüência de quadro das imagens de vídeo NTSC - 29,97 Hz.

Após a filtragem, as trajetórias foram deslocadas para indicar posições relativas à posição inicial de repouso. O valor da posição de repouso foi calculado pela média dos valores dos 10 quadros iniciais e dos 10 quadros finais da produção de todos os logatomas.

5.2.5 Identificação dos alvos articulatórios

Baseado na premissa de que durante a produção de uma seqüência de fones, os articuladores se movem de uma postura articulatória alvo para outra, implementou-se um procedimento para a de-

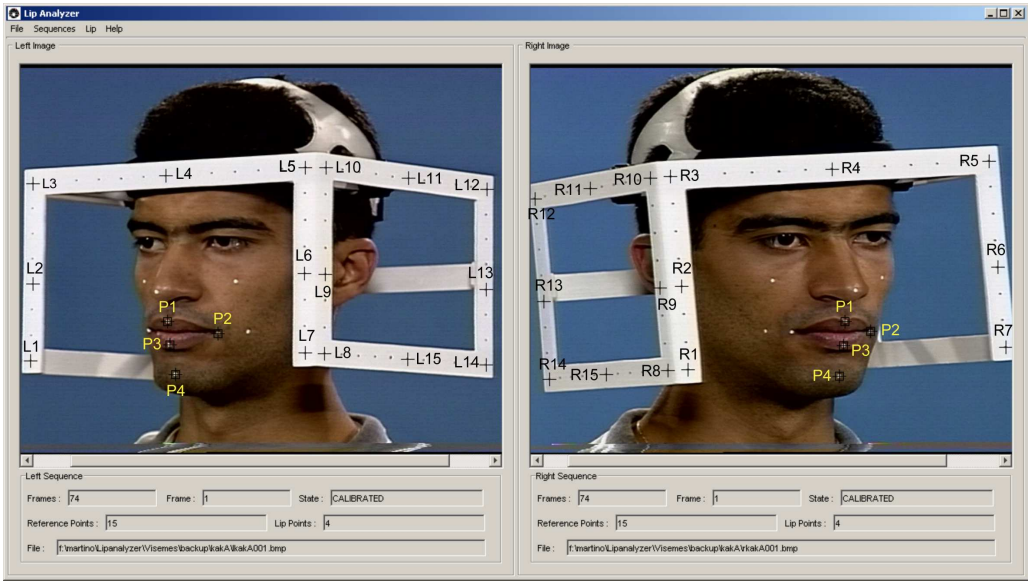

Fig. 5.6: Pontos de referência (L_i e R_i , $i = 1, 2, ..., 15$) utilizados para a calibração das câmeras e os pontos de interesse (P_1 a P_4).

Fig. 5.7: Pontos utilizados para o controle e avaliação da precisão das medidas.

terminação do primeiro ponto estacionário na trajetória dos pontos de interesse durante a produção dos logatomas. Entende-se como ponto estacionário, o instante na trajetória que apresenta a derivada em relação ao tempo igual a zero (velocidade nula). O ponto estacionário sinaliza o instante em que ocorre uma parada, com eventual mudança de direção, na trajetória do ponto de interesse, caracterizando o instante da realização da postura articulatória alvo do segmento. Os articuladores, uma vez atingido o alvo articulatório, alteram a trajetória movendo-se para realizar a postura articulatória alvo do próximo segmento. A determinação do ponto estacionário foi realizada tendo com referência a coordenada cartesiana (x, y, z) com maior variação durante a produção do logatoma. Esta direção de maior variação foi denominada de direção dominante.

A título de exemplo, as Figuras 5.8 a 5.10 apresentam o comportamento do ponto P_4, localizado no queixo do falante, durante a produção do logatoma [ˈpapɐ]. Nas figuras a abscissa representa o número do quadro do sinal de vídeo, sendo um quadro gerado a cada $(1/29,97)$ segundos. Neste exemplo, a direção Y é a direção dominante. Na Figura 5.9, as marcas circulares pretas indicam os instantes em que ocorrem os pontos estacionários da trajetória. Estes mesmos instantes são apresentados nas Figuras 5.8 e 5.10 com asteriscos. Os valores dos deslocamentos a partir do repouso das coordenadas x, y e z nos pontos estacionários definem os alvos articulatórios dos respectivos segmentos. Nas figuras, as linhas tracejadas verticais indicam as fronteiras da produção acústica dos segmentos, como determinado na segmentação do áudio. Observa-se que a busca dos instantes de inflexão foi efetuada dentro do intervalo de produção articulatória do segmento. O intervalo de produção articulatória do segmento sempre maior ou igual do que o intervalo de produção acústica, foi estabelecido através da análise das imagens de vídeo próximas das fronteiras do intervalo de produção acústica do segmento estabelecido pela segmentação do áudio. Em se observando movimentação significativa associada à produção do segmento próximo às fronteiras, o intervalo foi alargado estabelecendo-se o intervalo articulatório para abranger esta movimentação. Em particular para o fone [p], a análise das imagens do vídeo revelou que o instante de fechamento dos lábios, que caracteriza o segmento e a postura articulatória alvo, ocorre antes da produção acústica, como é possível observar nas figuras imediatamente antes dos quadros 28 e 38. Este fato se justifica ao se observar que o som característico da plosiva bilabial [p] é resultante da abertura forçada (plosiva) dos lábios e, portanto, só ocorre após o fechamento destes.

O logatoma [ˈpapɐ] caracteriza-se por ser uma seqüência do tipo ˈCVCV, onde o segmento consonantal, bilabial plosivo, está associado a uma forte oclusão de região visível do trato vocal e o segmento vocálico, baixo e central, apresenta uma forte abertura do trato. Assim, é de se esperar que a seqüência [ˈpapɐ] apresente uma movimentação marcadamente pulsada da boca entre as posições fechada/aberta/fechada/aberta, como implícito nas Figuras 5.8 a 5.10, corroborando as premissas adotadas na caracterização da postura articulatória alvo e apresentadas no início da Seção 5.2. Tais

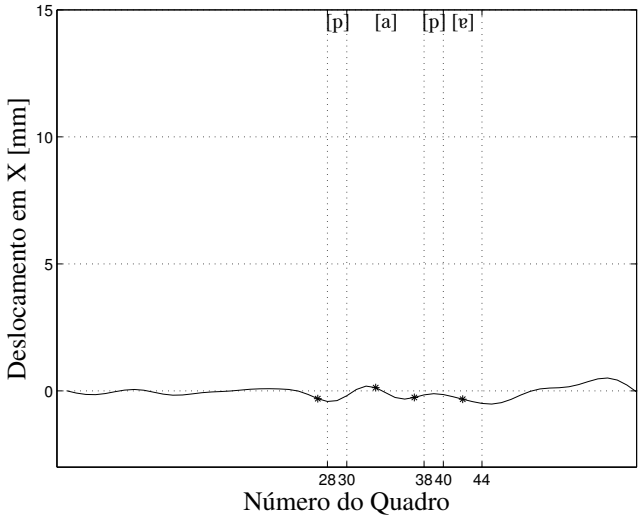

Fig. 5.8: Deslocamento na direção X de P_4 no plano sagital durante a produção de [ˈpapɐ].

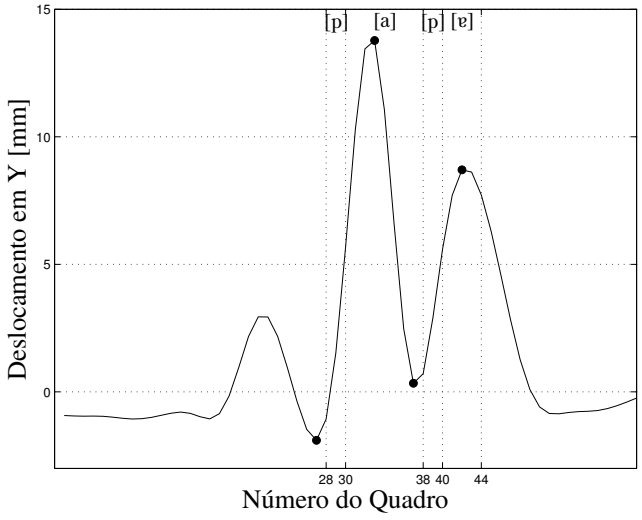

Fig. 5.9: Deslocamento na direção Y de P_4 no plano sagital durante a produção de [ˈpapɐ].

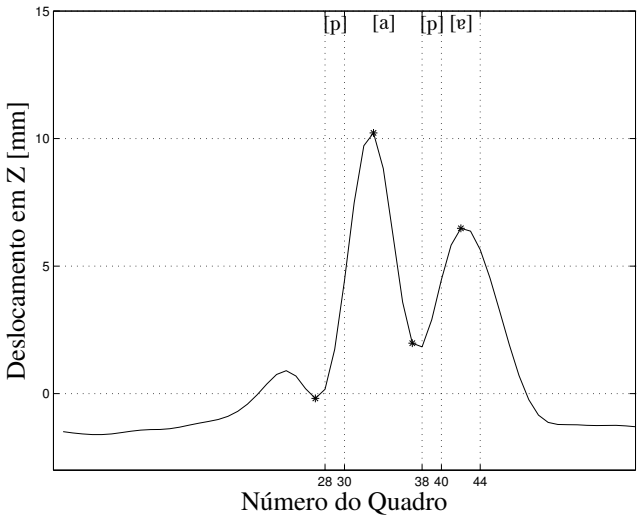

Fig. 5.10: Deslocamento na direção Z de P_4 no plano sagital durante a produção de [ˈpapɐ].

premissas são ainda mais fortalecidas ao se observar que foi possível identificar pontos estacionários nas trajetórias de todos os pontos de interesse e em todos os logatomas considerados. Mesmo para casos potencialmente menos nítidos como, por exemplo, no logatoma [ˈaɐ], composto pela seqüência de dois segmentos vocálicos abertos, foi possível identificar com clareza os pontos estacionários e, portanto, os alvos articulatórios. As Figuras 5.11 a 5.13 apresentam a trajetória do ponto P_4 e os pontos de estacionários determinados pelo método exposto acima para o logatoma [ˈaɐ].

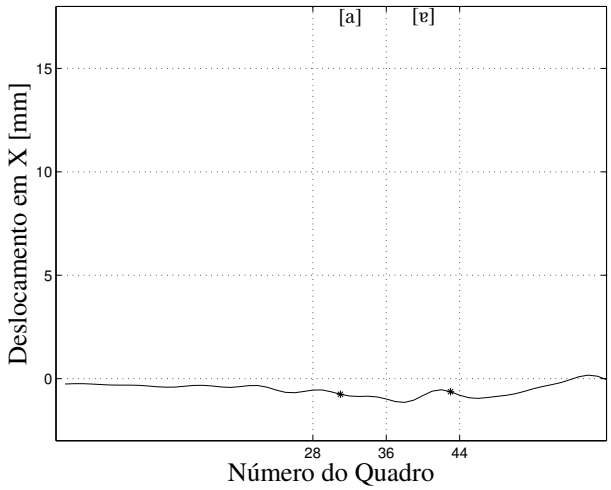

Fig. 5.11: Deslocamento na direção X de P_4 no plano sagital durante a produção de [ˈaɐ].

Para cada segmento foram tomados os valores dos deslocamentos a partir do repouso das coordenadas x, y e z nos pontos estacionários. Estes parâmetros foram utilizados para caracterizar a postura articulatória alvo e utilizados no agrupamento por similaridade efetuado e discutido a seguir.

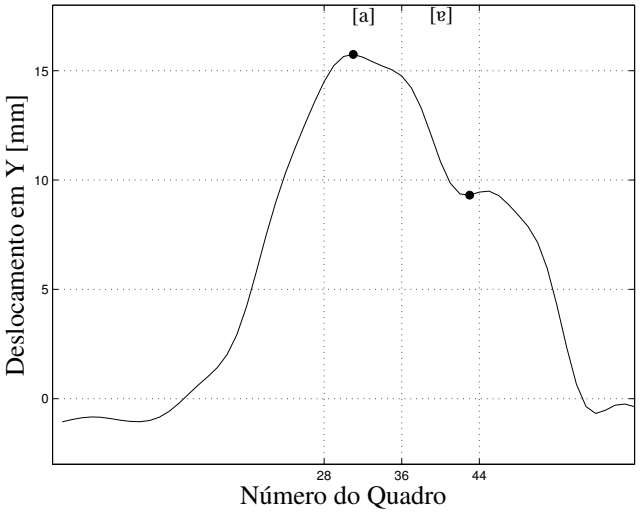

Fig. 5.12: Deslocamento na direção Y de P_4 no plano sagital durante a produção de [ˈaɐ].

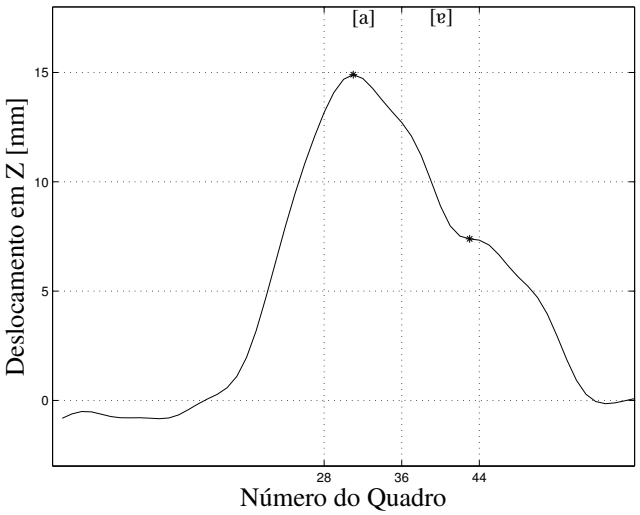

Fig. 5.13: Deslocamento na direção Z de P_4 no plano sagital durante a produção de [ˈaɐ].

5.2.6 Agrupamento dos alvos articulatórios

Conceitualmente, o agrupamento dos alvos articulatórios tem por objetivo detetar e agrupar por critérios de similaridade as posturas articulatórias alvo associadas à produção de um mesmo segmento em contextos fonéticos diferentes. É assumido que o conjunto de alvos articulatórios agrupados por critério de similaridade revela um padrão de movimentação articulatória similar, e que, portanto, é apropriado para a caracterização visual do segmento nos contextos fonéticos contemplados pelo agrupamento.

Assim, por exemplo, para o fone [f] (Tabela 5.10), foram considerados todos os logatomas do corpus que o continham: [f]V_1[f]V_2, com V_1 = {[i, a, u]} e V_2 = {[ɪ, ɐ, ʊ]}. Neste conjunto de 9 logatomas é possível identificar 18 produções do fone [f]. Considerando os efeitos da coarticulação adjacente é possível identificar 12 contextos distintos: [fi] (3 produções); [fa] (3 produções); [fu] (3 produções); [ifɪ]; [ifɐ]; [ifʊ]; [afɪ]; [afɐ]; [afʊ]; [ufɪ]; [ufɐ]; e [ufʊ]. Os alvos articulatórios, ou seja, os deslocamentos em X, Y, e Z nos pontos estacionários, de cada ponto de interesse associado às 18 produções, foram utilizados como parâmetros na análise de agrupamento por similaridade. A busca do agrupamento foi efetuado no espaço $\Re^{12}$, onde cada ponto deste espaço foi formado pelos alvos articulatórios dos quatro pontos de interesse - P_1 a P_4, resultando em dois grupos para o fone [f] (Tabela 5.10). Os outros segmentos e seus respectivos contextos fonéticos foram processados de forma análoga (Seção 5.3).

O procedimento de agrupamento por critério de similaridade adotado baseia-se no método *K-Means* (JAIN; DUBES, 1988). Este método, especificada a quantidade desejada de grupos, procura agrupar o conjunto de dados de entrada de maneira a minimizar o somatório do quadrado da distância euclidiana de cada ponto ao centróide do grupo (minimizar uma estimativa da variância intra-grupo). No processamento realizado, foram efetuadas 40 simulações a partir de posições iniciais diferentes. Em cada simulação foi efetuada a variação sistemática e exaustiva do número de grupos procurados, indo de 2 até o número total de contexto fonéticos diferentes considerados para o segmento. Dentre as soluções encontradas foram descartadas aquelas que alocaram em grupos diferentes realizações fonéticas consideradas iguais. Por exemplo, as três realizações de [fa] nos contextos [fafɐ], [fafɪ], e [fafʊ]. As diferenças nas produções do fonema /f/ nestes três contextos são tratadas como manifestações da variabilidade inerente da movimentação articulatória na produção de um mesmo fonema e que, por não alterar significativamente a informação visual, não devem ser tratadas como padrões articulatórios diferentes. Dentre os agrupamentos restantes, a solução final adotada foi o agrupamento com o menor índice de Davies-Bouldin (JAIN; DUBES, 1988). O Índice de Davies-Bouldin é uma medida da qualidade do agrupamento e estabelece uma relação entre a coesão intra-grupos e separação entre grupos. Menor o índice, melhor a qualidade do agrupamento.

O resultado do processo de agrupamento produz, para cada segmento, grupos de contextos fonéti-

cos que apresentam o mesmo padrão visual articulatório. O centróide do grupo foi considerado como o representante deste padrão articulatório e, portanto, associado a um visema.

5.2.7 Instante relativo de realização dos visemas

Para a identificação do instante de realização do alvo articulatório associado a um visema, adotou-se o instante percentual médio das realizações de todos alvos articulatórios pertencentes ao grupo que estabelece o visema. Para tanto, foi identificado, para cada realização do segmento fonético no grupo, o instante percentual de sua realização em relação ao intervalo da produção acústica do segmento. O instante de realização do alvo articulatório representante é dado pela média aritmética dos instantes relativos de realização de todos os segmentos do grupo de contextos fonéticos.

5.2.8 Representação paramétrica dos visemas

Com o conhecimento dos deslocamentos nas direções X, Y e Z associados aos alvos articulatórios de um visema e dos instantes de realização destes alvos, é possível aproximar a trajetória dos pontos de interesse em uma cadeia de fones pela interpolação dos deslocamentos dos alvos articulatórios. É importante que a curva de interpolação preserve a continuidade geométrica G^0 entre os vários segmentos de uma cadeia de visemas/fones e garanta derivada temporal igual a zero nos instantes de realização do alvo articulatório. Para atender a estas restrições adotou-se, para a representação da movimentação, uma curva paramétrica cúbica de Hermite (FOLEY et al., 1990). A equação 5.1 apresenta o modelo paramétrico adotado.

$$\begin{bmatrix} x(t) \\ y(t) \\ x(t) \end{bmatrix} = \begin{bmatrix} I_x & F_x \\ I_y & F_y \\ I_z & F_z \end{bmatrix} \begin{bmatrix} 2 & -3 & 0 & 1 \\ -2 & 3 & 0 & 0 \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix} \qquad 0 \leq t \leq 1 \tag{5.1}$$

onde:

- $x(t)$, $y(t)$ e $z(t)$ são as coordenadas do ponto de interesse;
- I_x, I_y e I_z são os deslocamentos nas direção X, Y, e Z do ponto de interesse no alvo articulatório do visema inicial;
- F_x, F_y e F_z são os deslocamentos nas direções X, Y e Z do ponto de interesse no alvo articulatório do visema final; e

- t é variável independente da representação paramétrica, normalizada em relação ao intervalo de tempo entre dois alvos articulatórios.

Para ilustrar o resultado do processo de interpolação, a Figura 5.14 apresenta as trajetórias medida e aproximada pelo modelo para a coordenada y do ponto P_4 durante a produção dos logatomas [ˈpapɐ] e [ˈaɐ].

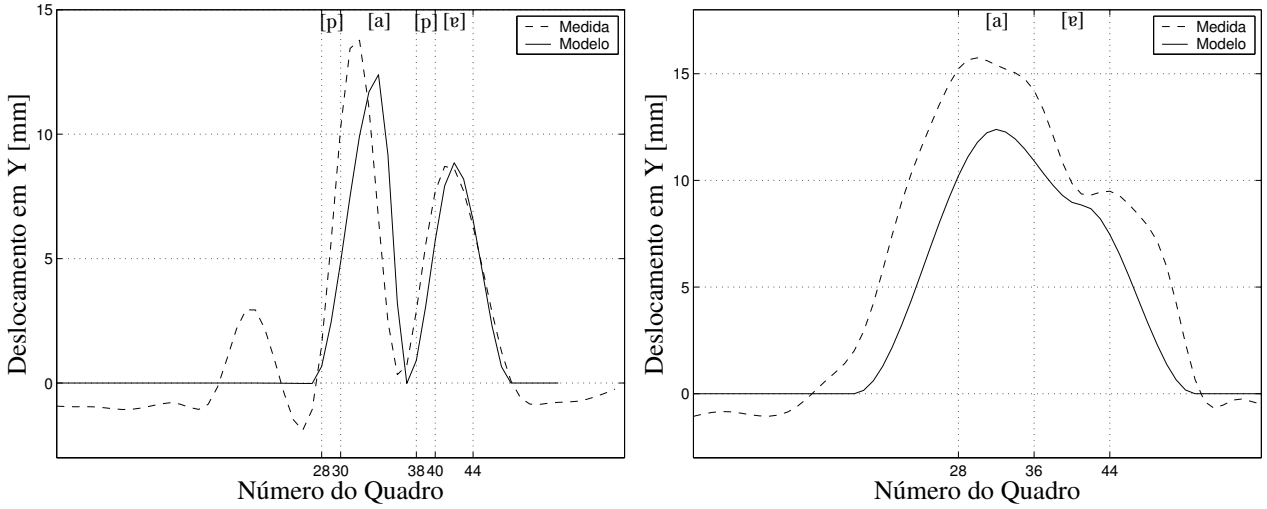

Fig. 5.14: Deslocamento na direção Y de ponto P_4: esquerda) logatoma [ˈpapɐ]; direita) logatoma [ˈaɐ].

5.3 Resultados

5.3.1 Estimativas da precisão das medidas

Para a caracterização da precisão das medidas efetuadas, simultaneamente ao cálculo das trajetórias dos pontos de interesse e utilizando o mesmo procedimento, foram efetuadas medidas de 27 pontos de controle com posições fixas e conhecidas a priori. A Figura 5.7 apresenta e identifica os pontos de controle utilizados. Tendo por base as medidas das posições dos pontos de controle são apresentadas a seguir estimativas dos erros máximos associados à medida de uma coordenada (x, y ou z) e à medida da posição tridimensional do ponto, considerando o intervalo de confiança de 95%, usualmente utilizado em análises estatísticas. O intervalo de confiança de 95% indica que um erro menor ou igual ao valor máximo estatisticamente ocorre em 95% das medidas, ou, em outras palavras, que a probabilidade de um erro maior ocorrer é de 0,05.

Para a estimativa do erro máximo associado à medida de uma coordenada, tomou-se o erro associado à medida de uma coordenada, denominado erro de coordenada, e_c, dado por:

$$e_c = v_m - v_r \tag{5.2}$$

onde v_m é o valor medido e v_r é o valor de referência conhecido a priori.

Para o erro de coordenada foram calculadas as seguintes estatísticas: valor médio μ, variância s^2 e a tolerância Δ do erro para um intervalo de confiança de 95%. A Tabela 5.4 apresenta os valores destes parâmetros para cada coordenada de cada ponto de controle.

Para o cálculo dos valores apresentados na Tabela 5.4 as seguintes relações foram utilizadas:

- Valor médio (μ):

$$\mu = \frac{1}{n}\sum_{i=1}^{n} e_i \tag{5.3}$$

- Variância amostral (s^2):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (e_i - \mu)^2 \tag{5.4}$$

- Tolerância para intervalo de confiança de 95% (Δ):

$$\Delta = t_{0,025}\sqrt{\frac{s^2}{n}} \tag{5.5}$$

onde n = 6724 é o número de medidas efetuadas e $t_{0,025} = 1,9604$ é o valor da distribuição de Student que deixa 2,5% da probabilidade na cauda superior, com n - 1 graus de liberdade. Valores no intervalo de confiança $\mu \pm \Delta$ possuem 95% de probabilidade de ocorrerem.

É possível observar na Tabela 5.4 que os valores máximos absolutos para o erro de coordenada são, respectivamente, para as coordenadas x, y e z: $0,8505$ (ponto P_8), $0,5411$ (ponto P_{25}) e $0,6238$ (Ponto P_5). Assim, ao se considerar o pior caso em que o máximo erro de coordenada seja igual a $0,9$ mm ($0,8505$ arredondado para uma casa decimal) é assegurado que no mínimo 95% das medidas apresentarão um erro de coordenada menor ou igual a este valor.

A Tabela 5.5 apresenta o erro de distância tridimensional, e_d, entre o vetor posição medido e o conhecido a priori. Na tabela são apresentados o erro de distância médio μ, a variância s^2 e a tolerância Δ deste erro para um intervalo de confiança de 95%. O erro de distância associado a uma dada medida é dado pela equação:

$$e_d = \sqrt{(x_m - x_r)^2 + (y_m - y_r)^2 + (z_m - z_r)^2} \tag{5.6}$$

onde (x_m, y_m, z_m) são os valores medidos das coordenadas e (x_r, y_r, z_r) são os valores de referência das coordenadas do ponto.

