UNIVERSIDADE FEDERAL DE MINAS GERAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

UM ESTUDO SOBRE AS RELAÇÕES DE PADRÕES DO MOVIMENTO FACIAL COM A ACÚSTICA DA FALA E COM A IDENTIDADE DO LOCUTOR

por

Kétia Soares Moreira

Tese de Doutorado submetida à banca examinadora designada pelo Programa de Pós Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Engenharia Elétrica

Orientadores: Hani Camille Yehia
Carmen Déa Moraes Pataro

Belo Horizonte - MG
Maio de 2008

A meu Pai,
Geraldo Alves Moreira.

Agradecimentos

Minha sincera gratidão

a Deus, por esta grande lição;

ao meu pai, Geraldo A. Moreira, por ter sido um grande pai;

à minha mãe, Flora M. S. Moreira, pelo afeto, dedicação e compreensão;

aos meus irmãos, Kátia e Heitor, pela grande amizade;

ao meu marido, André, pela paciência, amor e carinho;

aos professores Hani Yehia e Carmen Déa Pataro, por transmitirem seus conhecimentos;

aos amigos do CPDEE.

"Eu tudo posso naquele que me fortalece".

(Felipense 4:13)

Resumo

O estudo do acoplamento entre o movimento facial e a acústica da fala é importante para a compreensão do processo de produção da fala. Além disso, a relação do movimento facial com a identidade do locutor durante a fala é importante no processo de identificação com base em biometria. O objetivo deste trabalho é analisar movimentos faciais de forma a: (i) avaliar a variabilidade dos parâmetros relacionados ao movimento facial durante a produção da fala; (ii) verificar se tais parâmetros são dependentes ou independentes do contexto; e (iii) avaliar até que ponto tais movimentos são específicos de cada pessoa. Durante a produção da fala, a geometria do trato vocal determina suas freqüências de ressonância (formantes) e influencia no movimento da face ocorrido simultaneamente. Como resultado, padrões acústicos da fala e movimentos faciais são acoplados. A relação entre o movimento de partes da face pode ser modelada eficientemente por meio de Análise em Componentes Principais, enquanto o acoplamento entre a acústica da fala e o movimento facial pode ser modelado por meio de componentes do movimento facial alinhadas com parâmetros LSP extraídos da acústica da fala. Um dos objetivos deste estudo é avaliar como esse alinhamento varia com o tempo. Os resultados obtidos mostram que apenas a primeira componente do movimento facial é estável, independentemente do conteúdo acústico da fala, e concentrando até 55% da variância do movimento facial. Para a primeira componente acusticamente alinhada do movimento facial, esta estabilidade é menor. Entretanto, uma maior estabilidade é observada quando os parâmetros LSP, usados na representação acústica da fala, são ordenados em função de sua cavidade de afiliação, ao invés de simplesmente ordenados em ordem crescente. No estudo de padrões do movimento facial aplicados à identificação de indivíduos, utiliza-se o primeiro autovetor da matriz de covariância do movimento facial, pois este apresenta informações específicas de cada locutor. Neste sentido, foram realizados testes utilizando uma rede neural MLP na tarefa de identificação de locutores com base no autovetor associado ao maior autovalor da matriz de covariância do movimento facial. A taxa de acerto foi de 86,7%, indicando que apenas informações do movimento facial não são suficientes para um processo de identificação eficiente. Porém, tais informações podem ser usadas em conjunto com outras, tais como imagens estáticas ou a voz do indivíduo, tornando o processo de identificação mais robusto, especialmente em condições adversas.

Abstract

The study of the coupling between facial motion and speech acoustics is important for the comprehension of the speech production process. Moreover, the relation of facial motion with speaker identity during speech is important in the process of identification based on biometry. The objective of this work is to evaluate facial motion in order to: (i) evaluate the variability of parameters related to facial motion during speech production; (ii) verify whether such parameters are context dependent or independent; and (iii) evaluate to which degree facial motion is an individual characteristic. During speech, the geometry of the vocal tract determines its resonant frequencies (formants) and strongly influences the facial motion that occurs simultaneously. As a result, speech acoustic patterns and facial motion are coupled. The relation between regions of the face can be efficiently modeled by means of Principal Component Analysis, whereas the coupling between speech acoustics and facial motion can be modeled by facial motion components aligned with LSP parameters extracted from speech acoustics. One of the objectives of this study is to evaluate how that alignment varies with time. The results obtained show that only the first facial motion component is stable during speech, independently of the speech contents, and concentrates up to 55% of the facial motion variance. For the first acoustically aligned facial motion component, this stability is smaller. However, a larger stability is observed when LSP parameters, used to represent speech acoustically, are ordered based on their vocal tract cavity affiliation, rather than simply put in increasing order. In the study of facial motion patterns applied to person identification, the first eigenvector of the facial motion covariance matrix is used, as it exhibits speaker specific information. In this direction, tests using an MLP neural network were carried out for the task of person identification based on the eigenvector associated to the largest eigenvalue of the facial motion covariance matrix. An identification rate of 86,7% was attained, indicating that facial motion information alone is not enough for person identification. Nevertheless, this information can be used together with other pieces of information, such as static images or the speaker's voice, to improve the robustness of the identification process, specially under adverse conditions.

Lista de Símbolos

N	Número de adesivo marcadores.
M	Número de quadros.
K	Número de componentes.
μ_X	Vetor médio dos marcadores ao longo dos quadros.
C_{XX}	Matriz de covariância dos marcadores.
U	Matriz cujas colunas contêm os autovetores de C_{XX}.
S_{XX}	Matriz diagonal com os autovalores dos marcadores facial.
C_{XF}	Matriz de correlação cruzada.
U_{XF}	Matriz cujas colunas contêm os autovetores de $C_{XF}C_{XF}^T$.
μ_F	Vetor médio dos parâmetros LSP.
S_{XF}	Matriz diagonal com os autovalores das componentes acústicamente alinhadas.
V_{XF}	Matriz cujas colunas contêm os autovetores de $C_{XF}^T C_{XF}$.
α_p	Coeficientes de predição.
p	Ordem do filtro usada na análise LPC.
w_i, θ_i	Parâmetros LSP que variam $1, ..., p/2$.
$G(z)R(z)$	Transformada z da contribuição conjunta do fluxo de volume glotal e da radiação labial.
K_1	Constante relacionada com a amplitude do fluxo glotal.
z_a, z_b	Pólos relacionados com o fluxo glotal localizados no eixo real dentro do círculo unitário.
K_2	Constante relacionada com a amplitude do fluxo de volume nos lábios e a distância dos lábios ao microfone.
$\hat{s}(j)$	Sinal predito.
$s(j-i)$	Valores passados observados.
α_i	I-ésimo coeficiente de predição do filtro LPC.
$u(j)$	Entrada do filtro LPC.

p	Número de coeficientes de predição do filtro LPC (ordem das análises LPC e LSP).
(fi, gi)	Frequências LSP (em Hz).
fs	Frequência de amostragem.
$f(j)$	Vetor de frequências LSP no instante j.
$x(j)$	Vetor de posições dos marcadores faciais (vetor facial) no instante j.
$p(j)$	Vetor das componentes principais no instante j.
F	matriz cujas colunas são os vetores f(j).
X	matriz cujas colunas são os vetores x(j).
P	matriz cujas colunas são os vetores p(j).
P_X	Componentes principais do movimento facial.
P_F	Componentes principais dos parâmetros LSP.
P_{XF}	Componentes dos marcadores faciais acusticamente alinhadas.
T_{XF}	Estimador linear do erro médio quadrático mínimo (MMSE).
σ_l	Desvio-padrão do locutor l.
$\bar{\sigma}$	Desvio-padrão médio.
σ	Desvio-padrão dos valores de desvio-padrão médio obtidos para l locutores.
j	Janela deslizante de 3 segundos deslocada a cada 0,2 segundos.
k	Número de autovetores, varia de 1...6. .
q	Número de observações, varia de 1...Q.
l	Número de locutores, varia de 1...L.
L	Número total locutores.
Q	Número total observações.
u_{ik}	I-ésimo marcador do autovetor de referência k.
u_{ijk}	I-ésimo marcador do autovetor k calculado com base na j-ésima janela analisada.
d_{jk}	Distância entre os marcadores do k-ésimo autovetor calculado com base na j-ésima janela analisada e o autovetor de referência correspondente.
$\mathbf{u}_{lq}$	Primeiro autovetor $k = 1$ obtido a partir da q-ésima elocução do locutor l.
μ	Autovetor médio de todas as elocuções de todos os locutores.
τ_l	Características próprias de cada locutor.

$\mu + \tau_l$	Autovetor médio de cada locutor.
$\mathbf{e}_{lq}$	Componente de cada autovetor $\mathbf{u}_{lq}$ específica de cada elocução de cada locutor.
$\mathbf{M}_l$	Matriz composta por características próprias, τ_l, adicionada a informações específicas de cada elocução $\mathbf{e}_{lq}$.
σ_T^2	Variabilidade interlocutores dos marcadores.
σ_L^2	Variabilidade intralocutor dos marcadores.
$\mathbf{u}_{lq}$	Primeiro autovetor correspondentes ao locutor l do trecho proferido q.
$\bar{\mathbf{u}}$	Média global do primeiro autovetor.
$\bar{\mathbf{u}}_l$	Média global do locutor l do primeiro autovetor.
Ur	matriz de rotação usada na compensação do movimento da cabeça.
$r0$	vetor de translação usado na compensação do movimento da cabeça.
$s(t)$	pressão sonora no microfone (saída do filtro LPC).

Abreviaturas

HMM	Modelo oculto de markov.
LPC	Codificação por predição linear.
LSP	Pares de linhas espectrais.
MLP	Perceptron multicamadas.
$MPEG$	Grupo de especialistas em figura em movimento.
MSE	Erro médio quadrático
PCA	Análise em componentes principais.
RBF	Função de base radial.
RGB	Vermelho, verde, azul.
$RMSE$	Raiz do erro quadrático médio.
RNA	Rede neural artificial.
SVD	Decomposição em valores singulares.
$YCbCr$	Luminância, crominância azul, crominância vermelha.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Os seres humanos utilizam a informação acústica na comunicação. Ao produzirem som através do trato vocal, simultaneamente ocorrem deformações na face. Tais deformações constituem a informação visual e contribuem para a comunicação, dando um caráter bimodal à fala humana.

O mecanismo de produção da fala é conseqüência do ar oriundo dos pulmões, da vibração das pregas vocais (laringe) e da ressonância acústica definida pela configuração geométrica do trato vocal. O trato vocal, que vai desde as pregas vocais até os lábios e narinas, tem a função de moldar o som glótico, funcionando como um filtro acústico que, em função de sua forma e comprimento atenua a energia do som em certas freqüências, reforçando a energia em outras freqüências, definidas como formantes. A Figura 1.1 mostra uma representação gráfica do esquema fonte-filtro para produção da fala. Os movimentos dos articuladores no trato vocal afetam a freqüência de todos os formantes (Furui e Sondhi, 1991; Rabiner e Juang, 1993b; Rabiner e Shafer, 1978; Zemlin, 2000).

O trato vocal, composto pela cavidade faríngea, cavidade nasal e cavidade bucal, altera sua forma durante a produção da fala. Em cada som produzido, ocorrem deformações em partes da face correlacionadas à configuração do trato vocal. Dentre as partes do esqueleto facial ligadas à produção da fala que contribuem para os movimentos na face, encontram-se a mandíbula, a maxila e os ossos zigomáticos (Zemlin, 2000). A mandíbula auxilia na produção da fala pelo próprio movimento proporcionando os pontos de ligação para grande parte da musculatura da língua. A maxila tem um papel importante, pois contribui para a formação do teto da boca. Por sua vez, os ossos zigomáticos ou bochechas, estão ligados a músculos importantes para o movimento facial.

Os músculos da face têm tamanho, forma e tonicidade diferentes para cada pessoa. Suas propriedades variam em função da idade, sexo, dentição e de características individuais intrínsecas (Zemlin, 2000). Na face, a parte mais maleável são os lábios, devido aos inúmeros músculos existentes. Estes músculos possuem fibras que pertencem somente aos lábios e também fibras oriundas de outros músculos faciais. Assim, ao movimentar os músculos dos

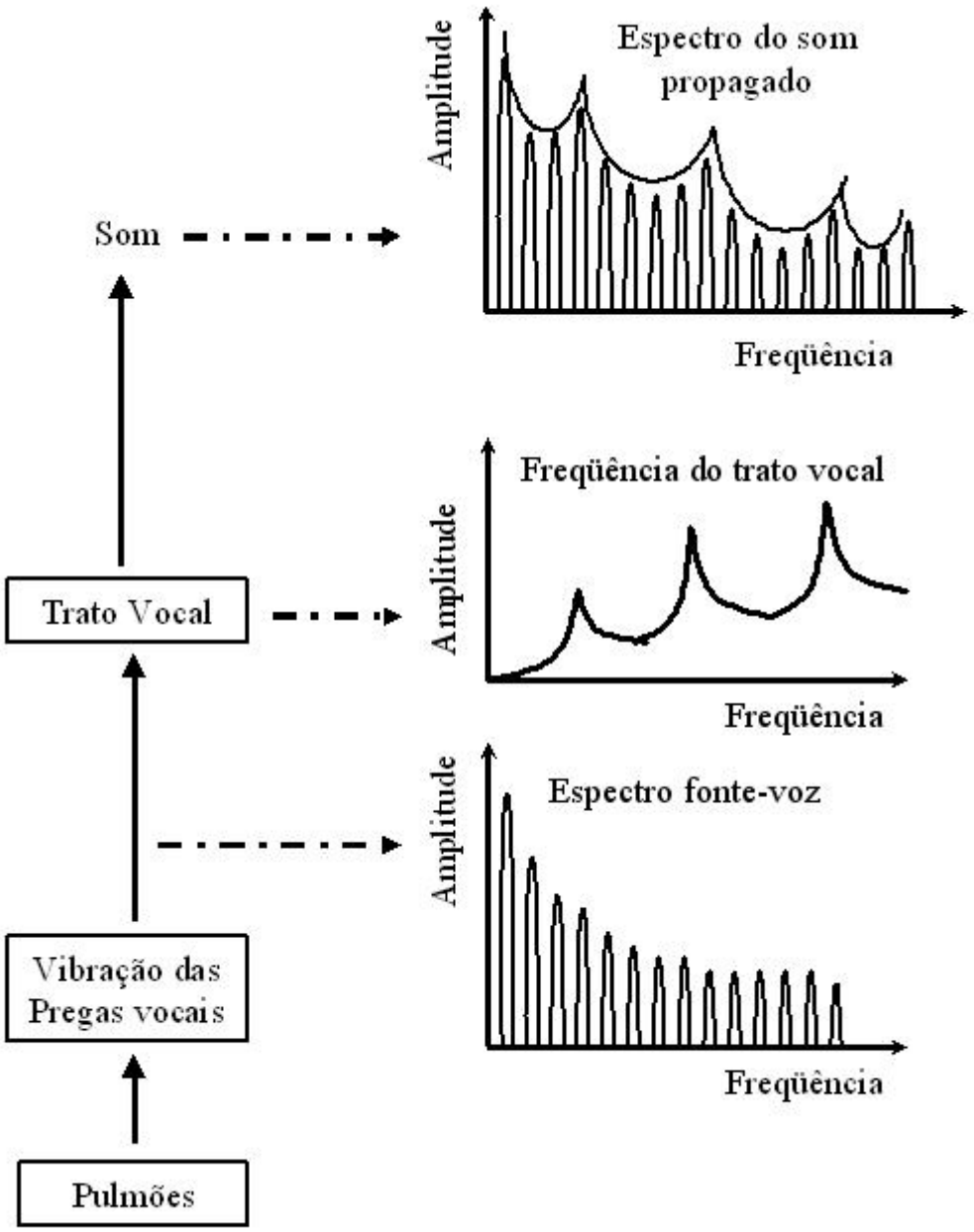


Figura 1.1: Diagrama em bloco da teoria fonte-filtro na produção da fala. Esboço do espectro do som glótico, da resposta acústica do trato vocal e do som propagado. O trato vocal tem a função de moldar o som glótico, funcionando como um filtro acústico que, em função de sua forma e comprimento atenua a energia do som em certas freqüências, e reforça a energia em outras freqüências.

lábios outros músculos da face também serão movimentados. De modo similar, as bochechas têm músculos que as ligam a músculos importantes da articulação e da mastigação. Por estes motivos, os movimentos entre partes da face estão relacionados entre si.

A acústica da fala é a principal fonte de informação na comunicação entre pessoas. Porém, diversos trabalhos vêm demonstrando que a informação visual é um facilitador no entendimento do conteúdo falado, principalmente quando no ambiente existem condições adversas à compreensão, tais como ruído e reverberação. A relação bimodal existente entre a acústica e os movimentos que acontecem durante a fala, sobretudo na região da face, é amplamente estudada. Diversos estudos comprovam que partes do movimento facial podem ser preditas pela acústica da fala e parte da acústica pode ser predita a partir dos movimentos faciais.

Desta forma, as deformações da face estão ligadas à acústica da fala. Os acoplamentos entre as partes visual e acústica podem ser parcialmente determinados linearmente pelos autovetores da matriz de correlação cruzada entre a posição de pontos da face e parâmetros extraídos da acústica da fala. Na face, acontecem movimentos correlacionados entre si em virtude de sua estrutura muscular e anatômica. Os acoplamentos entre as partes faciais podem ser determinados pelos autovetores da matriz de covariância da posição de pontos da face durante a produção da fala.

Os pontos centrais de análise deste trabalho são as componentes visuais originadas durante a produção da fala e suas relações com as componentes acústicas da fala e com as características específicas do locutor, com conseqüente utilização para reconhecimento de faces.

1.1 Percepção do movimento e identificação do locutor

A identificação automática de objetos e pessoas pelos recursos computacionais é importante para o desenvolvimento tecnológico do mundo moderno. Existem diversos métodos de identificação pelas características biométricas do indivíduo. Uma análise biométrica é projetada para verificar ou reconhecer a identidade de uma pessoa fundamentada em suas características fisiológicas, tais como, configuração facial, impressões digitais, retina, íris dos olhos e formato da mão.

Neste contexto, muitos trabalhos de identificação têm sido desenvolvidos utilizando a extração das características faciais. Alguns utilizam a geometria da face como característica discriminatória dos indivíduos, determinando as posições e dimensões da boca e do nariz, a distância entre os olhos, etc (Cox, Ghosn, e Yianilos, 1996; Brunelli e Poggio, 1991). Outra linha de pesquisa utiliza métodos que propõem empregar as representações de uma face, a partir de uma extração automática das características faciais da imagem de entrada (Er, Wu, Lu, e Toh, 2002). Esses métodos consideram todos os *pixels* da imagem, implicando alta dimensionalidade dos dados, o que aumenta consideravelmente o custo computacional. Para

atenuar esse problema podem ser utilizados métodos estatísticos de redução de dimensionalidade, como Análise em Componentes Principais (PCA) (Draper, Baek, e Bartlett, 2003; Brunelli e Poggio, 1991; Chen, Liao, e Lin, 2001; Iwano, Tamura, e Furui, 2001; Arandjelovi e Cipolla, 2004; Lamar, Bhuiyan, e Iwata, 1999a; Vatikiotis-Bateson, Yehia, e Kuratate, 2002; Barbosa e Yehia, 2001; Vatikiotis-Bateson e Yehia, 2000; Vatikiotis-Bateson, Kuratate, e Yehia, 1998b; Yehia, Rubin, e Vatikiotis-Bateson, 1998; Vatikiotis-Bateson e Yehia, 1997). A PCA é usada para extrair as características mais relevantes da imagem. Sua vantagem consiste no fato de pequenas variações locais não prejudicarem muito o reconhecimento, porém, variações de iluminação, expressão facial e outros fatores implicam aumento das dificuldades no reconhecimento (Er et al., 2002). O discriminante linear de Fisher também pode ser usado para extrair características da face (Er et al., 2002). As razões para reduzir a dimensionalidade de uma imagem são o custo de medição e a precisão do classificador. Se o espaço característico contiver somente as características mais claras, o classificador será mais robusto e mais rápido. Além disso, é necessário fazer uma escolha das características, pois padrões arbitrários podem tornar-se ambíguos se forem codificados com um número grande de características similares (Watanabe, 1985).

O ser humano utiliza vários recursos para identificar uma pessoa. As pessoas podem ser identificadas por meio de características físicas, voz, forma de comunicação ou gestos. A percepção do movimento (gestos) para o sistema visual é muito importante e fácil de ser constatada. Quando uma imagem é visualizada, os objetos que se movem são alvo de maior atenção. Assim, alguns animais, para fugir de predadores utilizam camuflagem e não fazem nenhum movimento.

Diversas áreas têm se beneficiado com os estudos do movimento. Como exemplo, a psicologia utiliza as características do movimento e da expressão para detectar o estado emocional de uma pessoa (Nordstrand, Svanfeldt, Granström, e House, 2004; Cowell e Ayesh, 2004; Bazzo e Lamar, 2004; Pantic e Rothkrantz, 1999). A emoção pode ser comunicada especialmente pelas expressões faciais e corporais, postura e sinais fisiológicos, tais como a variação da taxa de freqüência cardíaca (Cédras e Shah, 1995). A análise do movimento também pode ser utilizada no treinamento de atletas e na fisioterapia de pacientes com problemas de locomoção. Nestes casos, movimentos de indivíduos sadios são obtidos por meio de marcas em locais específicos do corpo, como as articulações. As informações coletadas destes movimentos, como ângulo de rotação e deslocamento, são examinadas e comparadas com as informações de pacientes com alguma anomalia em seus padrões de locomoção.

Dentro do conjunto dos movimentos corporais, os gestos da face e a linguagem de sinais são importantes durante a comunicação em condições adversas ou entre indivíduos portadores de deficiência auditiva (Goldschen, Garcia, e Petajan, 1994; Chiou e Hwang, 1997; Martin, 1992; Bregler e Konig, 1994; Lamar, Bhuiyan, e Iwata, 1999b; Lamar et al., 1999a). Em outro contexto, existem sistemas de reconhecimento automático da fala que empregam

informações bimodais *(informação visual + informação acústica)* para obter melhores taxas de reconhecimento na presença de ruído (Chibelushi, Deravi, e Mason, 2002; Iwano et al., 2001; Potamianos, Neti, Luettin, e Matthews, 2004; Chibelushi, Mason, e Deravi, 1997; Meiel, Hurst, e Duchnowski, 1996).

Uma outra situação em que é possível avaliar a importância dos movimentos é durante a exibição de uma *talking face*[1]. Quando os movimentos realizados por uma *talking face* não são os de uma pessoa real ou foram retirados de alguém com biotipo diferente, há uma sensação de falta de naturalidade. Por outro lado, *talking faces* são bem aceitas como locutores de informações quando exibem movimentos realísticos e sincronizados com a fala. Especificamente, os movimentos da cabeça e da face, realizados pela mandíbula, lábios, olhos e sobrancelhas, são importantes para o realismo destes locutores artificiais.

Durante a fala, existem movimentos básicos na face provenientes de necessidades biológicas (i.e. piscar dos olhos), de conseqüências do movimento do trato vocal (i.e. oscilação da mandíbula) e de formas auxiliares de comunicação (i.e. erguimento das sobrancelhas e balançar da cabeça na vertical e na horizontal). As correlações existentes entre os movimentos da cabeça e da face e a acústica da fala produzida simultaneamente foram estudadas em Yehia, Kuratate, e Vatikiotis-Bateson (2002); Barbosa (2000); Yehia et al. (1998); Kuratate, Jones, Callan, Kuratate, e Vatikiotis-Bateson (2004); Graf, Cosatto, Strom, e Huang (2002). Estes estudos mostram que a informação contida no movimento da cabeça tanto realçam, quanto complementam a informação contida na componente acústica da fala. Além da face, existem outros movimentos do corpo que têm o objetivo de enfatizar trechos da fala ou de deter a atenção do ouvinte, tais como gestos dos braços, ombros e mãos. Estes movimentos podem ser vistos como parte integrante da comunicação, adicionando informações à comunicação falada.

Finalmente, existem movimentos específicos de cada locutor, que expressam suas características individuais. Em Chen et al. (2001) utiliza-se o movimento facial para melhorar a identificação de indivíduos em imagens em que não há boa qualidade, devido a variações nas condições de iluminação. Assim, mesmo quando a face não está completamente visível, seu movimento pode ser um facilitador na identificação do locutor. Indo além da face, informações dos movimentos realizados durante a caminhada de indivíduos podem ser utilizadas para sua identificação (Nixon, Carter, Nash, Huang, Cunado, e Stevenage, 1999; Huang, Harris, e Nixon, 1998; Shutler, Nixon, e Harris, 2000; BenAbdelkader e Cutler, 2002).

[1]Faces falantes que são simulações com características de faces humanas durante a fala.

1.2 Modelos

O estudo da relação existente entre o conteúdo acústico e o movimento facial tem aplicação em faces falantes (*talking heads*), em leitura labial, estudos comportamentais do ser humano e estudos fonéticos e fonológicos. A análise de como estas informações se relacionam pode ser feita por modelos lineares ou não-lineares, estáticos ou dinâmicos e variantes ou invariantes no tempo (Barbosa, 2004).

O movimento facial é fortemente acoplado à acústica da fala. A componente linear deste acoplamento pode ser modelada por meio dos autovetores da matriz de correlação cruzada entre a posição de pontos da face do locutor e parâmetros extraídos da acústica da fala. De forma similar, o acoplamento entre regiões diferentes da própria face pode ser modelado pelos autovetores da matriz de covariância das posições de pontos distribuídos sobre a face. Observa-se, entretanto que, tanto os padrões de movimento da face, quanto o acoplamento entre o movimento facial e a acústica variam com o tempo. Este trabalho avalia, quantitativamente, a variabilidade dos autovetores utilizados para representar estes acoplamentos ao longo do tempo.

Independentemente do sexo e da idade, a anatomia da face e sua composição muscular são as mesmas para o ser humano. Porém, os músculos da face têm sua morfologia variando de pessoa para pessoa. Além disso, parte estrutural óssea da face (crânio), assim como a elasticidade da pele, também variam de indivíduo para indivíduo. Isto nos leva a considerar a hipótese de que os movimentos da face sejam capazes de diferenciar pessoas. Entretanto, para que esta informação seja de fato útil, é necessário medir a forma e a intensidade com que as diferenças existentes nos padrões de movimento acontecem e são percebidas. Neste sentido, este trabalho busca medir a variabilidade entre locutores dos autovetores da matriz de covariância das posições de pontos distribuídos sobre a face. A análise verifica o que é invariante e característico de cada locutor durante a produção da fala.

Resumindo, este trabalho tem como motivação: (i) a caracterização de um modelo de produção da fala que leva em conta a variabilidade dos autovetores usados para representar os padrões de movimento facial e seu acoplamento à acústica da fala ao longo do tempo; e (ii) a caracterização do locutor através do estudo da variabilidade desses autovetores de indivíduo para indivíduo.

1.3 Objetivo

O objetivo deste trabalho é analisar quantitativamente os movimentos que acontecem na face do ser humano de forma a: (i) verificar a variabilidade de parâmetros correspondentes ao movimento durante a produção da fala; (ii) verificar se tais movimentos são dependentes ou independentes do contexto; e (iii) verificar se tais movimentos são específicos de cada

pessoa. Desta forma, investiga-se até que ponto os movimentos da face são inerentes à anatomia humana e se podem ser um facilitador na identificação de indivíduos.

Especificamente, no contexto do estudo e caracterização do movimento da face e da acústica da fala, são objetivos deste trabalho:

- avaliar quantitativamente a variabilidade ao longo do tempo dos autovetores utilizados para representar ao acoplamentos entre as regiões da face;
- avaliar quantitativamente a variabilidade ao longo do tempo dos autovetores utilizados para representar o acoplamento entre o movimento facial e a acústica da fala; e
- fazer um estudo dos autovetores utilizados para representar padrões de movimento no processo de identificação de locutor.

1.4 Organização do texto

Este estudo está organizado da seguinte forma: no Capítulo 2 é relatada uma revisão bibliográfica de estudos sobre o movimento humano, sobre ferramentas de parametrização de dados usados para representação do movimento e para a identificação de indivíduos. Estas ferramentas são utilizadas ao longo do trabalho para compreensão dos fenômenos analisados. O Capítulo 3 apresenta os experimentos e o processamento dos sinais do movimento facial e da acústica da fala. No Capítulo 4, é feita uma análise dos resultados dos experimentos realizados para medir padrões de movimento facial e seu acoplamento à acústica da fala. O Capítulo 5 apresenta resultados e análise dos experimentos realizados para medir a variabilidade entre locutores. Finalmente, o capítulo 6 sintetiza as conclusões e propostas para a continuidade do trabalho.

Capítulo 2

Apreciação das informações visuais e acústicas

Neste capítulo são abordados temas como informação acústica, informação visual, movimento e classificadores, necessários para a compreensão dos fenômenos avaliados nos capítulos subseqüentes. Desta forma é fornecida aqui a base teórica para a explicação dos resultados experimentais obtidos e para sua análise.

Inicialmente, são mostrados alguns trabalhos que estudam a relação entre a informação visual e a acústica da fala. Após, elabora-se uma descrição contextualizada sobre a informação visual da fala e sobre como os movimentos da face se relacionam entre si e podem ser utilizados como um facilitador na identificação de pessoas.

2.1 Relações entre as informações acústicas e visuais

O ser humano utiliza a informação sonora contida na componente acústica da fala durante sua comunicação. Porém, a informação percebida pelo interlocutor, além da componente acústica, possui também uma componente visual. A informação visual representada pelo movimento da cabeça, lábios, mandíbula e bochechas facilita e complementa a percepção da fala (Vatikiotis-Bateson, Munhall, Hirayama, Lee, e Terzopoulos, 1996a; Vatikiotis-Bateson, Eigsti, Yano, e Munhall, 1998a). Vatikiotis-Bateson et al. (1998a) observaram que as pessoas, quando submetidas a ambientes ruidosos, necessitam mais da parcela visual da comunicação. Espectadores, ao assistir ao vídeo de um locutor falando na presença de ruído progressivamente mais intenso, dividiam sua atenção entre a região dos olhos e da parte inferior da face do locutor (onde se localiza a maior parte da informação visual da fala). À medida que o ruído aumentava, crescia também a atenção sobre a região inferior da face do locutor. Porém, mesmo na presença de níveis de ruído elevados, parte da atenção mantinha-se sobre os olhos do locutor. Concluíram, desta forma, que a detecção da informação visual contida

na fala acontece em baixa resolução temporal, que tal informação não se restringe à região da boca e que a atenção do ouvinte varia de acordo com a necessidade do ambiente.

A relação existente entre as componentes visual e a acústica da fala vem sendo explorada no meio científico, sendo aplicada na criação de movimentos realísticos de *talking faces* (Blanz, Basso, Poggio, e Vetter, 2003; Kuratate e Vatikiotis-Bateson, 2004; Yehia et al., 2002; Cohem e Massaro, 1990; Lucero e Munhallb, 1999), na estimação do movimento da face por meio da acústica da fala e vice-versa (Yehia et al., 1998; Yehia, Kuratate, e Vatikiotis-Bateson, 1999; Yehia et al., 2002; Barbosa, 2004; Vatikiotis-Bateson et al., 2002; Barbosa e Yehia, 2001; Jiang, Alwan, Keating, Auer, e Bernstein, 2002; Jiang, Alwan, Bernstein, Keating, e Auer, 2000), e em sistemas de comunicação audiovisual (Vatikiotis-Bateson et al., 1998b).

Os trabalhos apresentados acima demonstram que a informação visual da fala contida na face possui elementos lingüísticos importantes. Porém, além dos elementos lingüísticos, também existem informações paralingüísticas, que podem ser definidas pelo modo com que as frases são expressas. Por meio da paralingüística, pode-se identificar, por exemplo, o estado emocional do locutor (Pantic e Rothkrantz, 1999). Os movimentos da face, em conjunto com o movimento da cabeça, também indicam informações de prosódia, tais como entonação e tonicidade. Segundo Graf et al. (2002), na produção da fala, a direção e a intensidade do movimento da cabeça variam de locutor para locutor e estão sincronizadas com o conteúdo da fala. O movimento da cabeça, em conjunto com o movimento de partes específicas da face, tais como as sobrancelhas, relaciona-se com a freqüência fundamental e com a intensidade da voz do locutor (Kuratate et al., 2004; Graf et al., 2002; Yehia et al., 2002). Neste trabalho, o foco está limitado a aspectos acústicos visíveis do movimento facial e não a aspectos mais detalhados, tais como expressões de emoção e prosódia do discurso.

Segundo Yehia et al. (1998), existe uma associação entre o trato vocal, a acústica da fala e o movimento facial. Parte do movimento facial é determinada pelo movimento do trato vocal e parte do movimento do trato vocal é determinada em menor grau pelo movimento facial. Da mesma forma, existe um mapeamento não-linear entre a configuração do trato vocal e a acústica da fala. Como resultado, a acústica da fala pode ser parcialmente inferida a partir do movimento facial e vice-versa. Isto se deve ao fato da acústica da fala e grande parte do movimento facial durante a fala serem determinados pela configuração do trato vocal em movimento (Vatikiotis-Bateson et al., 2002).

Yehia, Rubin, e Vatikiotis-Bateson (1997); Yehia et al. (1998, 1999, 2002); Barbosa (2000, 2004); Jiang et al. (2000) e Jiang et al. (2002) verificaram e quantificaram as relações entre a configuração do trato vocal, a acústica da fala e o movimento facial. Nesses trabalhos, frequentemente estima-se o movimento facial, a partir da acústica da fala, com base em sentenças individuais. Em Vatikiotis-Bateson e Yehia (2000); Vatikiotis-Bateson et al. (2002); Yehia et al. (1998); Barbosa (2000); Yehia et al. (2002); Barbosa (2004) verifica-se que, com

o aumento do número de sentenças ou com a utilização de grupos de sentenças diferentes para treinamento e validação, o desempenho dos estimadores decresce.

Em Barbosa (2000), o movimento facial é estimado a partir da acústica da fala. Os estimadores apresentam um resultado inferior quando o conjunto utilizado para validar o sistema contém sentenças que não fazem parte do conjunto usado para treinamento. Em Barbosa (2004) utilizam-se de técnicas de identificação de sistemas para descobrir mapeadores que relacionam a acústica da fala e o movimento facial. Estes mapeadores, lineares e não-lineares, estáticos e dinâmicos, resultam em desempenhos bastante diferentes para predição de sentenças de curto e longo prazo. Para sentenças curtas, modelos dinâmicos e não-lineares apresentam um melhor resultado. Para sentenças longas, modelos estáticos proporcionam melhores resultados. Entretanto, algumas vezes os mapeadores lineares fornecem melhores resultados. Uma das hipóteses propostas por Barbosa (2004) é que a relação entre o movimento facial e a acústica da fala varia com o tempo, pois depende do conteúdo acústico falado. Neste contexto, este trabalho analisa como esta variabilidade acontece ao longo do tempo.

Além da relação existente do movimento da face e a acústica da fala, os próprios movimentos que acontecem na face são acoplados entre si, pois muitos dos músculos existentes na face são conectados a outros músculos da face, e a pele constitui um único tecido elástico (Zemlin, 2000). Por isso, os movimentos de partes da face são relacionados aos movimentos de outras partes da face. Nos movimentos da face existem muitas redundâncias, por exemplo, quando abrimos a boca, o centro do lábio inferior e o queixo apresentam movimentos acoplados. Estas redundâncias dão origem a um grande número de dados que podem ser representados por um conjunto de dados menor. Em Yehia et al. (1998, 2002); Vatikiotis-Bateson et al. (2002); Vatikiotis-Bateson e Yehia (2000); Kuratate, Munhall, Rubin, Vatikiotis-Bateson, e Yehia (1999); Barbosa (2000, 2004); Kroos, Kuratate, e Vatikiotis-Bateson (2002); Barbosa e Yehia (2001) os movimentos faciais são parametrizados utilizando análise em componentes principais (PCA). Análise em componentes principais é um método de transformação ortogonal, em que as variáveis são representadas em um novo sistema de coordenadas orientadas nas direções de máxima variabilidade dos dados sob análise. Estas novas variáveis contêm a mesma informação existente anteriormente, porém não são mais correlacionadas entre si e podem ser eficientemente representadas em um espaço cujo número de dimensões é igual ou próximo ao número de graus de liberdade do sistema analisado.

Em Kuratate e Vatikiotis-Bateson (2004), uma face é animada em três dimensões baseadas em fotografias e nas componentes principais de um modelo de face extraído de uma base de dados com informações em 3D. Utilizam-se informações dos autovetores de diferentes sujeitos para criar autovetores padrões e assim transferir o movimento facial para uma face animada.

2.2 Estudo do movimento para caracterização do indivíduo

Além da informação relativa ao conteúdo lingüístico, a acústica da fala e o movimento da face contêm informação relativa à identidade do locutor. Nesta seção analisa-se como é possível inferir informação inerente ao indivíduo com base nos seus padrões de movimento.

A principal forma de identificação dos seres humanos é a visão. Pela visão, é possível obter informações sobre a forma e sobre os padrões de movimento de uma pessoa. Tais informações, eventualmente somadas às características da voz permitem a identificação eficiente de indivíduos.

Nem sempre todas as características específicas de um indivíduo são acessíveis e, mesmo assim, seres humanos são capazes de reconhecer uma pessoa. Por exemplo, quando não se pode enxergar a face é relativamente fácil distinguir um conhecido através do modo de falar ou da forma de caminhar. De fato, é possível demonstrar que os movimentos da caminhada de uma pessoa podem ser utilizados para seu reconhecimento, quando padrões da face não são visíveis (Huang et al., 1998; Huang, Harris, e Nixon, 1999; Yam, Nixon, e Carter, 2002; Nixon et al., 1999; Shutler et al., 2000; BenAbdelkader e Cutler, 2002; Rangarajan, Allen, e Shah, 1992).

Assim como é possível identificar uma pessoa por seus movimentos realizados na caminhada, podem-se também associar os movimentos específicos da face para o processo de caracterização do indivíduo. Em Chen et al. (2001) é desenvolvido um sistema de reconhecimento de indivíduo baseado no movimento da face, estimado por fluxo óptico, (*optical flow*). O método descrito apresenta menor sensibilidade a variações de iluminação, possibilitando uma melhoria no reconhecimento facial. Segundo Knight e Johnston (1997); O'Toole, Roark, e Abdi (2002), informações dinâmicas contribuem mais significativamente para o reconhecimento em condições adversas de visão, como iluminação deficiente, resolução baixa e reconhecimento a distância.

O'Toole et al. (2002) descrevem duas hipóteses de como o ser humano reconhece pessoas pelas características do seu movimento facial: a primeira é a chamada hipótese suplementar da informação, que consiste na representação dos movimentos ou gestos faciais característicos dos indivíduos, somada à estrutura invariante da face. Esta hipótese tem como base a possibilidade de que, quando as informações estáticas e dinâmicas do indivíduo estão disponíveis, o ser humano confia primeiramente na informação estática para o reconhecimento. Isto ocorre porque as características faciais dinâmicas fornecem, provavelmente, menos informação confiável na identificação do que a estrutura facial estática. A segunda é a hipótese da representação, que consiste na possibilidade de o movimento facial contribuir para o reconhecimento, facilitando a percepção da estrutura tridimensional de uma face. Esta hipótese supõe que o movimento tenha adicionado qualidade à informação acessível da

estrutura da face humana e que este benefício transcende a visão da face estática. Assim do ponto de vista da psicologia, os seres humanos podem reconhecer uma face familiar dependendo não somente da geometria facial, mas também dos movimentos faciais.

Nesta mesma direção, porém, através de outra abordagem, Kuratate e Vatikiotis-Bateson (2004) mostraram que o movimento facial expressa não somente informações lingüísticas e paralingüísticas, mas também características pessoais. Desse modo, durante a produção da fala, cada pessoa possui características originais de deformação da face, determinadas pelas configurações geométricas do crânio e da mandíbula, estrutura muscular e propriedades da pele, tais como espessura e rigidez. Esta especificidade faz do movimento da face uma peculiaridade de cada indivíduo.

O primeiro passo para que se possa estudar o movimento da face no processo de reconhecimento do indivíduo é a medição deste movimento. Isto pode ser feito a partir de uma seqüência de quadros de vídeo da qual o movimento pode ser extraído. Em uma segunda etapa, os dados obtidos podem ser utilizados diretamente ou parametrizados para o processo do reconhecimento. Este processo consiste em uma comparação entre dados de classes conhecidas e dados de classes que se deseja classificar. Cada conjunto de dados não classificados é associado ao grupo conhecido mais próximo dele.

A medição do movimento pode ser utilizada tanto para reconhecer diferentes padrões de movimento humano, quanto para reconhecer pessoas diferentes realizando os mesmos padrões de movimento. Assim, o reconhecimento do movimento humano implica habilidade de discriminar ações diferentes, enquanto o reconhecimento do indivíduo pelo seu movimento implica habilidade de diferenciar suas características específicas.

O movimento pode ser analisado como um todo quando a percepção é global, ou de forma relativa, quando um movimento é analisado em relação a outro movimento. O movimento relativo é mais apropriado para o reconhecimento de objetos articulados, tais como a face humana, objeto de estudo deste trabalho (Cédras e Shah, 1995).