Ponto	Coordenada								
	x			y			z		
	μ [mm]	s^2 [mm^2]	Δ [mm]	μ [mm]	s^2 [mm^2]	Δ [mm]	μ [mm]	s^2 [mm^2]	Δ [mm]
P_5	-0,0525	0,0028	0,0013	0,3190	0,0022	0,0011	-0,6220	0,0057	0,0018
P_6	0,2881	0,0060	0,0018	0,4717	0,0079	0,0021	-0,4551	0,0163	0,0031
P_7	0,5726	0,0057	0,0018	-0,0117	0,0050	0,0017	-0,3295	0,0146	0,0029
P_8	0,8490	0,0037	0,0015	-0,1247	0,0038	0,0015	-0,2055	0,0088	0,0022
P_9	0,4760	0,0057	0,0018	-0,2400	0,0056	0,0018	-0,0959	0,0147	0,0029
P_{10}	0,2865	0,0062	0,0019	-0,1418	0,0066	0,0019	-0,0102	0,0173	0,0031
P_{11}	-0,0900	0,0054	0,0018	-0,1949	0,0051	0,0017	-0,0182	0,0150	0,0029
P_{12}	-0,1140	0,0026	0,0012	0,0471	0,0024	0,0012	0,0459	0,0059	0,0018
P_{13}	-0,3685	0,0044	0,0016	-0,1732	0,0074	0,0021	0,0988	0,0109	0,0025
P_{14}	-0,1623	0,0042	0,0016	-0,3059	0,0073	0,0020	0,1878	0,0105	0,0024
P_{15}	0,1105	0,0042	0,0016	0,0170	0,0069	0,0020	0,2160	0,0101	0,0024
P_{16}	0,1826	0,0084	0,0022	0,2766	0,0059	0,0018	0,1956	0,0211	0,0035
P_{17}	0,0966	0,0061	0,0019	0,0053	0,0075	0,0021	0,2040	0,0131	0,0027
P_{18}	-0,2423	0,0024	0,0012	0,1318	0,0041	0,0015	0,1611	0,0061	0,0019
P_{19}	-0,0124	0,0079	0,0021	-0,0601	0,0043	0,0016	0,1159	0,0186	0,0033
P_{20}	-0,0432	0,0051	0,0017	0,1733	0,0044	0,0016	0,0869	0,0116	0,0026
P_{21}	0,1729	0,0054	0,0018	0,2035	0,0042	0,0016	-0,0134	0,0139	0,0028
P_{22}	0,3287	0,0066	0,0019	0,1833	0,0039	0,0015	-0,1182	0,0145	0,0029
P_{23}	0,2035	0,0054	0,0018	0,0448	0,0045	0,0016	-0,2305	0,0126	0,0027
P_{24}	0,2556	0,0027	0,0013	0,0522	0,0020	0,0011	-0,4009	0,0067	0,0020
P_{25}	-0,2095	0,0071	0,0020	-0,5397	0,0034	0,0014	-0,3153	0,0138	0,0028
P_{26}	0,0335	0,0057	0,0018	-0,2644	0,0038	0,0015	-0,1048	0,0140	0,0028
P_{27}	-0,3961	0,0053	0,0017	-0,1896	0,0033	0,0014	0,0528	0,0125	0,0027
P_{28}	-0,3391	0,0033	0,0014	-0,0189	0,0021	0,0011	0,2173	0,0079	0,0021
P_{29}	-0,0490	0,0045	0,0016	-0,0779	0,0035	0,0014	0,2938	0,0105	0,0024
P_{30}	-0,1380	0,0049	0,0017	-0,4277	0,0049	0,0017	0,4114	0,0119	0,0026
P_{31}	0,3322	0,0029	0,0013	-0,0305	0,0012	0,0008	0,3994	0,0078	0,0021

Tab. 5.4: Estatística do erro de coordenada dos pontos de controle P_5 a P_{31}.

Ponto	Distância		
	μ	$\mathbf{s}^2$	$\mathbf{\Delta}$
$\mathbf{P}_5$	0,7053	0,0047	0,0016
$\mathbf{P}_6$	0,7300	0,0100	0,0024
$\mathbf{P}_7$	0,6751	0,0062	0,0019
$\mathbf{P}_8$	0,8895	0,0038	0,0015
$\mathbf{P}_9$	0,5600	0,0057	0,0018
$\mathbf{P}_{10}$	0,3540	0,0071	0,0020
$\mathbf{P}_{11}$	0,2583	0,0052	0,0017
$\mathbf{P}_{12}$	0,1587	0,0030	0,0013
$\mathbf{P}_{13}$	0,4398	0,0049	0,0017
$\mathbf{P}_{14}$	0,4126	0,0070	0,0020
$\mathbf{P}_{15}$	0,2701	0,0075	0,0021
$\mathbf{P}_{16}$	0,4189	0,0081	0,0021
$\mathbf{P}_{17}$	0,2626	0,0086	0,0022
$\mathbf{P}_{18}$	0,3331	0,0037	0,0014
$\mathbf{P}_{19}$	0,1979	0,0089	0,0023
$\mathbf{P}_{20}$	0,2350	0,0054	0,0018
$\mathbf{P}_{21}$	0,3008	0,0045	0,0016
$\mathbf{P}_{22}$	0,4155	0,0080	0,0021
$\mathbf{P}_{23}$	0,3317	0,0091	0,0023
$\mathbf{P}_{24}$	0,4859	0,0041	0,0015
$\mathbf{P}_{25}$	0,6737	0,0050	0,0017
$\mathbf{P}_{26}$	0,3175	0,0047	0,0016
$\mathbf{P}_{27}$	0,4601	0,0050	0,0017
$\mathbf{P}_{28}$	0,4150	0,0035	0,0014
$\mathbf{P}_{29}$	0,3228	0,0091	0,0023
$\mathbf{P}_{30}$	0,6200	0,0086	0,0022
$\mathbf{P}_{31}$	0,5241	0,0081	0,0021

Tab. 5.5: Estatística do erro de distância dos pontos P_5 a P_{31}.

Para o cálculo dos valores apresentados na Tabela 5.5 foram utilizadas as definições expressas pelas Equações 5.3, 5.4 e 5.5.

A análise dos valores do erro de distância apresentados na Tabela 5.5 revela que o máximo erro de distância é de $0,8909$ mm (ponto P_8) para um intervalo de confiança de 95%. Assim, um valor máximo de erro de distância igual a $0,9$ mm assegura que 95% do erro de distância serão menores ou igual a este valor.

A análise do erro de coordenada conjuntamente com a do erro de distância permite assumir que, dentro de um intervalo de confiança de 95%, o erro das medidas realizadas é menor ou igual a $0,9$ mm para ambos os erros. Assim, ao se avaliar o erro de medida associado a uma coordenada, o valor $0,9$ mm deve ser considerado para um intervalo de confiança de 95%. Este mesmo valor deve ser considerado nas avaliações envolvendo a posição tridimensional dos pontos.

5.3.2 Posição de repouso

A posição de repouso é a posição em que a boca encontra-se fechada e os dentes cerrados. O falante foi instruído a realizar esta postura antes e depois da locução de cada logatoma. Após as medidas das trajetórias, a posição de repouso de cada ponto de interesse, P_1 a P_4, foi calculada pela média dos valores dos 10 quadros iniciais e dos 10 quadros finais da produção de todos os logatomas. A Tabela 5.6 apresenta os valores das coordenadas x, y e z dos pontos de interesse na condição de repouso.

Ponto	**x** [mm]	**y** [mm]	**z** [mm]
$\mathbf{P}_1$	121,1	113,0	2,6
$\mathbf{P}_2$	150,0	121,2	23,2
$\mathbf{P}_3$	120,0	131,0	6,9
$\mathbf{P}_4$	118,8	152,6	15,4

Tab. 5.6: Posição de repouso dos pontos de interesse.

5.3.3 Visemas consonantais

A Tabela 5.7 apresenta os 22 visemas consonantais e respectivos contextos fonéticos determinados pela aplicação da metodologia descrita na Seção 5.2. A primeira coluna da tabela apresenta a simbologia adotada para a representação dos visemas. Neste trabalho, os visemas são representados

por símbolos IPA entre "< >". Na representação de um visema é utilizado o símbolo do segmento analisado para a sua identificação. Os vários visemas associados a contextos diferentes de um mesmo segmento são identificados por um índice numérico. Na segunda coluna da Tabela 5.7 têm-se os contextos fonéticos associados a cada visema. A terceira coluna apresenta as denominações dadas aos visemas. A quarta e última apresenta o conjunto de fones associado a cada visema.

Símbolo	Contextos	Denominação	Homofemas
<p_1>	[pi] [pa] [ipɪ] [ipɐ] [ipʊ] [apɪ] [apɐ] [apʊ] [upɐ]	Bilabiais	[p, b, m]
<p_2>	[pu] [upɪ] [upʊ]		
<f_1>	[fi] [fa] [ifɪ] [ifɐ] [ifʊ] [afɪ] [afɐ]	Labiodentais	[f, v]
<f_2>	[fu] [afʊ] [ufɪ] [ufɐ] [ufʊ]		
<t_1>	[ti] [tu] [itɪ] [itɐ] [itʊ] [atɪ] [atʊ] [utɪ] [utɐ] [utʊ]	Alveolares Plosivos/Nasais	[t, d, n]
<t_2>	[ta] [atɐ]		
<s_1>	[si] [sa] [isɪ] [isɐ] [asɪ] [asɐ]	Alveolares Fricativos	[s, z]
<s_2>	[su] [isʊ] [asʊ] [usɪ] [usɐ] [usʊ]		
<l_1>	[li] [ilɪ] [alʊ] [ulɪ] [ulɐ]	Alveolares Laterais	[l]
<l_2>	[la] [ilɐ] [alɪ] [alɐ]		
<l_3>	[lu]		
<l_4>	[ilʊ] [ulʊ]		
<$ʃ_1$>	[ʃi] [ʃa] [iʃɪ] [iʃɐ] [iʃʊ] [aʃɪ] [aʃɐ] [aʃʊ] [uʃɪ] [uʃɐ]	Pós-Alveolares	[ʃ, ʒ]
<$ʃ_2$>	[ʃu] [uʃʊ]		
<$ʎ_1$>	[ʎi] [ʎa] [iʎɪ] [iʎɐ] [aʎɪ] [aʎɐ]	Palatais	[ʎ, ɲ]
<$ʎ_2$>	[ʎu] [uʎɪ] [uʎɐ]		
<$ʎ_3$>	[iʎʊ] [aʎʊ] [uʎʊ]		
<k_1>	[ki] [ikɪ] [ikɐ] [akɪ] [ukɪ] [ukɐ]	Velares Plosivos	[k, g]
<k_2>	[ka] [akɐ]		
<k_3>	[ku] [ikʊ] [akʊ] [ukʊ]		
<$ɾ_1$>	[ɣi] [ɣa] [iɾɪ] iɾɐ] [aɾɪ] [aɾɐ] [uɾɐ]	Velares Fricativos/Tepes	[ɣ], [ɾ]
<$ɾ_2$>	[ɣʊ] [iɾʊ [aɾʊ] [uɾɪ] [uɾʊ]		

Tab. 5.7: Visemas consonantais e respectivos contextos fonéticos.

Nas Subseções 5.3.3.1 a 5.3.3.9 são apresentados os resultados do processo de agrupamento em visemas e os valores numéricos dos alvos articulatórios associados a cada um dos 22 visemas consonantais. Em cada subseção são apresentadas duas tabelas. A primeira apresenta o resultado do processo de agrupamento e o índice de Davies-Bouldin associado à solução. Na segunda tabela são apresentados os alvos articulatórios (centróides) representantes dos grupos encontrados e o valor do instante relativo de realização destes alvos. Os valores dos alvos articulatórios são deslocamentos

relativos à posição de repouso.

5.3.3.1 Visemas bilabiais

Logatoma	Contexto [p]V		Contexto V[p]V	
	Seqüência	Grupo	Seqüência	Grupo
[pipɪ]	[pi]	1	[ipɪ]	1
[pipɐ]	[pi]	1	[ipɐ]	1
[pipʊ]	[pi]	1	[ipʊ]	1
[papɪ]	[pa]	1	[apɪ]	1
[papɐ]	[pa]	1	[apɐ]	1
[papʊ]	[pa]	1	[apʊ]	1
[pupɪ]	[pu]	2	[upɪ]	2
[pupɐ]	[pu]	2	[upɐ]	1
[pupʊ]	[pu]	2	[upʊ]	2
Índice Davies-Bouldin: 0,96				

Tab. 5.8: Agrupamento em visemas das diferentes produções do fone [p].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-1,0	6,1	1,0	-0,03
	P2	-1,5	2,8	0,8	-0,07
	P3	0,4	0,3	-1,6	-0,14
	P4	0,5	0,0	2,0	-0,14
Grupo 2	**P1**	-1,6	4,6	-0,8	-0,84
	P2	-2,9	2,7	-5,1	-0,14
	P3	0,4	1,5	-2,6	-0,86
	P4	1,0	-0,1	1,4	-0,56

Tab. 5.9: Alvos Articulatórios dos visemas bilabiais.

5.3.3.2 Visemas labiodentais

Logatoma	Contexto [f]V		Contexto V[f]V	
	Seqüência	Grupo	Seqüência	Grupo
[**fifɪ**]	[fi]	1	[ifɪ]	1
[**fifɐ**]	[fi]	1	[ifɐ]	1
[**fifʊ**]	[fi]	1	[ifʊ]	1
[**fafɪ**]	[fa]	1	[afɪ]	1
[**fafɐ**]	[fa]	1	[afɐ]	1
[**fafʊ**]	[fa]	1	[afʊ]	2
[**fufɪ**]	[fu]	2	[ufɪ]	2
[**fufɐ**]	[fu]	2	[ufɐ]	2
[**fufʊ**]	[fu]	2	[ufʊ]	2
Índice Davies-Bouldin: 0,76				

Tab. 5.10: Agrupamento em visemas das diferentes produções do fone [f].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	0,4	-3,6	-1,3	0,67
	P2	-0,6	0,2	-1,3	0,58
	P3	0,4	0,1	2,5	0,50
	P4	-0,2	1,3	3,9	0,42
Grupo 2	**P1**	0,4	-4,4	-4,4	0,59
	P2	-2,9	3,1	-6,0	0,69
	P3	0,1	1,8	0,1	0,40
	P4	0,0	1,8	4,7	0,36

Tab. 5.11: Alvos Articulatórios dos visemas labiodentais.

5.3.3.3 Visemas alveolares plosivos/nasais

Logatoma	Contexto [t]V		Contexto V[t]V	
	Seqüência	Grupo	Seqüência	Grupo
[titɪ]	[ti]	1	[itɪ]	1
[titɐ]	[ti]	1	[itɐ]	1
[titʊ]	[ti]	1	[itʊ]	1
[tatɪ]	[ta]	2	[atɪ]	1
[tatɐ]	[ta]	2	[atɐ]	2
[tatʊ]	[ta]	2	[atʊ]	1
[tutɪ]	[tu]	1	[utɪ]	1
[tutɐ]	[tu]	1	[utɐ]	1
[tutʊ]	[tu]	1	[utʊ]	1
Índice Davies-Bouldin: 0,57				

Tab. 5.12: Agrupamento em visemas das diferentes produções do fone [t].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,3	-4,7	-3,2	0,70
	P2	-1,9	1,8	-5,3	0,72
	P3	-0,9	11,1	3,7	0,62
	P4	-0,2	3,8	2,0	0,67
Grupo 2	**P1**	-0,4	-2,1	1,8	0,16
	P2	2,4	-2,9	6,0	-0,53
	P3	-0,8	7,8	7,5	-0,41
	P4	0,3	5,2	2,7	-0,66

Tab. 5.13: Alvos articulatórios dos visemas alveolares plosivos/nasais.

5.3.3.4 Visemas alveolares fricativos

Logatoma	Contexto [s]V		Contexto V[s]V	
	Seqüência	Grupo	Seqüência	Grupo
[sisɪ]	[si]	1	[isɪ]	1
[sisɐ]	[si]	1	[isɐ]	1
[sisʊ]	[si]	1	[isʊ]	2
[sasɪ]	[sa]	1	[asɪ]	1
[sasɐ]	[sa]	1	[asɐ]	1
[sasʊ]	[sa]	1	[asʊ]	2
[susɪ]	[su]	2	[usɪ]	2
[susɐ]	[su]	2	[usɐ]	2
[susʊ]	[su]	2	[usʊ]	2
Índice Davies-Bouldin: 0,63				

Tab. 5.14: Agrupamento em visemas das diferentes produções do fone [s].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,2	-4,4	1,5	0,57
	P2	0,4	-1,7	2,5	0,47
	P3	-0,7	3,7	3,1	0,43
	P4	0,0	2,7	1,5	0,43
Grupo 2	**P1**	-0,5	-6,0	-5,3	0,67
	P2	-2,9	2,5	-7,5	0,72
	P3	-1,2	6,9	-0,3	0,66
	P4	-0,1	1,6	1,9	0,62

Tab. 5.15: Alvos articulatórios dos visemas alveolares fricativos.

5.3.3.5 Visemas alveolares laterais

Logatoma	Contexto [l]V		Contexto V[l]V	
	Seqüência	Grupo	Seqüência	Grupo
[lilɪ]	[li]	1	[ilɪ]	1
[lilɐ]	[li]	1	[ilɐ]	2
[lilʊ]	[li]	1	[ilʊ]	4
[lalɪ]	[la]	2	[alɪ]	2
[lalɐ]	[la]	2	[alɐ]	2
[lalʊ]	[la]	2	[alʊ]	1
[lulɪ]	[lu]	3	[ulɪ]	1
[lulɐ]	[lu]	3	[ulɐ]	1
[lulʊ]	[lu]	3	[ulʊ]	4
Índice Davies-Bouldin: 0,49				

Tab. 5.16: Agrupamento em visemas das diferentes produções do fone [l].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,8	-2,2	-0,3	0,52
	P2	-0,8	2,6	0,9	0,57
	P3	-1,8	10,9	7,3	0,74
	P4	-0,3	6,2	3,5	0,54
Grupo 2	**P1**	-0,4	-1,9	2,9	0,85
	P2	2,9	-1,6	9,5	0,61
	P3	-1,2	10,3	11,7	0,62
	P4	0,2	7,1	6,6	0,59
Grupo 3	**P1**	0,0	-5,3	-6,1	0,57
	P2	-3,5	4,1	-6,3	0,85
	P3	-1,4	15,0	4,8	0,52
	P4	-0,6	5,4	3,9	0,63
Grupo 4	**P1**	0,5	-1,6	-3,2	0,31
	P2	-1,3	2,3	-5,3	0,57
	P3	-0,3	7,0	0,4	0,64
	P4	0,0	1,1	1,4	0,53

Tab. 5.17: Alvos articulatórios dos visemas alveolares laterais.

5.3.3.6 Visemas pós-alveolares

Logatoma	Contexto [ʃ]V		Contexto V[ʃ]V	
	Seqüência	Grupo	Seqüência	Grupo
[ʃiʃɪ]	[ʃi]	1	[iʃɪ]	1
[ʃiʃɐ]	[ʃi]	1	[iʃɐ]	1
[ʃiʃʊ]	[ʃi]	1	[iʃʊ]	1
[ʃaʃɪ]	[ʃa]	1	[aʃɪ]	1
[ʃaʃɐ]	[ʃa]	1	[aʃɐ]	1
[ʃaʃʊ]	[ʃa]	1	[aʃʊ]	1
[ʃuʃɪ]	[ʃu]	2	[uʃɪ]	1
[ʃuʃɐ]	[ʃu]	2	[uʃɐ]	1
[ʃuʃʊ]	[ʃu]	2	[uʃʊ]	2
Índice Davies-Bouldin: 1,05				

Tab. 5.18: Agrupamento em visemas das diferentes produções do fone [ʃ].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,6	-6,9	-3,0	0,77
	P2	-1,3	0,9	-3,6	0,60
	P3	-2,0	12,7	3,9	0,60
	P4	-0,6	3,7	0,3	0,55
Grupo 2	**P1**	-1,2	-7,7	-6,2	0,85
	P2	-4,2	3,5	-9,2	0,69
	P3	-2,5	11,5	0,9	0,60
	P4	-0,7	3,2	0,9	0,58

Tab. 5.19: Alvos articulatórios dos visemas pós-alveolares.

5.3.3.7 Visemas palatais

Logatoma	Contexto [λ/V		Contexto V[λ/V	
	Seqüência	Grupo	Seqüência	Grupo
[λiλɪ]	[λi]	1	[iλɪ]	1
[λiλɐ]	[λi]	1	[iλɐ]	1
[λiλʊ]	[λi]	1	[iλʊ]	3
[λaλɪ]	[λa]	1	[aλɪ]	1
[λaλɐ]	[λa]	1	[aλɐ]	1
[λaλʊ]	[λa]	1	[aλʊ]	3
[λuλɪ]	[λu]	2	[uλɪ]	2
[λuλɐ]	[λu]	2	[uλɐ]	2
[λuλʊ]	[λu]	2	[uλʊ]	3
Índice Davies-Bouldin: 0,60				

Tab. 5.20: Agrupamento em visemas das diferentes produções do fone [λ].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,9	-2,5	0,8	0,51
	P2	0,7	-0,1	3,5	0,41
	P3	-0,9	8,0	5,9	0,36
	P4	0,1	4,7	1,1	0,34
Grupo 2	**P1**	-0,5	-5,6	-4,4	0,63
	P2	-1,6	3,0	-4,0	0,44
	P3	-1,4	13,5	5,4	0,29
	P4	-0,3	4,9	1,7	0,81
Grupo 3	**P1**	-0,9	-4,7	-3,7	0,46
	P2	-2,4	2,3	-7,1	0,38
	P3	-2,1	8,7	-0,4	0,58
	P4	-0,8	1,3	0,6	0,58

Tab. 5.21: Alvos articulatórios dos visemas palatais.

5.3.3.8 Visemas velares plosivos

Logatoma	Contexto [k]V		Contexto V[k]V	
	Seqüência	Grupo	Seqüência	Grupo
[kikɪ]	[ki]	1	[ikɪ]	1
[kikɐ]	[ki]	1	[ikɐ]	1
[kikʊ]	[ki]	1	[ikʊ]	3
[kakɪ]	[ka]	2	[akɪ]	1
[kakɐ]	[ka]	2	[akɐ]	2
[kakʊ]	[ka]	2	[akʊ]	3
[kukɪ]	[ku]	3	[ukɪ]	1
[kukɐ]	[ku]	3	[ukɐ]	1
[kukʊ]	[ku]	3	[ukʊ]	3
Índice Davies-Bouldin: 0,49				

Tab. 5.22: Agrupamento em visemas das diferentes produções do fone [k].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,4	-3,0	0,2	0,65
	P2	-0,6	1,8	2,9	0,79
	P3	-0,8	12,8	9,2	0,74
	P4	0,1	8,6	5,0	0,66
Grupo 2	**P1**	-0,5	-3,1	3,0	0,67
	P2	4,1	-3,1	11,0	0,75
	P3	-2,0	12,5	13,2	0,92
	P4	-0,8	10,5	7,6	0,93
Grupo 3	**P1**	0,0	-2,9	-4,9	0,58
	P2	-3,8	3,1	-7,7	0,73
	P3	-0,9	8,2	0,7	0,69
	P4	-0,5	3,2	3,2	0,56

Tab. 5.23: Alvos articulatórios dos visemas velares plosivos.

5.3.3.9 Velares fricativos/tepes

Logatoma	Contexto [ɣ]V		Contexto V[ɾ]V	
	Seqüência	Grupo	Seqüência	Grupo
[ɣiɾɪ]	[ɣi]	1	[iɾɪ]	1
[ɣiɾɐ]	[ɣi]	1	[iɾɐ]	1
[ɣiɾʊ]	[ɣi]	1	[iɾʊ]	2
[ɣaɾɪ]	[ɣa]	1	[aɾɪ]	1
[ɣaɾɐ]	[ɣa]	1	[aɾɐ]	1
[ɣaɾʊ]	[ɣa]	1	[aɾʊ]	2
[ɣuɾɪ]	[ɣu]	2	[uɾɪ]	2
[ɣuɾɐ]	[ɣu]	2	[uɾɐ]	1
[ɣuɾʊ]	[ɣu]	2	[uɾʊ]	2
Índice Davies-Bouldin: 0,68				

Tab. 5.24: Agrupamento em visemas das diferentes produções dos fones [ɣ] e [ɾ].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,4	-1,6	1,7	0,59
	P2	0,9	0,3	6,5	0,72
	P3	-1,2	12,5	10,8	0,71
	P4	0,2	9,2	6,4	0,71
Grupo 2	**P1**	0,0	-2,6	-4,6	0,72
	P2	-3,3	3,7	-6,3	0,59
	P3	-1,2	10,6	2,2	0,45
	P4	-0,7	4,3	4,1	0,39

Tab. 5.25: Alvos articulatórios dos visemas velares fricativos/tepes.

5.3.4 Visemas vocálicos

A Tabela 5.26 apresenta os 7 visemas vocálicos e respectivos contextos fonéticos identificados. Esta tabela, organizada da mesma maneira que a Tabela 5.7, apresenta na primeira coluna o símbolo do visema. Na segunda têm-se os contextos fonéticos associados aos visema. A terceira coluna apresenta as denominações destes. E a quarta e última coluna apresenta o conjunto de fones associado aos visemas.

Símbolo	Contextos	Denominação	Homofemas
<i_1>	todos os contextos exceto [tit] e [ʃiʃ]	Alto Anterior	[i, ĩ]
<i_2>	[tit] [ʃiʃ]		
<a>	todos os contextos	Baixo Central	[a, ɛ, ɐ̃]
<u>	todos os contextos	Alto Posterior	[u, o, õ, ũ]
<ɪ>	todos os contextos	Postônico Alto Anterior	[ɪ]
<ɐ>	todos os contextos	Postônico Baixo Central	[ɐ, e, ɔ, ẽ]
<ʊ>	todos os contextos	Postônico Alto Posterior	[ʊ]

Tab. 5.26: Visemas vocálicos e respectivos contextos fonéticos.

Nas Subseções 5.3.4.1 a 5.3.4.6 são apresentados os resultados do processo de agrupamento em visemas e os valores numéricos das posturas articulatórias alvo associadas a cada um dos 7 visemas vocálicos. Em cada subseção são apresentadas duas tabelas. A primeira apresenta o resultado do processo de agrupamento. Em todos os casos, exceto para o fone [i], a solução do processo de agrupamento levou a apenas um grupo. Para o caso especial do fone [i], a tabela apresenta também o índice de Davies-Bouldin associado à solução. Na segunda tabela são apresentados os alvos articulatórios (centróides) representantes dos grupos encontrados e o valor do instante relativo de realização destes alvos.

Nas Subseções 5.3.4.7 a 5.3.4.10 são apresentados os dados numéricos associados aos fones [e, ɛ, ɔ, o]. Em cada uma destas subseções também têm-se duas tabelas. A primeira apresenta o centróide das posturas articulatórias alvo das realizações do fone. Na segunda tabela são apresentadas as distâncias euclidianas deste centróide a cada um dos 7 visemas vocálicos. O inverso desta distância foi utilizado como critério de similaridade para a associação do fone a um dos visemas vocálicos. Esta avaliação de similaridade permitiu a clara associação dos fones [e, ɛ, o] a um dos visemas. Já o fone [ɔ] apresentou similaridades aproximadamente iguais em relação aos visemas <i_1>, <i_2> e <ɐ>. Neste caso, optou-se pela associação ao visema <ɐ> para permitir a associação do fone a um único visema válido para todos os contextos.

5.3.4.1 Visemas altos anteriores

Logatoma	Contexto C[i]C	
	Seqüência	Grupo
[**pipɪ**]	[pip]	1
[**pipɐ**]	[pip]	1
[**pipʊ**]	[pip]	1
[**fifɪ**]	[fif]	1
[**fifɐ**]	[fif]	1
[**fifʊ**]	[fif]	1
[**titɪ**]	[tit]	2
[**titɐ**]	[tit]	2
[**titʊ**]	[tit]	2
[**sisɪ**]	[sis]	1
[**sisɐ**]	[sis]	1
[**sisʊ**]	[sis]	1
[**lilɪ**]	[lil]	1
[**lilɐ**]	[lil]	1
[**lilʊ**]	[lil]	1
[**ʃiʃɪ**]	[ʃiʃ]	2
[**ʃiʃɐ**]	[ʃiʃ]	2
[**ʃiʃʊ**]	[ʃiʃ]	2

Logatoma	Contexto C[i]C	
	Seqüência	Grupo
[**λiλɪ**]	[λi λ]	1
[**λiλɐ**]	[λi λ]	1
[**λiλʊ**]	[λi λ]	1
[**kikɪ**]	[kik]	1
[**kikɐ**]	[kik]	1
[**kikʊ**]	[kik]	1
[**ɣirɪ**]	[ɣiɾ]	1
[**ɣirɐ**]	[ɣiɾ]	1
[**ɣirʊ**]	[ɣiɾ]	1

Logatoma	Contexto [i]V	
	Seqüência	Grupo
[**iɪ**]	[iɪ]	1
[**iɐ**]	[iɐ]	1
[**iʊ**]	[iʊ]	1

Índice Davies-Bouldin: 0,94

Tab. 5.27: Agrupamento em visemas das diferentes produções do fone [i].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,6	-3,5	-0,4	0,50
	P2	-0,5	1,6	2,5	0,49
	P3	-0,9	12,4	8,3	0,53
	P4	0,7	8,2	4,1	0,48
Grupo 2	**P1**	-0,4	-6,2	-3,0	0,71
	P2	-1,2	1,3	-2,8	0,55
	P3	-2,3	16,1	7,0	0,46
	P4	-0,4	5,8	2,1	0,38

Tab. 5.28: Alvos articulatórios dos visemas altos anteriores.

5.3.4.2 Visema baixo central

Logatoma	Contexto C[a]C	
	Seqüência	Grupo
[**papɪ**]	[pap]	1
[**papɐ**]	[pap]	1
[**papʊ**]	[pap]	1
[**fafɪ**]	[faf]	1
[**fafɐ**]	[faf]	1
[**fafʊ**]	[faf]	1
[**tatɪ**]	[tat]	1
[**tatɐ**]	[tat]	1
[**tatʊ**]	[tat]	1
[**sasɪ**]	[sas]	1
[**sasɐ**]	[sas]	1
[**sasʊ**]	[sas]	1
[**lalɪ**]	[lal]	1
[**lalɐ**]	[lal]	1
[**lalʊ**]	[lal]	1
[**ʃaʃɪ**]	[ʃaʃ]	1
[**ʃaʃɐ**]	[ʃaʃ]	1
[**ʃaʃʊ**]	[ʃaʃ]	1

Logatoma	Contexto C[a]C	
	Seqüência	Grupo
[**λaλɪ**]	[λa λ]	1
[**λaλɐ**]	[λa λ]	1
[**λaλʊ**]	[λa λ]	1
[**kakɪ**]	[kak]	1
[**kakɐ**]	[kak]	1
[**kakʊ**]	[kak]	1
[**ɣarɪ**]	[ɣaɾ]	1
[**ɣarɐ**]	[ɣaɾ]	1
[**ɣarʊ**]	[ɣaɾ]	1

Logatoma	Contexto [a]V	
	Seqüência	Grupo
[**aɪ**]	[aɪ]	1
[**aɐ**]	[aɐ]	1
[**aʊ**]	[aʊ]	1

Tab. 5.29: Agrupamento em visemas das diferentes produções do fone [a].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,6	-3,0	2,2	0,48
	P2	2,3	-0,6	10,4	0,45
	P3	-2,3	15,5	13,5	0,49
	P4	-0,6	12,4	9,1	0,46

Tab. 5.30: Alvo articulatório do visema baixo central.

5.3.4.3 Visema alto posterior

Logatoma	Contexto C[u]C	
	Seqüência	Grupo
[pupɪ]	[pup]	1
[pupɐ]	[pup]	1
[pupʊ]	[pup]	1
[fufɪ]	[fuf]	1
[fufɐ]	[fuf]	1
[fufʊ]	[fuf]	1
[tutɪ]	[tut]	1
[tutɐ]	[tut]	1
[tutʊ]	[tut]	1
[susɪ]	[sus]	1
[susɐ]	[sus]	1
[susʊ]	[sus]	1
[lulɪ]	[lul]	1
[lulɐ]	[lul]	1
[lulʊ]	[lul]	1
[ʃuʃɪ]	[ʃuʃ]	1
[ʃuʃɐ]	[ʃaʃ]	1
[ʃuʃʊ]	[ʃaʃ]	1

Logatoma	Contexto C[u]C	
	Seqüência	Grupo
[ʎuʎɪ]	[ʎa ʎ]	1
[ʎuʎɐ]	[ʎa ʎ]	1
[ʎuʎʊ]	[ʎa ʎ]	1
[kukɪ]	[kak]	1
[kukɐ]	[kak]	1
[kukʊ]	[kak]	1
[ɣurɪ]	[ɣar]	1
[ɣurɐ]	[ɣar]	1
[ɣurʊ]	[ɣar]	1

Logatoma	Contexto [u]V	
	Seqüência	Grupo
[uɪ]	[uɪ]	1
[uɐ]	[uɐ]	1
[uʊ]	[uʊ]	1

Tab. 5.31: Agrupamento em visemas das diferentes produções do fone [u].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,2	-3,4	-5,4	0,48
	P2	-4,5	4,1	-8,2	0,53
	P3	-1,0	11,3	2,2	0,53
	P4	-0,4	4,0	4,2	0,50

Tab. 5.32: Alvo articulatório do visema alto posterior.

5.3.4.4 Visema postônico alto anterior

Logatoma	Contexto C[ɪ]	
	Seqüência	Grupo
[pipɪ]	[pɪ]	1
[papɪ]	[pɪ]	1
[pupɪ]	[pɪ]	1
[fifɪ]	[fɪ]	1
[fafɪ]	[fɪ]	1
[fufɪ]	[fɪ]	1
[titɪ]	[tɪ]	1
[tatɪ]	[tɪ]	1
[tutɪ]	[tɪ]	1
[sisɪ]	[sɪ]	1
[sasɪ]	[sɪ]	1
[susɪ]	[sɪ]	1
[lilɪ]	[lɪ]	1
[lalɪ]	[lɪ]	1
[lulɪ]	[lɪ]	1
[ʃiʃɪ]	[ʃɪ]	1
[ʃaʃɪ]	[ʃɪ]	1
[ʃuʃɪ]	[ʃɪ]	1
[λiλɪ]	[λɪ]	1
[λaλɪ]	[λɪ]	1
[λuλɪ]	[λɪ]	1

Logatoma	Contexto C[ɪ]	
	Seqüência	Grupo
[kakɪ]	[kɪ]	1
[kikɪ]	[kɪ]	1
[kukɪ]	[kɪ]	1
[ɣiɾɪ]	[ɾɪ]	1
[ɣaɾɪ]	[ɾɪ]	1
[ɣuɾɪ]	[ɾɪ]	1

Logatoma	Contexto V[ɪ]	
	Seqüência	Grupo
[iɪ]	[iɪ]	1
[eɪ]	[eɪ]	1
[ɛɪ]	[ɛɪ]	1
[aɪ]	[aɪ]	1
[ɔɪ]	[ɔɪ]	1
[oɪ]	[oɪ]	1
[uɪ]	[uɪ]	1

Tab. 5.33: Agrupamento em visemas das diferentes produções do fone [ɪ].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,3	-2,5	-0,4	0,64
	P2	-0,3	0,8	1,1	0,66
	P3	-1,1	9,4	6,0	0,72
	P4	0,0	5,5	2,9	0,63

Tab. 5.34: Alvo articulatório do visema postônico alto anterior.

5.3.4.5 Visema postônico baixo central

Logatoma	Contexto C[ɐ]	
	Seqüência	Grupo
[pipɐ]	[pɐ]	1
[papɐ]	[pɐ]	1
[pupɐ]	[pɐ]	1
[fifɐ]	[fɐ]	1
[fafɐ]	[fɐ]	1
[fufɐ]	[fɐ]	1
[titɐ]	[tɐ]	1
[tatɐ]	[tɐ]	1
[tutɐ]	[tɐ]	1
[sisɐ]	[sɐ]	1
[sasɐ]	[sɐ]	1
[susɐ]	[sɐ]	1
[lilɐ]	[lɐ]	1
[lalɐ]	[lɐ]	1
[lulɐ]	[lɐ]	1
[ʃiʃɐ]	[ʃɐ]	1
[ʃaʃɐ]	[ʃɐ]	1
[ʃuʃɐ]	[ʃɐ]	1
[λiλɐ]	[λɐ]	1
[λaλɐ]	[λɐ]	1
[λuλɐ]	[λɐ]	1

Logatoma	Contexto C[ɐ]	
	Seqüência	Grupo
[kakɐ]	[kɐ]	1
[kikɐ]	[kɐ]	1
[kukɐ]	[kɐ]	1
[ɣiɾɐ]	[ɾɐ]	1
[ɣaɾɐ]	[ɾɐ]	1
[ɣuɾɐ]	[ɾɐ]	1

Logatoma	Contexto V[ɐ]	
	Seqüência	Grupo
[iɐ]	[iɐ]	1
[eɐ]	[eɐ]	1
[ɛɐ]	[ɛɐ]	1
[aɐ]	[aɐ]	1
[ɔɐ]	[ɔɐ]	1
[oɐ]	[oɐ]	1
[uɐ]	[uɐ]	1

Tab. 5.35: Agrupamento em visemas das diferentes produções do fone [ɐ].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	-0,6	-1,7	0,5	0,68
	P2	-0,4	2,0	4,8	0,72
	P3	-1,5	12,6	10,1	0,57
	P4	-0,6	8,8	6,8	0,60

Tab. 5.36: Alvo articulatório do visema postônico baixo central.

5.3.4.6 Visema postônico alto posterior

Logatoma	Contexto C[ʊ]	
	Seqüência	Grupo
[**pipʊ**]	[pʊ]	1
[**papʊ**]	[pʊ]	1
[**pupʊ**]	[pʊ]	1
[**fifʊ**]	[fʊ]	1
[**fafʊ**]	[fʊ]	1
[**fufʊ**]	[fʊ]	1
[**titʊ**]	[tʊ]	1
[**tatʊ**]	[tʊ]	1
[**tutʊ**]	[tʊ]	1
[**sisʊ**]	[sʊ]	1
[**sasʊ**]	[sʊ]	1
[**susʊ**]	[sʊ]	1
[**lilʊ**]	[lʊ]	1
[**lalʊ**]	[lʊ]	1
[**lulʊ**]	[lʊ]	1
[**ʃiʃʊ**]	[ʃʊ]	1
[**ʃaʃʊ**]	[ʃʊ]	1
[**ʃuʃʊ**]	[ʃʊ]	1
[**λiλʊ**]	[λʊ]	1
[**λaλʊ**]	[λʊ]	1
[**λuλʊ**]	[λʊ]	1

Logatoma	Contexto C[ʊ]	
	Seqüência	Grupo
[**kakʊ**]	[kʊ]	1
[**kikʊ**]	[kʊ]	1
[**kukʊ**]	[kʊ]	1
[**ɣiɾʊ**]	[ɾʊ]	1
[**ɣaɾʊ**]	[ɾʊ]	1
[**ɣuɾʊ**]	[ɾʊ]	1

Logatoma	Contexto V[ʊ]	
	Seqüência	Grupo
[**iʊ**]	[iʊ]	1
[**eʊ**]	[eʊ]	1
[**ɛʊ**]	[ɛʊ]	1
[**aʊ**]	[aʊ]	1
[**ɔʊ**]	[ɔʊ]	1
[**oʊ**]	[oʊ]	1
[**uʊ**]	[uʊ]	1

Tab. 5.37: Agrupamento em visemas das diferentes produções do fone [ʊ].

		Centróide			Instante
		x [mm]	**y** [mm]	**z** [mm]	**t** [%]
Grupo 1	**P1**	0,1	-1,0	-3,3	0,79
	P2	-2,0	2,1	-6,4	0,74
	P3	-0,2	5,2	-0,2	0,91
	P4	-0,1	0,7	1,8	0,83

Tab. 5.38: Alvo articulatório do visema postônico alto posterior.

5.3.4.7 Classificação do fone [e]

	Centróide			Instante
	x [mm]	**y** [mm]	**z** [mm]	**t** [%]
P1	-0,3	-1,0	0,9	0,37
P2	-0,5	1,6	5,2	0,36
P3	-1,9	13,5	10,4	0,39
P4	-1,5	10,2	6,5	0,36

Tab. 5.39: Alvo articulatório do fone [e].

<i_1>	<i_2>	<a>	<u>	<ɪ>	<ɐ>	<ʊ>
6,0	12,9	8,4	19,1	9,8	2,2	21,3

Tab. 5.40: Distâncias em milímetros do alvo articulatório do fone [e] aos visemas vocálicos.

5.3.4.8 Classificação do fone [ɛ]

	Centróide			Instante
	x [mm]	**y** [mm]	**z** [mm]	**t** [%]
P1	-0,7	-1,2	1,9	0,48
P2	0,5	1,9	9,5	0,49
P3	-2,1	16,4	14,8	0,43
P4	-1,0	12,9	10,9	0,51

Tab. 5.41: Alvo articulatório do fone [ɛ].

<i_1>	<i_2>	<a>	<u>	<ɪ>	<ɐ>	<ʊ>
13,8	19,8	4,5	26,7	18,1	9,7	29,6

Tab. 5.42: Distâncias em milímetros do alvo articulatório do fone [ɛ] aos visemas vocálicos.

	Centróide			Instante
	x [mm]	**y** [mm]	**z** [mm]	**t** [%]
P1	0,0	-4,3	-3,4	0,26
P2	-3,1	5,5	-1,3	0,42
P3	-1,3	19,0	9,7	0,39
P4	-0,2	9,8	9,4	0,42

Tab. 5.43: Alvo articulatório do fone [ɔ].

5.3.4.9 Classificação do fone [ɔ]

<i_1>	<i_2>	<a>	<u>	<ɪ>	<ɐ>	<ʊ>
11,1	10,7	16,5	15,3	14,7	11,3	21,9

Tab. 5.44: Distâncias em milímetros do alvo articulatório do fone [ɔ] aos visemas vocálicos.

5.3.4.10 Classificação do fone [o]

	Centróide			Instante
	x [mm]	**y** [mm]	**z** [mm]	**t** [%]
P1	0,1	-3,1	-4,6	0,22
P2	-4,7	5,9	-6,1	0,70
P3	-0,6	15,5	6,0	0,17
P4	0,0	6,9	7,3	0,52

Tab. 5.45: Alvo articulatório do fone [o].

<i_1>	<i_2>	<a>	<u>	<ɪ>	<ɐ>	<ʊ>
12,5	9,5	22,4	7,7	13,2	14,5	15,5

Tab. 5.46: Distâncias em milímetros do alvo articulatório do fone [o] aos visemas vocálicos.

5.4 Comentários Finais

Para a caracterização de um conjunto de visemas para o português do Brasil, procurou-se identificar padrões visíveis de movimentação articulatória durante a produção dos segmentos da língua em diversos contextos fonéticos. Para tanto, foram efetuadas medidas da trajetória de pontos visíveis ao redor da boca e no queixo de um falante durante a produção de um conjunto de logatomas. A identificação dos padrões visíveis baseou-se na identificação de similaridade utilizando o algoritmo de agrupamento *K-Means*.

A análise articulatória, considerando os efeitos da coarticulação, apresenta inerentemente o desafio da explosão combinatória das possibilidades a serem tratadas. Para contornar tal dificuldade, adotou-se um corpus que permitisse a caracterização dos efeitos da coarticulação adjacente perseveratória e antecipatória entre segmentos consonantais e vocálicos. A análise foi realizada utilizando um corpus formado por logatomas paroxítonos do tipo 'CVCV. Os segmentos consonantais e vocálicos utilizados procuram explorar a similaridade articulatória visível dos segmentos fonéticos, baseado nos homofemas apresentados nas Tabelas 5.1 e 5.2. Adicionalmente, a análise dos segmentos vocálicos foi norteada pela busca de visemas para os casos extremos alto-anterior, baixo-central e alto-posterior do diagrama das vogais, garantida a diferenciação dos segmentos postônicos. O tratamento dos 102 casos que constituem o corpus adotado permitiu a análise de doze contextos fonéticos diferentes para nove consoantes, dezesseis contextos para os três segmentos vocálicos postônicos e doze contextos para os segmentos vocálicos [i, a, u].

Os visemas determinados pela metodologia são expressos por um modelo paramétrico que incorpora os alvos articulatórios estabelecidos para os segmentos e o instante relativo de realização destes alvos dentro do intervalo da produção acústica, contemplando ainda os efeitos das coarticulação entre segmentos adjacentes na cadeia da produção da fala.

Os visemas estabelecidos pela metodologia desenvolvida, ao lado da própria metodologia, constituem as principais contribuições deste capítulo. A partir dos homofemas identificados no Capítulo 4 foram obtidos os visemas dependentes do contexto fonético apresentados nas Tabelas 5.7 e 5.26.

O conceito de alvo articulatório característico de um segmento, oriundo da análise fonética tradicional, foi caracterizado neste trabalho através dos pontos estacionários das trajetórias de pontos de interesse localizados ao redor da boca e no queixo. Estes alvos articulatórios foram associados a visemas através de processo de agrupamento por similaridade baseado no algoritmo *K-Means*. Observa-se que os visemas determinados são dependentes do contexto fonético da realização do fonema. O modelo paramétrico desenvolvido descreve a trajetória dos pontos de interesse durante a produção da cadeia de segmentos que constitui uma locução.

No próximo capítulo, as trajetórias dos pontos de interesse descritas pelo modelo paramétrico de visemas são analisadas e decompostas em transformações e deformação geométricas mais apropriadas

para a animação facial.

Capítulo 6

Modelagem da movimentação facial

6.1 Introdução

Neste capítulo, são derivados, a partir das trajetórias dos pontos de interesse, modelos e procedimentos para a manipulação da geometria de uma cabeça virtual tridimensional. As transformações geométricas a serem aplicadas à cabeça podem ser classificadas em duas categorias: transformações de corpo rígido e deformações. As transformações de corpo rígido estão associadas ao movimento da articulação temporomandibular que permite a elevação/depressão (movimento para cima/baixo) e a protrusão/retração (movimento para frente/trás) da mandíbula. As deformações estão associadas à movimentação voluntária do tecido da face, na região em torno e incluindo os lábios, para a realização das posturas labiais necessárias à produção dos segmentos da fala.

Parte das pistas visuais presentes na face de um falante, principalmente na região envolvendo o queixo e o lábio inferior, advém da movimentação da mandíbula. A movimentação desta estrutura óssea é determinada pelo comportamento de sua articulação com o osso temporal. Na Seção 6.2 é apresentado um modelo geométrico, derivado das trajetórias dos pontos de interesse, para representar a movimentação da articulação temporomandibular. Este modelo descreve os movimentos de rotação e translação da articulação durante a fala.

As deformações faciais associadas à movimentação voluntária do tecido labial estão embutidas nas trajetórias dos pontos de interesse P1, P2 e P3, localizados no lábios superior, canto da boca e lábio inferior, respectivamente (Seção 5.2.4). Destas três localizações, apenas o ponto P3 sofre influência significativa da movimentação da mandíbula. Assim, na trajetória de P3, além da movimentação voluntária do tecido labial associada a posturas específicas, como, por exemplo, protrusão, está embutida a movimentação da mandíbula. A decomposição da trajetória de P3 nas componentes associadas à movimentação voluntária do tecido do lábio e à movimentação da mandíbula é apresentada na Seção 6.3. As trajetórias dos pontos de interesse P1 e P2, por não sofrerem influência

significativa da movimentação da mandíbula, expressam o comportamento devido à movimentação labial voluntária. As Seções 6.4 e 6.5 tratam destes comportamentos.

Em todas as seções mencionadas são apresentados gráficos derivados do modelo paramétrico dos visemas para ilustrar e fornecer evidências da coerência e conformidade com a realidade dos modelos e simplificações adotados. Em particular, é apresentado um conjunto de diagramas envolvendo as vogais, cada qual enfatizando um parâmetro tradicionalmente utilizado para a classificação desta classe de segmentos. Os parâmetros adotados são: abertura da boca, protrusão do lábio inferior, protrusão do lábio superior e extensão/arrendondamento da boca. Tais parâmetros, por serem tradicionalmente utilizados na análise fonética, permitem a comparação das características derivadas das medidas e dos modelos adotados com o comportamento previsto pela fonética. A opção pela análise das vogais baseia-se na consideração de que tais segmentos são menos suscetíveis aos efeitos da coarticulação, apresentando, portando, características mais nítidas e menos dependentes do contexto fonético de sua realização, facilitando, assim, a comparação com as características descritas pela fonética.

6.2 Comportamento da articulação temporomandibular

A mandíbula é o único osso móvel da face. Este osso serve de base para os dentes inferiores e proporciona pontos de ligação para músculos da face e para uma grande parte da musculatura da língua. A principal função da mandíbula é a mastigação, contribuindo para a produção da fala ao modificar as características de ressonância do trato vocal.

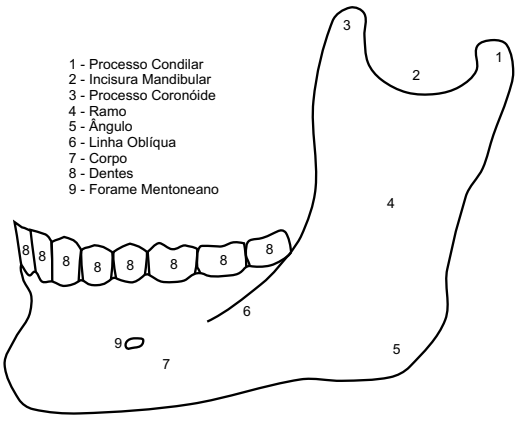

Fig. 6.1: Visão lateral da mandíbula - figura adaptada de Gray (2000).

A mandíbula é uma estrutura óssea rígida constituída de uma parte horizontal, denominada corpo, e duas verticais, denominadas ramos, que se unem de maneira contínua ao corpo, em sua parte posterior, em um ângulo próximo a 90 graus (Figura 6.1). A parte superior do ramo possui duas proeminências, o processo coronóide à frente e o processo condilar atrás, separados por uma cavidade,

denominada incisura mandibular. O processo condilar se insere na fossa mandibular do osso temporal, que se encontra à frente do canal auditivo, formando a articulação temporomandibular ou ATM (Figura 6.2). A ATM pode ser apalpada colocando-se um dedo bem à frente e um pouco abaixo da abertura do meato auditivo externo, enquanto a mandíbula é movimentada.

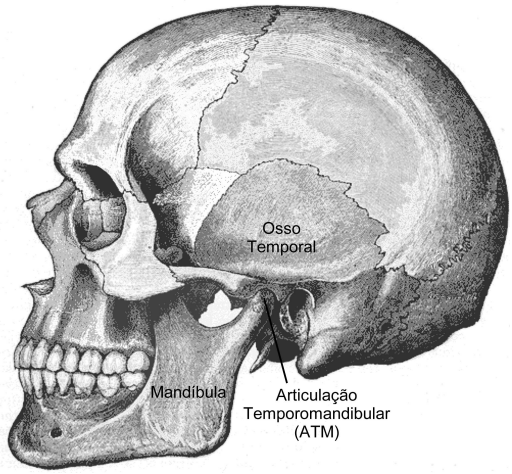

Fig. 6.2: Articulação Temporomandibular - figura adaptada de Gray (2000).

Na articulação temporomandibular, a parte anterior da fossa mandibular do osso temporal é lisa e revestida por fibrocartilagem e se articula indiretamente com o côndilo da mandíbula. Um delgado envelope de tecido fibroso, denominado cápsula articular (*"Capsula Articularis"*), envolve completamente a articulação, ligando-se às faces articulares da fossa mandibular e ao colo do côndilo. O côndilo é separado da fossa por um disco articular (Figura 6.3).

O teto da fossa mandibular do osso temporal é fino e transparente, sugerindo que a fossa, que abriga parte do disco articular e parte do côndilo da mandíbula, normalmente não é um elemento funcional que suporta esfoço mecânico. A tensão mecânica é absorvida entre o côndilo e o disco articular, e entre o disco e a parte anterior espessada da fossa, denominada de eminência articular. Por causa do disco articular interposto, a ATM é uma articulação dupla, uma entre o disco e a eminência articular (articulação superior), e outra entre o côndilo da mandíbula e o disco (articulação inferior). Devida à natureza da articulação e à disposição dos músculos que sobre ela atuam, o movimento mandibular é complexo. As direções de forças de ação muscular sobre a mandíbula são apresentadas esquematicamente na Figura 6.4. Os principais movimentos ocasionados pela ação muscular podem ser resumidos em:

1. **Elevação** (movimento para cima): pterigóideo medial, masseter e temporal;

2. **Depressão** (movimento para baixo): pterigóideo lateral, gênio-hióideo, digástrico (ventre anterior), milo-hióideo, genioglosso;

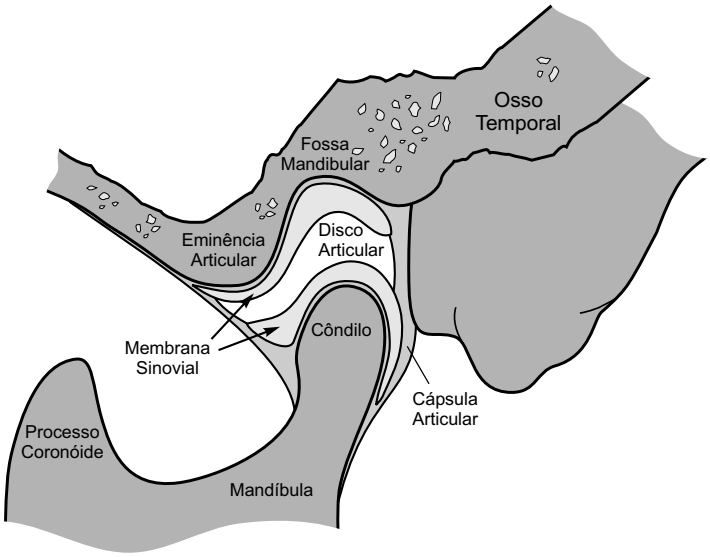

Fig. 6.3: Seção sagital da Articulação Temporomandibular - figura adaptada de Gray (2000).