2.3 Técnicas de aquisição do movimento facial

A extração de parâmetros a partir do sinal de vídeo é um passo importante para o estudo do movimento, conseqüentemente, para o reconhecimento baseado em informações nele contidas. A aquisição pode ser realizada em 3 dimensões por equipamentos como o OPTOTRAK e o QUALISY (Yehia et al., 1997, 2002; Barbosa, 2004; Nordstrand et al., 2004); ou em duas dimensões, pela técnica conhecida como *Moving Light Display (MLD)*, em que adesivos são colocados sobre a face e rastreados ao longo do tempo (Barbosa, 2004, 2000; BenAbdelkader e Cutler, 2002; Cédras e Shah, 1995). Segundo Knight e Johnston (1997) um dos problemas desta técnica de aquisição é a produção de representações das amostras na velocidade de geração dos quadros do filme, implicando degradação das informações espaciais sobre a

estrutura facial e degradação das informações dinâmicas sobre mudanças na estrutura facial.

Uma outra técnica de medição de movimento é o *optical flow* (fluxo óptico) (Chen et al., 2001; Rekleitis, 1996; Black e Yacoob, 1997; Iwano et al., 2001; Huang et al., 1998; Cédras e Shah, 1995; Lee e Yang, 2006; Vatikiotis-Bateson, Munhall, Kasahara, Garcia, e Yehia, 1996b; Horn e Rhunck, 1981). Este método consiste no cálculo do campo vetorial que descreve os deslocamentos ocorridos entre dois quadros consecutivos de uma seqüência de vídeo. O *optical flow* é a aproximação em duas dimensões do campo de fluxo com base na intensidade da imagem (Huang et al., 1998; Cédras e Shah, 1995).

Neste trabalho, são usados marcadores adesivos para salientar pontos de interesse sobre a face. Tais marcadores são localizados ao longo dos quadros, resultando em trajetórias cartesianas bidimensionais como explicado na Seção 3.1.2.

2.4 Técnicas de aquisição da acústica da fala

O sinal acústico da fala pode ser conseguido sem maiores dificuldades por meio de um microfone e de um processo de conversão A/D (análogico-digital). Após a conversão A/D, o sinal é reamostrado a uma taxa apropriada para sua análise. Neste trabalho, seguem-se os passos descritos em Barbosa (2004); Yehia et al. (1998); Barbosa e Yehia (2001); Vatikiotis-Bateson et al. (1998b); Kuratate et al. (1999); Vatikiotis-Bateson e Yehia (1997, 2000); Jiang et al. (2002), em que o sinal acústico da fala é transformado em parâmetros LSP (Sugamura e Itakura, 1986), que são fortemente relacionados com a geometria do trato vocal, como detalhado na seção 3.2.2.

2.5 Classificadores

Conforme mencionado na seção anterior, um sistema de reconhecimento de movimento pode ser resumido na localização e isolamento do movimento, na extração de características discriminatórias e na classificação de padrões. Um processo de classificação consiste basicamente em mapear um conjunto de observações em um conjunto de classes.

O método utilizado para classificar pode ser um simples cálculo de distância entre modelos existentes e uma entrada não-conhecida, sendo selecionado o modelo que apresentar a menor distância (Cédras e Shah, 1995). Existe uma ampla gama de definições de distância que podem ser usadas no processo de classificação, destacando entre elas a distância Euclidiana e a distância de Mahalanobis, como mostram Huang et al. (1998); Brunelli e Poggio (1991); Huang et al. (1999).

Um exemplo prático de classificador é o algoritmo das k-médias, em que se particiona um espaço N-dimensional em M células e se associa um vetor quantizado ao centróide de

cada célula (Braga, Carvalho, e Ludermir, 2000; Huang, Acero, e Hon, 2001; Haykin, 2001; BenAbdelkader e Cutler, 2002; Nixon et al., 1999; Yam et al., 2002).

Outros exemplos de classificadores são os Modelos Ocultos de Markov (HMM) e as Redes Neurais Artificiais (RNA). Os Modelos Ocultos de Markov (HMM) são métodos estatísticos que calculam a probabilidade de um vetor de observações gerado por uma seqüência não-conhecida ter sido gerado por uma seqüência particular de estados. Nos Modelos Ocultos de Markov, os padrões são representados por uma rede com Q estados, caracterizada por uma matriz de probabilidades de transição entre estados e por um conjunto de funções de probabilidade de observações. A classificação é feita pela determinação dos parâmetros que maximizam a probabilidade de que uma seqüência de observações tenha sido gerada pelo modelo de referência (Huang et al., 2001; Rabiner e Juang, 1993a; Gavrila, 1999; Aggarwal e Cai, 1999). As Redes Neurais Artificiais são mapeadores universais de funções multi-variáveis (Braga et al., 2000). Em sua fase inicial, o problema passa por um processo de aprendizagem, em que as amostras são apresentadas à rede. Em seguida, a rede extrai as características necessárias para representar a informação desejada. Após a extração, as redes são capazes de generalizar, dando respostas coerentes para dados não-conhecidos (Braga et al., 2000; Haykin, 2001).

Em Yamato, Ohya, e Ishii (1992); Goldschen et al. (1994); Iwano et al. (2001); Kashi, Hu, Nelson, e Turin (1998); Cédras e Shah (1995); Manhews, Bangham, e Cox (1996); Chiou e Hwang (1997); Bregler e Konig (1994), modelos ocultos de Markov (HMM) são utilizados como reconhecedores. Já em Er et al. (2002); Wang, Tan, e Zhu (1991); Brunelli e Poggio (1991), o reconhecedor é uma rede neural do tipo RBF (Função de base radial). Já em Barbosa e Yehia (2001); Sehad, Hadid, Hocini, Djeddi, e Ameur (2000); Perelmuter, E., Vellasco, e Pacheco (1995); Fadzil e H. (1994); Lamar et al. (1999a); Vatikiotis-Bateson e Yehia (2000); Chibelushi et al. (1997); Bregler e Konig (1994), a rede utilizada foi uma MLP (Perceptron multi camadas).

2.6 Sumário

Neste capítulo foi apresentada uma visão global das técnicas usadas para a medição e a análise das componentes visual e acústica da fala humana. No próximo capítulo, são apresentados os experimentos realizados para a aquisição de sinais do movimento facial e da acústica da fala. Os dados obtidos são analisados nos capítulos seguintes com a finalidade de determinar as mudanças ocorridas nos parâmetros utilizados na representação acústica e visual da fala e do locutor.

Capítulo 3

Experimentação e construção da base de dados

A aquisição e o processamento dos sinais originados do movimento facial e da acústica da fala são mostrados ao longo deste capítulo. Nos capítulos subseqüentes, os dados adquiridos são analisados. O movimento facial é adquirido por meio do rastreamento da trajetória de adesivos marcadores afixados em pontos específicos da face (Barbosa, 2004). Simultaneamente, o sinal acústico da fala é adquirido por meio de um microfone do tipo condensador.

3.1 Aquisição dos sinais do movimento facial e da acústica da fala

Na obtenção da base de dados, sinais provenientes da acústica da fala e do movimento facial foram adquiridos simultaneamente. Os experimentos foram realizados em duas sessões distintas: a primeira com a participação de oito locutores e a segunda com três locutores pertencentes ao grupo de 8 locutores da primeira sessão. Os locutores proferiram trechos de aproximadamente três minutos da crônica, *O Popular*, publicada Veríssimo (1984). Cada sessão do experimento foi constituída de duas repetições de parte do texto, afixado atrás da filmadora. Após a aquisição, o texto foi dividido em três trechos de aproximadamente um minuto, mostrados na Tabela 3.1.

Tabela 3.1: Composição do corpus de frases extraído da crônica *OPopular* (Veríssimo, 1984). O texto foi dividido em três partes com aproximadamente um minuto cada.

Parte I
Um número recente da Veja trazia fotografias sensacionais das (como diria um inglês) "incomodações"na Irlanda do Norte. Todas eram de ganhar prêmio, mas uma me impressionou especialmente. Nela aparecia a versão irlandesa do Popular. É uma figura que sempre me intrigou. A foto da Veja mostra um soldado inglês espichado na calçada, protegido pela quina de um prédio, o rosto tapado por uma máscara de gás, fazendo pontaria contra um franco-atirador local. Atrás dele, agachados no vão de uma porta, dois ou três dos seus companheiros, também em plena parafernália de guerra, esperam tensamente para entrar no tiroteio. Há fumaça por todos os lados, um clima de medo e drama. Mas ao lado do soldado que atira, em primeiro plano, está o Popular. De pé, olhando com algum interesse o que se passa, com as mãos nos bolsos e um embrulho embaixo do braço. O Popular foi no armazém e na volta parou para ver a guerra.
Parte II
Sempre pensei que o Popular fosse uma figura exclusivamente brasileira. Nas nossas incomodações políticas, no tempo em que ainda havia política no Brasil, o Popular não perdia uma. Os jornais mostravam tanques na Cinelândia protegidos por soldados de baioneta calada e lá estava o Popular, com um embrulho embaixo do braço, examinando as correias de um dos tanques. Pancadaria na Avenida? Corria polícia, corria manifestante, corria todo mundo, menos o Popular. O Popular assistia. Cheguei a imaginar, certa vez, uma série de cartuns em que o Popular aparecia assistindo ao Descobrimento do Brasil, à Primeira Missa, ao Grito da Independência, à Proclamação da República... Sempre com seu embrulho debaixo do braço. E de camisa esporte clara para fora das calças.
Parte III
Não se deve confundir o Popular com o Transeunte, também conhecido como o Passante. O Transeunte ou Passante às vezes leva uma bala perdida, o Popular nunca. O Transeunte às vezes vai preso por engano, o Popular é que fica assistindo à sua prisão. O Transeunte, não raro, se compromete com os acontecimentos. Aplaude o visitante ilustre que passa, por exemplo. O Popular fica com as mãos nos bolsos e quase sempre presta mais atenção ao motociclo dos batedores do que à figura ilustre. O Transeunte pode se entusiasmar momentaneamente com uma frase de comício ou um drama na rua, e aí o Popular é que fica olhando para o Transeunte. O Popular não tem opinião sobre as coisas. Quando o rádio ou a televisão resolvem ouvir "a opinião de um popular"na rua, sempre se enganam. O Popular nunca é o entrevistado, é o sujeito que está atrás do entrevistado, olhando para a câmara.

As filmagens foram realizadas a uma taxa de 30 quadros/s gravadas em formato Mini-DV. No primeiro experimento foi utilizada uma filmadora digital Sony DCR TRV-110 NTSC, operando com uma resolução de 720×480 *pixels*. No segundo experimento foi utilizada uma filmadora digital Panasonic PV-GS300, também operando com uma resolução 720×480 *pixels*. Buscando melhorar a precisão na aquisição da posição dos marcadores, o segundo experimento usou a câmera posicionada com uma rotação de 90º, ou seja, 480 *pixels* de largura e 720 de altura, de forma a conseguir uma melhor resolução vertical, uma vez que é a principal direção em que ocorre o movimento durante a fala.

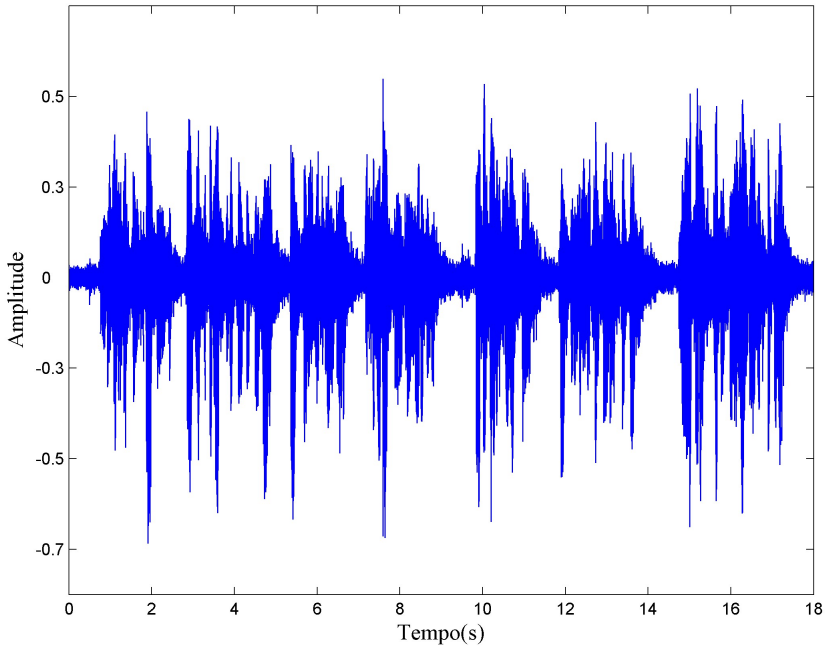


Figura 3.1: Primeiros 18 s do sinal de áudio adquirido. Os locutores proferiram trechos de aproximadamente três minutos da crônica *O Popular* (Veríssimo, 1984).

O sinal de áudio foi adquirido por meio de um microfone de condensação, a uma taxa de 48.000 amostras/s e 16 bits/amostra e reamostrado a 8.040 amostras/s. Nos experimentos realizados ao longo do trabalho, a gravação foi feita com uma relação sinal/ruído mínima de aproximadamente $30dB$. A Figura 3.1 mostra um trecho do sinal de áudio de um dos experimentos.

Os dados de áudio e vídeo adquiridos foram transferidos para um computador por meio de um software de captura de vídeo dvgrab e da interface gráfica Kino. A posição dos marcadores sobre a face foi então extraída das imagens que formam o sinal de vídeo, enquanto

que parâmetros acústicos foram extraídos do sinal de áudio como ilustrado na Figura 3.2 e descrito em detalhe nas próximas seções.

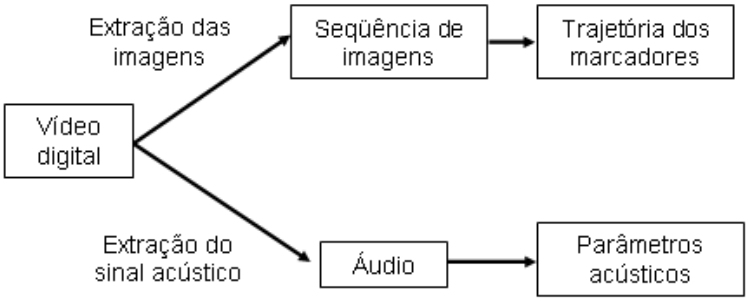


Figura 3.2: Diagrama que representa a aquisição de dados da acústica da fala e do movimento facial nos vídeos digitais. Os dados de áudio e vídeo foram adquiridos, sendo a posição dos marcadores sobre a face extraída das imagens que formam o sinal de vídeo e os parâmetros acústicos extraídos do sinal de áudio.

3.1.1 Posição dos marcadores na face

Ao longo dos experimentos, 28 marcadores adesivos foram colocados sobre a face dos locutores conforme mostra as figuras 3.3 e 3.4. Para que seja possível uma comparação consistente entre a posição de marcadores colocados sobre diferentes indivíduos, é necessário que sejam seguidas referências determinadas pela anatomia humana, conforme descrito a seguir.

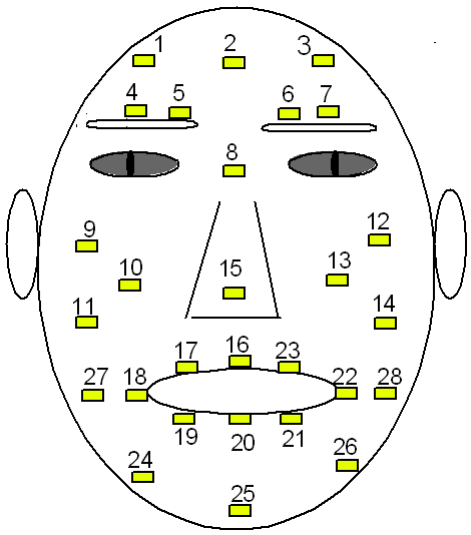


Figura 3.3: Localização dos adesivos marcadores posicionados em partes específicas da face dos locutores. Para cada locutor foram colocados vinte oito adesivos marcadores na cor azul, resultando assim, em uma maior distinção entre os adesivos e a pele do locutor.

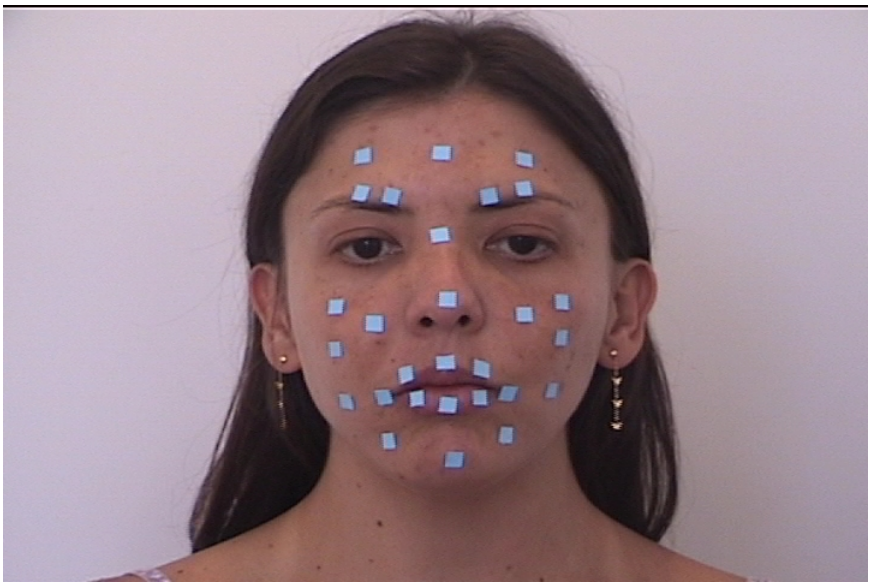


Figura 3.4: Localização dos adesivos marcadores na face de uma locutora. Foram colocados vinte oito adesivos marcadores na cor azul, resultando em uma maior distinção entre os adesivos e a pele do locutor.

A Figura 3.5a ilustra o procedimento seguido para a colocação dos marcadores sobre a face. Inicialmente, o marcador 15 é posicionado no ápice do nariz. A seguir, posicionam-se os marcadores 4 e 7 sobre as sobrancelhas, na direção das pupilas oculares. Os marcadores 5 e 6 são então posicionados também sobre as sobrancelhas, na direção dos cantos internos dos olhos. Os marcadores 18 e 22 são colocados sobre os cantos dos lábios. Os marcadores 16 e 20 são colocados no centro dos lábios superior (tubérculo) e inferior. Finalmente, o marcador 25 é colocado no centro da protuberância mentoniana.

Uma vez colocados os primeiros marcadores com base em referências anatômicas, os demais marcadores são posicionados com base nos marcadores já colocados como ilustra a Figura 3.5b. O marcador 8 é posicionado acima do marcador da ponta do nariz (15) e entre os olhos. Continuando no mesmo sentido e direção, no centro da testa encontra-se o marcador 2, posicionado de forma que o marcador 8 esteja no ponto médio entre os marcadores 15 e 2. A seguir, os marcadores 1 e 3 são posicionados nas interseções da reta horizontal que passa sobre o marcador 2 com as retas verticais que passam sobre os marcadores 4 e 7, respectivamente.