3. **Protrusão** (movimento para frente): pterigóideo lateral, pterigóideo medial;

4. **Retração** (movimento para trás): temporal (parte posterior), milo-hióideo, genio-hióideo, digástrico (ventre anterior);

5. **Lateralização** (movimento para os lados): pterigóideo lateral e temporal (parte posterior).

Os movimentos da mandíbula influenciam a postura dos lábios, a posição da língua e a configuração da cavidade oral, além de alterar as dimensões da cavidade faríngea. Para a produção da fala, o complexo movimento mandibular pode ser decomposto em translação e rotação (ZEMLIN, 2000). A translação é realizada pela articulação superior, entre o disco e o osso temporal. Na protrusão e retração, os discos articulares e a mandíbula deslizam para a frente e para baixo, ou para cima e para trás, com os côndilos em contato firme com as eminências articulatórias. A ação de deslizamento ocorre entre o disco articular e a eminência articular e, em conseqüência, o côndilo é levado para frente e para trás. As articulações inferiores, entre o disco e a mandíbula, realizam um movimento de rotação em torno de um eixo horizontal que passa no centro dos côndilos da mandíbula.

A movimentação da ATM durante a fala foi analisada em outros trabalhos, como (EDWARDS; HARRIS, 1990), (WESTBURY, 1994) e (VATIKIOTIS-BATESON; OSTRY, 1995). Em (WESTBURY, 1994) e em (EDWARDS; HARRIS, 1990) a análise da movimentação da mandíbula foi realizada no plano sagital médio. Já em (VATIKIOTIS-BATESON; OSTRY, 1995) a análise foi efetuada no espaço tridimensional. Os resultados destes trabalhos indicam que, durante a fala, a articulação temporomandibular apresenta fundamentalmente os movimentos de rotação e translação.

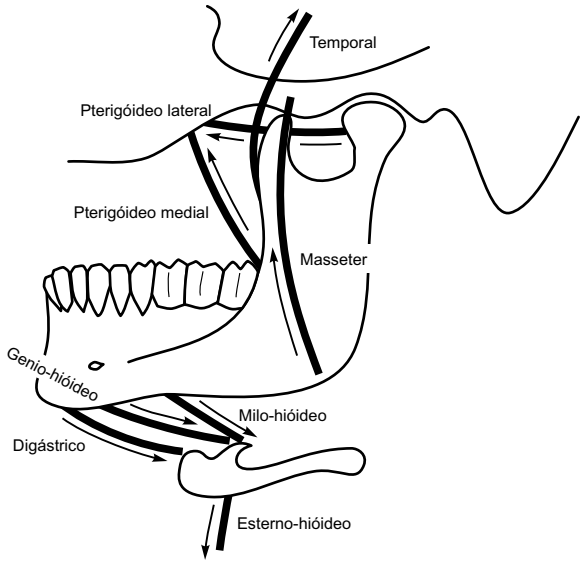

Fig. 6.4: Principais músculos responsáveis pela movimentação da ATM - figura adaptada de Zemlin (2000).

No presente trabalho, com base nas trajetórias dos pontos de interesse P1 a P4 (Seção 5.2.4), o movimento da ATM foi modelado no plano sagital médio. O modelo adotado assume que a ATM, durante a fala, executa movimentos de rotação e translação neste plano. A rotação é realizada em torno do centro da ATM, definido pela interseção do eixo que passa pelos centros dos côndilos com o plano sagital. O centro da ATM sofre translação devido ao deslizamento do disco articular. Uma vez que não há movimentação significativa da pele na ponta do queixo durante a fala, foi assumido que a trajetória do ponto de interesse P_4, localizado no queixo do falante, reflete o movimento da mandíbula e conseqüentemente permite estimar as componentes de rotação e translação da ATM.

A trajetória de P_4 no plano sagital médio pode ser expressa por (Figura 6.5):

$$\overline{P}_4(t) = y_4(t)\ \ \widehat{y} + z_4(t)\ \ \widehat{z} \tag{6.1}$$

onde:

- $y_4(t)$ e $z_4(t)$ são as coordenadas y e z da trajetória de P_4, calculadas como na Seção 5.2.4;
- $\widehat{y}$ e $\widehat{z}$ são vetores unitários nas direções Y e Z, respectivamente.

Considerando as convenções e referenciais apresentados na Figura 6.5, a trajetória de P_4 no plano sagital médio também pode ser expressa por:

$$\overline{P}_4(t) = \overline{R}_4(t) + \overline{C}(t) \tag{6.2}$$

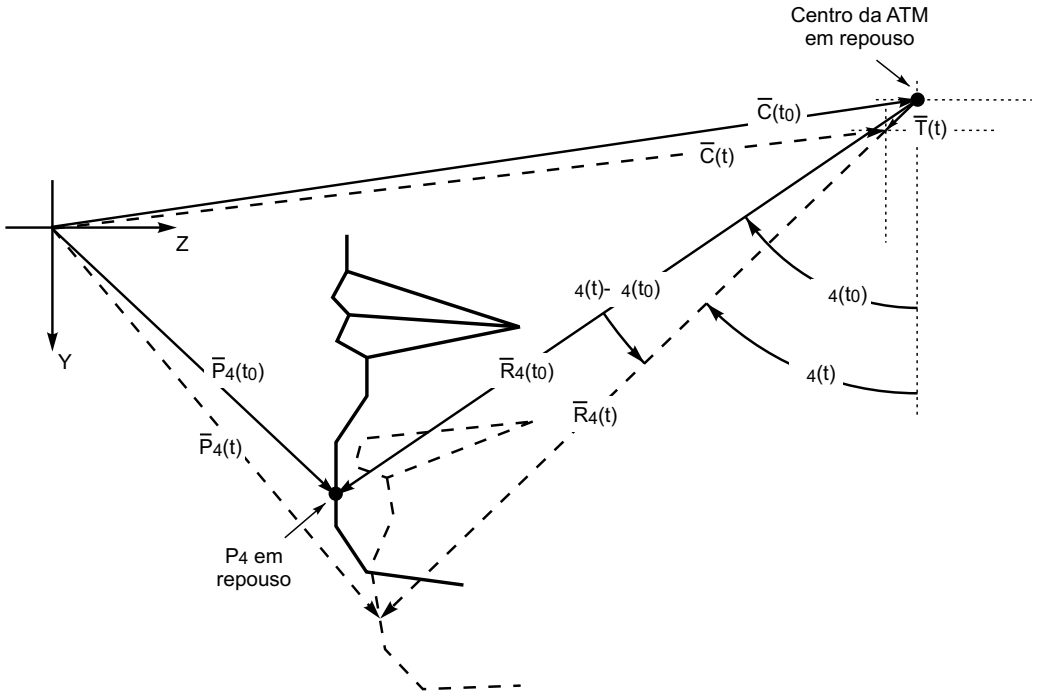

Fig. 6.5: Movimentação da ATM no plano sagital médio.

onde:

- $\overline{R}_4\,(t)$ é um vetor associado ao movimento de rotação da ATM;
- $\overline{C}\,(t)$ é o vetor posição do centro da ATM que incorpora o movimento de translação.

O vetor $\overline{R}_4\,(t)$, que descreve o movimento de rotação em torno do centro da ATM, tem magnitude constante, r_4, definida pela distância entre P_4 e o centro da ATM. $\overline{R}_4\,(t)$ é dado por:

$$\overline{R}_4\,(t) = r_4\,\cos(\theta_4\,(t))\;\widehat{y} + r_4\,\sin(\theta_4\,(t))\;\widehat{z} \tag{6.3}$$

onde:

- r_4 é a distância de P_4 ao centro da ATM;
- $\theta_4\,(t)$ é o ângulo de rotação de P_4 em relação ao referencial adotado (Figura 6.5). Por convenção, ângulos no sentido anti-horário são positivos;
- $\widehat{y}$ e $\widehat{z}$ são vetores unitários nas direções Y e Z, respectivamente.

O vetor posição do centro da ATM é dado por:

$$\overline{C}\,(t) = \overline{C}\,(t_0) + \overline{T}\,(t) \tag{6.4}$$

onde:

- $\overline{C}\,(t_0)$ é a posição do centro da ATM em repouso, no instante inicial t_0, quando a boca encontra-se fechada e os dentes cerrados;
- $\overline{T}\,(t)$ é a translação do centro da ATM durante a fala.

A posição do centro da ATM na situação de repouso é definida por:

$$\overline{C}\,(t_0) = y_C\,(t_0)\ \ \widehat{y} + z_C\,(t_0)\ \ \widehat{z} \tag{6.5}$$

onde:

- $y_C\,(t_0)$ e $z_C\,(t_0)$ são as coordenadas x e y do centro da ATM no instante t_0;
- $\widehat{y}$ e $\widehat{z}$ são vetores unitários nas direções Y e Z, respectivamente.

A translação $\overline{T}(t)$ do centro da ATM é dada por:

$$\overline{T}\,(t) = y_T\,(t)\ \ \widehat{y} + z_T\,(t)\ \ \widehat{z} \tag{6.6}$$

onde:

- $y_T\,(t)$ e $z_T\,(t)$ são os deslocamentos nas direções Y e Z do centro da ATM;
- $\widehat{y}$ e $\widehat{z}$ são vetores unitários nas direções Y e Z, respectivamente.

Manipulando as Equações 6.1 a 6.6 é possível estabelecer a seguinte formulação para as componentes $y_T(t)$ e $z_T(t)$ da translação do centro da ATM no plano sagital médio:

$$\begin{cases} y_T\,(t) = y_4\,(t) - y_C\,(t_0) - r_4\,\cos\left(\theta_4\,(t)\right) \\ z_T\,(t) = z_4\,(t) - z_C\,(t_0) - r_4\,\sin\left(\theta_4\,(t)\right) \end{cases} \tag{6.7}$$

Os parâmetros r_4 e θ_4 das Equações 6.3. 6.7 podem ser calculados a partir da posição do centro da ATM em repouso e da trajetória de P_4 pelas relações:

$$r_4 = \sqrt{\left[y_4\,(t_0) - y_C\,(t_0)\right]^2 + \left[z_4\,(t_0) - z_C\,(t_0)\right]^2} \tag{6.8}$$

$$\theta_4\,(t) = \arctan\left(\frac{z_4\,(t) - z_C\,(t_0)}{y_4\,(t) - y_C\,(t_0)}\right) \tag{6.9}$$

onde:

- $y_4\,(t)$ e $z_4\,(t)$ são as coordenadas y e z do ponto P_4 medidas no instante t;
- $y_C\,(t)$ e $z_C\,(t)$ são as coordenadas do centro de rotação da ATM no instante t;
- t_0 é o instante inicial de repouso, quando a boca encontra-se fechada e os dentes cerrados.

As Equações 6.3, 6.8 e 6.9 descrevem o movimento de rotação da ATM em função da trajetória de P_4 no plano sagital médio e da posição do centro da ATM no repouso. As Equações 6.6, 6.7 e 6.9 descrevem o movimento de translação da ATM em função dos mesmos parâmetros.

As coordenadas $y_4(t)$ e $z_4(t)$ de P_4 foram medidas pelo procedimento fotogramétrico como discutido no Capítulo 5. Utilizando este mesmo procedimento, as coordenadas $y_C\,(t_0)$ e $z_C\,(t_0)$ do centro da ATM em repouso, foram estimadas através da medida da localização presumida do centro da ATM. Para tanto, foi localizado com a ponta do dedo indicador o côndilo da ATM na região anterior ao ouvido do informante, durante a movimentação lateral voluntária e forçada da mandíbula. Na região assim identificada, foi marcada com tinta branca, para posterior medida, a localização presumida do centro da ATM. A Tabela 6.1 apresenta a localização do centro da ATM no plano sagital médio do informante.

$\mathbf{y_C(t_0)}$ [mm]	$\mathbf{z_C(t_0)}$ [mm]
63,8	82,3

Tab. 6.1: Posição no plano sagital médio do centro da articulação temporomandibular em repouso.

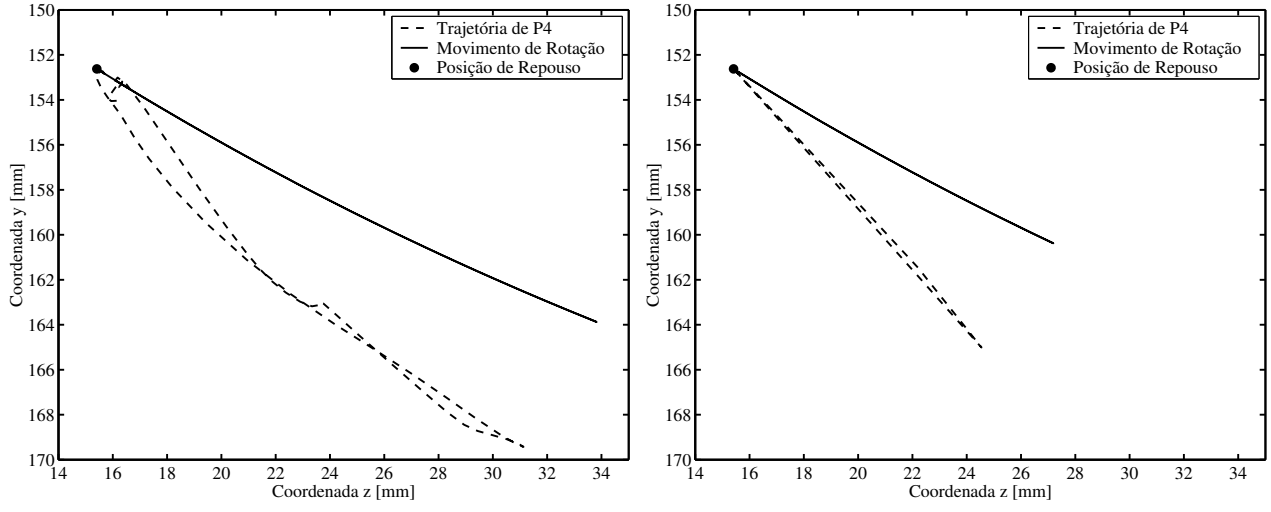

Fig. 6.6: Rotação da ATM durante ['aɐ]: esquerda) medidas; direita) modelo paramétrico dos visemas.

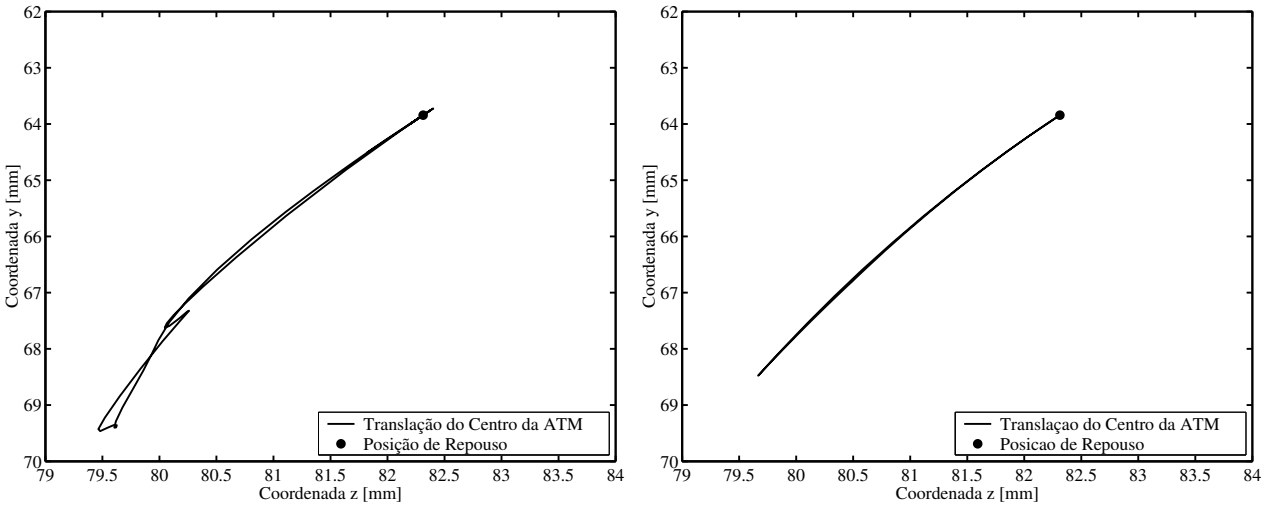

Fig. 6.7: Translação da ATM durante ['aɐ]: esquerda) medidas; direita) modelo paramétrico dos visemas.

As Figuras 6.6 e 6.7 apresentam as componentes de rotação e translação do movimento da ATM derivadas da trajetória de P_4 durante a produção do logatoma ['aɐ] consoante o equacionamento apresentado acima. O gráfico no lado esquerdo da Figura 6.6 apresenta a trajetória medida de P_4 e a componente de rotação embutida nesta trajetória; no gráfico da direta são apresentadas a trajetória de P_4 como descrita pelo modelo paramétrico dos visemas (Seção 5.2.8) e a componente de rotação da ATM derivada desta trajetória. O gráfico no lado esquerdo da Figura 6.7 apresenta a translação do centro da ATM calculada a partir da trajetória de P_4; no gráfico da direita tem-se a translação estimada baseada no modelo paramétrico dos visemas. Nos gráficos, a cabeça do informante está orientada como na Figura 6.5.

As Figuras 6.8 e 6.9 apresentam, de forma análoga às Figuras 6.6 e 6.7, as componentes de rotação e translação do movimento da ATM derivadas da trajetória de P_4 durante a produção do logatoma ['papɐ].

Utilizando os alvos articulatórios dos visemas vocálicos (Seção 5.3.4) e a modelagem do comportamento da ATM apresentada, foi gerado o diagrama da Figura 6.10. Este diagrama apresenta o ângulo de rotação da mandíbula para os alvos articulatórios destes segmentos (Seção 5.3.4). Observa-se no diagrama que o visema <a> possui maior abertura que os visemas <i, u>, sendo que estes dois últimos apresentam abertura mais próximas. Estas observações são compatíveis com o diagrama idealizado das vogais e com a teoria fonética, em que o fone [a] é uma vogal baixa, e portanto com grande abertura, e as vogais [i, u] são vogais altas, e portanto com uma abertura menor.

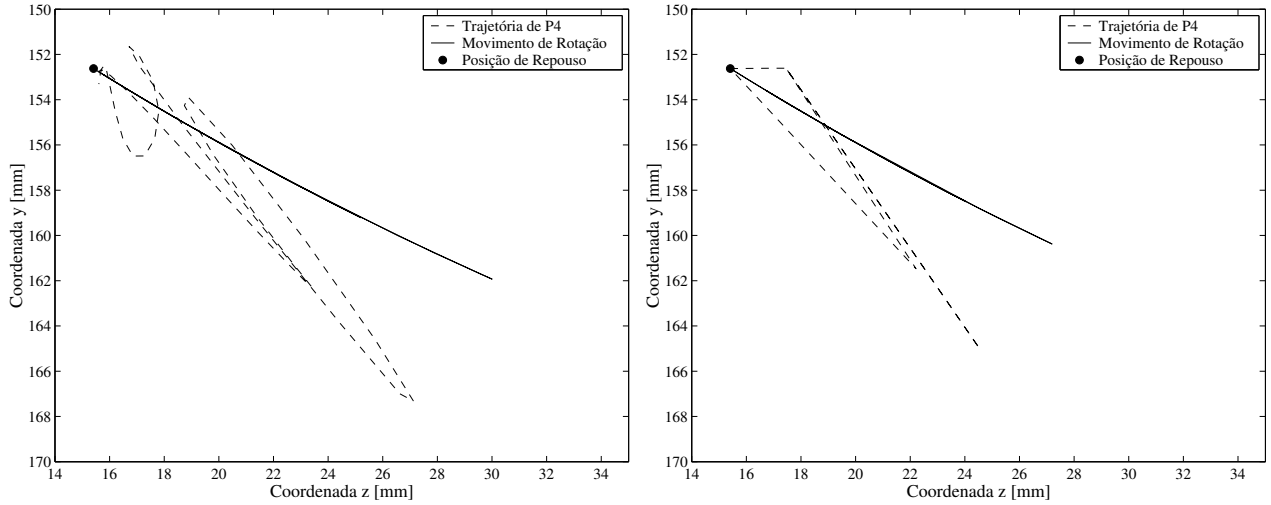

Fig. 6.8: Rotação da ATM durante [ˈpapɐ]: esquerda) medidas; direita) modelo paramétrico.

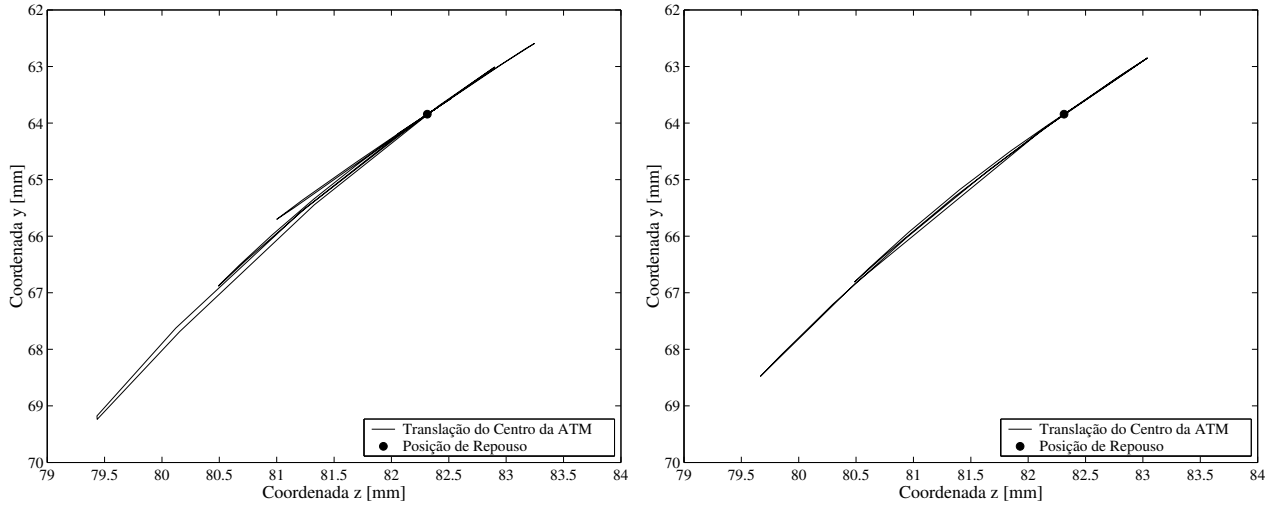

Fig. 6.9: Translação da ATM durante [ˈpapɐ]: esquerda) medidas; direita) modelo paramétrico.

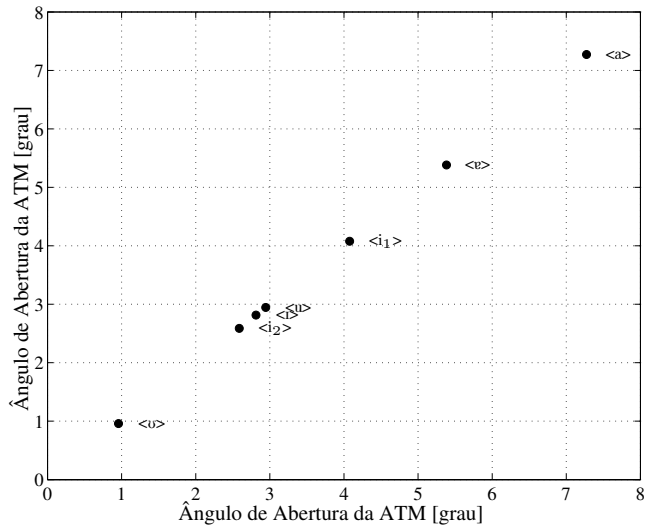

Fig. 6.10: Diagrama das vogais - ângulo de abertura.

6.3 Comportamento do lábio inferior

A movimentação do lábio inferior, capturada pela trajetória do ponto P_3, pode ser desmembrada em duas componentes principais. A primeira componente advém da movimentação da mandíbula que afeta a abertura da boca e, portanto, o posicionamento do lábio inferior. A segunda componente refere-se à movimentação voluntária do tecido labial necessária à produção de certas posturas articulatórias, como protrusão e extensão labial. Esta segunda componente é denominada neste trabalho de comportamento do lábio inferior.

O vetor $\overline{L}\,(t)$, que descreve o comportamento do lábio inferior, é definido por:

$$\overline{L}\,(t) = y_L\,(t)\;\;\widehat{y} + z_L\,(t)\;\;\widehat{z} \tag{6.10}$$

É possível estimar $\overline{L}\,(t)$, a partir da trajetória de P_3 e dos movimentos de translação e rotação da ATM pela relação (Figura 6.11):

$$\overline{L}\,(t) = \overline{P}_3\,(t) - \overline{R}_3\,(t) - \overline{C}\,(t) \tag{6.11}$$

onde:

- $\overline{P}_3\,(t)$ é a trajetória de P_3;
- $\overline{R}_3\,(t)$ é o movimento de rotação de P_3 devido à rotação da ATM;
- $\overline{C}\,(t)$ é a posição do centro da ATM no instante t.

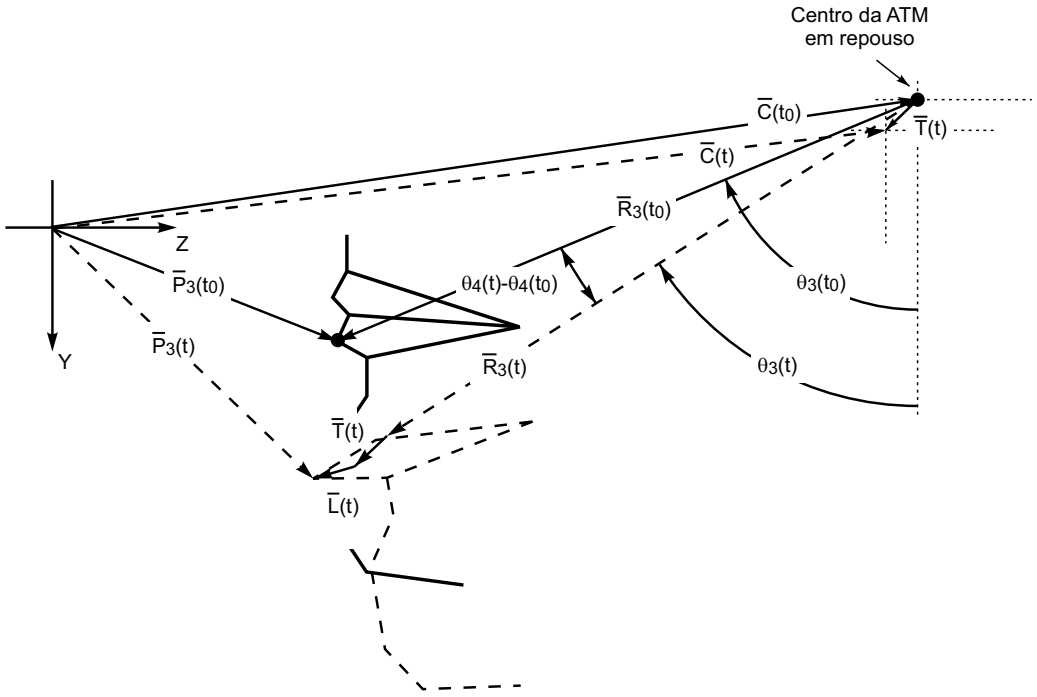

Fig. 6.11: Movimentação do lábio inferior no plano sagital médio.

A trajetória de P_3 no plano sagital médio é expressa por

$$\overline{P}_3(t) = y_3(t)\ \widehat{y} + z_3(t)\ \widehat{z} \tag{6.12}$$

onde:

- $y_3(t)$ e $z_3(t)$ são as coordenadas y e z da trajetória de P_3, calculadas como descrito no Capítulo 5;
- $\widehat{y}$ e $\widehat{z}$ são vetores unitários nas direções Y e Z, respectivamente.

A rotação, $\overline{R}_3(t)$, sofrida por P_3 devido à rotação da ATM é dada por:

$$\overline{R}_3(t) = r_3\ \cos(\theta_3(t))\ \widehat{y} + r_3\ \sin(\theta_3(t))\ \widehat{z} \tag{6.13}$$

onde:

- r_3 é a distância de P_3 ao centro da ATM;
- $\theta_3(t)$ é o ângulo de rotação de P_3 em relação ao referencial adotado (Figura 6.11);
- $\widehat{y}$ e $\widehat{z}$ são vetores unitários nas direções Y e Z, respectivamente.

A variação do ângulo de rotação de P_3 devido à ATM, considerando a estrutura óssea rígida da mandíbula, é igual à de P_4. Portanto, $\theta_3(t)$ pode ser expresso por:

$$\theta_3(t) = \theta_3(t_0) + \theta_4(t) - \theta_4(t_0) \tag{6.14}$$

onde:

- $\theta_3(t_0)$ é o ângulo inicial de P_3;
- $\theta_4(t) - \theta_4(t_0)$ é a variação de ângulo sofrida por P_4.

A partir das Equações 6.10 a 6.14 e Equações 6.4 a 6.7 é possível derivar:

$$\begin{cases} y_L(t) = y_3(t) - y_4(t) - r3\cos(\theta_3(t_0) + \theta_4(t) - \theta_4(t_0)) - r4\cos(\theta_4(t)) \\ z_L(t) = z_3(t) - z_4(t) - r3\sin(\theta_3(t_0) + \theta_4(t) - \theta_4(t_0)) - r4\sin(\theta_4(t)) \end{cases} \tag{6.15}$$

sendo que r_3 e $\theta_3(t_0)$ são dados por

$$r_3 = \sqrt{[y_3(t_0) - y_C(t_0)]^2 + [z_3(t_0) - z_C(t_0)]^2} \tag{6.16}$$

$$\theta_3(t_0) = \arctan\left(\frac{z_3(t_0) - z_C(t_0)}{y_3(t_0) - y_C(t_0)}\right) \tag{6.17}$$

onde:

- $y_3(t_0)$ e $z_3(t_0)$ são as coordenadas y e z do ponto P_3 em repouso, no instante t_0;
- $y_C(t_0)$ e $z_C(t_0)$ são as coordenadas do centro de rotação da ATM em repouso.

As Figuras 6.12 a 6.20 apresentam as trajetórias de P_3, estabelecidas pelo modelo paramétrico dos visemas (Seção 5.2.8), para os logatomas [ˈiɪ], [ˈiɐ], [ˈiʊ], [ˈaɪ], [ˈaɐ], [ˈaʊ], [ˈuɪ], [ˈuɐ] e [ˈuʊ/, decompostas nas componentes associadas à ATM e à movimentação do tecido labial utilizando a modelagem proposta. Nas figuras, a linha tracejada representa a movimentação associada à ATM e a linha cheia o comportamento do lábio inferior. Nas figuras, o posicionamento da cabeça do informante segue a orientação apresentada na Figura 6.11.

Nos gráficos das figuras, sempre que a linha cheia está à esquerda do ponto de repouso tem-se a protrusão do lábio inferior. Observa-se que os gráficos e, portanto, os modelos adotados, indicam que há protrusão sempre que os fones [u] ou [ʊ] são produzidos. Tal observação evidencia a consistência dos modelos e aproximações adotados, uma vez que a produção destes fones está associada à protrusão labial.

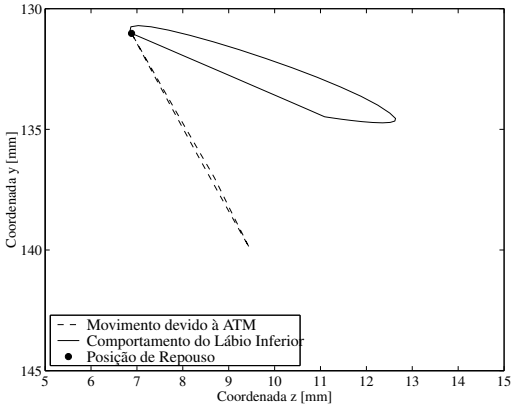

Fig. 6.12: Protrusão do lábio inferior durante a produção do logatoma [ˈiɪ].

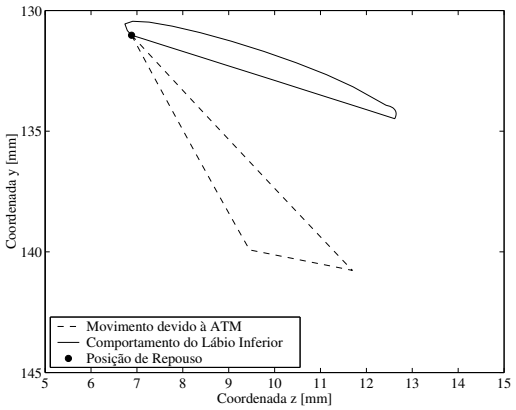

Fig. 6.13: Protrusão do lábio inferior durante a produção do logatoma [ˈiɐ].

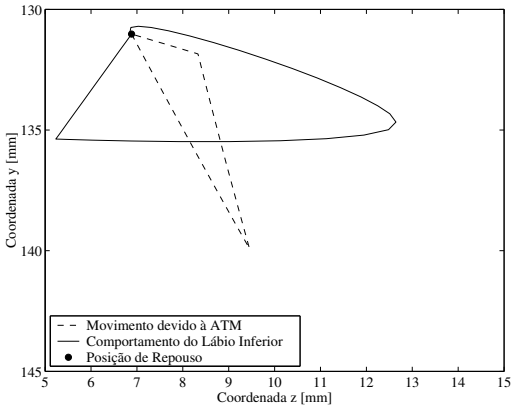

Fig. 6.14: Protrusão do lábio inferior durante a produção do logatoma [ˈiʊ].

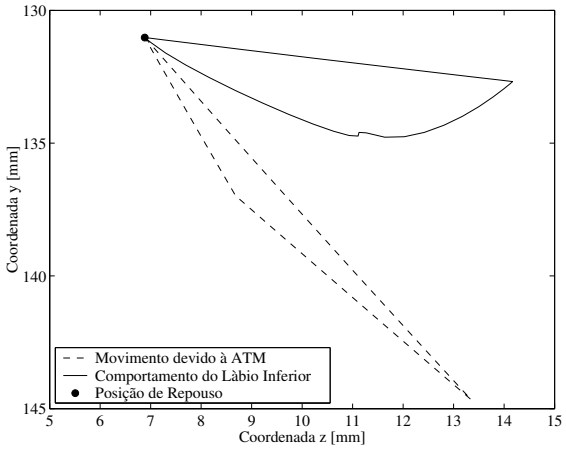

Fig. 6.15: Protrusão do lábio inferior durante a produção do logatoma [ˈaɪ].

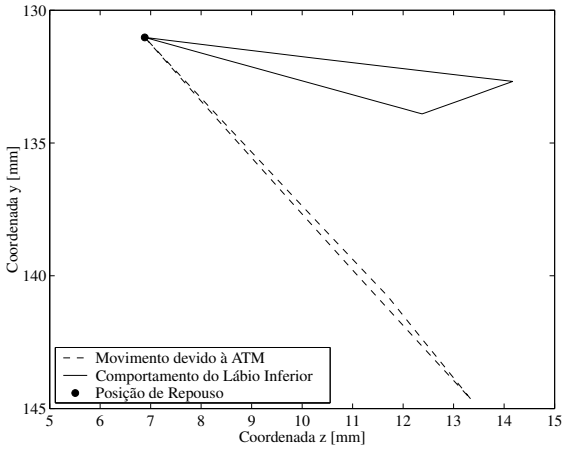

Fig. 6.16: Protrusão do lábio inferior durante a produção do logatoma [ˈaɐ].

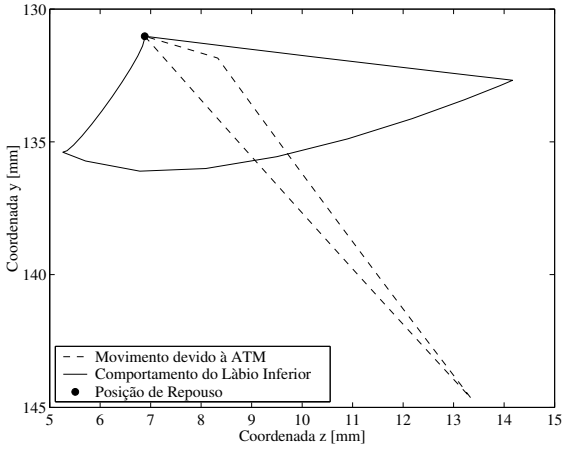

Fig. 6.17: Protrusão do lábio inferior durante a produção do logatoma [ˈaʊ].

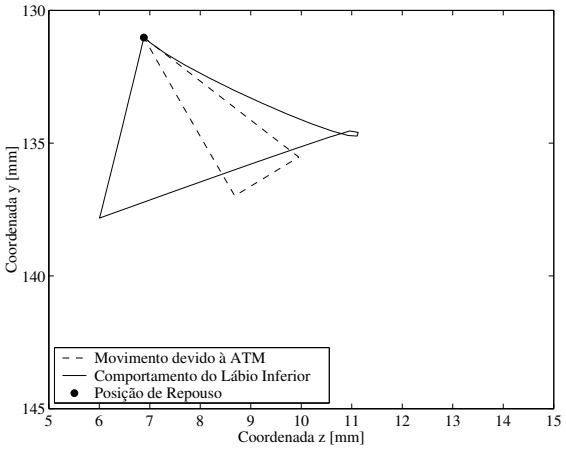

Fig. 6.18: Protrusão do lábio inferior durante a produção do logatoma [ˈuɪ].

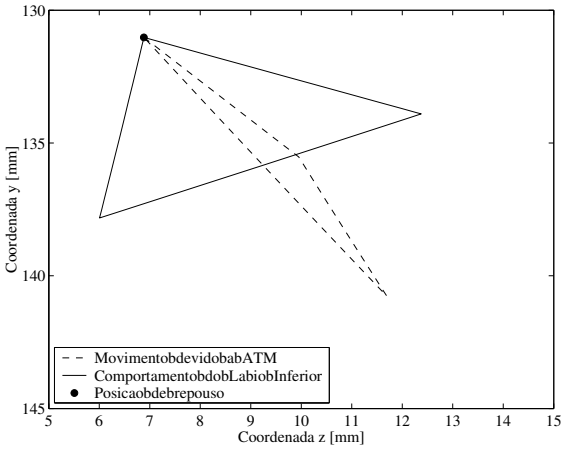

Fig. 6.19: Protrusão do lábio inferior durante a produção do logatoma [ˈuɑ].

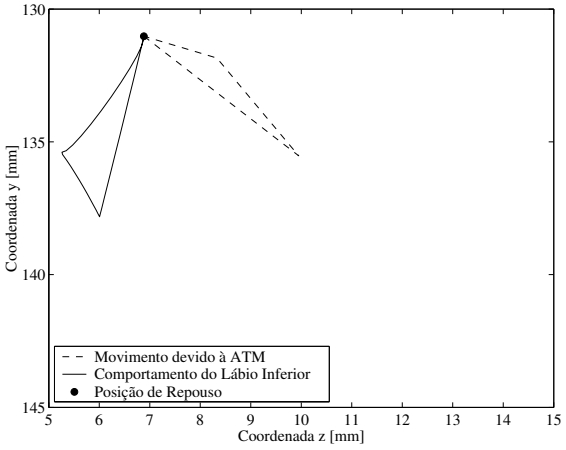

Fig. 6.20: Protrusão do lábio inferior durante a produção do logatoma [ˈuʊ].

O diagrama da Figura 6.21 apresenta a estimativa da protrusão do lábio inferior para os alvos articulatórios dos segmentos vocálicos calculados no Capítulo 5 - Seção 5.3.4. Observa-se no diagrama, como esperado pelo conhecimento fonético, que somente os visemas <u, ʊ> apresentam protrusão.

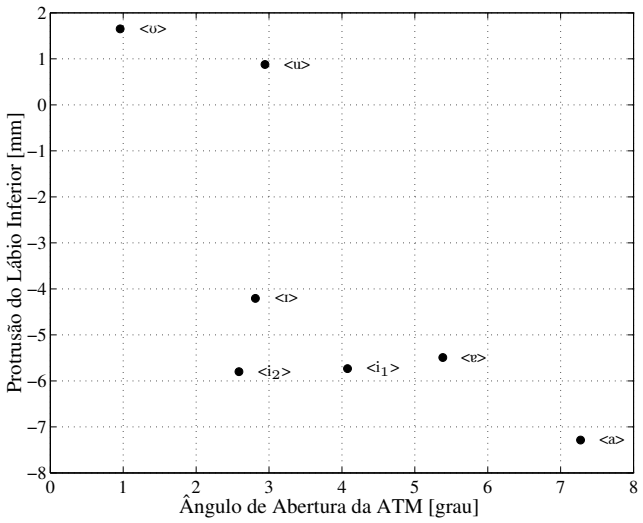

Fig. 6.21: Diagrama das vogais - protrusão lábio inferior.

6.4 Comportamento do lábio superior

De forma análoga ao comportamento do lábio inferior, o comportamento do lábio superior refere-se à movimentação voluntária do tecido do lábio superior para a produção dos sons da fala. Diferentemente do lábio inferior, a movimentação do lábio superior expressa pela trajetória de P_1 não sofre influência significativa da movimentação da articulação temporomandibular. Assim, é considerado que o comportamento do lábio superior é dado diretamente pela trajetória de P_1.

O diagrama da Figura 6.22 apresenta a estimativa da protrusão do lábio superior para os alvos articulatórios (deslocamento z de P_1 - Seção 5.3.4). De mesma forma que na análise da protrusão do lábio inferior para os segmentos vocálicos da seção anterior, observa-se pelo diagrama que somente os visemas <u, ʊ> apresentam protrusão.

6.5 Comportamento do canto da boca

O comportamento do canto da boca é dado pela trajetória do ponto de interesse P_2. Este comportamento expressa a movimentação voluntária do tecido labial. É assumido tacitamente que a movimentação da mandíbula não afeta significativamente a trajetória de P_2.

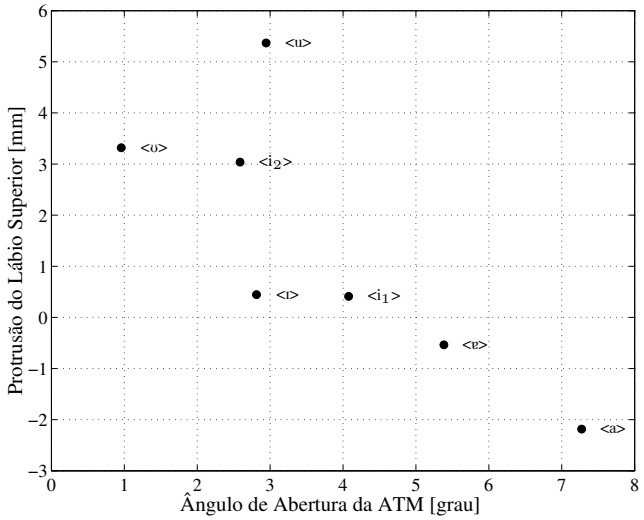

Fig. 6.22: Diagrama das vogais - protrusão lábio superior.

É possível estimar com os valores medidos a extensão da boca para os visemas vocálicos. Entende-se como extensão da boca a distância entre os cantos dos lábios. A extensão da boca também pode ser entendida como um parâmetro que reflete de forma inversa o arredondamento dos lábios. Na Figura 6.23 é apresentado um diagrama da extensão da boca. Observa-se no diagrama que os visemas <u, ʊ, i_2> apresentam uma diminuição da extensão da boca em relação à posição de repouso. Esta diminuição está associada ao arredondamento dos lábios necessário à produção destes visemas. É interessante observar que o visema <i_2> também apresenta arredondamento. Este visema foi identificado no processamento do fonema nos contextos /ʃiʃ/ e /tit/. O arredondamento do visema <i_2> é coerente com a observação de que, devido ao efeito de coarticulação, o arredondamento típico do segmento consonantal /ʃ/ afeta o segmento vocálico. Adicionalmente, no contexto /tit/, houve um processo de palatalização da oclusiva alveolar e a produção da africada [tʃ] (BARBOSA; ALBANO, 2004) (SILVA, 2002).

6.6 Comentários Finais

Os modelos desenvolvidos neste capítulo descrevem características-chave da movimentação articulatória visível. Um aspecto importante que norteou a modelagem foi a preocupação em desacoplar os efeitos do movimento da articulação temporomandibular da movimentação voluntária do tecido labial. Considerando os pontos de interesse analisados, localizados no queixo (P_4), no lábio inferior (P_3), no canto da boca (P_2) e no lábio superior (P_1), foi considerado que apenas as trajetórias de P_4 e P_3 são significativamente influenciadas pela movimentação da ATM. Em particular, foi assumido que a trajetória de P_4 é exclusivamente ocasionada pela movimentação da mandíbula, servindo de

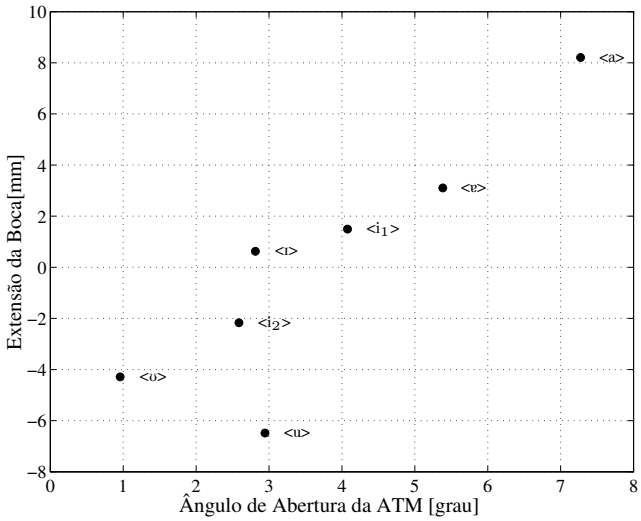

Fig. 6.23: Diagrama das vogais - extensão da boca.

referência para a modelagem do comportamento da ATM. O modelo assim derivado representa o comportamento da ATM através de transformações de rotação e translação. A estimativa do comportamento da ATM permitiu, por sua vez, a extração da trajetória de P_3 da componente associada à movimentação voluntária do lábio inferior. Já a movimentação labial associada a P_1 e P_2 é descrita pelas trajetórias destes pontos definidas pelo modelo paramétrico dos visemas (Seção 5.2.8). Os comportamentos dos pontos labiais P_1, P_2, e P_3, excluída a componente associada à ATM, refletem as deformações sofridas pela região ao redor da boca pela atuação de músculos faciais necessárias a produção de diferentes posturas labiais.

Ao longo do capítulo, características articulatórias, tais como protrusão, abertura e extensão da boca, expressas pelos modelos propostos, foram comparadas com os comportamentos esperados por conhecimento fonético consolidado (Capítulo 3). As figuras 6.10, 6.12 a 6.23 apresentam as características previstas pelos modelos. A análise comparativa indica que os modelos desenvolvidos reproduzem de maneira consistente as características articulatórias apontadas pela fonética.

Tanto as transformações de rotação e translação associadas à ATM como a movimentação dos pontos ao redor da boca devem ser incorporadas e implementadas na animação facial. O mapeamento dos modelos em estratégias de manipulação da geometria tridimensional de uma cabeça virtual admite variantes, podendo ir desde uma solução puramente geométrica à sofisticação de uma modelagem física que leve em conta as características biomecânicas do tecido facial. Entretanto, independente da estratégia, os modelos aqui desenvolvidos oferecem referências que descrevem características importantes da movimentação articulatória visível. No capítulo que se segue é apresentado o protótipo desenvolvido utilizando os modelos aqui apresentados no intuito de demonstrar e permitir a avaliação dos conceitos e modelos adotados.

Capítulo 7

Implementação Piloto

7.1 Introdução

Neste capítulo é apresentado o protótipo de sistema de animação facial sincronizado com a fala, implementado com o propósito de validar os conceitos e desenvolvimentos discutidos em capítulos anteriores. Esta implementação piloto permite a geração de animações a partir de uma trilha de áudio contendo um locução. Além do áudio, o sistema possui como entrada a transcrição fonética temporizada da locução. O sistema foi utilizado para gerar animações faciais sincronizadas com a fala a partir tanto de fala natural quanto de fala sintética gerada por um conversor texto-fala.

Na Seção 7.2 tem-se a descrição da arquitetura do sistema piloto implementado. Na Seção 7.3 é discutida a conversão da descrição fonética de entrada na seqüência de visemas que comanda a animação facial. A Seção 7.4 aborda a estratégia adotada para a manipulação da geometria da face virtual. Na Seção 7.5 tem-se os comentários finais.

7.2 Arquitetura do sistema

A Figura 7.1 apresenta o diagrama de blocos do sistema implementado. O sistema é uma implementação piloto que objetiva validar os conceitos e desenvolvimentos discutidos em capítulos anteriores.

O sistema possui como entrada uma trilha de áudio e a transcrição fonética temporizada da locução contida nesta trilha. No protótipo implementado foi adotado o formato RIFF Waveform (RIFF, 1991) para a entrada de áudio. A transcrição fonética temporizada é composta pela seqüência de fones que compõem a locução e suas respectivas durações. Esta transcrição fonética temporizada pode ser gerada de forma manual, automática ou ainda por uma combinação destas duas técnicas. Na geração manual, um operador com conhecimentos fonéticos deve analisar, segmentar e rotular a seqüência

de fones que compõem a locução, identificando os instantes de ínício e a duração de cada fone. A geração automática baseia-se em algoritmos de reconhecimento de fala. Observa-se, entretanto, que os algoritmos existentes para o reconhecimento da fala contínua de vocabulário ilimitado e independente do locutor ainda necessitam de refinamentos para aumentar a confiabilidade (RABINER; JUANG, 1993) (YNOGUTI, 1999) (ALBRECHT; HABER; SEIDEL, 2002). Alternativamente a estes métodos, é possível gerar a transcrição fonética temporizada como subproduto do processo de síntese da fala a partir de texto. O sistema de animação facial implementado foi testado com fala natural, transcrita manualmente, assim como integrado a um conversor texto-fala capaz de gerar o áudio e a transcrição fonética temporizada a partir do texto.

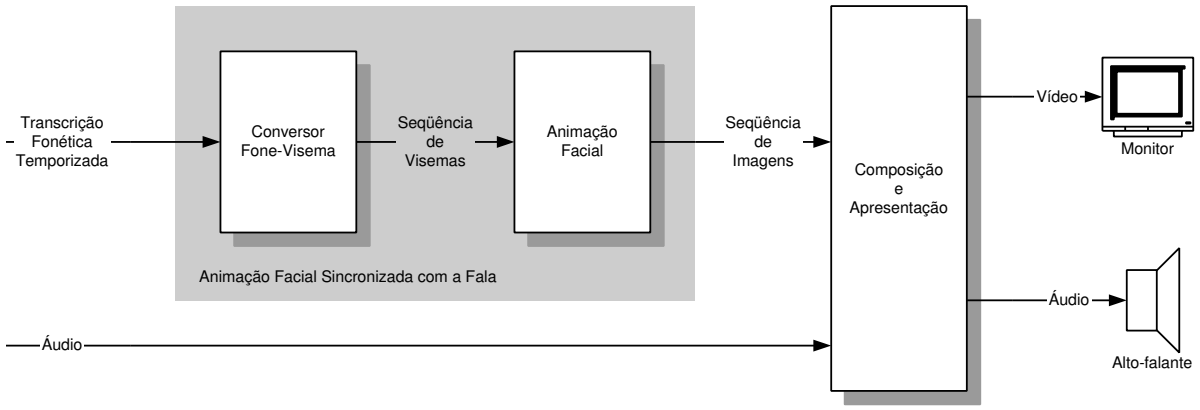

Fig. 7.1: Diagrama de blocos do sistema de animação facial.