A posição dos marcadores na região da boca pode ser vista na Figura 3.5c. No lábio superior, entre os marcadores 16 e 18, encontra-se o marcador 17; e entre os marcadores 16 e 22, encontra-se o marcador 23. No lábio inferior entre os marcadores 18 e 20, localiza-se o marcador 19; e entre os marcadores 20 e 22, o marcador 21. A seguir, os marcadores 10 e 13 são posicionados sobre as interseções da reta horizontal que passa sobre o marcador 15 com as retas verticais que passam sobre os marcadores 18 e 22, respectivamente. De forma

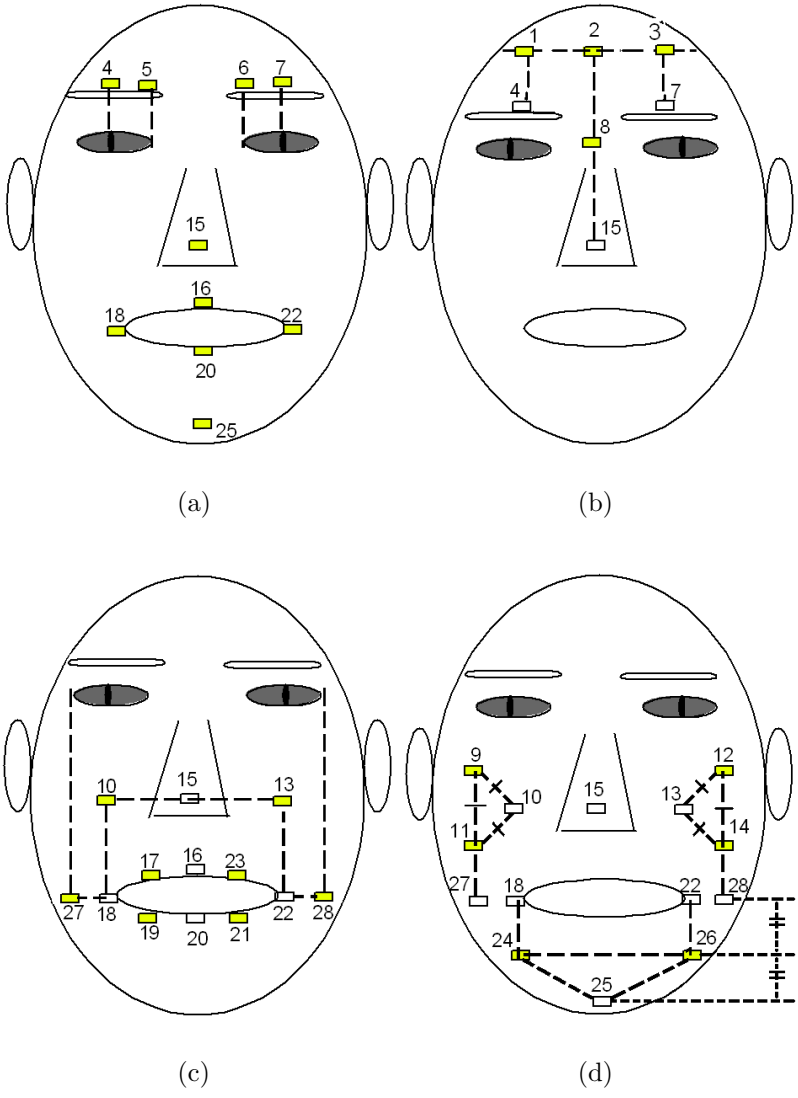


Figura 3.5: Localização dos adesivos marcadores posicionados em partes específicas da face dos locutores. (a) Primeiro passo na localização dos marcadores colocados observando partes anatômicas da face. (b) Segunda etapa na localização dos marcadores com base nos marcadores anteriores: região da fronte. (c) Terceira etapa na localização dos marcadores: região da boca. (d) Quarta etapa na localização dos marcadores: queixo e zigomáticos.

semelhante, os marcadores 27 e 28 são posicionados sobre as interseções da reta horizontal que passa sobre os marcadores 18 e 22 com as retas verticais que passam sobre os cantos externos dos olhos.

A última etapa é ilustrada na Figura 3.5d. Os marcadores 24 e 26 são posicionados sobre os pontos médios dos segmentos determinados pelos marcadores 18 e 25 e pelos marcadores 22 e 25, respectivamente. Finalmente, os marcadores 9 e 11 são posicionados sobre a reta vertical que passa sobre o marcador 27, de maneira a formar um triângulo eqüilátero com o marcador 10. Da mesma forma, os marcadores 12 e 14 são posicionados sobre a reta vertical que passa sobre o marcador 28 de maneira a formar um triângulo eqüilátero com o marcador 13.

3.1.2 Rastreamento de marcadores

Como resultados do primeiro experimento foram obtidos dois filmes com aproximadamente três minutos cada, para cada um dos oito locutores, enquanto para o segundo experimento, foram gerados dois filmes com aproximadamente dois minutos cada, para cada um dos três locutores. Considerando um filme como uma seqüência de imagens, um procedimento para aquisição do movimento facial consiste na medição dos deslocamentos ao longo do tempo da posição dos marcadores definidos na Seção 3.1.1. Os adesivos usados para salientar os pontos de interesse são localizados ao longo dos quadros, resultando em trajetórias bidimensionais de cada um dos marcadores. O procedimento aqui adotado é o mesmo descrito em Barbosa (2004). Nesse trabalho, a primeira etapa na localização de um marcador consiste na definição de uma região de busca, que é uma pequena região retangular onde o marcador será procurado no próximo quadro da seqüência de vídeo. O centro da região de busca de um marcador em um dado quadro é calculado tomando como referência as posições conhecidas de outros dois marcadores previamente selecionados. Os dois marcadores de referência e o marcador a ser localizado formam um triângulo que tem como característica o fato de variar o mínimo possível entre os quadros consecutivos. Este procedimento é chamado de triangulação e evita perdas nas localizações dos marcadores que variam mais bruscamente.

Dentro da região de busca, a localização dos marcadores é feita pela cor. Nos experimentos, foram colocados na face dos locutores adesivos com uma cor que se destaca da cor da pele, facilitando sua localização. As componentes RGB das imagens são combinadas linearmente de forma a realçar as regiões dos marcadores e, assim, possibilitar sua detecção na imagem de forma robusta.

Nos experimentos realizados ao longo do trabalho, locutores foram posicionados em frente à câmera em que movimentos da face e da cabeça foram adquiridos. Para observar a qualidade do sinal obtido, as trajetórias dos marcadores ao longo do tempo foram representadas

graficamente, como mostra a Figura 3.6, em que os gráficos das letras a e b, correspondem, respectivamente, ao movimento horizontal e vertical dos marcadores para uma dada sentença. Observa-se que o movimento vertical na região da boca e do queixo representa maior variação do movimento em relação aos demais marcadores. Observa-se também que o movimento horizontal dos marcadores possui uma menor variabilidade.

3.1.3 Compensação do movimento da cabeça

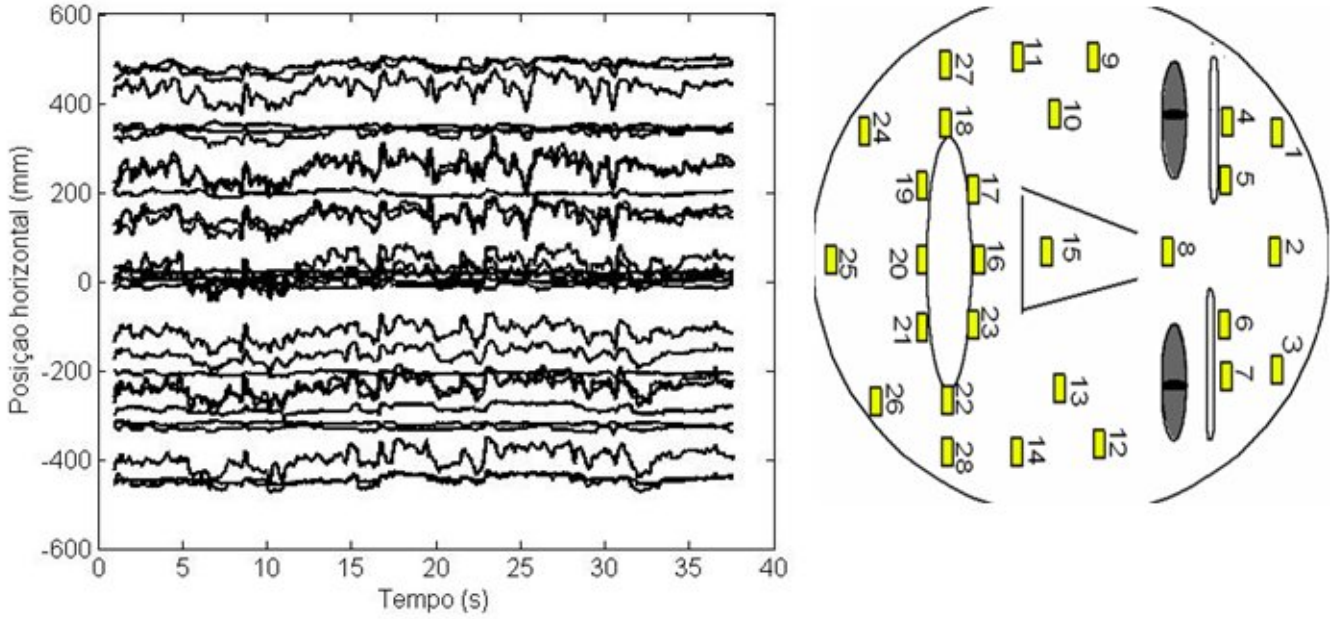
Durante a comunicação, é natural que as pessoas movimentem a cabeça e exibam expressões faciais. O movimento da cabeça está relacionado em maior ou menor grau com a freqüência fundamental e com a intensidade da voz do locutor durante a produção da fala (Kuratate et al., 2004; Graf et al., 2002; Yehia et al., 2002). Segundo Graf et al. (2002), a direção e a intensidade do movimento da cabeça variam de locutor para locutor e estão acopladas com ao texto falado.

Desta forma, nos sinais adquiridos dos marcadores, existem contribuições de movimentos das regiões da face e contribuições de movimentos oriundos da cabeça. Especificamente neste trabalho, o foco são os movimentos que acontecem na face de cada locutor. Assim, é necessário remover o movimento da cabeça para que seja possível analisar os movimentos da face.

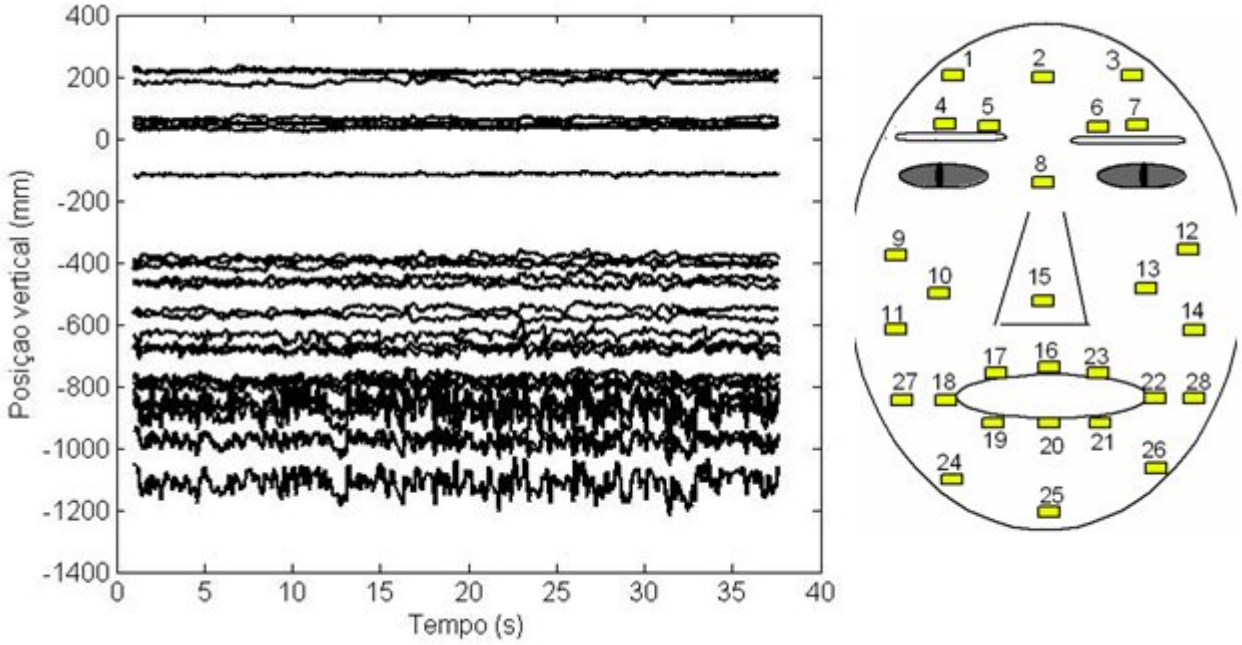
Na condição de corpo rígido, a cabeça possui 6 graus de liberdade (3 de rotação e 3 de translação), que devem respeitar as restrições impostas pela ligação da cabeça ao pescoço e do pescoço ao corpo. A Figura 3.7 ilustra os principais graus de liberdade da cabeça: (a) rotação em torno de eixo vertical; (b) rotação em torno de eixo horizontal; (c) translação ao longo do eixo x (horizontal); (d) rotação em torno de eixo perpendicular ao plano da face; (e) translação ao longo do eixo definido pelo eixo perpendicular ao plano da face; e (f) translação ao longo do eixo definido pelo eixo vertical. Ao longo dos experimentos observou-se que apenas os movimentos ilustrados na Figura 3.7 (c), (d) e (f) ocorreram significativamente, sendo as únicas componentes compensadas neste trabalho.

A compensação do movimento da cabeça é feita utilizando cinco marcadores da face fortemente acoplados à estrutura fixa da cabeça, localizados sobre a testa e o nariz: marcadores 1, 2, 3, 8 e 15. Inicialmente, o ponto médio destes cinco marcadores é utilizado como origem de um novo sistema de coordenadas. A seguir, para compensar o movimento de rotação em torno do eixo z, determina-se o ângulo de rotação como o ângulo entre as retas definidas pelos marcadores 8 e 2 em um quadro de referência e no quadro atual.

As posições dos marcadores em relação ao novo sistema de coordenadas são dadas pela translação para a nova origem e da multiplicação por uma matriz de rotação.



(a)



(b)

Figura 3.6: Movimento dos marcadores ao longo do tempo para um locutor proferindo parte de um dos trechos da crônica, após a remoção do movimento da cabeça. (a) Movimento horizontal dos marcadores. (b) Movimento vertical dos marcadores.

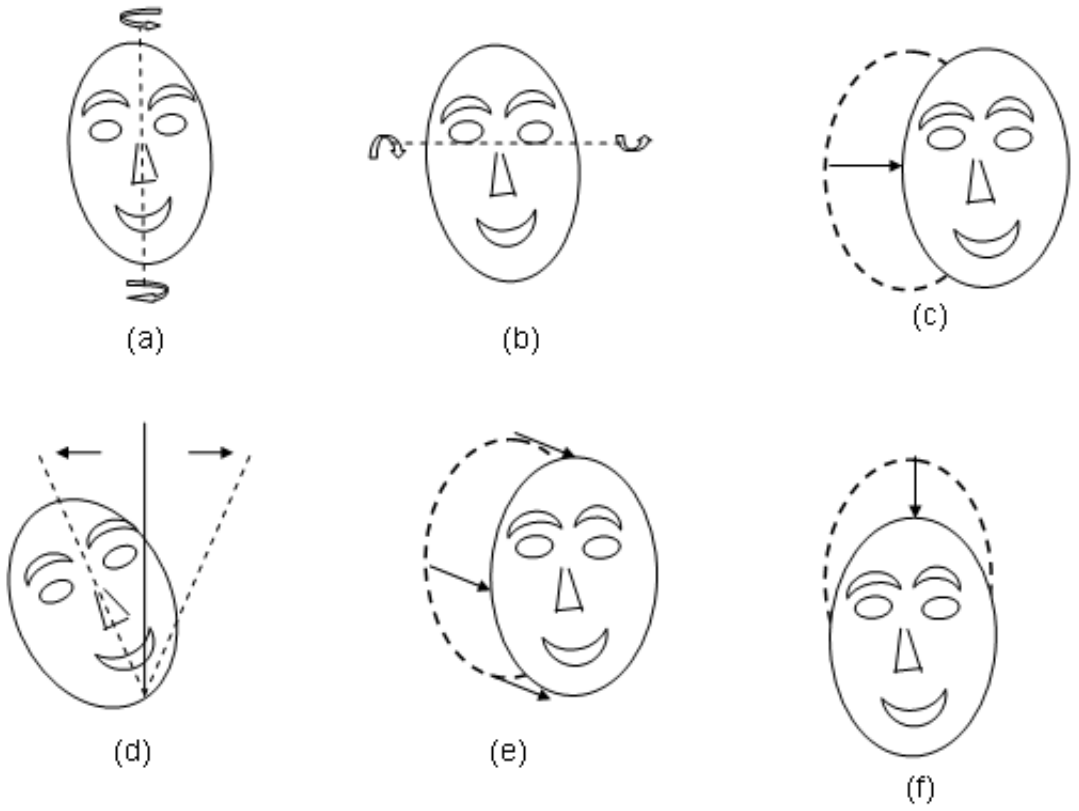


Figura 3.7: Tipos de movimentos que acontecem na cabeça do locutor. (a) Movimento de rotação do eixo y (vertical). (b) Movimento de rotação do eixo x (horizontal). (c) Movimento de translação ao longo do eixo x (horizontal). (d) Movimento de rotação do eixo z (perpendicular ao plano da face). (e) Movimento de translação ao longo do eixo z (perpendicular ao plano da face). (f) Movimento de translação ao longo do eixo y (vertical).

$$\begin{bmatrix} X' \\ Y' \end{bmatrix} = \begin{bmatrix} cos(\alpha) & sen(\alpha) \\ -sen(\alpha) & cos(\alpha) \end{bmatrix} \cdot \begin{bmatrix} X - O_X \\ Y - O_Y \end{bmatrix}. \tag{3.1}$$

Após a compensação do movimento da cabeça, obtém-se uma aproximação da componente facial do movimento de cada marcador.

A seguir, mudanças bruscas nas posições dos marcadores, possivelmente derivadas de erros de medição, são eliminadas por um filtro *Butterworth* passa-baixas de oitava ordem, com uma freqüência de corte de 10Hz (Barbosa, 2004), pois praticamente toda a energia do movimento facial encontra-se abaixo de 8Hz (Yehia et al., 1999). Assim, a freqüência de corte de 10 Hz é baixa o bastante pra atenuar erros de medição de alta freqüência e alta o suficiente para não afetar a medição do movimento facial. A Figura 3.9 mostra o movimento dos marcadores faciais antes e após o procedimento de compensação do movimento da cabeça para uma sentença.

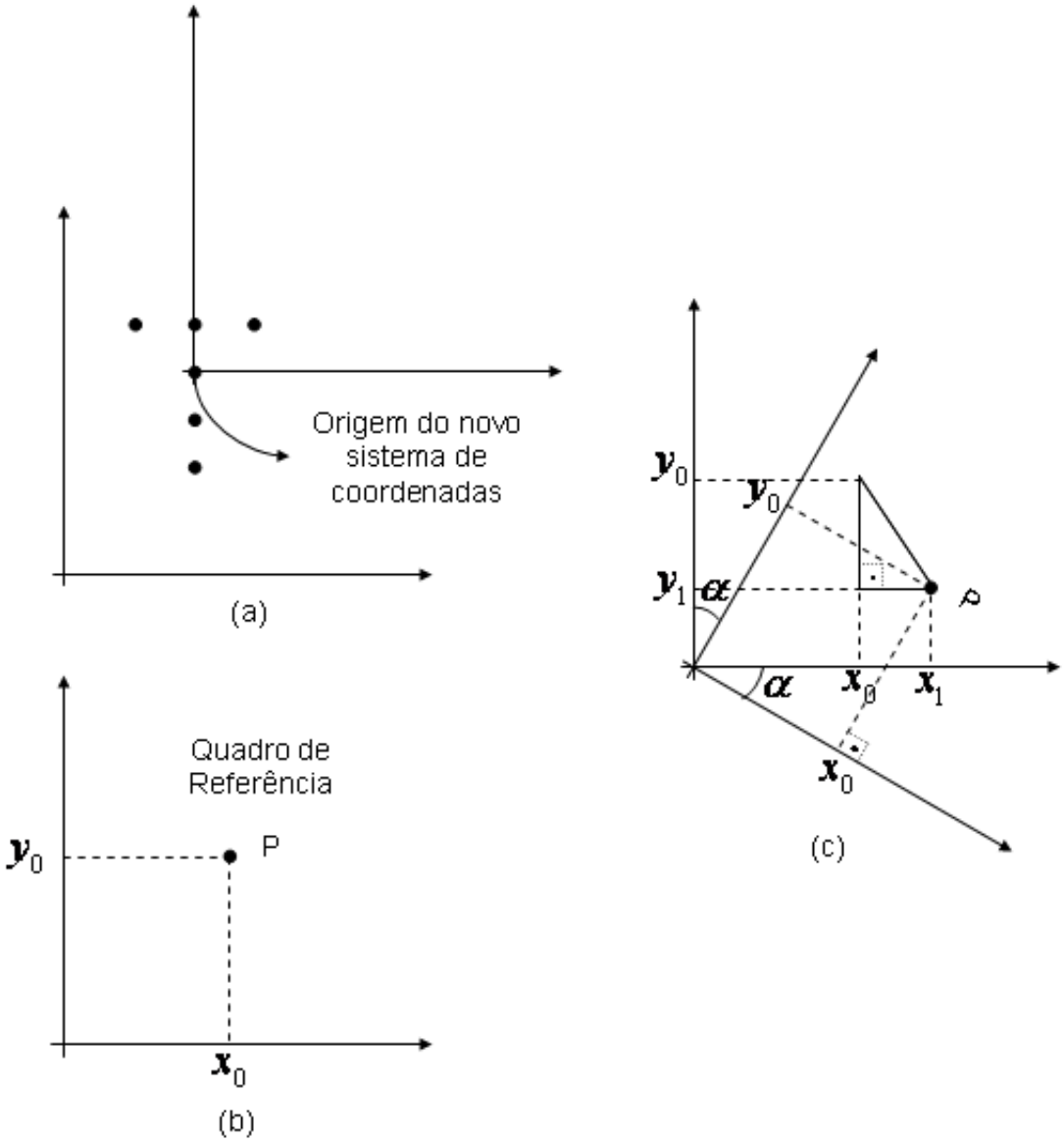


Figura 3.8: Compensação da rotação do movimento da cabeça em torno do eixo z. (a) Novo sistema de coordenadas calculado a partir de cinco marcadores da face fortemente acoplados à estrutura fixa da cabeça, localizados sobre a testa e o nariz. (b) Referência do sistema. (c) Compensação do ângulo de rotação da cabeça.