Independentemente da estratégia utilizada para a geração do áudio e da transcrição fonética, é tarefa do sistema gerar a animação facial em sincronismo com a seqüência de fones que compõem a locução. O núcleo do sistema implementado, como mostrado na Figura 7.1 pela área mais escura, é formado pelos módulos *Conversor Fone-Visema* e *Animação Facial*. O módulo Conversor Fone-Visema é responsável pela conversão da descrição fonética temporizada em uma seqüência de visemas. Detalhes da implementação deste módulo são discutidos na Seção 7.3. O módulo de Animação Facial utiliza a seqüência de visemas gerada pelo Conversor Fone-Visema para controlar a movimentação da superfície da face virtual. A principal tarefa do sistema de animação facial é reproduzir na face virtual a movimentação articulatória descrita pelos visemas. Além disso, também está embutido no processamento associado a este módulo a geração da seqüência de imagens que

compõem a animação. O módulo de animação facial é discutido na Seção 7.4.

A etapa final do processamento envolve a composição do áudio e da seqüência de imagens em um vídeo sonorizado seguida de sua apresentação. Admitem-se duas grandes estratégias para esta etapa final: a composição "off-line" e posterior apresentação, ou a composição e apresentação simultaneamente aos processos de conversão e animação. Neste último caso, taxas de geração de imagens e latência apropriados que permitam que o áudio e o vídeo sejam percebidos de forma contínua, sem atrasos e descontinuidades incômodas, resultam em um sistema de animação facial sincronizado com a fala em tempo-real. No escopo deste trabalho, não foram tratados os problemas associados à geração em tempo-real. Na implementação realizada a composição é efetuada "off-line", utilizando um sistema de edição não-linear iFinish V60 - Media 100.

7.3 Conversor fone-visema

Considerando o conjunto de visemas estabelecido no Capítulo 5, Tabelas 5.7 e 5.26, foi adotado o seguinte procedimento para a conversão da seqüência de fones na seqüência de visemas que controla a animação facial:

```
(1)     WHILE existir proximo fone {
(2)         ler proximo fone (F = fone lido)
(3)         IF fone é vogal {
(4)             substituir F por fone vocálico representante (F = fone representante)
(5)             IF F == [i] {
(6)                 buscar fone anterior (F1 = fone anterior)
(7)                 buscar proximo fone (F2 = proximo fone)
(8)                 buscar visema associado ao contexto [F1 F F2]
(9)             } ELSE {
(10)                buscar visema associado a F
(11)            }
(12)        } ELSE IF fone é consoante {
(13)            substituir F por fone consonantal representante (F = fone representante)
(14)            buscar a ultima vogal ou silencio produzido antes de F (V1 = ultima vogal ou silencio)
(15)            buscar a primeira vogal ou silencio produzido após F (V2 = primeira vogal ou silencio )
(16)            IF V1 é vogal {
(17)                IF V1 é vogal postônica
(18)                    buscar e substituir V1 por vogal correspondente
(19)                IF V2 não é vogal postônica
(20)                    buscar e substituir V2 por vogal postônica correspondente
(21)            } ELSE IF V1 é silencio {
(22)                IF V2 é vogal postônica
(23)                    buscar e substituir V2 por vogal correspondente
(24)            }
(25)                buscar visema associado ao contexto [V1 C V2]
(26)        }
(27)    }
```

O procedimento percorre cada um dos fones da seqüência de entrada (linhas 1 e 2). Caso o fone a ser analisado seja uma vogal (linha 3), ele é substituído pelo fone representante de seu grupo de homofemas. Para esta substituição é utilizada a Tabela 7.1 (derivada da Tabela 5.26). Nesta tabela, a primeira coluna especifica o fone e a segunda o fone representante associado. Caso o fone representante seja [i], é efetuado o processamento necessário para o tratamento dos contextos [tit[e [ʃiʃ] (linhas 5 a 8). As buscas indicadas nas linhas 6 e 7 são efetuadas na seqüência de fones que compõem a locução. A busca do visema vocálico (linhas 8 e 10) é efetuada com o auxílio da Tabela 7.1.

Representante	**Contextos**	**Visema**
[i/	todos os contextos exceto [tit] e [ʃiʃ]	<i_1>
	[tit] [ʃiʃ]	<i_2>
[a]	todos os contextos	<a>
[u]	todos os contextos	<u>
[ɪ]	todos os contextos	<ɪ>
[ɐ]	todos os contextos	<ɐ>
[ʊ]	todos os contextos	<ʊ>

Tab. 7.1: Segmentos vocálicos, contextos fonéticos e respectivos visemas.

Caso o fone a ser analisado seja uma consoante (linha 12), é efetuada a substituição pelo fone representante de seu grupo de homofemas (linha 13), utilizando a Tabela 7.2 (derivada da Tabela 4.24). Na linha 14 é efetuada uma busca, para trás, na seqüência de fones de entrada até ser encontrado um segmento vocálico ou um silêncio. No contexto do procedimento, o silêncio é tratado como um fone. A busca da linha 14 identifica, portanto, o último segmento vocálico, ou silêncio, produzido antes do segmento consonantal em análise. Também é efetuada uma busca para frente para identificar o próximo segmento vocálico, ou silêncio, produzido após o segmento consonantal (linha 15). As buscas para trás e para frente permitem o tratamento dos contextos fonético: #CV, #CCV, VC#, VCV, VCCV, VCCCV e VCCCCV (confira Seção 3.3.1) . Estes contextos estão associados à ocorrência de encontros consonantais em uma palavra e entre palavras. Nos encontros consonantais são considerados apenas o efeitos da coarticulação na produção dos segmentos consonantais advindos dos segmentos vocálicos produzidos antes e após o encontro. Esta estratégia considera que os efeitos da coarticulação perseveratória e antecipatória entre as consoantes dos encontros consonantais são menos visíveis que os efeitos de coarticulação dos segmentos vocálicos que antecedem e sucedem o encontro consonantal.

O processamento realizado nas linhas 16 a 24 é necessário para associar as diferentes possibilidades de produção fonética ao conjunto de visemas disponível. A estratégia adotada consistiu em

substituir os segmentos vocálicos postônicos pelos não postônicos e vice-versa. Tal estratégia forma os contextos #CV e VC# com segmentos não postônicos, e os contextos VCV com o primeiro segmento não postônico e o segundo postônico. Este procedimento faz-se necessário uma vez que a busca do visema realizado na linha 15 utiliza a Tabela 7.3 derivada da Tabela 5.7. Os visemas associados aos contextos VC#, onde C={[ɾ, s]}, foram estimados a partir dos visemas associados aos diferentes contextos analisados para estas consoantes. Estes fones apresentam dois visemas distintos, sendo que um deles está fortemente associado à coarticulação com o segmento [u]. Assim, foi feita a associação dos contextos [uɾ] e [us] aos visemas associados à coarticulação com [u]. Os contextos [aɾ, iɾ] e [as, is] foram associados ao outro visema.

Segmento	Representante
[p]	[p]
[b]	[p]
[m]	[p]
[f]	[f]
[v]	[f]
[t]	[t]
[d]	[t]
[n]	[t]
[s]	[s]
[z]	[s]

Segmento	Representante
[l]	[l]
[ʃ]	[ʃ]
[ʒ]	[ʃ]
[ʎ]	[ʎ]
[ɲ]	[ʎ]
[k]	[k]
[g]	[k]
[ɾ]	[ɾ]
[ɣ]	[ɾ]

Tab. 7.2: Segmentos consonantais e respectivos representantes.

Para ilustrar, são apresentados a seguir a transcrição fonética e a respectiva temporização da locução de uma lista de frases foneticamente balanceada, que reproduz a estatística de ocorrência dos fones do português (ALCAIM; SOLEWICZ; MORAES, 1992). A segmentação fonética e a temporização foram realizadas utilizando o conversor texto-fala Aiuruetê (BARBOSA et al., 1999). A informação de temporização apresentada está associada à produção das frases em uma seqüência contínua, sendo contabilizados os silêncios produzidos entre as frases e nos instantes inicial e final da produção do conjunto total de frases. A informação da temporização dos fones indica a duração de cada fones em microsegundos. Na transcrição em visemas, o símbolo "/" representa o visema associado ao silêncio, quando a face encontra-se na posição de repouso. Na informação de temporização o símbolo "-" é utilizado para separar as várias frases e facilitar a associação com os fones da transcrição fonética. Deve-se observar ainda que após o símbolo "-" tem-se a duração do silêncio que separa duas frases consecutivas.

Representante	Contextos	Visemas
[p]	[pi] [pa] [ipɪ] [ipɐ] [ipʊ] [apɪ] [apɐ] [apʊ] [upɐ]	<p_1>
	[pu] [upɪ] [upʊ]	<p_2>
[f]	[fi] [fa] [ifɪ] [ifɐ] [ifʊ] [afɪ] [afɐ]	<f_1>
	[fu] [afʊ] [ufɪ] [ufɐ] [ufʊ]	<f_2>
[t]	[ti] [tu] [itɪ] [itɐ] [itʊ] [atɪ] [atʊ] [utɪ] [utɐ] [utʊ]	<t_1>
	[ta] [atɐ]	<t_2>
[s]	[si] [sa] [isɪ] [isɐ] [asɪ] [asɐ] [is] [as]	<s_1>
	[su] [isʊ] [asʊ] [usɪ] [usɐ] [usʊ] [us]	<s_2>
[l]	[li] [ilɪ] [alʊ] [ulɪ] [ulɐ]	<l_1>
	[la] [ilɐ] [alɪ] [alɐ]	<l_2>
	[lu]	<l_3>
	[ilʊ] [ulʊ]	<l_4>
[ʃ]	[ʃi] [ʃa] [iʃɪ] [iʃɐ] [iʃʊ] [aʃɪ] [aʃɐ] [aʃʊ] [uʃɪ] [uʃɐ]	<$ʃ_1$>
	[ʃu] [uʃʊ]	<$ʃ_2$>
[ʎ]	[ʎi] [ʎa] [iʎɪ] [iʎɐ] [aʎɪ] [aʎɐ]	<$ʎ_1$>
	[ʎu] [uʎɪ] [uʎɐ]	<$ʎ_2$>
	[iʎʊ] [aʎʊ] [uʎʊ]	<$ʎ_3$>
[k]	[ki] [ikɪ] [ikɐ] [akɪ] [ukɪ] [ukɐ]	<k_1>
	[ka] [akɐ]	<k_2>
	[ku] [ikʊ] [akʊ] [ukʊ]	<k_3>
[ɾ]	[ɾi] [ɾa] [iɾɪ] iɾɐ] [aɾɪ] [aɾɐ] [uɾɐ] [iɾ] [aɾ]	<$ɾ_1$>
	[ɾʊ] [iɾʊ [aɾʊ] [uɾɪ] [uɾʊ] [uɾ]	<$ɾ_2$>

Tab. 7.3: Segmentos consonantais, contextos fonéticos e respectivos visemas.

- Representação ortográfica das frases:
  "*Nosso telefone quebrou.*
  Desculpe se magoei o velho.
  Queremos discutir o orçamento.
  Ela tem muita fome.
  Uma índia andava na mata.
  Zé, vá mais rápido!
  Hoje dormirei bem.
  João deu pouco dinheiro.
  Ainda são seis horas.
  Ela saía discretamente."

- Transcrição fonética:
  [nɔsʊtelefonɪkebroʊ
  deskuʊpɪsimagoeɪuvɛlʊ
  keremʊsdiskutiruorsamẽtʊ
  ɛlɐtẽmuĩtɐfomɪ
  umɑ̃ĩdɾɐẽdavɐnamatɐ
  zɛ vamaɪsrapɪdʊ
  oʒɪdormireɪbẽ
  ʒoẽʊdeʊpoʊkʊdiɲeɪɾʊ
  aĩdɐsẽʊseɪsɔɾɐs
  ɛlɐsaiɐdiskɾɛtamẽtɪ]

- Temporização dos fones:
  276750 56375 131812,5 91937,5 54687,5 79500 112437,5 43812,5 105250 73375 119625 26562,5 43250 81250 122125 49437,5 51500 69500 69437,5 - 530750 59500 99125 58125 73250 104500 31750 63937,5 56000 104312,5 107187,5 59062,5 117937,5 62312,5 107500 69937,5 55812,5 102687,5 54187,5 130312,5 55750 73500 - 521937,5 75687,5 119375 60312,5 97625 56000 54375 43875 66625 76625 58125 73250 97562,5 72437,5 87125 57062,5 118437,5 109750 37875 79312,5 142937,5 60875 112562,5 78000 74312,5 - 527250 129375 47437,5 44125 101687,5 192375 67375 75312,5 111812,5 61500 64437,5 81250 143750 48000 62500 - 524687,5 125937,5 37687,5 62687,5 124937,5 53812,5 86312,5 28187,5 97687,5 63125 124062,5 43812,5 55812,5 49812,5 152000 57062,5 147125 74062,5 62500 - 529500 87812,5 177875 101000 74875 172937,5 70750 100125 74375 72000 91625 146562,5 74687,5 48062,5 44375 71187,5 - 534437,5 133812,5 53500 42250 53812,5 105125 53875 55562,5 83812,5 54937,5 68000 54500 68125 184562,5 - 528750 54250 107375 69312,5 69250 68625 92625 67000 77312,5 95250 74125 56750 48437,5 62875 72250 83562,5 65437,5 54000 61000 64187,5 - 532687,5

134687,5 118125 56062,5 60000 105625 96750 96750 110500 109125 88000 74937,5 133375 54500 59750 79937,5 - 525250 117625 43250 61937,5 83750 149000 75500 50937,5 52625 92875 58250 69750 59625 115375 68125 135500 52437,5 129437,5 55812,5 64625 246312,5

- Transcrição em visemas:
  </t2ɐs2ʊt1ɐl2ɐf2ut1ɪk1ɐp1r2uʊ/
  t2ɐs2k3uʊp2ɪs1i1p1ak3uɐɪuf2aʎ3ʊ/
  k2ɐr1ɐp1ʊs2t1i1s2k3ut1i1r2uur2s2ap1ɐt1ʊ/
  al2ɐt2ɐp1uɪt1 ɐf2up2ɪ/
  up1ɐi1t1ɪɐat2af1ɐt2ap1at2ɐ/
  s1a/f1ap1aɪs1r1ap1ɪt1ʊ /
  uʃ1ɪt1ur2p2i1r1ɐɪp1ɐ/
  ʃ2uaʊt1ɐʊp2uʊk3ʊt1i1ʎ1ɐɪr2ʊ/
  ai1t1ɐs1aʊs2ɐɪs1ɐr1ɐs1/
  al2ɐs1ai1ɐt1i1s1k1r1at2ap1ɐt1ɪ/>

7.4 Animação facial

No sistema implementado foi utilizado um modelo geométrico de cabeça virtual adaptado do modelo MiraFace desenvolvido na Universidade de Genebra, cedido à International Organization for Standardization-ISO e por esta instituição publicado como parte do software de referência do padrão MPEG-4. As superfícies que compõem a cabeça são representadas por malhas de triângulos. As Figuras 7.2 e 7.3 apresentam o modelo utilizado.

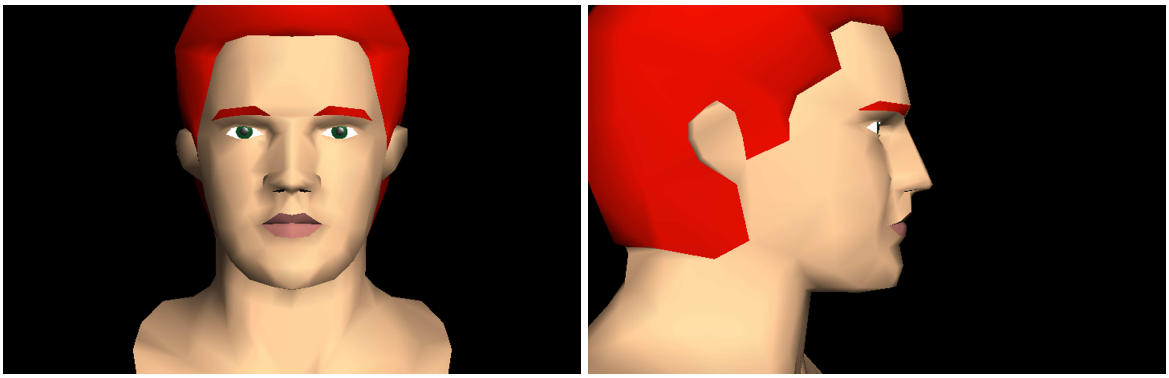

Fig. 7.2: Modelo da cabeça virtual - visualização tonalizada.

A animação da face virtual é controlada pela seqüência de visemas gerados pelo módulo Conversão Fone-Visema discutido na Seção 7.3. Os visemas e respectivas temporizações permitem o

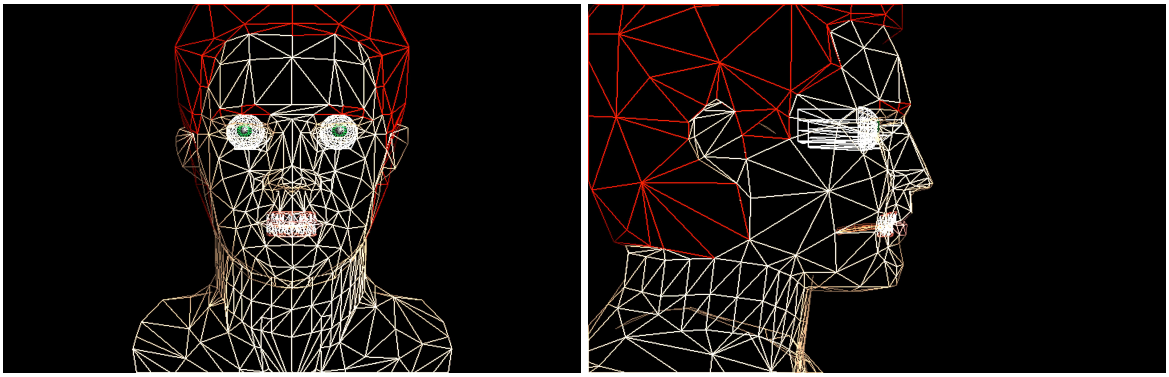

Fig. 7.3: Modelo da cabeça virtual - visualização aramada.

cálculo da trajetória dos pontos da face virtual correspondentes aos pontos de interesse, conforme discutido no Capítulo 5. Para utilizar na face virtual a representação paramétrica dos visemas descrita na Seção 5.2.8, os deslocamentos do pontos de interesse apresentados nas tabelas da Seção 5.3 foram multiplicadas por um fator de escala proporcional à razão entre as dimensões da cabeça do informante e a do modelo geométrico. O fator de escala, fs, adotado é a razão entre a largura da boca do modelo e do informante definida por:

$$fs = \frac{\left|\overline{P}_2\left(t_0\right) - \overline{P}'_2\left(t_0\right)\right|_v}{\left|\overline{P}_2\left(t_0\right) - \overline{P}'_2\left(t_0\right)\right|_r} \tag{7.1}$$

onde:

- $\left|\overline{P}_2\left(t_0\right) - \overline{P}'_2\left(t_0\right)\right|_v$ é a distância entre os P_2 e P'_2 localizados nos cantos da boca (Figura 7.4) da face virtual na posição de repouso (instante t_0);

- $\left|\overline{P}_2\left(t_0\right) - \overline{P}'_2\left(t_0\right)\right|_r$ é a distância entre os P_2 e P'_2 localizados nos cantos da boca (Figura 7.4) da face do informante na posição de repouso (instante t_0). Esta distância é igual a 57,7 mm para o informante deste trabalho.

A manipulação da geometria da face virtual foi desmembrada em dois procedimentos. O primeiro envolve as transformações de corpo rígido associadas ao comportamento da articulação temporomandibular discutidas na Seção 6.2. O segundo trata das deformações da superfície da face virtual derivada da movimentação do tecido da face ocasionada pela atuação dos músculos faciais. A Seção 7.4.1 apresenta o procedimento adotado para reproduzir na cabeça virtual a movimentação da articulação temporomandibular e, conseqüentemente, da mandíbula. Na Seção 7.4.2 é apresentada a estratégia implementada para aproximar a movimentação do tecido facial ao redor da boca durante a fala.

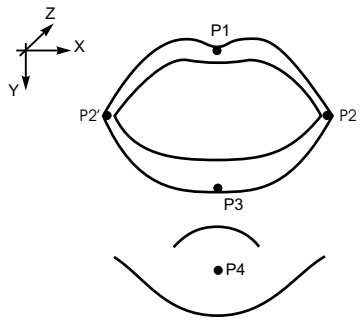

Fig. 7.4: Pontos de interesse estendido.

7.4.1 Comportamento da articulação temporomandibular

As transformações de corpo rígido de rotação e translação que descrevem o comportamento da junta temporomandibular apresentadas na Seção 6.2 são aplicadas aos vértices dos triângulos que compõem o modelo geométrico da face virtual. O vértices transformados são aqueles localizados ao longo do osso da mandíbula, mais precisamente, na região abaixo e incluindo o lábio inferior e a região lateral de cada lado da face abaixo do plano imaginário que contém o eixo da ATM e o canto do lábio. O limite inferior desta região é definido pelo pescoço. A Figura 7.5 ressalta, através de pontos brancos, os vértices manipulados pela movimentação da mandíbula. O eixo de rotação da ATM é representado na figura pelo cilindro que transpassa a cabeça em frente ao ouvido. A transformação de rotação é efetuada em torno deste eixo, de ângulo dado por

$$\Delta\theta_4\left(t\right) = \theta_4\left(t\right) - \theta_4\left(t_0\right) \tag{7.2}$$

onde:

- $\theta_4\left(t\right)$ é o ângulo de rotação de P_4 no instante t dado pela Equação 6.9;
- $\theta_4\left(t_0\right)$ é o ângulo de rotação de P_4 em repouso dado pela Equação 6.9.

A transformação de translação aplicada aos vértices é dada pela translação da ATM no plano sagital descrita pela Equação 6.7.

7.4.2 Comportamento dos lábios inferior e superior

A movimentação voluntária do tecido facial ao redor da boca, incluindo os lábios, não relacionados diretamente à rotação da mandíbula, é derivada dos comportamentos dos lábios superior e inferior apresentados nas Seções 6.4 e 6.3. A estratégia de deformação a ser aplicada na face sintética para reproduzir esta movimentação foi norteada por três considerações. Primeiro, a movimentação

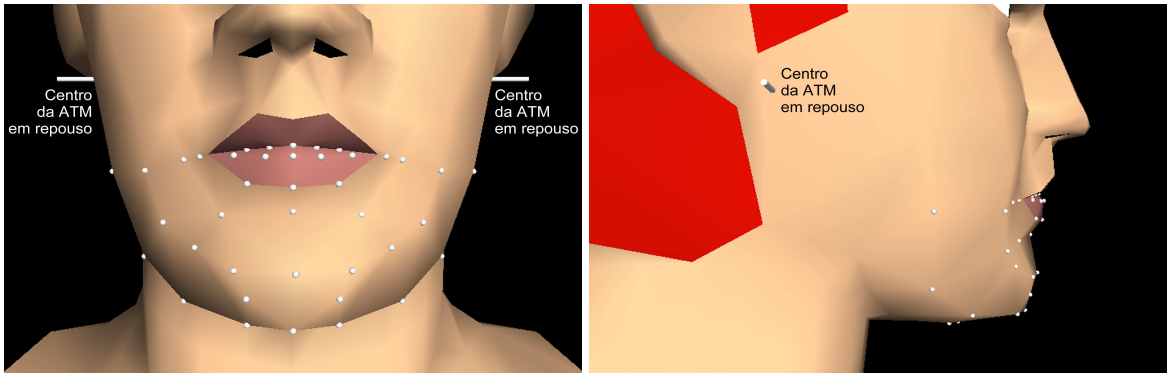

Fig. 7.5: Vértices transformados devido à movimentação da articulação temporomandibular.

dos pontos na face virtual correspondentes aos pontos de interesse, P_1, P_2 (P'_2), P_3 e P_4, deve seguir as trajetórias modeladas (Seções 6.4 e 6.3). Segundo, durante a fala, o tecido facial ao redor da boca, incluindo os lábios, sofre deformações atribuídas principalmente ao comportamento do músculo *Orbicular da Boca*. Terceiro, outros músculos que também influenciam a pele ao redor da boca estão distribuídos assimetricamente em relação ao plano transversal, uma vez que diferentes grupos de músculos se inserem no tecido da região superior e inferior da boca. A Figura 7.6 apresenta a localização e forma dos principais músculos ao redor da boca.

Baseado nas considerações mencionadas, foi estabelecido um modelo para expressar a movimentação visível do tecido facial ao redor da boca durante a fala. Este modelo estabelece uma região de influência esferoidal ao redor da boca, inspirada na constituição elíptica do músculo Orbicular da Boca (PARKE; WATERS, 1996). Adicionalmente, para acomodar a assimetria da distribuição muscular ao redor da boca em relação ao plano transversal, a região de influência foi dividida em duas: uma região superior, influenciada pelo comportamento do lábio superior(Seção 6.4); e uma região inferior, influenciada pelo comportamento do lábio inferior(Seção 6.3). Na estratégia adotada, cada uma destas regiões é definida por dois esferóides concêntricos. O esferóide externo, o qual é uma versão em escala maior do interno, define o limite de influência do comportamento do lábio associado. O esferóide interno define os pontos de máxima influência do comportamento. Esta influência é atenuada ao se afastar da superfície do esferóide interno, cessando totalmente na superfície do esferóide externo.