3.2 Representação paramétrica dos dados

Inicialmente, para a análise dos dados obtidos nos experimentos (Tabela 3.1), foram eliminados os trechos de silêncio, uma vez que, em tais trechos, não é possível analisar a relação entre o movimento facial e a acústica da fala. A seguir, o movimento facial e a acústica da fala são representados por parâmetros que refletem tão diretamente quanto possível o acoplamento entre as componentes visual e a acústica da fala, conforme detalhado nas seções 3.2.1 e 3.2.2.

3.2.1 Representação paramétrica do movimento dos marcadores

A posição de cada marcador é representada por suas componentes cartesianas x e y. Assim, um conjunto de $N = 28$ marcadores é representado por $2N = 56$ dimensões. Entretanto, o número de graus de liberdade do movimento facial é significativamente inferior a este valor. Desta forma, torna-se importante encontrar um sistema de coordenadas cujo número de dimensões reflita o número de graus de liberdade (i.e. a dimensionalidade) do movimento facial. Esta tarefa é realizada através de Análise em Componentes Principais (PCA) (Yehia et al., 1998, 1999, 2002; Vatikiotis-Bateson et al., 2002; Vatikiotis-Bateson e Yehia, 2000; Kuratate et al., 1999; Barbosa, 2000, 2004; Kroos et al., 2002; Barbosa e Yehia, 2001; da Silva, 2001).

Análise em Componentes Principais - PCA

Inicialmente uma matriz $2N \times M$, em que N representa o número de marcadores e M o número de quadros, contém as medidas realizadas durante uma elocução:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,M} \\ \vdots & \vdots & \ddots & \vdots \\ x_{2N,1} & x_{2N,2} & \cdots & x_{2N,M} \end{bmatrix}. \tag{3.2}$$

Cada quadro é representado por $2N$ parâmetros, sendo os N primeiros a representação do eixo horizontal, e o restante, a representação do eixo vertical. Cada linha da matriz acima contém a trajetória de um marcador ao longo dos M quadros. Uma vez que os movimentos de diferentes marcadores são fortemente correlacionados, é possível trabalhar em uma base ortogonal em que os movimentos faciais são adequadamente caracterizados a partir de um conjunto de $K < 2N$ componentes principais descorrelacionadas. O cálculo das componentes principais inicia-se com a remoção da média de cada uma das linhas da matriz X de movimento:

$$X_0 = X - \mu_X; \tag{3.3}$$

seguida do cálculo da matriz de covariância entre as linhas de X:

$$C_{XX} = \frac{1}{M-1} X_0 X_0^T; \tag{3.4}$$

em que μ_X representa o vetor médio dos marcadores ao longo dos quadros e M é o número de quadros. A técnica de Decomposição em Valores Singulares (SVD) é então usada para expressar a matriz de covariância C_{XX} na forma

$$C_{XX} = U S_{XX} U^T, \tag{3.5}$$

em que U é uma matriz cujas colunas contêm os autovetores de C_{XX} normalizados e S_{XX} é uma matriz diagonal com os autovalores correspondentes em ordem decrescente ao longo da diagonal principal (Horn e Johnson, 1985, pp.411–455). O valor da soma dos K primeiros autovalores de S_{XX} é igual à variância correspondente às K primeiras componentes principais.

Uma matriz X, que contém o movimento dos marcadores, é aproximada por meio da Equação 3.7 de acordo com a variância total observada. As K primeiras colunas de U são utilizadas para definir as componentes principais, Equação 3.6, como se segue:

$$P = U_K^T X_0, \tag{3.6}$$

$$X \approx U_K P + \mu_X. \tag{3.7}$$

A matriz U_K contém os autovetores das K primeiras componentes principais que representam os N marcadores do movimento facial. Assim, o movimento facial é representado pelas componentes principais do movimento facial, P, existindo uma redução da dimensão de X ($2N \times M$) para a dimensão de P ($K \times M$).

3.2.2 Representação paramétrica da acústica da fala

Para a representação do sinal acústico da fala utiliza-se de coeficientes LSP (*Line Spectrum Pairs*) (Sugamura e Itakura, 1986) que são eficientes por estarem ligados às freqüências de ressonância do trato vocal, os formantes. Esta representação é justificada porque os formantes são determinados pela geometria do trato vocal, e o formato do trato vocal tem forte influência sobre os movimentos realizados na face (Yehia et al., 1998). Os parâmetros LSP são fundamentados no princípio de conservação da envoltória do espectro da fala que, segundo Sugamura e Itakura (1986), é suficiente para assegurar a inteligibilidade do sinal.

O sinal acústico da fala pode ser modelado como a saída de um filtro linear variante no tempo excitado por pulsos quase-periódicos no caso de fala sonora ou por ruído branco no caso de fala surda (Sugamura e Itakura, 1986), (Barbosa, 2004), (Flanagan, 1972). Em pequenos segmentos, o sinal da fala pode ser representado por um modelo fonte-filtro em que o sinal da pressão sonora é o produto: da velocidade do volume de ar gerado pela fonte, da característica de propagação dos lábios e da configuração do trato vocal. Assumindo que o processo de produção da fala seja estacionário em um curto intervalo de tempo, pode-se definir uma função de transferência para o trato vocal dentro deste intervalo. Na fala sonora, a função de transferência do trato vocal possui somente pólos, entretanto em sons surdos e nasais, normalmente, a função de transferência possui zeros e pólos, porém os zeros podem ser aproximados por pólos. Assim, o sinal da fala pode ser visto aproximadamente como um sinal de saída de um filtro de polos (Sugamura e Itakura, 1986), (Barbosa, 2004), (Flanagan, 1972).

Além da filtragem executada pelo trato vocal, a radiação labial e o fluxo glotal também contribuem para o processo de filtragem. Contudo, o fluxo de volume glotal durante um único período de vibração possui apenas pólos; e a radiação do som ao sair da boca possui zeros que por sua vez pode ser aproximado em pólos (Barbosa, 2004). Assim, a função de transferência, no domínio z, do fluxo de volume glotal e da radiação labial pode ser representada aproximadamente como:

$$G(z)R(z) = \frac{K_1 K_2 (1 - z_{-1})}{(1 - z_a z_{-1})(1 - z_b z_{-1})}, \tag{3.8}$$

onde $G(z)R(z)$ é transformada z da contribuição conjunta do fluxo de volume glotal e da radiação labial. K_1 é uma constante relacionada com a amplitude do fluxo glotal e z_a e z_b são pólos relacionados com o fluxo glotal localizados no eixo real dentro do círculo unitário para que o filtro seja estável. K_2 é uma constante relacionada com a amplitude do fluxo de volume nos lábios e a distância dos lábios ao microfone.

Um modelo funcional do processo de produção da fala com base no modelo de pólos, em que as contribuições conjuntas do fluxo glotal, do trato vocal e da radiação labial são representadas por um único filtro auto-regressivo, linear de ordem p no instante j-ésimo, fundamentado na predição linear do sinal da fala (LPC), é descrito por:

$$\hat{s}(j) = -\sum_{i=1}^{p} \alpha_i s(j - i), \tag{3.9}$$

em que $\hat{s}(j)$ é o valor do sinal predito, $s(j - i)$ são os valores passados observados e α_i, $i = 1, ..., p$, são os coeficientes de predição linear que respondem pela ação de filtragem executada pelo trato vocal, pela radiação labial e pelo fluxo glotal. E a função de transferência do filtro de pólos é:

$$H(z) = \frac{1}{1 + \sum_{i=1}^{p} \alpha_i z^{-i}}. \tag{3.10}$$

A análise por predição linear (LPC) objetiva uma boa estimação das propriedades espectrais do sinal. Para isto, os coeficientes de predição (α_1, ...,α_p) são obtidos em cada fala correspondente aos quadros, em que p é ordem do filtro usada na análise. Para uma taxa de amostragem de $f_s = 8$ kHz, há aproximadamente 4 freqüências de ressonância até a freqüência de Nyquist (4 kHz), implicando necessidade de 8 coeficientes de predição linear (2 coeficientes para cada par de pólos conjugados). Além disso, verificou-se ser útil empregar um par extra de coeficientes para representar a inclinação espectral determinada pela influência do pulso glotal e pela carga de irradiação nos lábios. Assim, utiliza-se um filtro de predição de ordem $p = 10$ (Flanagan, 1972), (Atal e Hanauer, 1971).

Equivalente aos parâmetros LPC, no domínio da freqüência, um novo conjunto de parâmetros chamados de LSP *(Line Spectrunm Pairs)* é definido $(w_1, \theta_1, ..., w_{p/2}, \theta_{p/2})$. Estes parâmetros LSP (w_i, θ_i) são obtidos a partir de um filtro de pólos estável:

$$A_p(z^{-1}) = 1 + \sum_{i=1}^{p} a_i z^{-i}. \tag{3.11}$$

O objetivo dos parâmetros LSP é de representar o polinômio $A_p(z^{-1})$ por meio de dois outros polinômios cujos zeros estão sobre a circunferência unitária:

$$P(z^{-1}) = A_p(z^{-1}) - z^{-(p+1)}A_p(z) = 1 + (a_1 - a_p)z^{-1} + \cdots + (a_p - a_1)z^{-p} - z^{-(p+1)}, \tag{3.12}$$

$$Q(z^{-1}) = A_p(z^{-1}) + z^{-(p+1)}A_p(z) = 1 + (a_1 + a_p)z^{-1} + \cdots + (a_p + a_1)z^{-p} + z^{-(p+1)}, \tag{3.13}$$

reconstruindo:

$$A_p(z^{-1}) = \frac{1}{2}[P(z^{-1}) + Q(z^{-1})]. \tag{3.14}$$

Considerando que todas as raízes do polinômio, (i.e e^{+jw} e e^{-jw}), estejam sobre o círculo unitário, ou seja, o filtro LSP é estável, e expressando o polinômio como um produtório, obtêm-se:

Para p par:

$$P(z^{-1}) = (1 - z^{-1}) \prod_{i=1}^{p/2} (1 - 2\cos w_i z^{-1} + z^{-2}), \tag{3.15}$$

$$Q(z^{-1}) = (1 + z^{-1}) \prod_{i=1}^{p/2} (1 - 2\cos \theta_i z^{-1} + z^{-2}). \tag{3.16}$$

Para p ímpar:

$$P(z^{-1}) = (1 - z^{-1}) \prod_{i=1}^{(p-1)/2} (1 - 2\cos w_i z^{-1} + z^{-2}), \tag{3.17}$$

$$Q(z^{-1}) = \prod_{i=1}^{(p+1)/2} (1 - 2\cos \theta_i z^{-1} + z^{-2}). \tag{3.18}$$

Assim, um conjunto de parâmetros (w_i, θ_i) é obtido e convertido em (f_i, g_i) em Hertz:

$$f_i = w_i(2\pi T) \tag{3.19}$$

e

$$g_i = \theta_i(2\pi T). \tag{3.20}$$

em que T é o período de amostragem. Os parâmetros LSP são:

$$\mathbf{f} = [f_1, g_1, f_2, g_2, \cdots, f_{\frac{p}{2}}, g_{\frac{p}{2}}]. \tag{3.21}$$

Neste trabalho, o sinal de voz é amostrado a uma taxa de 8040 amostras/s e dividido em quadros de 134 amostras (i.e. 60 quadros/s). Cada quadro é multiplicado por uma janela de Hamming, para reduzir efeitos causados pelo janelamento. A cada quadro foi aplicada análise LPC de ordem 10. Os coeficientes LPC foram convertidos em coeficientes LSP, usados para representar cada quadro acusticamente. A Figura 3.10 mostra a parametrização de um trecho do sinal de voz, em que a envoltória espectral é a resposta em freqüência do filtro LPC. Os parâmetros LSP são representados nessa figura pelas linhas verticais. Por sua vez, a Figura 3.11 mostra as trajetórias dos parâmetros LSP ao longo do tempo.

3.3 Acoplamento entre a acústica da fala e o movimento facial

O acoplamento linear entre o vetor da face e os coeficientes LSP pode ser expresso como Yehia et al. (1998)

$$X_0 \approx T_{XF} F_0, \tag{3.22}$$

$$X_0 = X - \mu_X, \tag{3.23}$$

$$F_0 = F - \mu_F, \tag{3.24}$$

em que μ_X e μ_F representam, respectivamente, o vetor médio dos marcadores e o vetor médio dos coeficientes LSP. E T_{XF} é o estimador linear do erro médio quadrático mínimo (MMSE) definido por

$$T_{XF} = X_0 F_0^T (F_0 F_0^T)^{-1}. \tag{3.25}$$

Se as principais componentes alinhadas são analisadas, aplica-se Decomposição em Valores Singulares (SVD), com a finalidade de expressar a matriz de correlação cruzada entre os movimentos dos marcadores facial e os parâmetros LSP extraídos da fala (Horn e Johnson, 1985, pp.411–455).

$$C_{XF} = \frac{1}{M-1} X_0 F_0^T, \tag{3.26}$$

$$C_{XF} = U_{XF} S_{XF} V_{XF}^T, \tag{3.27}$$

em que S_{XF} é uma matriz diagonal com os autovalores correspondentes em ordem decrescente ao longo da diagonal, enquanto U_{XF} é uma matriz unitária cujas colunas contêm os autovetores normalizados da componentes acusticamente alinhadas de $C_{XF}C_{XF}^T$, e V_{XF} é uma matriz unitária cujas colunas contêm os autovetores normalizados da componentes acusticamente alinhadas de $C_{XF}^T C_{XF}$. Assim,

$$P_X = U_{XF}^T X_0, \tag{3.28}$$

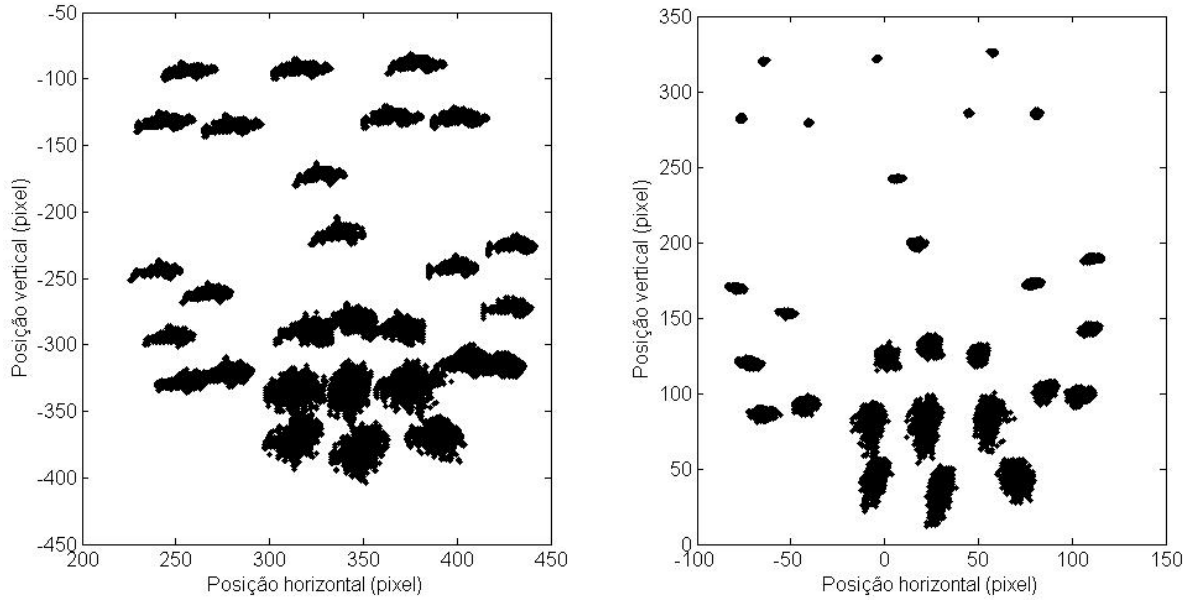
e

$$P_F = V_{XF}^T F_0, \tag{3.29}$$

representam o movimento facial e a componente acústica em um sistema de coordenadas que são otimamente alinhados. Isto é, que maximiza a correlação existente entre cada componente de posição facial e a componente acústica correspondente.

3.4 Sumário

Este capítulo descreve os experimentos nos quais os movimentos dos marcadores e o sinal acústico da fala são adquiridos. Após a aquisição e remoção do movimento da cabeça, os dados são representados parametricamente, na forma mais adequada para representar o acoplamento entre movimento facial e acústica da fala. Em uma primeira etapa, a análise em componentes principais (PCA) é usada para reduzir o número de parâmetros necessários para representar o movimento facial. Assim, os autovetores da matriz de covariância das posições dos marcadores modelam o acoplamento entre regiões diferentes da face. Por outro lado, dos sinais de áudio obtêm-se os coeficientes LPC, que são convertidos em parâmetros LSP, diretamente ligados às freqüências de ressonância do trato vocal, as quais, por sua vez, são determinadas pela geometria do trato vocal. Como a configuração do trato vocal tem forte influência sobre os movimentos que ocorrem na face, a componente linear do acoplamento entre o movimento facial e acústica da fala é modelada pelos autovetores obtidos a partir da matriz de correlação cruzada entre a posição dos marcadores e os parâmetros LSP. No próximo capítulo, são analisados os dados parametrizados, representados pelas componentes principais do movimento facial e pelas componentes do movimento facial acusticamente alinhadas com os parâmetros LSP.



(a) (b)

Figura 3.9: Posição dos marcadores durante uma elocução. (a) Marcadores faciais antes da compensação do movimento da cabeça; (b) marcadores faciais depois da compensação do movimento da cabeça.

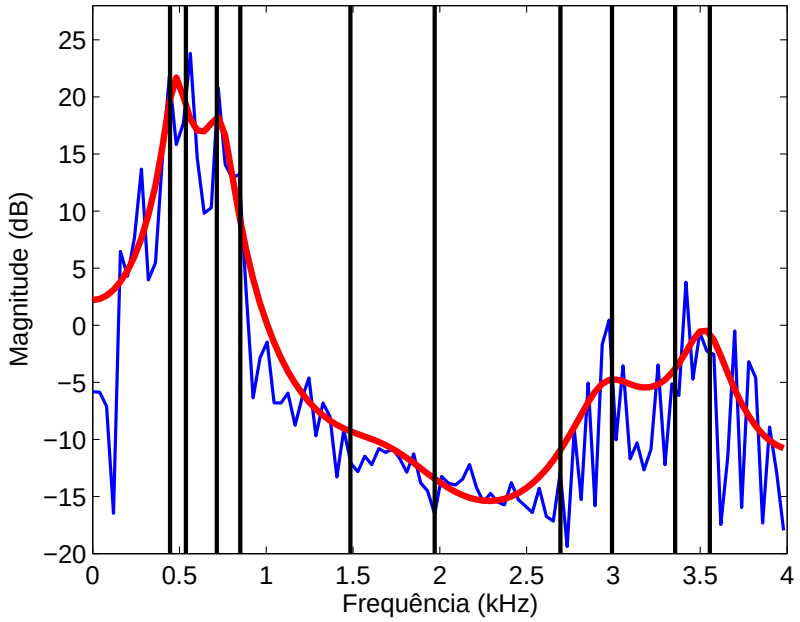


Figura 3.10: Parametrização do sinal acústico, em que a envoltória espectral é a resposta em freqüência do filtro LPC. Os parâmetros LSP são representados pelas linhas verticais. Os pares dos parâmetros LSP são usados para representar cada quadro acusticamente.