O esferóide interno, assumindo um sistema cartesiano de referência centrado na boca, com a mesma orientação da Figura 7.4, é dado por:

$$\frac{x^2}{a^2} + \frac{y^2}{b_i^2} + \frac{z^2}{b_i^2} = 1 \quad i = 1,3 \tag{7.3}$$

Os valores dos coeficientes a e b_i, $i = 1,3$ ($i = 1$ região superior; $i = 3$ região inferior) são influenciados pela geometria da face virtual. As seguintes considerações definem estes coeficientes:

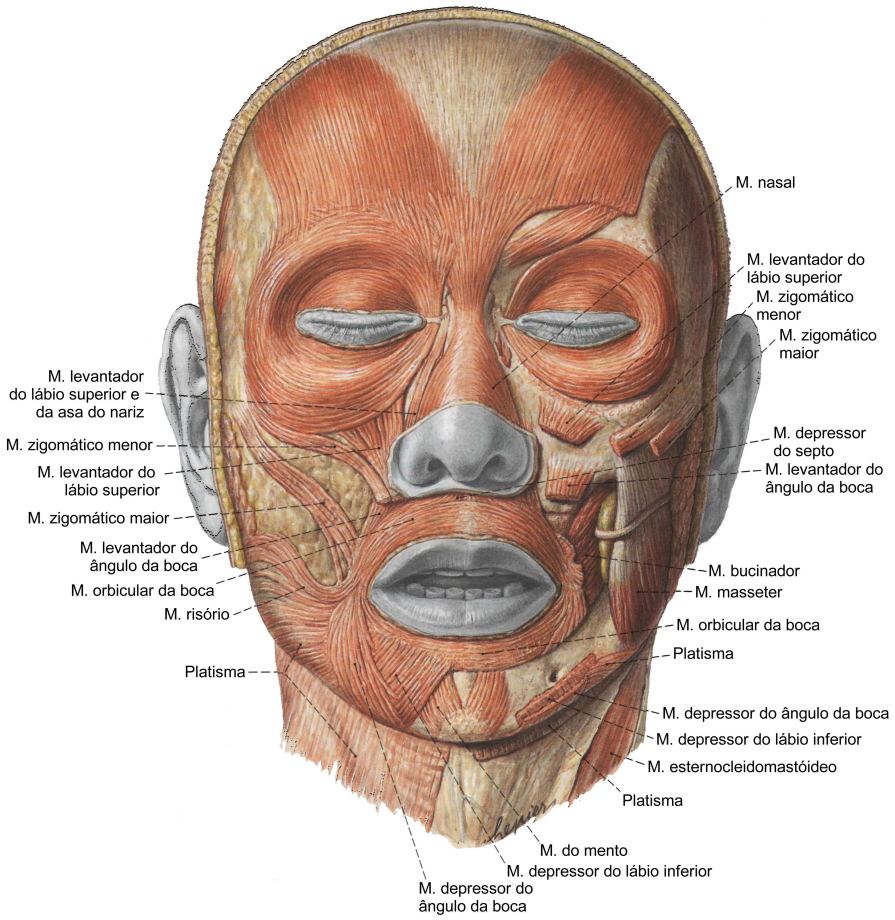

Fig. 7.6: Músculos da face - ilustração adaptada de Sobotta (1990)).

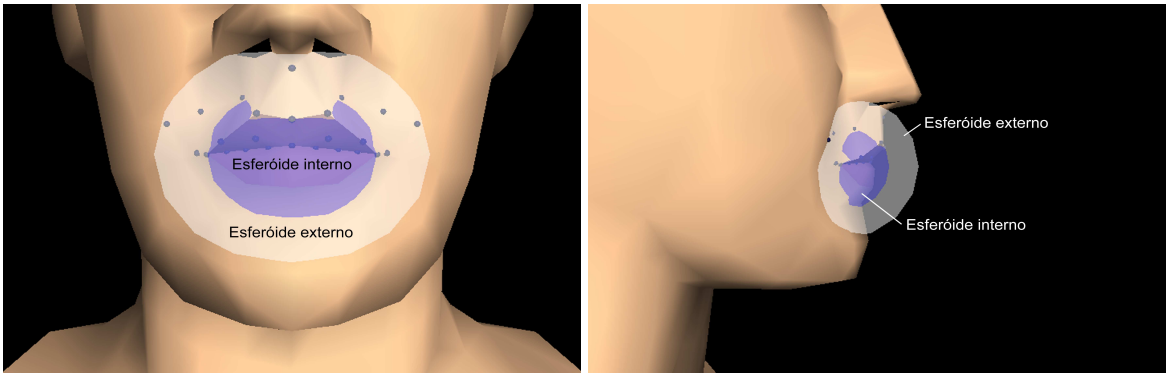

Fig. 7.7: Região de influência do comportamento do lábio superior.

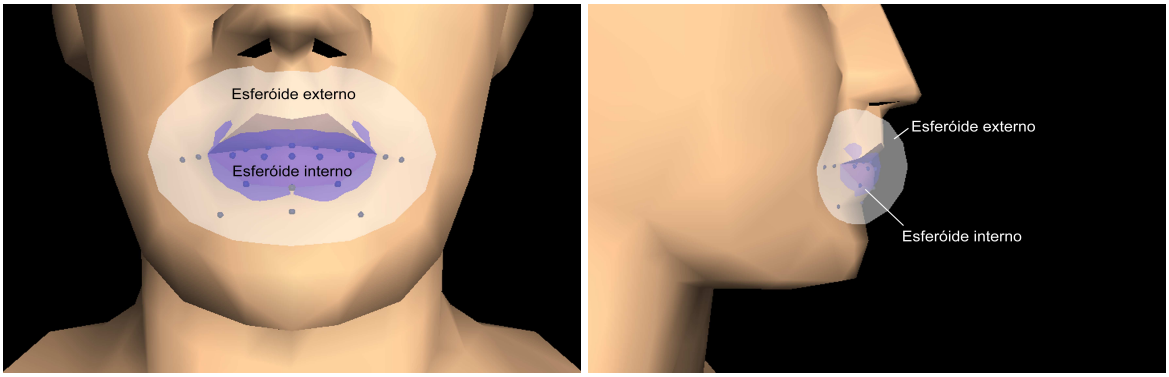

Fig. 7.8: Região de influência do comportamento do lábio inferior.

1. Os pontos nos cantos da boca (P_2 e o simétrico P'_2 - Figura 7.4) são os vértices do eixo maior do esferóide interno, ou seja, a distância entre P_2 e P'_2 é $2a$;

2. O ponto de interesse P_i está localizado na superfície do esferóide interno, ou seja, b_i é a distância entre o eixo maior do esferóide e o ponto de interesse P_i ($i = 1$ região superior, $i = 3$ região inferior).

O esferóide externo é definido por:

$$\frac{x^2}{a^2} + \frac{y^2}{b_i^2} + \frac{z^2}{b_i^2} = F_i^2 \quad i = 1, 3 \tag{7.4}$$

O fator de escala F_1 (região superior) é definido como a distância entre o lábio superior e a base do nariz. Para a região inferior, o fator F_3 limita a região de influência ao ponto médio entre o começo do queixo e a ponta deste. Esta última definição garante que parte da pele que recobre o queixo, sem ultrapassar a ponta deste, sofra deformação. A Figura 7.8 apresenta os esferóides que definem a região de influência do comportamento do lábio inferior e os vértices dos triângulos da face virtual que são afetados por este comportamento. De forma análoga, a Figura 7.7 apresenta os esferóides e os vértices associados ao comportamento do lábio superior. Os vértices que sofrem deformação são identificados nas figuras pelas esferas na superfície da face e dos lábios.

O deslocamento $\Delta\overline{V}$ de um vértice dentro da região de influência é calculado pela seguinte relação:

$$\Delta\overline{V} = R_i \left[K_i \, \Delta\overline{P}_2 + (1 - K_i) \, \Delta\overline{P}_i \right] \quad i = 1, 3 \tag{7.5}$$

onde:

- K_i, com $0 \leq K_i \leq 1$, é um fator de interpolação que pondera as distâncias entre o vértice e os pontos P_2 e P_i (i = 1 região superior, i = 3 região superior);

- R_i, com $0 \leq R_i \leq 1$, é um fator de decaimento esferoidal que leva em conta a distância do vértice ao esferóide interno.

O fator de interpolação K_i é dado por:

$$K_i = \left[cos\left(\frac{d_2}{d_2 + d_i}\pi\right) + 1 \right] /2 \tag{7.6}$$

onde:

- d_j, $j = 1, 2, 3$, é a distância entre o vértice e o ponto de interesse P_j na posição de repouso.

O fator de decaimento R_i depende se o vértice está dentro ou fora do esferóide interno, sendo dado por

$$\begin{cases} R_i = cos\left((1 - S_i)\ (\pi/2)\right) & \text{dentro (incluindo a superfície)} \\ R_i = cos\left[((S_i - 1)\ /\ (F_i^2 - 1))\ (\pi/2)\right] & \text{fora} \end{cases} \tag{7.7}$$

O fator S_i, que atenua R_i em função da distância do vértice à superfície do esferóide interno, é dado por:

$$S_i = \frac{x^2}{a^2} + \frac{y^2}{b_i^2} + \frac{z^2}{b_i^2} \quad i = 1, 3 \tag{7.8}$$

Para ilustrar o efeito do comportamento do lábio superior e inferior, a Figura 7.9 apresenta a simulação do deslocamento como modelado acima, aplicado a uma grade regular planar. A localização dos pontos de interesse também é apresentada na figura.

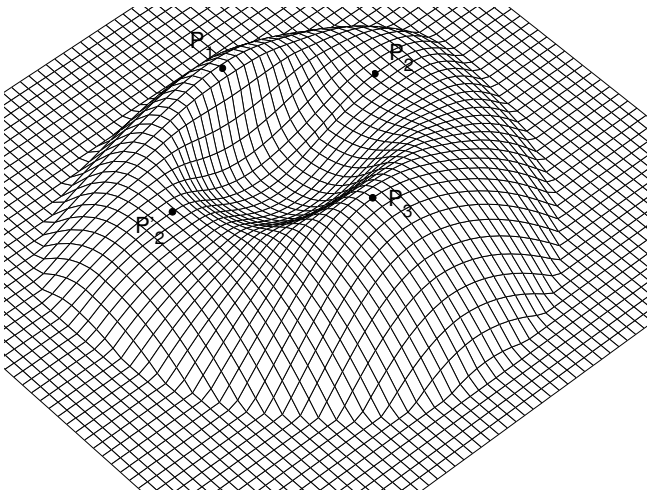

Fig. 7.9: Efeito da estratégia de deformação.

7.5 Comentários Finais

O sistema implementado anima a face virtual baseado em uma estratégia puramente geométrica. Diferentemente de outros sistemas que também utilizam tal filosofia como Parke (1982) e Kalra et al. (1992), o sistema implementado baseia-se em conjunto de visemas derivados da análise da produção de fala. Também é importante destacar o tratamento mais acurado da movimentação da mandíbula que foi negligenciado nos sistemas mencionados.

Outro aspecto inovador da solução desenvolvida é o tratamento dos efeitos da coarticulação com o auxílio do conceito de visemas dependentes do contexto fonético. Tradicionalmente as soluções adotadas, tais como Öhman (1967), Pelachaud (1991) e Cohen e Massaro (1993), procuram tratar a coarticulação através da deformação de visemas "puros" independentes do contexto. Em particular, a solução proposta em Cohen e Massaro (1993), baseada no conceito de função de dominância, deixa para o animador a árdua tarefa de especificar tais funções.

Os vídeos demonstrativos que acompanham este trabalho foram gerados utilizando o sistema de animação facial discutido acima e servem como ponto de partida para a avaliação do potencial da solução proposta neste trabalho.

Capítulo 8

Conclusão

A animação facial possui uma ampla gama de aplicações abrangendo cinema, jogos eletrônicos e educação. Além disso, constitui-se em uma alternativa promissora para a compressão de dados em ambientes de vídeo conferência. A animação facial é uma tecnologia estratégica para a construção de novas interfaces que suportem de maneira natural, amigável e divertida, a interação homem-máquina. Espera-se que no futuro, agentes virtuais personificados sejam incorporados em máquinas de atendimento automático, em computadores pessoais e em eletrodomésticos. O aperfeiçoamento das tecnologias de animação facial, síntese e reconhecimento da fala, assim como da inteligência artificial, são necessárias para viabilizar este futuro, talvez não tão distante.

O presente trabalho procurou contribuir com o desenvolvimento das técnicas associadas à animação facial. Dentre os vários aspectos que compõem o complexo problema da produção de animações faciais realistas, o trabalho concentrou-se no tratamento da movimentação articulatória visível. Visema é o conceito fundamental para este tipo de animação. Visema é o padrão de movimentação articulatória visível associado a um ou mais segmentos da fala. Os visemas estão sujeitos aos efeitos da coarticulação. Tradicionalmente, procura-se modelar os efeitos da coarticulação através da fusão de visemas que independem do contexto fonético de sua produção. Visemas independentes do contexto fonético são visemas "puros" que representam uma idealização da movimentação associada a produção do segmento. Supõe-se nesta idealização que o segmento é produzido de forma isolada, sem sofrer interferência da movimentação associada à produção de segmentos próximos na seqüência que define a locução. Tanto a definição dos visemas "puros", quanto a caracterização da função de fusão de visemas, exigem um criterioso trabalho de coleta e análise de material lingüístico. Entretanto, atualmente não há uma concordância consolidada dos pesquisadores na área sobre a forma, dinâmica e peso da função de fusão. Neste contexto, destacam-se dois modelos que têm sido explorados com mais freqüência pelos pesquisadores: o proposto por Öhman (1967) e o proposto por Cohen e Massaro (1993).

A abordagem do presente trabalho perseguiu uma linha alternativa e inovadora: visemas dependentes do contexto fonético. O conceito central é a caracterização de um visema, não como a representação visual da produção de um segmento isolado, mas levando em consideração as interferências dos segmentos que o antecedem e sucedem na seqüência da locução. Em particular, a análise efetuada no trabalho abrigou os contextos trifônicos CVC e VCV.

No presente trabalho foi elaborada uma metodologia para a medida e caracterização de visemas dependentes do contexto fonético. Esta caracterização procura identificar padrões visíveis de movimentação articulatória durante a produção dos fonemas da língua em diversos contextos fonéticos. Para tanto, foram efetuadas medidas da trajetória de pontos visíveis ao redor da boca e no queixo de um falante durante a produção de um conjunto de logatomas. A identificação dos padrões visíveis baseou-se na identificação de similaridade utilizando o algoritmo de agrupamento *K-Means*.

A caracterização de um conjunto de visemas para o português do Brasil foi realizada utilizando corpus formado por logatomas paroxítonos do tipo 'CVCV. O tratamento do corpus adotado permitiu a análise de doze contextos fonéticos diferentes para nove consoantes, dezesseis contextos para os três segmentos vocálicos postônicos e doze contextos para os segmentos vocálicos extremos /i, a, u/. Os visemas determinados são expressos por um modelo paramétrico constituído de alvos articulatórios, pelo instante relativo de realização destes alvos dentro do intervalo da produção acústica e por um mecanismo de interpolação entre os alvos de segmentos adjacentes na cadeia da produção da fala.

A representação dos visemas é utilizada por um conjunto de modelos geométricos para manipular uma face virtual, reproduzindo nesta a movimentação da mandíbula e dos lábios. Este modelos geométricos também constituem uma contribuição do presente trabalho. Os modelos desenvolvidos descrevem características marcantes da movimentação articulatória visível, tais como protrusão, rotação e translação da articulação temporomandibular (ATM). Um aspecto importante que norteou a modelagem foi a preocupação em separar os efeitos da movimentação da articulação temporomandibular dos da movimentação voluntária do tecido labial. O modelo desenvolvido representa o comportamento da ATM através de transformações de rotação e translação. A estimativa do comportamento da ATM permitiu, por sua vez, o cálculo da movimentação voluntária do tecido dos lábios, principalmente do lábio inferior. A movimentação da mandíbula e dos pontos dos lábios descritas pelos modelos estão em concordância com o comportamento qualitativo aceito e consolidado pelo conhecimento fonético.

Tanto as transformações de rotação e translação associadas à ATM como a movimentação do tecido facial ao redor da boca, incluindo os lábios, devem ser incorporadas e implementadas na animação facial. O mapeamento dos modelos em estratégias de manipulação da geometria tridimensional de uma face virtual admite variantes, podendo ir desde uma solução puramente geométrica, como a adotada, até uma sofisticada modelagem física que leve em consideração a estrutura e características biomecânicas da face e do tecido que a recobre. Não obstante, independente da estratégia de mani-

pulação da geometria da face virtual, os modelos desenvolvidos no presente trabalho reproduzem aspectos da movimentação articulatória visível, tais como abertura da boca, protrusão, arredondamento e extensão labiais, observados em faces reais, estabelecendo uma sólida base para refinamentos futuros. Neste sentido, o sistema piloto implementado permitiu validar os conceitos e resultados gerados durante o trabalho, além de estabelecer uma plataforma inicial de avaliação e desenvolvimento.

Como desenvolvimento futuro, está planejado experimento de avaliação do impacto da face virtual na inteligibilidade da fala, principalmente em situação de deterioração do sinal de áudio por ruído. Adicionamente pretende-se melhorar a qualidade da animação gerada. Já se encontra em desenvolvimento a implementação de um modelo biomecânico de manipulação da face, inspirado na abordagem de Terzopoulos e Waters (1990). A utilização de textura também faz parte de desenvolvimentos a serem realizados em um futuro próximo. A observação, tanto das imagens de vídeo do informante, como da animação gerada, revela uma rápida movimentação associada a determinados fonemas, como por exemplo /p/. Tal observação enseja a incorporação de "motion blur" (POTMESIL; CHAKRAVARTY, 1983) no processo de geração das imagens da animação para reproduzir com maior realismo os efeitos desta rápida movimentação articulatória. A reprodução de emoções, sinais conversacionais e movimentação fisiológica são outras linhas de desenvolvimento a serem perseguidas.

Há ainda muito a fazer para aumentar o grau de realismo da face virtual. Assim, o presente trabalho representa, no contexto local, o estabelecimento de uma linha de pesquisa, fértil e promissora, voltada ao estabelecimento de soluções que reproduzam as características lingüísticas e culturais da comunicação face-a-face praticada no país, contribuindo também, em caráter mais amplo, nos esforços de busca de soluções para a animação facial, que representem de forma realista a movimentação articulatória visível, incluindo os efeitos da coarticulação.

Referências Bibliográficas

ALBRECHT, I.; HABER, J.; SEIDEL, H.-P. Speech synchronization for physics-based facial animation. In: SKALA, V. (Ed.). *Proceedings of the 10th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision - WSCG'2002*. Plzen, Czech Republic: UNION Agency, 2002. p. 9–16.

ALCAIM, A.; SOLEWICZ, J. A.; MORAES, J. A. de. Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no Rio de Janeiro. *Revista da Sociedade Brasileira de Telecomunicações*, v. 7, n. 1, p. 23–41, dezembro 1992.

ALEXA, M.; BEHR, J.; MüLLER, W. The morph node. In: *The Web3D-VRML 2000 Conference*. Monterey, CA, USA: [s.n.], 2000. p. 29–34.

AYACHE, N. *Artificial Vision for Mobile Robots: Stereo Vision and Multisensory Perception*. [S.l.]: Edinburgh University Press, 1991.

BADIN, P. et al. Three-dimensional linear articulatory modeling of tongue, lips and face, based on mri and video images. *Journal of Phonetics*, v. 30, n. 3, p. 533–553, July 2002.

BAILLY, G. Audiovisual speech synthesis. In: *Proceedings of ESCA European Tutorial and Research Workshop on Speech Synthesis*. Perthshire, Scotland: ESCA, 2001. p. 1–10.

BARBOSA, P. A.; ALBANO, E. C. Illustrations of the IPA - Brazilian Portuguese. *Journal of the International Phonetic Association*, v. 34, n. 02, p. 227–232, December 2004.

BARBOSA, P. A. et al. Aiuruetê: A high-quality concatenative texto-to-speech system for Brazilian Portuguese with demisyllabic analysis-based units and a hierarchical model of rhythm production. In: *Proceedings of the 6th European Conference on Speech Communication and Technology - Eurospeech'99*. Budapest, Hungary: ISCA, 1999. p. 2059–2062.

BELL-BERTI, F.; HARRIS, K. S. Anticipatory coarticulation: Some implications from a study of lip rounding. *Journal of the Acoustical Society of America*, v. 65, n. 5, p. 449–454, May 1979.

BELL-BERTI, F.; HARRIS, K. S. Temporal patterns of coarticulation: Lip rounding. *Journal of the Acoustical Society of America*, v. 71, n. 2, p. 449–454, February 1982.

BENGUEREL, A.-P.; COWAN, H. A. Coarticulation of upper lip protrusion in French. *Phonetica*, v. 30, p. 41–55, 1974.

BENGUEREL, A.-P.; PICHORA-FULLER, M. K. Coarticulation effects in lipreading. *Journal of Speech and Hearing Research*, v. 25, p. 600–607, December 1982.

BENOÎT, C.; Le Goff, B. Audio-visual speech synthesis from french text: Eight years of models, designs and evaluation at the ICP. *Speech Communication*, v. 26, n. 1-2, p. 117–129, October 1998.

BENOÎT, C. et al. A set of French visemes for visual speech synthesis. In: BAILLY, G.; BENOîT, C.; SAWALLIS, T. R. (Ed.). *Talking Machines: Theories, Models and Designs*. North Holland, Amsterdam: Elsevier Science, 1992. p. 485–504.

BESKOW, J. Rule-based visual speech synthesis. In: *Proceedings of the 4th European Conference on Speech Communication and Technology - EUROSPEECH'95*. Madrid, Spain: ESCA, 1995. p. 299–302.

BESKOW, J. Animation of talking agents. In: BENOîT, C.; CAMPBELL, R. (Ed.). *Proceedings of the ESCA/ESCOP Workshop on Audio-Visual Speech Processing - AVSP'97*. Rhodes, Greece: [s.n.], 1997. p. 149–152.

BESKOW, J. *Talking Heads: Models and Applications for Multimodal Speech Synthesis*. Tese (Doutorado) — Departament of Speech, Music and Hearing - Royal Institute of Technology, 2003.

BESKOW, J. Trainable articulatory control models for visual speech synthesis. *International Journal of Speech Technology*, v. 7, n. 4, p. 335–349, October 2004.

BINNIE, C. A.; JACKSON, P.; MONTGOMERY, A. Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation. *Journal of Speech and Hearing Disorders*, v. 41, p. 530–539, 1976.

BONDY, M. D. et al. Model-based face and lip animation for interactive virtual reality applications. In: *Proceedings of the 9th ACM international conference on Multimedia 2001*. Ottawa, ON, Canada: ACM Press, 2001. p. 559–563.

BREGLER, C.; COVELL, M.; SLANEY, M. Video rewrite: driving visual speech with audio. In: *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997. p. 353–360.

BUTTFIELD, A. A new approach to rapid image morphing for lip motion synthesis. In: OUDSHOORN, M. J. (Ed.). *26th Australasian Computer Science Conference - ACSC'03*. Adelaide, South Australia: Australian Computer Society, 2003. p. 79–86.

CAGLIARI, L. C. *Elementos de Fonética do Português Brasileiro*. Monografia (Tese de Livre Docência) — Instituto de Estudos da Linguagem, UNICAMP, 1981.

CHABANAS, M.; PAYAN, Y. A 3D Finite Element model of the face for simulation in plastic and maxillo-facial surgery. In: DELP, S. L.; DIGIOIA, A. M.; JARAMAZ, B. (Ed.). *Proceedings of the Third International Conference on Medical Image Computing and Computer-Assisted Interventions - MICCAI'2000*. Pittsburgh, USA: Springer, 2000. (Lecture Notes in Computer Science), p. 1068–1075.

COHEN, M. M.; MASSARO, D. W. Synthesis of visible speech. *Behavioral Research Methods, Instrumentation and Computer*, v. 22, n. 2, p. 260–263, 1990.

COHEN, M. M.; MASSARO, D. W. Modeling coarticulation in synthetic visual speech. In: MAGNENAT-THALMANN, N.; THALMANN, D. (Ed.). *Models and Techniques in Computer Animation*. Tokyo: Springer-Verlag, 1993. p. 139–156.

COSATTO, E.; GRAF, H. P. Sample-based synthesis of photo-realistic talking heads. In: *Proceedings Computer Animation '98*. Philadelphia, Pennsylvania: IEEE Computer Society, 1998. p. 103–110.

COSI, P. et al. Labio coarticulation modeling for realistic facial animation. In: *Fourth IEEE International Conference on Multimodal Interfaces - ICMI'02*. Pittsburgh, PA, USA: IEEE Press, 2002. p. 505–510.

DE MARTINO, J. M.; MAGALHÃES, L. P. Um conjunto de visemas para uma cabeça falante do português do Brasil. In: *III Congresso Iberoamericano IBERDISCAP 2004 - Tecnología de apoyo a la discapacidad*. Jan Jose, Costa Rica: Universidad Estatal a Distancia, 2004. p. 198–203.

DE MARTINO, J. M.; VIOLARO, F. Um talking head para o português do Brasil que objetiva suportar a leitura labial. In: GARCIA, J. V. et al. (Ed.). *I Jornadas CYTED de Tecnologías de apoyo a la discapacidad - Programa Iberoamericano de Ciencia e Tecnologia para el Desarrollo*. Natal, RN, Brasil: Universidad Estatal a Distancia, 2003. p. 123–127.

EDWARDS, J.; HARRIS, K. Rotation and translation of the jaw during speech. *Journal of Speech and Hearing Research*, v. 33, p. 550–562, 1990.

EKMAN, P.; FRIESEN, W. V. *Manual of Facial Action the Coding System*. Palo Alto, CA: Consulting Psychologists Press, 1978.

ELISEI, F. et al. Creating and controlling video-realistic talking heads. In: MASSARO, D. W.; LIGHT, J.; GERACI, K. (Ed.). *Proceedings of the International Auditory-Visual Speech Processing Workshop - AVSP'01*. Aalborg, Denmark: ISCA, 2001. p. 90–97.

ENGWALL, O. *Tongue Talking*. Tese (Doutorado) — Departament of Speech, Music and Hearing - Royal Institute of Technology, 2002.

ERBER, N. P. Auditory, visual, and auditory-visual recognition of consonants by children with normal and impaired hearing. *Journal of Speech and Hearing Disorders*, v. 15, p. 413–422, 1972.

ERBER, N. P. Discussion: Lipreading skills. In: STARK, R. E. (Ed.). *Sensory capabilities of hearing impaired children*. Baltimore: University Park Press, 1974. p. 69–73.

ERBER, N. P. Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*, v. 40, p. 481–492, 1975.