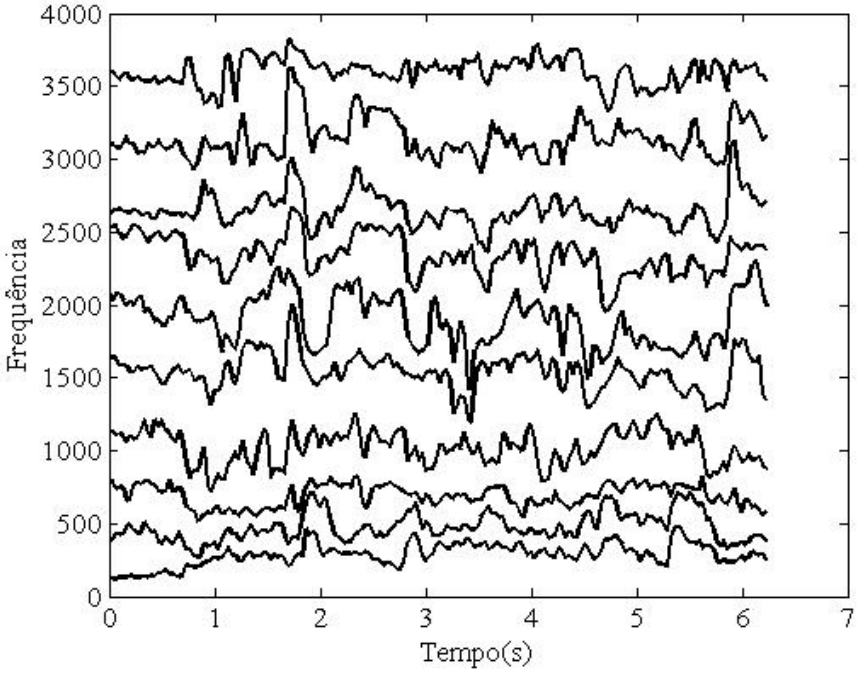


Figura 3.11: Exemplo das trajetórias dos dez parâmetros LSP ao longo do tempo. Estes parâmetros LSP representam o sinal acústico em cada quadro.

Capítulo 4

Análise dos resultados experimentais

A componente acústica e a componente visual da fala são parcialmente relacionadas, uma vez que ambas são conseqüências da configuração do trato vocal. Portanto, partes do movimento facial podem ser modeladas em função da acústica da fala e parte da acústica da fala pode ser modelada em função do movimento facial.

Este capítulo analisa o acoplamento entre os movimentos da face e a acústica da fala (Moreira e Yehia, 2006). Esta análise é realizada por meio de componentes do movimento facial acusticamente alinhadas obtidas a partir da matriz de correlação cruzada entre as posições dos adesivos marcadores e os parâmetros LSP (*Line Spectrum Pairs*), conforme descrito no Capítulo 3. Além disso, é também analisado o acoplamento entre os movimentos de diferentes regiões da face durante a produção da fala. Este acoplamento é modelado pelos autovetores da matriz de covariância das posições dos marcadores. Os autovetores do movimento dos marcadores da face representam as direções ortogonais de máxima variabilidade do movimento facial dentro do sistema de coordenadas de $2N$ dimensões, definido pelas componentes x e y dos N marcadores faciais. Essas direções mudam ao longo do tempo. Examina-se, aqui, como essas mudanças ocorrem. Da mesma forma, por meio das componentes acusticamente alinhadas, são analisadas as variações temporais das direções de máximo alinhamento entre as componentes de movimento facial e acústica durante a fala.

4.1 Análise dos autovetores do movimento facial ao longo do tempo

O movimento dos marcadores foi adquirido no experimento descrito no Capítulo 3, em que locutores leram em voz alta um texto com aproximadamente 3 minutos de duração. Foram utilizados dados faciais representados por vetores de $2N$ dimensões, correspondendo às componentes x e y dos $N = 28$ marcadores fixados sobre a face.

Uma janela deslizante de 3 segundos foi deslocada a cada 0,2 segundos sobre a matriz

correspondente ao movimento dos marcadores para um trecho com duração de aproximadamente 1 minuto de gravação. As componentes principais foram calculadas para cada deslocamento, como mostra a Figura 4.1. O objetivo é analisar mudanças ocorridas nos autovetores da matriz de covariância do movimento facial ao longo do tempo.

Nos experimentos, a análise dos dados utiliza as seis primeiras componentes principais, que representam aproximadamente 95% da variabilidade do movimento facial. A Figura 4.2 ilustra a variância média e o desvio-padrão acumulados pelos autovetores do movimento facial.

A Figura 4.3 mostra os seis primeiros autovetores e o desvio-padrão dos autovetores das amostras adquiridas ao longo do trecho analisado para três locutores diferentes. A Tabela 4.1 mostra para cada um dos oitos locutores do primeiro experimento (ver Seção 3.1) o desvio-padrão médio dos autovetores do movimento facial. Similarmente, a Tabela 4.2 mostra o desvio-padrão médio para os três locutores participantes do segundo experimento.

Na Figura 4.4 verifica-se a posição dos marcadores durante uma elocução estimada por meio da Equação 3.7. Cada autovetor é utilizado separadamente com o objetivo de verificar a relação entre autovetor e movimento da face. Os resultados obtidos e visualizados nas figuras 4.4 e 4.3 indicam que o primeiro autovetor captura o movimento vertical do queixo e do lábio inferior e em menor intensidade o movimento horizontal desta mesma região, sendo constante ao longo do tempo. Os outros autovetores capturam os acoplamentos restantes entre as regiões faciais. Como pode ser observado nas tabelas 4.1 e 4.2, eles apresentam variâncias maiores evidenciando que os demais autovetores são muito mais variáveis ao longo do tempo. Esta variação indica que o acoplamento entre as diversas regiões da face modifica de acordo com o conteúdo falado.

A Figura 4.5 mostra a posição média dos marcadores durante uma elocução adicionada e subtraída pelo desvio-padrão ponderado. Nesta figura observa-se também que a primeira componente do movimento relaciona-se ao movimento da mandíbula, a segunda componente do movimento relaciona-se ao movimento horizontal da região da boca da parte esquerda da

Tabela 4.1: Desvio padrão médio ($\overline{\sigma_l}$) dos autovetores do movimento dos marcadores para os oitos locutores participantes do primeiro experimento. $\overline{\sigma}$ e σ são, respectivamente, a média e o desvio-padrão dos valores de desvio-padrão médio obtidos para os 8 locutores.

	Locutor $\overline{\sigma_L}$									
Autovetor	AS	BG	GF	KM	LL	LA	MC	TS	Média ($\overline{\sigma}$)	σ
1	0,015	0,054	0,032	0,056	0,038	0,049	0,025	0,026	0,036	0,015
2	0,120	0,113	0,133	0,140	0,129	0,136	0,131	0,145	0,131	0,010
3	0,116	0,131	0,132	0,137	0,119	0,121	0,120	0,138	0,127	0,009
4	0,113	0,142	0,116	0,137	0,110	0,129	0,115	0,125	0,123	0,012
5	0,101	0,129	0,106	0,137	0,109	0,113	0,119	0,114	0,116	0,012
6	0,076	0,122	0,102	0,128	0,103	0,098	0,095	0,099	0,103	0,016

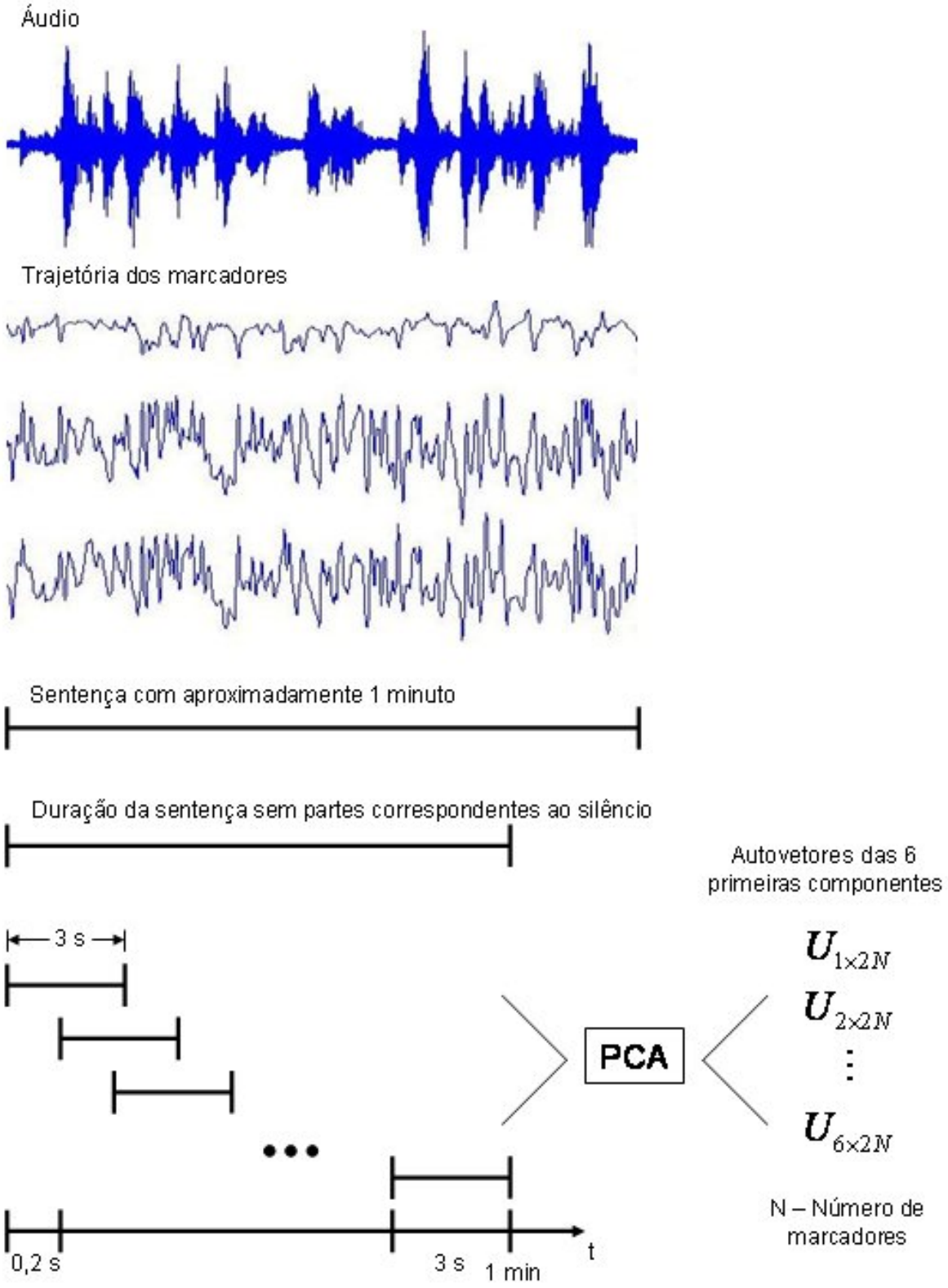


Figura 4.1: Processamento dos dados para análise dos autovetores do movimento facial ao longo do tempo. Os autovetores das $K = 6$ primeiras componentes principais dos movimentos dos marcadores foram calculados para janelas deslizantes de 3 segundos deslocadas a cada $0,2$ segundos sobre a matriz correspondente ao movimento dos $N = 28$ marcadores.

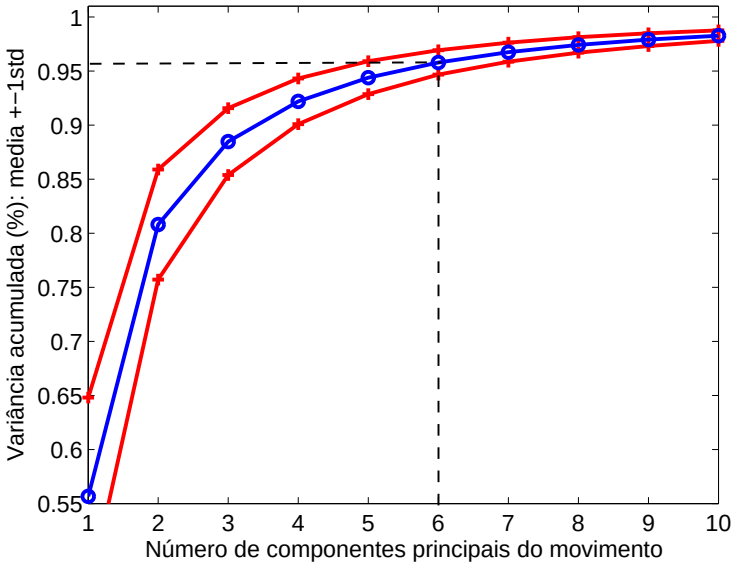


Figura 4.2: Variância acumulada pelos autovetores do movimento facial. As seis primeiras componentes principais representam cerca de 95% da variabilidade do movimento facial.

Tabela 4.2: Desvio padrão médio ($\overline{\sigma_l}$) dos autovetores do movimento dos marcadores para os três locutores participantes do segundo experimento. $\overline{\sigma}$ e σ são, respectivamente, a média e o desvio-padrão dos valores de desvio-padrão médio obtidos para os 3 locutores.

Autovetor	Locutor $\overline{\sigma_L}$			Média ($\overline{\sigma}$)	σ
	KM	LA	MC		
1	0,034	0,047	0,023	0,034	0,012
2	0,091	0,091	0,095	0,092	0,003
3	0,097	0,113	0,112	0,108	0,009
4	0,098	0,111	0,113	0,108	0,008
5	0,121	0,117	0,120	0,119	0,002
6	0,126	0,126	0,128	0,127	0,001

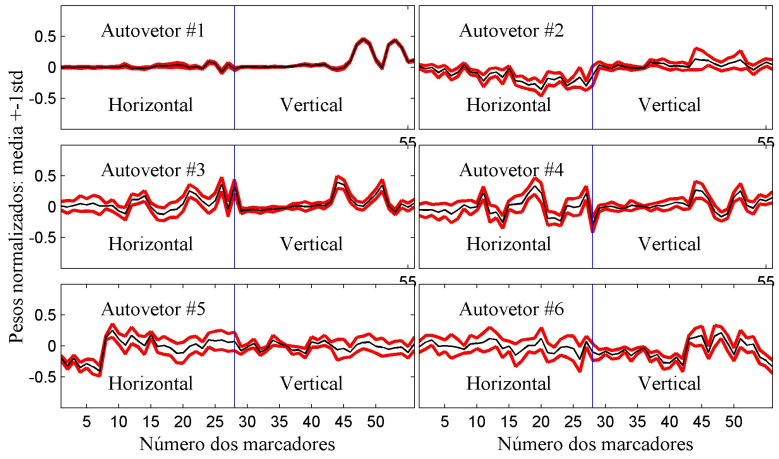
face, e a terceira componente do movimento relaciona-se ao movimento horizontal da região da boca da parte direita da face, e ao movimento do lábio superior.

As últimas colunas das Tabelas 4.1 e 4.2 mostram a média e o desvio-padrão médio dos autovetores entre locutores:

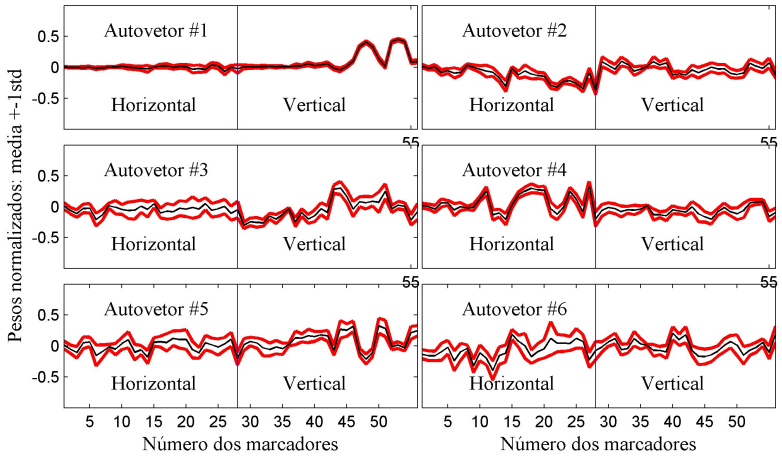
$$\bar{\sigma} = \frac{1}{L}\sum_{l=1}^{L}\bar{\sigma}_l, \tag{4.1}$$

$$\sigma = \sqrt{\frac{\sum_{l=1}^{L}(\bar{\sigma}_l - \bar{\sigma})^2}{L}}, \tag{4.2}$$

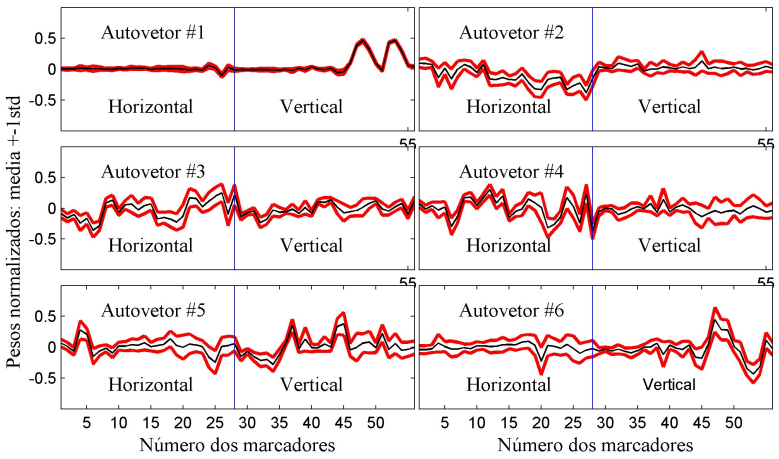
em que $\overline{\sigma_l}$ é o desvio-padrão do locutor l, $\bar{\sigma}$ é o desvio-padrão médio e σ é o desvio-padrão dos valores de desvio-padrão médio obtidos para os $L = 8$ locutores do primeiro experimento



(a)

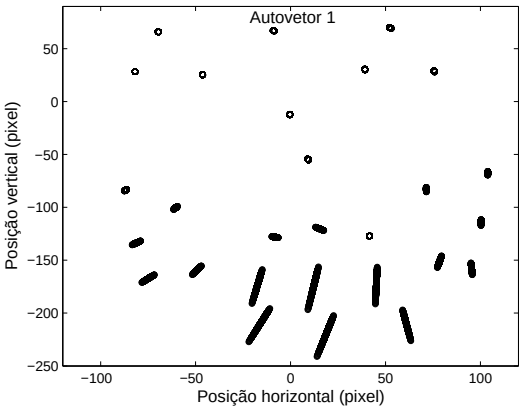


(b)

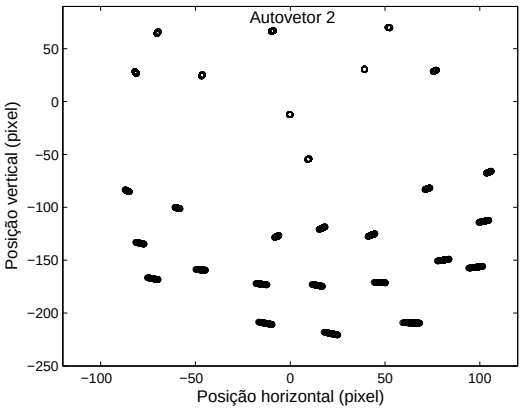


(c)

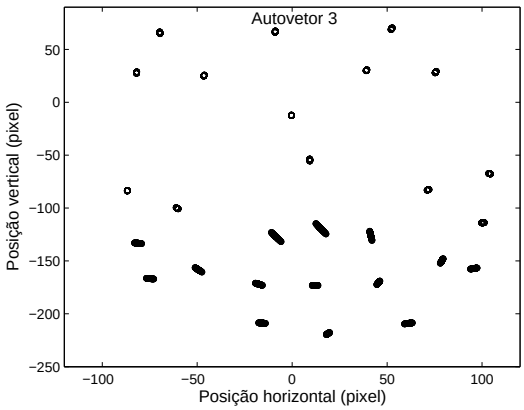
Figura 4.3: Seis primeiros autovetores do movimento facial de 3 locutores em uma narração de aproximadamente 1 minuto. O primeiro autovetor mostra o movimento do lábio inferior e da mandíbula e é constante. Os outros autovetores mostram os movimentos acoplados das outras regiões da face. (a) Locutor AS; (b) locutor GF; (c) locutor MC.



(a)

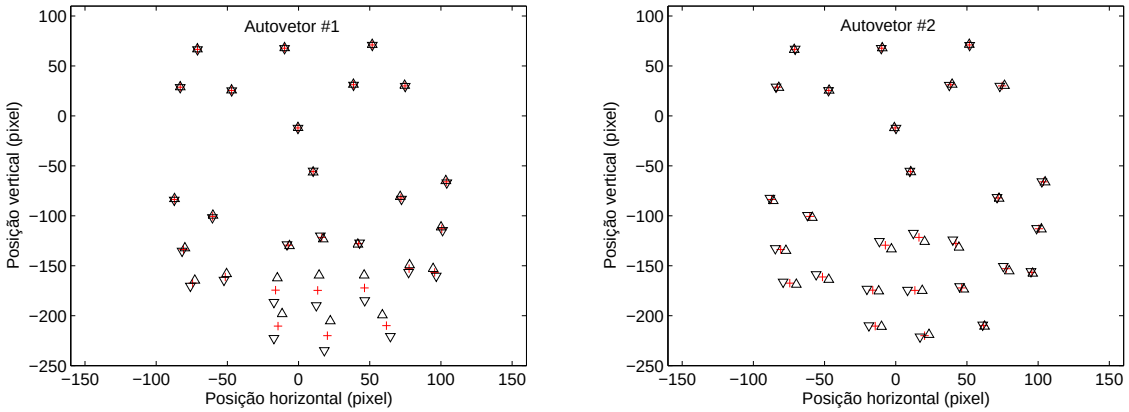


(b)



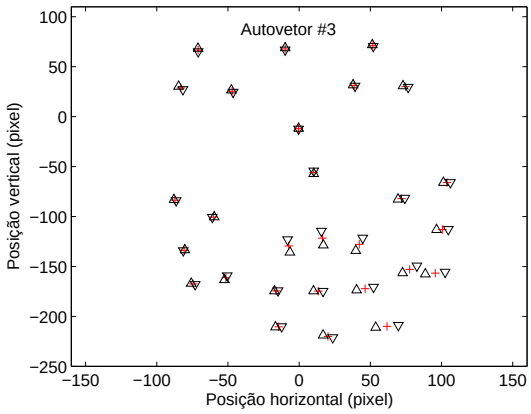
(c)

Figura 4.4: Posição dos marcadores faciais durante uma elocução estimada separadamente por meio do (a) primeiro autovetor; (b) segundo autovetor; e (c) terceiro autovetor.