EZZAT, T.; GEIGER, G.; POGGIO, T. Trainable videorealistic speech animation. In: *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 2002. p. 388–398.

EZZAT, T.; POGGIO, T. Miketalk: A talking facial display based on morphing visemes. In: *Proceedings of the Computer Animation - CA'98*. Philadelphia, Pennsylvania: IEEE Computer Society, 1998. p. 96–102.

EZZAT, T.; POGGIO, T. Visual speech synthesis by morphing visemes. *International Journal of Computer Vision*, Kluwer Academic Publishers, v. 38, n. 1, p. 45–57, June 2000.

FOLEY, J. D. et al. *Computer Graphics Principles and Practice*. 2th. ed. Massachusetts: Addison-Wesley Publishing Company, 1990.

FUJIMURA, O. Modern methods of investigation in speech production. *Phonetica*, v. 37, p. 38–54, 1980.

GELFER, C.; BELL-BERTI, F.; HARRIS, K. S. Determining the extent of coarticulation: Effects of experimental design. *Journal of the Acoustical Society of America*, v. 86, n. 6, p. 2443–2445, December 1989.

GONZALEZ, R. C.; WINTZ, P. *Digital Image Processing*. Massachusetts: Addison-Wesley Publishing Company, 1987.

GRAY, H. *Anatomy of the human body*. On-line 20th ed. [S.l.]: Philadelphia: Lea and Febiger; Bartleby.com; www.bartleby.com/107/, 2000. Re-edited by Warren H. Lewis.

HILL, D. R.; PEARCE, A.; WYVILL, B. Animating speech: an automated approach using speech synthesised by rules. *The Visual Computer*, v. 3, n. 4, p. 277–289, 1988.

IPA. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. International Phonetic Association - Cambridge, 1999.

JACKSON, P. L. The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review*, v. 90, n. 5, p. 99–115, 1988.

JACKSON, P. L.; MONTGOMERY, A. A.; BINNIE, C. A. Perceptual dimensions underlying vowel lipreading performance. *Journal of Speech and Hearing Research*, v. 19, p. 796–812, 1976.

JAIN, A. K.; DUBES, R. C. *Algorithms for Clustering Data*. Englewood Cliffs NJ, USA: Prentice-Hall, 1988.

JEFFERS, J.; BARLEY, M. *Speechreading (Lipreading)*. Springfield, Illinois, USA: Charles C. Thomas Publisher, 1971.

JOBSON, J. D. *Applied Multivariate Data Analysis*. New York, NY, USA: Springer-Verlag, 1992.

KALRA, P. et al. SMILE: A multilayered facial animation system. In: KUNII, T. L. (Ed.). *Proceedings IFIP WG 5.10 Modeling in Computer Graphics*. Tokyo, Japan: Springer Verlag, 1991. p. 189–198.

KALRA, P. et al. Simulation of facial muscle actions based on rational free form deformations. *Computer Graphics Forum*, European Association for Computer Graphics, v. 11, n. 3, p. 59–69, May 1992.

KASS, M.; WITKIN, A.; TERZOPOULOS, D. Snakes: Active contour models. *International Journal of Computer Vision*, Kluwer Academic Publisher, Boston, USA, v. 1, n. 4, p. 321–331, 1987.

KEEVE, E. et al. Anatomy-based facial tissue modeling using the finite element method. In: *Proceedings of Seventh Annual IEEE Visualization Conferece - Visualization '96*. San Francisco, CA, USA: IEEE Computer Society, 1996. p. 21–28.

KäHLER, K.; HABER, J.; SEIDEL, H.-P. Geometry-based muscle modeling for facial animation. In: *Procedings Graphics interface 2001 - GI'01*. Toronto, Ont., Canada, Canada: Canadian Information Processing Society, 2001. p. 37–46.

KLEISER, J. A fast, efficient, accurate way to represent the human face. In: *SIGGRAPH'89 Tutorials: State of the Art in Facial Animation*. Boston, Massachusetts, USA: ACM Press, 1989. v. 22.

KOCH, R. M. et al. Simulating facial surgery using finite element models. In: *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 1996. p. 421–428.

KOZLOWSKI, L. *A Percepção visual da fala*. Rio de Janeiro, RJ: Livraria e Editora Revinter, 1997.

KRICOS, P. B.; LESNER, S. A. Differences in visual intelligibility across talkers. *The Volta Review*, v. 264, p. 219–225, May 1982.

KURATATE, T.; YEHIA, H.; VATIKIOTIS-BATESON, E. Kinematics-based synthesis of realistic talking faces. In: BURNHAM, D.; ROBERT-RIBES, J.; VATIKIOTIS-BATESON, E. (Ed.). *Proceedings of the International Conference on Auditory-Visual Speech Processing - AVSP'98*. Terrigal, Australia: ISCA, 1998. p. 185–190.

LE GOFF, B. Automatic modeling of coarticulation in text-to-visual speech synthesis. In: *Proceedings of the 5th European Conference on Speech Communication and Technology - EUROSPEECH'97*. Rhodes, Greece: ESCA, 1997. p. 1667–1670.

LEE, Y.; TERZOPOULOS, D.; WALTERS, K. Realistic modeling for facial animation. In: *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 1995. p. 55–62.

LöFQVIST, A. Speech as audible gesture. In: HARDCASTLE, W.; MARCHAL, A. (Ed.). *Speech Production and Speech Modeling*. Dordrecht, the Netherlands: Kluwer Academinc Publishers, 1990. p. 289–322.

LUBKER, J. Anticipatory labial coarticulation: Experimental, biological, and linguistic variables. *The Journal of the Acoustical Society of America*, v. 71, n. 2, p. 437–448, February 1982.

LUCERO, J. C.; MUNHALL, K. G. A model of facial biomechanics for speech production. *The Journal of the Acoustical Society of America*, v. 106, n. 5, p. 2834–2842, November 1999.

MASSARO, D. W. A computer-animated tutor for spoken and written language learning. In: *Proceedings of 5(th) Conference on Multimodal Interfaces - ICMI'03*. Vancouver, British Columbia, Canada. New York: ACM Press, 2003. p. 172–175.

MASSARO, D. W. From multisensory integration to talking heads and language learning. In: CALVERT, G.; SPENCE, C.; STEIN, B. (Ed.). *Handbook of Multisensory Processes*. Cambridge, MA, USA: MIT Press, 2004. p. 153–176.

MASSARO, D. W.; LIGHT, J. Read my tongue movements: Bimodal learning to perceive and produce non-native speech /r/ and /l/. In: *Proceedings of 8th European Conference on Speech Commmunication Technology - Eurospeech'03 / Interspeech'03*. Geneva, Switzerland: ISCA, 2003. p. 2249–2252.

MCGURK, H.; MACDONALD, J. Hearing lips and seeing voices. *Nature*, v. 264, p. 746–748, 1976.

MENACHE, A. *Understanding Motion Capture for Animation and Video Games*. San Francisco, CA, USA: Morgan Kaufmann, 2000.

MONARETTO, V. N. O.; QUEDNAU, L. R.; HORA, D. da. As consoantes do português. In: BISOL, L. (Ed.). *Introdução a estudos de fonologia do português brasileiro*. 3ª. ed. Porto Alegre, RS: EDIPUCRS, 2001. p. 195–249.

MONTGOMERY, A. A.; JACKSON, P. L. Physical characteristics of the lips underlying vowel lipreading performance. *The Journal of the Acoustical Society of America*, v. 73, n. 6, p. 2134–2144, June 1983.

MPEG4 AUDIO. *Information technology - Coding of audio-visual objects - Parte 3: Audio*. Geneva, Switzerland, December 2001. International Standard.

MPEG4 SYSTEM. *Information technology - Coding of audio-visual objects - Parte 1: System*. Geneva, Switzerland, October 2001. ISO/IEC 14496-1:2001(E). International Standard.

MPEG4 VISUAL. *Information technology - Coding of audio-visual objects - Parte 2: Visual*. Geneva, Switzerland, December 2001. ISO/IEC 14496-2:2001(E). International Standard.

NITCHIE, E. *New lessons in lipreading*. USA: J.B. Lippincott, 1950.

ÖHMAN, S. E. G. Coarticulation in vcv utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, v. 39, p. 151–168, 1966.

ÖHMAN, S. E. G. Numerical model of coarticulation. *The Journal of the Acoustical Society of America*, v. 41, n. 2, p. 310–320, 1967.

OLIVÈS, J.-L. et al. Towards a high quality finnish talking head. In: *Proceedings of 1999 IEEE 3rd Workshop on Multimedia Signal Processing*. Copenhagen, Denmark: IEEE Computer Society, 1999. p. 433–437.

OSTERMANN, J. Face animation in mpeg-4. In: PANDZIC, I. S.; FORCHHEIMER, R. (Ed.). *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. Hoboken, NJ, USA: John Wiley and Sons, 2002. p. 17–55.

OWENS, E.; BLAZEK, B. Visemes observed by hearing-impaired and normal adult viewers. *Journal of Speech and Hearing Research*, v. 28, p. 381–393, September 1985.

PAERSALL, J.; TRUMBLE, B. (Ed.). *The Oxford English Reference Dictionary*. 2nd. ed. Oxford, UK: Oxford University Press, 1996.

PARKE, F. I. *Computer generated animation of faces*. Dissertação (Mestrado) — University of Utah, June 1972.

PARKE, F. I. Parameterized models for facial animation. *IEEE Computer Graphics and Applications*, IEEE Computer Society, v. 2, n. 9, p. 61–68, November 1982.

PARKE, F. I. Measuring three-dimensional surfaces with a two-dimensional data tablet. In: *SIGGRAPH'90 Course Notes: State of the Art in Facial Animation*. Dallas Convention Center, Texas, USA: ACM Press, 1990. v. 26, p. 233–242.

PARKE, F. I.; WATERS, K. *Computer Facial Animation*. Wellesley, Massachusetts, USA: A K Peters, 1996.

PAYAN, Y. et al. Biomechanical models to simulate consequences of maxillofacial surgery. *Les Comptes Rendus de l'Académie des Sciences (CRAS), C.R. Biologies*, Académie des Sciences / Éditions scientifiques et médicales Elsevier SAS, v. 325, n. 4, p. 407–417, April 2002.

PELACHAUD, C. *Communication and Coarticulation in Facial Animation*. Tese (Doutorado) — University of Pennsylvania, 1991.

PELACHAUD, C. Visual Text-to-Speech. In: PANDZIC, I. S.; FORCHHEIMER, R. (Ed.). *MPEG-4 Facial Animation*. Hoboken, NJ, USA: John Wiley and Sons, 2002. p. 125–140.

PELACHAUD, C.; BADLER, N. I.; STEEDMAN, M. Generating facial expressions for speech. *Cognitive Science*, v. 20, n. 1, p. 1–46, January-March 1996.

PELACHAUD, C. et al. Modelling an Italian Talking Head. In: *Proceeding of the Visual Speech Processing 2001 - AVSP'01*. Aalborg, Denmark: ISCA, 2001. p. 72–77.

PERKELL, J. S. Testing models of speech production: Implications of some detailed analysis of variable articulatory data. In: HARDCASTLE, W.; MARCHAL, A. (Ed.). *Speech Production and Speech Modeling*. Dordrecht, the Netherlands: Kluwer Academinc Publishers, 1990. p. 263–288.

PIGHIN, F. et al. Synthesizing realistic facial expressions from photographs. In: *SIGGRAPH '98: Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 1998. p. 75–84.

PITERMANN, M.; MUNHALL, K. G. An inverse dynamics approach to face animation. *Journal of the Acoustical Society of America*, Acoustical Society of America, NY, USA, v. 110, n. 3, p. 1570–1580, 2001.

PLATT, S. M.; BADLER, N. I. Animating facial expressions. In: *SIGGRAPH '81: Proceedings of the 8th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 1981. p. 245–252.

POTMESIL, M.; CHAKRAVARTY, I. Modeling motion blur in computer-generated images. In: *SIGGRAPH '83: Proceedings of the 8th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 1983. v. 17, n. 3, p. 389–399.

RABINER, L. R.; JUANG, B.-H. *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall, 1993.

REISBERG; MCLENAY; GOLDFIELD. Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In: DOOD, B.; CAMPBELL, R. (Ed.). *Hearing by eye: the psychology of lip-reading*. London, UK: Lawrence Erlbaum Associates, 1987. p. 97–114.

REVÉRET, L.; BAILLY, G.; BADIN, P. Mother: A new generation of talking heads providing a flexible control for video-realistic speech animation. In: *Proceedings of the 6th International Conference on Spoken Language Processing - ICSLP'00*. Beijing, China: ISCA, 2000. p. 755–758.

RIFF. *Multimedia Programming Interface and Data Specifications 1.0*. IBM Corporation e Microsoft Corporation, August 1991.

SEDERBERG, T. W.; PARRY, S. R. Free-form deformation of solid geometric models. In: *SIGGRAPH '86: Proceedings of the 13th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 1986. p. 151–160.

SILVA, T. C. *Fonética e fonologia do português: roteiro de estudos e guia de exercícios*. 6ª. ed. São Paulo, SP: Editora Contexto (Editora Pinsky Ltda), 2002.

SOBOTTA, J. *Atlas de anatomia humana*. Rio de Janeiro, RJ: Guanabara Koogan, 1990. Tradução de Helcio Werneck do original Atlas der Anatomie des Menschen.

STURMAN, D. J. Computer puppetry. *IEEE Computer Graphics and Applications*, IEEE Computer Society, v. 18, n. 1, p. 38–45, 1998.

TERZOPOULOS, D.; WATERS, K. Physically-based facial modelling, analysis, and animation. *The Journal of Visualization and Computer Animation*, v. 1, n. 4, p. 73–80, March 1990.

TOMMASELLI, A.; TOZZI, C. Técnicas de calibração de câmaras em visão computacional. In: *Jornadas EPUSP/IEEE em Computação Visual - Tutorial*. São Paulo, SP: [s.n.], 1990. p. 13–36.

VATIKIOTIS-BATESON, E.; OSTRY, D. J. An analysis of the dimensionality of jaw motion in speech. *Journal of Phonetics*, v. 23, n. 1-2, p. 101–117, January-April 1995.

WALDEN, B. E. et al. Some effects of training on speech recognition by hearng-impaired adults. *Journal of Speech and Hearing Research*, v. 24, n. 1, p. 201–216, 1981.

WALDEN, B. E. et al. Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, v. 20, n. 1, p. 130–145, 1977.

WATERS, K. A muscle model for animation three-dimensional facial expression. In: *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques*. New York, NY, USA: ACM Press, 1987. p. 17–24.

WESTBURY, J. Mandible and hyoid bone movements during speech. *Journal of Speech and Hearing Research*, v. 31, p. 405–416, 1994.

WOZNIAK, V. D.; JACKSON, P. L. Vowel and diphthong perception from two horizontal viewing angles. *Journal of Speech and Hearing Research*, p. 354–365, June 1979.

YNOGUTI, C. A. *Reconhecimento de Fala Contínua Usando Modelos Ocultos de Markov*. Tese (Doutorado) — Universidade Estadual de Campinas, maio 1999.

ZEMLIN, W. R. *Princípios de Anatomia e Fisiologia em Fonoaudiologia*. 4ª. ed. Porto Alegre, RS: Artmed Editora, 2000. Tradução de Terezinha Oppido do original Speech and hearing science: Anatomy and Physiology.

Apêndice A

Técnica Fotogramétrica

A.1 Introdução

Neste anexo é detalhada a técnica fotogramétrica utilizada para a medida da movimentação dos pontos de interesse discutida no Capítulo 5. Esta técnica permite a determinação das coordenadas tridimensionais de pontos localizados no espaço a partir da análise de, pelo menos, duas imagens capturadas de ângulos diferentes. Na técnica discutida, utiliza-se um par de imagens, geradas simultaneamente por duas câmeras.

A técnica fotogramética assume, como hipótese fundamental, que o processo de geração de uma imagem por uma câmera pode ser representado por uma transformação linear, descrita por uma matriz com coeficientes constantes. A determinação destes coeficientes é efetuada através de um processo de calibração que necessita de, pelo menos, seis pontos, com *Coordenadas de Mundo* (coordenadas no espaço tridimensional) e *Coordenadas de Imagem* (coordenadas bidimensionais no plano imagem) conhecidas. Após o processo de calibração de cada câmera, é possível determinar as Coordenadas de Mundo de um ponto qualquer a partir do conhecimento de sua localização em cada uma das imagens.

Na Seção A.2 é discutido o processo de calibração, com a formulação adotada para o cálculo dos coeficientes da transformação de geração da imagem. Na Seção A.3 é apresentado o procedimento utilizado para o cálculo das Coordenadas de Mundo de um ponto a partir de suas Coordenadas de Imagem.

Para um maior aprofundamento sobre a técnica, consultar, entre outros (GONZALEZ; WINTZ, 1987), (PARKE, 1990), (TOMMASELLI; TOZZI, 1990) e (AYACHE, 1991).

A.2 Calibração da Câmera

A câmera pode ser considerada como um dispositivo que projeta em um plano pontos do espaço tridimensional. Esta projeção pode ser descrita pelas seguintes relações:

$$\begin{bmatrix} U \\ V \\ S \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{A.1}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} U/S \\ V/S \end{bmatrix} \tag{A.2}$$

onde:

- x, y, z são as Coordenadas de Mundo de um ponto;
- u e v são as Coordenadas de Imagem deste ponto.

Através da manipulação algébrica de Eq. A.1 e Eq. A.2 é possível obter o seguinte sistemas de equações:

$$\begin{cases} a_{11}\,x + a_{12}\,y + a_{13}\,z + a_{14} - a_{31}\,u\,x - a_{32}\,u\,y - a_{33}\,u\,z = a_{34}\,u \\ a_{21}\,x + a_{22}\,y + a_{23}\,z + a_{24} - a_{31}\,v\,x - a_{32}\,v\,y - a_{33}\,v\,z = a_{34}\,v \end{cases} \tag{A.3}$$

A divisão por a_{34} dos dois lados de cada equação deste sistema resulta em:

$$\begin{cases} t_{11}\,x + t_{12}\,y + t_{13}\,z + t_{14} - t_{31}\,u\,x - t_{32}\,u\,y - t_{33}\,u\,z = u \\ t_{21}\,x + t_{22}\,y + t_{23}\,z + t_{24} - t_{31}\,v\,x - t_{32}\,v\,y - t_{33}\,u\,z = v \end{cases} \tag{A.4}$$

onde:

- $t_{ij} = a_{ij}/a_{34} \quad i = 1,2,3 \quad j = 1,2,3,4 \quad (i,j) \neq (3,4)$.

O que equivale a dizer que a formulação expressa por Eq. A.1 e Eq. A.2 pode ser rescrita como

$$\begin{bmatrix} U \\ V \\ S \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} \\ t_{21} & t_{22} & t_{23} & t_{24} \\ t_{31} & t_{32} & t_{33} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{A.5}$$

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} U/S \\ V/S \end{bmatrix} \tag{A.6}$$

Ou ainda, de forma expandida, pelo sistema de equações

$$\begin{cases} t_{11}\,x + t_{12}\,y + t_{13}\,z + t_{14} - t_{31}\,u\,x - t_{32}\,u\,y - t_{33}\,u\,z = u \\ t_{21}\,x + t_{22}\,y + t_{23}\,z + t_{24} - t_{31}\,v\,x - t_{32}\,v\,y - t_{33}\,v\,z = v \end{cases} \tag{A.7}$$

Colocando-se em evidência os coeficientes t_{ij}, este sistema de equações pode ser expresso por

$$\begin{bmatrix} x & y & z & 1 & 0 & 0 & 0 & 0 & -u\,x & -u\,y & -u\,z \\ 0 & 0 & 0 & 0 & x & y & z & 1 & -v\,x & -v\,y & -v\,z \end{bmatrix} \begin{bmatrix} t_{11} \\ t_{12} \\ t_{13} \\ t_{14} \\ t_{21} \\ t_{22} \\ t_{23} \\ t_{24} \\ t_{31} \\ t_{32} \\ t_{33} \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix} \tag{A.8}$$

Para a determinação dos 11 parâmetros t_{ij} é necessário estabelecer, no mínimo, 11 equações. Faz-se, portanto, necessário o conhecimento das Coordenadas de Mundo e Coordenadas Imagem de, no mínimo, 6 pontos distintos ($5\frac{1}{2}$ mais precisamente). Uma estimação mais acurada dos coeficientes pode ser alcançada, entretanto, aplicando-se o método dos mínimos quadrados (pseudo-inversa) e calculando-se os coeficientes a partir de um conjunto n de pontos maior do que 6 . A resolução do sistema linear resultante pelo método dos mínimos quadrados é dada por

$$\mathbf{t} = \left(\mathbf{M^t}\,\mathbf{M}\right)^{-\mathbf{1}}\,\mathbf{M^t}\,\mathbf{p} \tag{A.9}$$

Com

$$\mathbf{t} = \begin{bmatrix} t_{11} \\ t_{12} \\ t_{13} \\ t_{14} \\ t_{21} \\ t_{22} \\ t_{23} \\ t_{24} \\ t_{32} \\ t_{33} \end{bmatrix} \tag{A.10}$$

$$\mathbf{M} = \begin{bmatrix} x_1 & y_1 & z_1 & 1 & 0 & 0 & 0 & 0 & -u_1\,x_1 & -u_1\,y_1 & -u_1\,z_1 \\ 0 & 0 & 0 & 0 & x_1 & y_1 & z_1 & 1 & -v_1\,x_1 & -v_1\,y_1 & -v_1\,z_1 \\ x_2 & y_2 & z_2 & 1 & 0 & 0 & 0 & 0 & -u_2\,x_2 & -u_2\,y_2 & -u_2\,z_2 \\ 0 & 0 & 0 & 0 & x_2 & y_2 & z_2 & 1 & -v_2\,x_2 & -v_2\,y_2 & -v_2\,z_2 \\ & & & & & \ddots & & & & & \\ x_n & y_n & z_n & 1 & 0 & 0 & 0 & 0 & -u_n\,x_n & -u_n\,y_n & -u_n\,z_n \\ 0 & 0 & 0 & 0 & x_n & y_n & z_n & 1 & -v_n\,x_n & -v_n\,y_n & -v_n\,z_n \end{bmatrix} \tag{A.11}$$

$$\mathbf{p} = \begin{bmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ \vdots \\ u_n \\ v_n \end{bmatrix} \tag{A.12}$$

O processo de estimação dos coeficientes do vetor **t** é denominado de processo de calibração da câmera. Observe-se que os coeficientes do vetor **t** definem a matriz que representa o processo de geração de imagem modelado por Eq. A.5 e Eq. A.6. Os coeficientes do vetor **t** são parâmetros que caracterizam a câmera utilizada no processo de geração da imagem. Estes parâmetros englobam características intrínsicas, como distância focal e abertura da lente, e extrínsicas da câmera, como sua posição e orientação.

A.3 Visão Estéreo

A partir de Eq. A.9, Eq. A.11, Eq. A.10 e Eq. A.12, através de manipulação algébrica, é possível estabelecer o seguinte sistema de equações

$$\begin{cases} (t_{11} - t_{31}\,u)\;x + (t_{12} - t_{32}\,u)\;y + (t_{13} - t_{33}\,u)\;z = u - t_{14} \\ (t_{21} - t_{31}\,v)\;x + (t_{22} - t_{32}\,v)\;y + (t_{23} - t_{33}\,v)\;z = v - t_{24} \end{cases} \tag{A.13}$$

Este sistema, após a calibração descrita na Seção A.2, é um sistema a duas equações com três incógnitas x, y e z - as Coordenadas de Mundo associadas às Coordenadas de Imagem u e v.

Ao se tomar duas imagens capturadas de posições diferentes é possível construir o sistema de equações

$$\begin{cases} (t'_{11} - t'_{31}\,u')\;x + (t'_{12} - t'_{32}\,u')\;y + (t'_{13} - t'_{33}\,u')\;z = u' - t'_{14} \\ (t'_{21} - t'_{31}\,v')\;x + (t'_{22} - t'_{32}\,v')\;y + (t'_{23} - t'_{33}\,v')\;z = v' - t'_{24} \\ (t''_{11} - t''_{31}\,u'')\;x + (t''_{12} - t''_{32}\,u'')\;y + (t''_{13} - t''_{33}\,u'')\;z = u'' - t''_{14} \\ (t''_{21} - t''_{31}\,v'')\;x + (t''_{22} - t''_{32}\,v'')\;y + (t''_{23} - t''_{33}\,v'')\;z = v'' - t''_{24} \end{cases} \tag{A.14}$$

onde:

- $t'_{ij} \quad i = 1,2,3 \quad j = 1,2,3,4 \quad (i,j) \neq (3,4)$, são os parâmetros da câmera 1;
- $t''_{ij} \quad i = 1,2,3 \quad j = 1,2,3,4 \quad (i,j) \neq (3,4)$, são os parâmetros da câmera 2;
- u' e v' são as coordenadas de um ponto na imagem gerada pela câmera 1;
- u'' e v'' são as coordenadas do mesmo ponto na imagem gerada pela câmera 2.

O sistema de equação Eq. A.14 pode ser resolvido pelo método da pseudo-inversa, levando à seguinte solução

$$\mathbf{c} = \left(\mathbf{Q^t}\,\mathbf{Q}\right)^{-1}\,\mathbf{Q^t}\,\mathbf{r} \tag{A.15}$$

Com

$$\mathbf{c} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \tag{A.16}$$

$$\mathbf{Q} = \begin{bmatrix} t'_{11} - t'_{31}\,u' & t'_{12} - t'_{32}\,u' & t'_{13} - t'_{33}\,u' \\ t'_{21} - t'_{31}\,v' & t'_{22} - t'_{32}\,v' & t'_{23} - t'_{33}\,v' \\ t''_{11} - t''_{31}\,u'' & t''_{12} - t''_{32}\,u' & t''_{13} - t''_{33}\,u'' \\ t''_{21} - t''_{31}\,v'' & t''_{22} - t''_{32}\,v'' & t''_{23} - t''_{33}\,v'' \end{bmatrix} \tag{A.17}$$

$$\mathbf{r} = \begin{bmatrix} u' - t'_{34} \\ v' - t'_{34} \\ u'' - t'_{34} \\ v'' - t'_{34} \end{bmatrix} \tag{A.18}$$