(a)

(b)



(c)

Figura 4.5: Posição média dos marcadores durante uma elocução adicionada e subtraída pelo desvio-padrão ponderado. (a) primeira componente do movimento ponderada por dois desvios-padrão de sua variância adicionada ($\triangle$) e subtraída ($\triangledown$) das posições médias dos marcadores ($+$); (b) segunda componente do movimento ponderada por três desvios-padrão de sua variância adicionada ($\triangle$) e subtraída ($\triangledown$) das posições médias dos marcadores ($+$); e (c) terceira componente do movimento ponderada por quatro desvios-padrão de sua variância adicionada ($\triangle$) e subtraída ($\triangledown$) das posições médias dos marcadores ($+$).

e $L = 3$ locutores do segundo experimento. O resultado indica que a variabilidade dos autovetores tem valores semelhantes para os diversos locutores.

4.2 Análise do alinhamento do movimento facial com a acústica da fala

A rotação e o alinhamento dos sistemas de coordenadas dos parâmetros da acústica da fala e do movimento facial originam as componentes do movimento facial acusticamente alinhadas. Nesta seção é analisada a variabilidade, ao longo do tempo, dos parâmetros do movimento facial obtidos, alinhados com os parâmetros acústicos (i.e. LSP rotacionados), como explicado na Seção 3.3.

Para o alinhamento acústico, foram utilizados dados faciais representados por vetores de $2N = 56$ dimensões, correspondendo às componentes x e y dos $N = 28$ marcadores fixados sobre a face, e dados da acústica da fala representados por $p = 10$ coeficientes LSP.

Da mesma maneira que na Seção 4.1, uma janela deslizante de 3 segundos foi deslocada a cada 0,2 segundos sobre a matriz de correlação cruzada entre as posições dos adesivos marcadores e os parâmetros LSP para um trecho com duração de aproximadamente 1 minuto de gravação. As componentes acusticamente alinhadas foram calculadas para cada deslocamento, com o objetivo de analisar mudanças ocorridas nos parâmetros obtidos a partir da matriz de correlação cruzada do movimento facial com a acústica da fala ao longo do tempo.

Nos experimentos, a análise dos dados utiliza as seis primeiras componentes principais, que representam aproximadamente 95% da variabilidade do movimento facial alinhado com a acústica da fala. A Figura 4.6 ilustra a variância média e o desvio-padrão acumulados pelos autovetores do movimento facial.

Os seis primeiros autovetores acusticamente alinhados usados para representar o movimento facial e o desvio-padrão dos parâmetros das amostras adquiridas ao longo do trecho analisado são mostrados na Figura 4.7, para três locutores diferentes. O autovetor 1 (1ª componente acusticamente alinhada) é mais constante ao longo do trecho, contrastando com os demais autovetores, que exibem uma variabilidade maior.

A Tabela 4.3 mostra, para cada um dos oitos locutores do primeiro experimento (ver Seção 3.1), o desvio-padrão médio das componentes do movimento facial acusticamente alinhadas obtido. Similarmente, a Tabela 4.4 mostra o desvio-padrão médio para os três locutores participantes do segundo experimento. Nestas tabelas, verifica-se que a parametrização do movimento por meio de componentes acusticamente alinhadas resulta no autovetor 1 (1ª componente acusticamente alinhada) menos constante em relação à parametrização do movimento dos marcadores por meio de componentes principais do movimento facial do trecho (Tabelas 4.1 e 4.2).

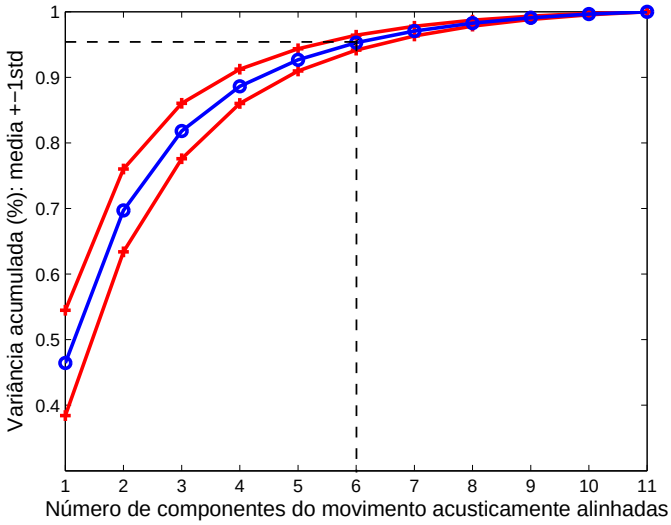


Figura 4.6: Variância acumulada pelas componentes acusticamente alinhadas do movimento facial. As seis primeiras componentes principais representam cerca de 95% da variância dos dados.

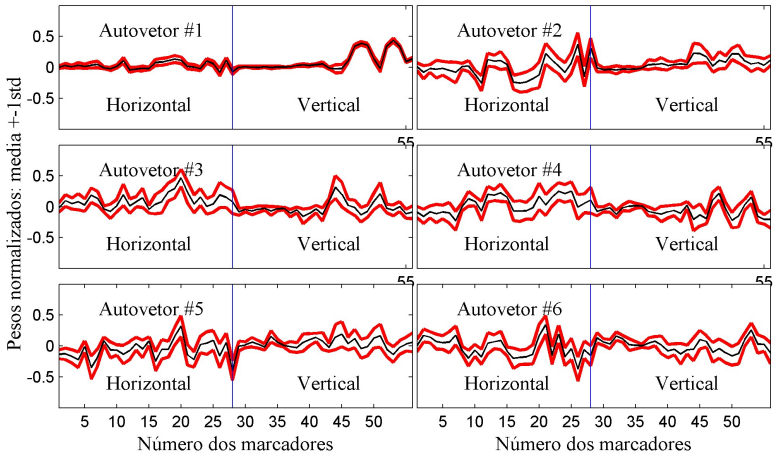
As últimas colunas das Tabelas 4.3 e 4.4 mostram a média (Equação 4.1) e o desvio-padrão médio (Equação 4.2) dos autovetores entre locutores. O resultado indica que a variabilidade dos autovetores tem valores semelhantes para os diversos locutores.

4.3 Distâncias entre autovetores ao longo do tempo e autovetores médios

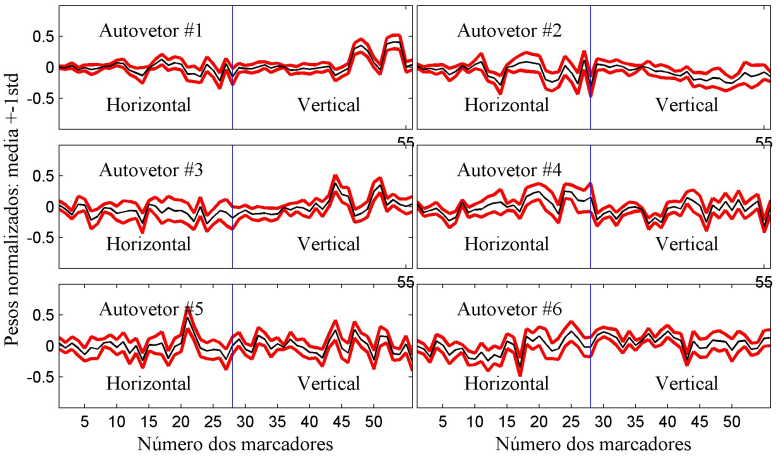
Com o objetivo de entender melhor o comportamento dos autovetores usados para representar o movimento facial, os autovetores obtidos com base em uma janela deslizante de 3 segundos, deslocada a cada 0,2 segundos, sobre a matriz correspondente ao movimento dos marcadores, para um trecho com duração de aproximadamente 1 minuto de gravação são comparados aos

Tabela 4.3: Desvio padrão médio ($\overline{\sigma_l}$) das componentes acusticamente alinhadas para os oitos locutores participantes do primeiro experimento. $\overline{\sigma}$ e σ são, respectivamente, a média e o desvio-padrão dos valores de desvio-padrão médio obtidos para os 8 locutores.

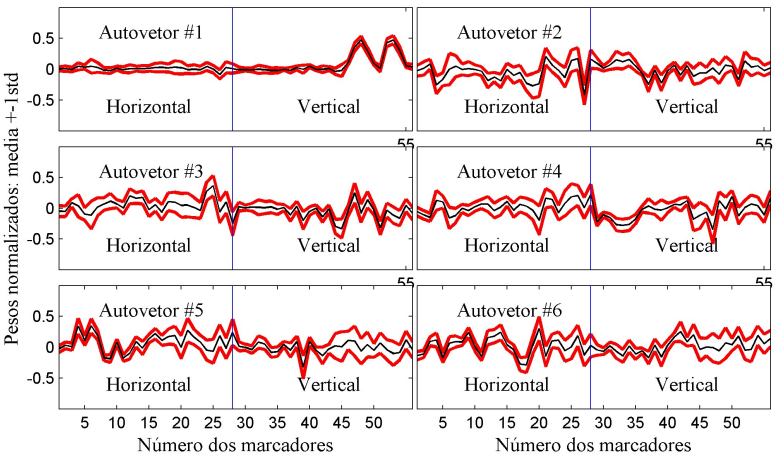
Comp. acust. alinhada	Locutor $\overline{\sigma_l}$								Média ($\overline{\sigma}$)	σ
	AS	BG	GF	KM	LL	LA	MC	TS		
1	0,046	0,079	0,077	0,066	0,052	0,061	0,050	0,044	0,059	0,014
2	0,120	0,143	0,129	0,146	0,134	0,137	0,121	0,121	0,131	0,010
3	0,108	0,131	0,131	0,133	0,105	0,121	0,124	0,125	0,122	0,011
4	0,118	0,127	0,129	0,126	0,116	0,109	0,114	0,117	0,119	0,007
5	0,104	0,132	0,127	0,123	0,099	0,116	0,107	0,118	0,115	0,012
6	0,093	0,124	0,119	0,117	0,103	0,105	0,102	0,096	0,107	0,011



(a)



(b)



(c)

Figura 4.7: Seis primeiros autovetores acusticamente alinhados usados para representar o movimento facial de 3 locutores em uma narração de aproximadamente 1 minuto. O primeiro autovetor (primeira componente acusticamente alinhada) é mais constante em contraste com os autovetores restantes. (a) Locutor AS; (b) locutor GF; (c) locutor MC.

Tabela 4.4: Desvio padrão médio ($\overline{\sigma_l}$) das componentes acusticamente alinhadas para os três locutores participantes do segundo experimento. $\overline{\sigma}$ e σ são, respectivamente, a média e o desvio-padrão dos valores de desvio-padrão médio obtidos para os 3 locutores.

Comp. acust. alinhada	Locutor $\overline{\sigma_L}$			Média ($\overline{\sigma}$)	σ
	KM	LA	MC		
1	0,067	0,090	0,033	0,064	0,029
2	0,108	0,105	0,106	0,106	0,002
3	0,101	0,123	0,118	0,114	0,012
4	0,115	0,119	0,122	0,119	0,004
5	0,126	0,123	0,123	0,124	0,002
6	0,130	0,130	0,129	0,130	0,001

autovetores de referência, obtidos com base no trecho completo. Para isto, a distância euclidiana é calculada:

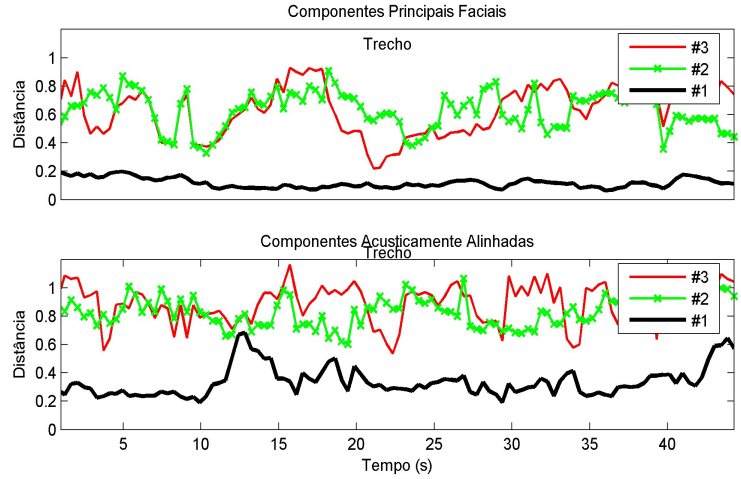
$$d_{jk} = \sqrt{\sum_{i=1}^{2N} (\bar{u}_{ik} - u_{ijk})^2}, \tag{4.3}$$

em que $\bar{u}_{ik}$ representa o i-ésimo marcador do k-ésimo autovetor de referência (i.e. calculado com base no trecho completo); u_{ijk} representa o i-ésimo marcador do k-ésimo autovetor calculado com base na j-ésima janela analisada; e d_{jk} é a distância do k-ésimo autovetor calculado com base na j-ésima janela analisada e o autovetor de referência correspondente.

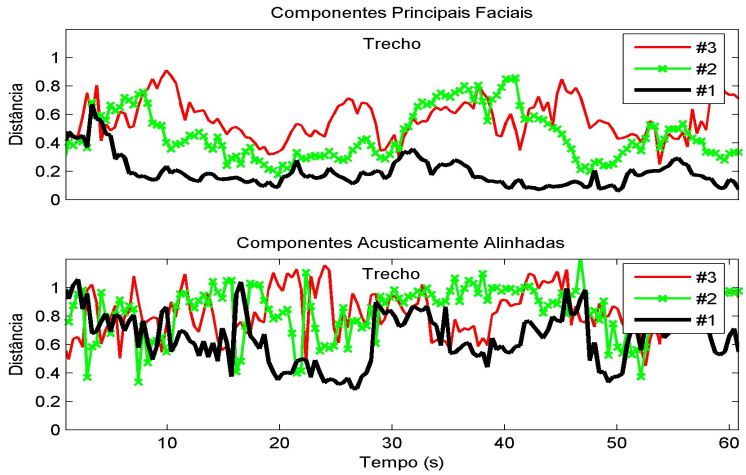
As distâncias dos três primeiros autovetores para as componentes principais do movimento facial e para as componentes acusticamente alinhadas do movimento facial são mostradas na Figura 4.8. A Tabela 4.5 mostra a distância média entre os autovetores do movimento facial ao longo do tempo e os autovetores de referência para cada um dos oitos locutores do primeiro experimento e, na última coluna, o valor médio da distância média entre locutores. Similarmente, a Tabela 4.6 mostra a distância média para os três locutores participantes do segundo experimento. A Tabela 4.7 mostra a distância média das componentes acusticamente alinhadas para cada um dos oitos locutores do primeiro experimento e, na última coluna, o valor médio da distância média. Da mesma maneira, a Tabela 4.8 mostra a distância para os três locutores participantes do segundo experimento.

Somente o primeiro autovetor exibe uma distância pequena de seu valor médio através do tempo. Para a primeira componente alinhada acusticamente, a distância é menor em relação às demais componentes, porém maior do que a primeira componente do movimento facial, demonstrando que a variabilidade do movimento facial é menor quando comparada à variabilidade do alinhamento acústico.

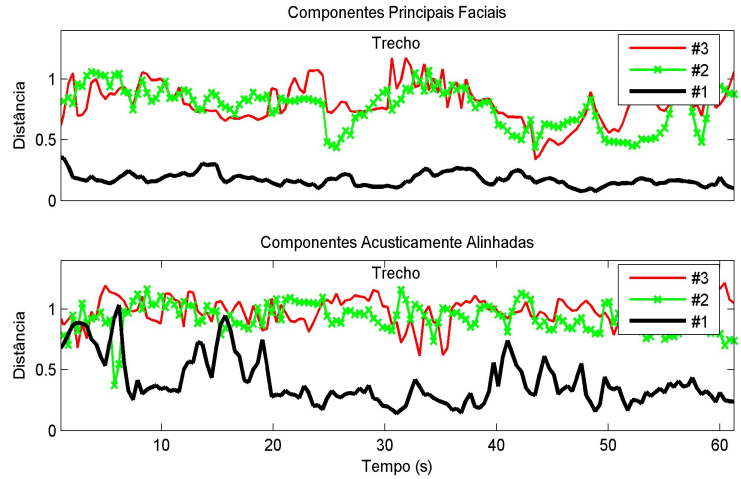
Dando seqüência à análise, compara-se agora a variabilidade dos autovetores para três situações distintas: *(i)* janela deslizante sobre o trecho de referência; *(ii)* janela deslizante sobre repetição do trecho de referência; e *(iii)* janela deslizante sobre trecho distinto do trecho



(a)



(b)



(c)

Figura 4.8: Distâncias entre os autovetores do trecho de referência e os autovetores das janelas analisadas ao longo do tempo. O 1º autovetor exibe uma distância pequena de seu valor médio através do tempo. (a) Locutor AS; (b) locutor GF; (c) locutor MC.

Tabela 4.5: Distâncias entre os autovetores do movimento dos marcadores, das janelas analisadas ao longo do tempo, e os autovetores do trecho de referência para os oitos locutores participantes do primeiro experimento.

	Locutor *distância (pixel)*								
Autovetor	AS	BG	GF	KM	LL	LA	MC	TS	Média
1	0,126	0,301	0,231	0,357	0,294	0,355	0,145	0,186	0,249
2	0,649	0,465	0,503	0,469	0,439	0,467	0,166	0,363	0,441
3	0,652	0,654	0,607	0,735	0,750	0,692	0,532	0,599	0,653
4	0,853	0,811	0,805	0,707	0,877	0,841	0,489	0,731	0,764
5	0,908	0,880	0,867	0,765	0,917	0,948	0,526	0,840	0,831
6	1,043	0,974	1,017	1,061	1,047	1,041	0,820	1,015	1,000

Tabela 4.6: Distâncias entre os autovetores do movimento dos marcadores, das janelas analisadas ao longo do tempo, e os autovetores do trecho de referência para os três locutores participantes do segundo experimento.

	Locutor *distância (pixel)*			
Autovetor	KM	LA	MC	Média
1	0,227	0,324	0,259	0,270
2	0,343	0,395	0,395	0,378
3	0,770	0,617	0,617	0,668
4	0,937	0,843	0,843	0,874
5	0,961	0,875	0,875	0,904
6	1,037	1,017	1,017	1,024

Tabela 4.7: Distâncias entre as componentes acusticamente alinhadas das janelas analisadas ao longo do tempo, e as componentes acusticamente alinhadas do trecho de referência para os oitos locutores participantes do primeiro experimento.

Comp. acust.	Locutor *distância (pixel)*								
alinhada	AS	BG	GF	KM	LL	LA	MC	TS	Média
1	0,352	0,577	0,576	0,392	0,349	0,379	0,296	0,307	0,403
2	0,422	0,571	0,545	0,573	0,610	0,461	0,470	0,464	0,515
3	0,387	0,533	0,589	0,603	0,431	0,485	0,345	0,518	0,486
4	0,429	0,491	0,464	0,663	0,396	0,411	0,829	0,498	0,523
5	0,399	0,489	0,493	0,608	0,404	0,399	0,305	0,469	0,446
6	0,392	0,456	0,434	0,558	0,362	0,343	0,392	0,504	0,430

Tabela 4.8: Distâncias entre as componentes acusticamente alinhadas das janelas analisadas ao longo do tempo, e as componentes acusticamente alinhadas do trecho de referência para os três locutores participantes do segundo experimento.

Comp. acust. alinhada	Locutor *distância (pixel)*			Média
	KM	LA	MC	
1	0,585	0,530	0,472	0,529
2	0,598	0,660	0,660	0,639
3	0,913	0,825	0,825	0,855
4	1,049	1,029	1,029	1,036
5	1,086	1,087	1,087	1,087
6	1,106	1,126	1,126	1,119

de referência. Nesta comparação, são verificadas as distâncias entre os autovetores retirados de uma sentença com duração de aproximadamente um minuto (trecho de referência) e os autovetores calculados com base em uma janela deslizante que percorre: *(i)* o trecho de referência; *(ii)* uma repetição do trecho de referência (trecho repetido); e *(iii)* um trecho com um novo conteúdo acústico, porém com aproximadamente a mesma duração do trecho de referência (trecho novo).

Assim, a janela deslizante de 3 segundos é deslocada de 0,2 em 0,2 segundos percorrendo os três trechos: referência, repetido e novo. A comparação ocorre entre os autovetores do trecho de referência com duração de um minuto (autovetores de referência) e os autovetores da janela que percorre os três grupos. A distância euclidiana, ao longo do tempo, calculada para dois locutores é mostrada na Figura 4.9 para os autovetores faciais e para os autovetores das componentes acusticamente alinhadas.

Observando as Figuras 4.9, verifica-se que a variabilidade é menor para o primeiro autovetor independentemente do conteúdo falado. Para a primeira componente acusticamente alinhada, a variabilidade também é menor em comparação com as demais componentes. Entretanto, a primeira componente acusticamente alinhada apresenta uma variabilidade maior quando comparada com o primeiro autovetor do movimento facial.

Com o objetivo de observar a variação do autovetor ao longo do tempo, a distância euclidiana foi calculada entre a janela atual e a janela seguinte, tanto para os autovetores faciais quanto para os autovetores das componentes acusticamente alinhadas. Na Figura 4.10, a distância de um passo à frente para um locutor é ilustrada para os três primeiros autovetores. O primeiro autovetor tem uma variação menor em comparação aos demais autovetores, o que não é surpreendente, dado que também é o que menos varia em relação ao autovetor de referência.

Outro fato a ser observado é que, para a componente acusticamente alinhada, existem variações mais intensas entre a janela atual e a janela subseqüente. Apesar de o primeiro autovetor ser mais constante que os demais, há pequenas diferenças dependendo do conteúdo

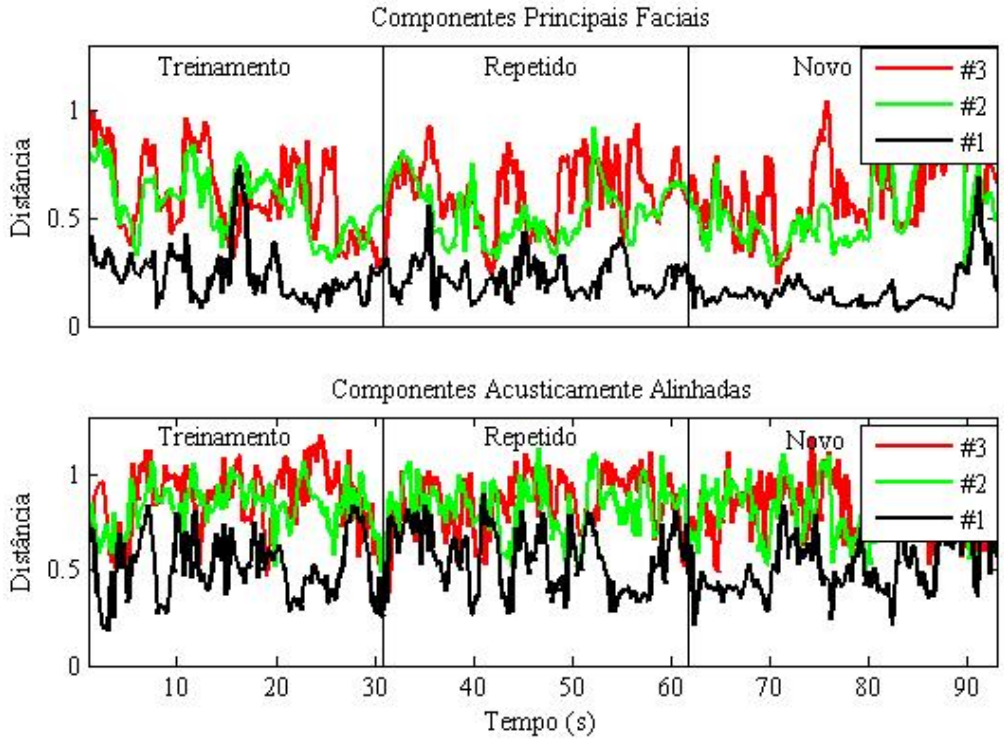


Figura 4.9: Distâncias entre os autovetores do trecho de referência e os autovetores dos trechos obtidos a partir da *(i)* janela deslizante sobre o trecho de referência; *(ii)* janela deslizante sobre repetição do trecho de referência; e *(iii)* janela deslizante sobre trecho distinto do trecho de referência. O 1º autovetor exibe uma distância pequena de seu valor médio ao longo do tempo. (Locutor AS)

acústico para um curto intervalo de tempo. Uma hipótese para esta variação em um curto intervalo de tempo é que o alinhamento acústico utiliza os parâmetros LSP e estes se relacionam com os formantes que variam sua afiliação na produção acústica da fala, como detalhado na Seção 4.6.

4.4 Porcentagem do movimento facial expressa pelas primeiras componentes ao longo do tempo

A PCA organiza os dados em ordem decrescente de variância, como visto nas Seções 4.1e 4.2, sendo que as seis primeiras componentes principais representam cerca de 95% da variância dos autovetores do movimento facial e da variância das componentes acusticamente alinhadas. Cada componente principal ou acusticamente alinhada representa uma porcentagem da variância do movimento facial. Esta porcentagem pode ser vista na Figura 4.11, em que a variância explicada por cada componente principal facial (em preto) e por cada componente acusticamente alinhada (em vermelho) varia ao longo do tempo.

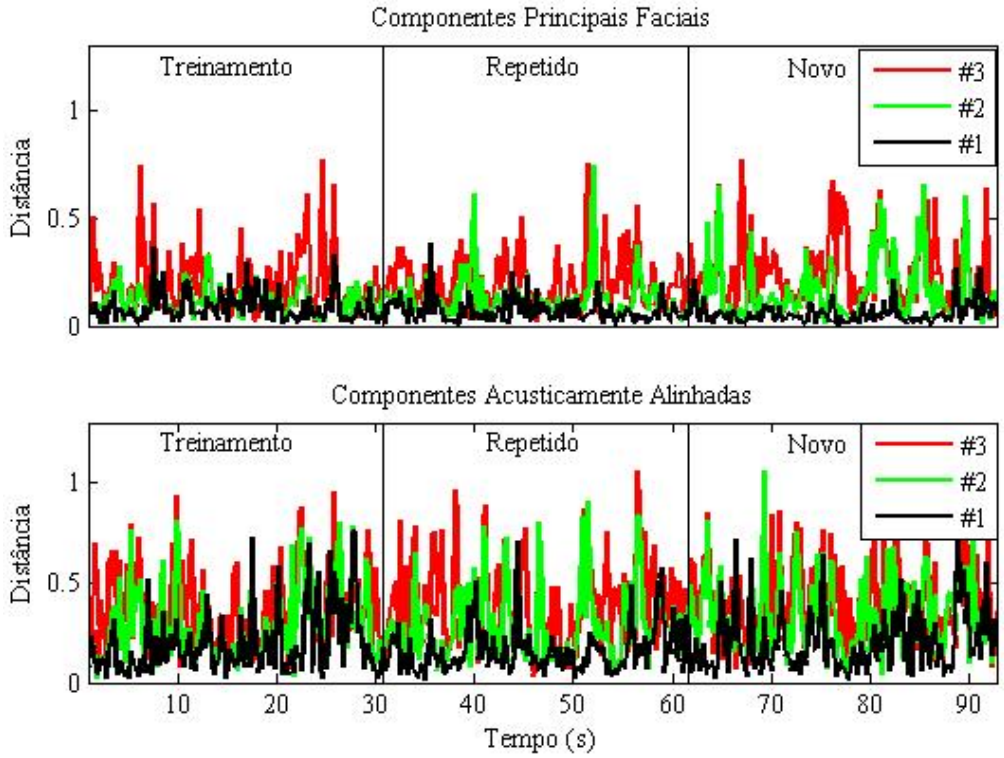
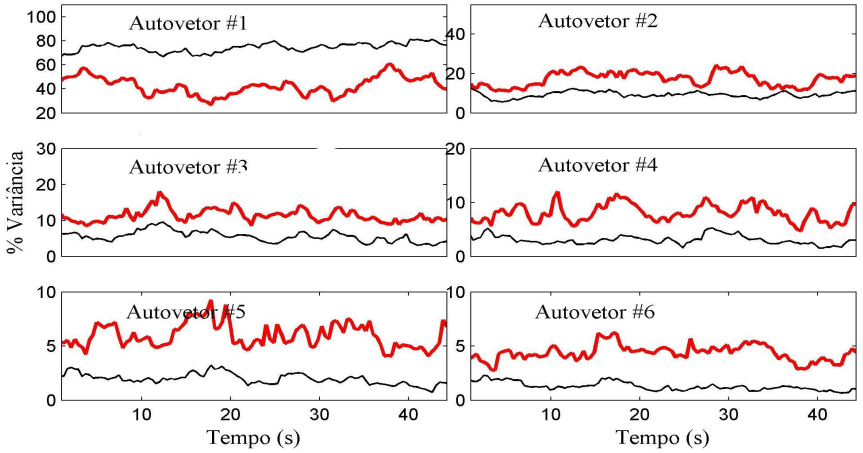
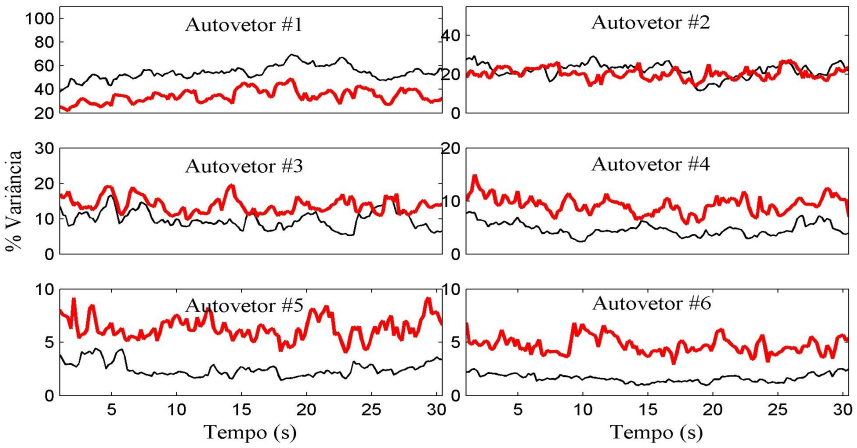


Figura 4.10: Distâncias entre os autovetores da janela analisada a um passo à frente. O 1º autovetor do movimento facial exibe uma distância pequena de seu valor médio através do tempo; enquanto que o 1º autovetor da componente acusticamente alinhada exibe variações mais intensas entre a janela atual e a janela subseqüente. (locutor AS)

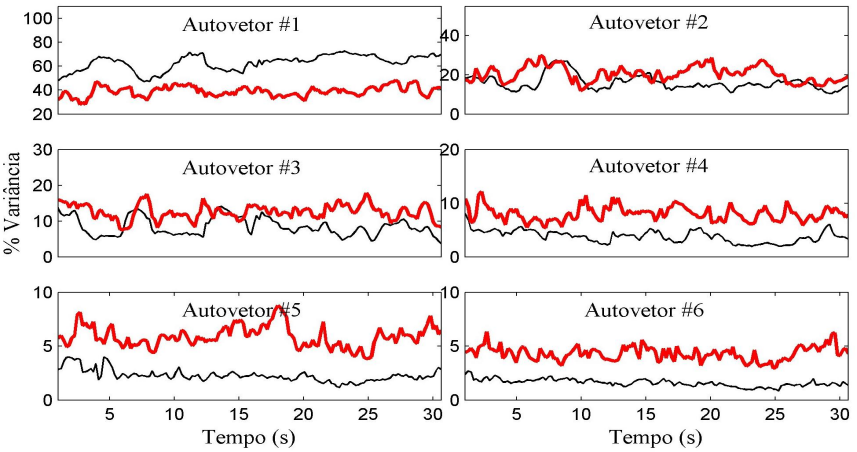
Nesta figura, verifica-se que a primeira componente principal facial, i.e. primeiro autovetor do movimento facial, representa aproximadamente 60% da variância presente nos dados. Assim, aproximadamente 60% dos movimentos que acontecem na face localizam-se na região do lábio inferior (ver Seção 4.1). Na representação da componente acusticamente alinhada esta porcentagem é menor, demonstrando que a representação do movimento da boca está diluída nas demais componentes acusticamente alinhadas. O contrário acontece no restante da variância explicada para os demais autovetores, em que as componentes acusticamente alinhadas, normalmente, representam uma maior proporção em relação aos autovetores dos marcadores faciais. É importante ressaltar que a porcentagem das componentes é influenciada pelo contexto e pelo modo como os locutores proferem o trecho.



(a)



(b)



(c)

Figura 4.11: Porcentagem da variância explicada por cada componente principal facial (linha preto) e componente acusticamente alinhada (linha vermelho) variando ao longo do tempo. (a) Locutor AS; (b) locutor GF; (c) locutor MC.

4.5 Padronização dos autovetores para diferentes durações de trechos

O estudo visto da variabilidade dos autovetores do movimento facial e das componentes acusticamente alinhadas utiliza trechos provenientes da janela deslizante. Cada trecho representa 5% da duração do trecho total (60 segundos), portanto, são pequenas elocuções. Nesta seção, é analisada a variabilidade dos autovetores para diferentes durações de trechos.

A fim de verificar a variabilidade dos autovetores, para elocuções de diferentes durações, foram criados três grupos de trechos com tempos de elocução diferentes. Destes trechos, intervalos correspondentes ao silêncio foram descartados. Na análise, verifica-se a distância euclidiana e a correlação entre as componentes de trechos com o mesmo conteúdo acústico, proferidos em aquisições diferentes, por um mesmo locutor. Analogamente, observam-se a distância euclidiana e a correlação entre as componentes de trechos com o mesmo conteúdo acústico, proferidos por dois locutores diferentes.

Assim, a variabilidade dos autovetores é testada para três grupos: no primeiro, foram calculadas as distâncias entre os autovetores para uma elocução com duração de aproximadamente 2 segundos, denominado trecho de curta duração. No segundo grupo, as distâncias são calculadas para um trecho de média duração, aproximadamente 5 segundos, e no terceiro grupo, para um trecho de longa duração, aproximadamente 1 minuto. Como as pessoas falam em diferentes velocidades ou a mesma pessoa pode falar o mesmo trecho com maior ou menor velocidade, existem pequenas variações de duração das elocuções. Assim, o mesmo conteúdo (texto) é mantido para todas as comparações. As Tabelas 4.9, 4.10 e 4.11 mostram as variações calculadas por meio da distância euclidiana e da correlação entre os autovetores do movimento facial para diferentes tamanhos de trechos e mesmo locutor.

Tabela 4.9: Distância entre os autovetores do movimento facial para repetições de elocuções de diferentes tamanhos e mesmo locutor. Em parênteses a correlação entre autovetores do movimento facial para diferentes tamanhos de trechos (locutor GF).

Autovetores	Trecho pequeno	Trecho médio	Trecho grande
1	0,65 (0,80)	0,34 (0,35)	0,06 (0,96)
2	0,63 (0,70)	0,43 (0,11)	0,10 (0,99)
3	1,02 (0,37)	0,69 (0,72)	0,21 (0,98)
4	1,09 (0,20)	0,64 (0,78)	0,56 (0,82)

Observa-se, por meio das tabelas, que o aumento do tamanho dos trechos implica reduções das distâncias e aumento das correlações. Similarmente, o mesmo é observado para as distâncias medidas por meio das componentes acusticamente alinhadas, como mostra a Tabela 4.12.

Da mesma maneira, os autovetores do grupo de trechos foram comparados para locutores

Tabela 4.10: Distância entre os autovetores do movimento facial para repetições de elocuções de diferentes tamanhos e mesmo locutor. Em parênteses a correlação entre autovetores do movimento facial para diferentes tamanhos de trechos (locutor AS).

Autovetores	Trecho pequeno	Trecho média	Trecho grande
1	0,34 (0,94)	0,16 (0,99)	0,09 (1,00)
2	0,67 (0,23)	0,97 (0,39)	0,26 (0,95)
3	0,64 (0,05)	1,09 (0,74)	0,23 (0,98)
4	1,09 (0,22)	1,24 (0,22)	0,46 (0,91)

Tabela 4.11: Distância entre os autovetores do movimento facial para repetições de elocuções de diferentes tamanhos e mesmo locutor. Em parênteses a correlação entre autovetores do movimento facial para diferentes tamanhos de trechos (locutor BG).

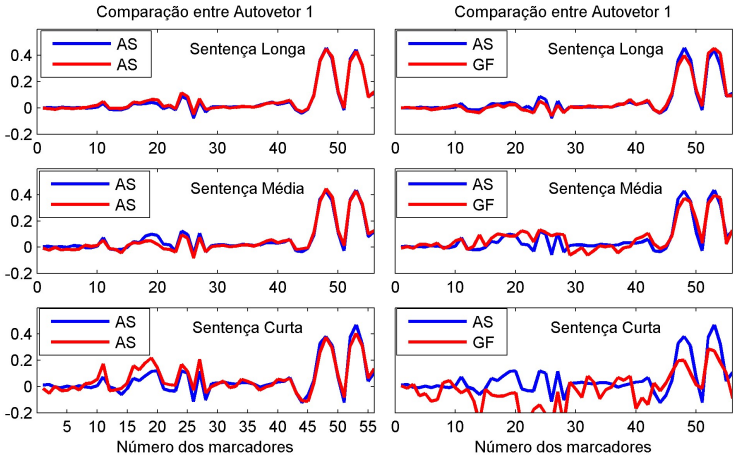
Autovetores	Trecho pequena	Trecho média	Trecho grande
1	0,43 (0,89)	0,38 (0,92)	0,24 (0,96)
2	0,48 (0,89)	0,38 (0,93)	0,24 (0,99)
3	0,85 (0,46)	0,46 (0,83)	0,25 (0,15)
4	0,70 (0,06)	0,60 (0,82)	0,22 (0,09)

diferentes, ou seja, dois locutores proferindo o mesmo trecho. Os valores correspondentes às distâncias dos autovetores para um trecho de curta, média e longa duração podem ser vistos nas Tabelas 4.13 e 4.14, em que a distância euclidiana e a correlação são observadas para os quatro primeiros autovetores. De forma semelhante à análise para um mesmo locutor, a distância diminui com o aumento da duração do trecho proferido. Entretanto, principalmente para o primeiro autovetor, as distâncias apresentam valores maiores para locutores diferentes em comparação com as distâncias para o mesmo locutor. A Figura 4.13 é a ilustração das semelhanças e diferenças existentes dos quatros primeiros autovetores do movimento facial em diferentes tamanhos de trechos para um mesmo locutor ou para locutores diferentes.

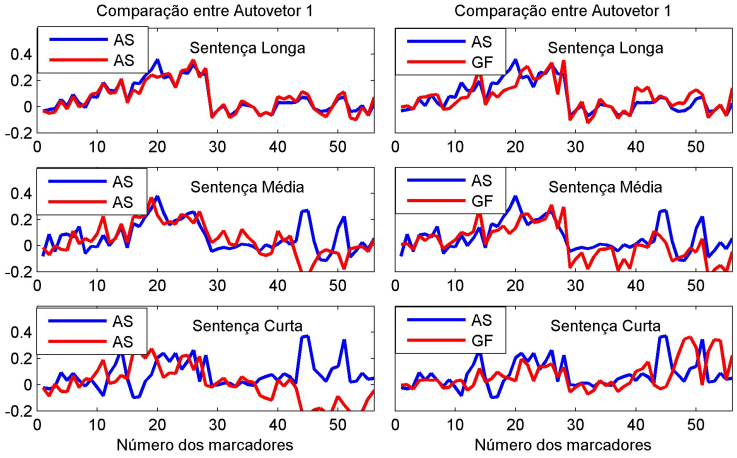
4.5.1 Padronização dos autovetores para conteúdos acústicos diferentes

Na seção anterior, verificou-se que para trechos pequenos os autovetores apresentam distâncias maiores e correlações menores em comparação com os valores apresentados para os autovetores de trechos maiores. Principalmente o primeiro autovetor, à medida que aumenta a duração dos trechos, cria um padrão de comportamento, tornando-se mais constante. Observa-se também que as distâncias entre locutores diferentes apresentam valores maiores para os trechos grandes quando comparados às distâncias obtidas pelos trechos proferidos pelo mesmo locutor, indicando que além do conteúdo acústico, os autovetores exibem informações inerentes à anatomia de cada locutor.

Na Tabela 4.15, observam-se, para um mesmo locutor, as distâncias para um trecho com

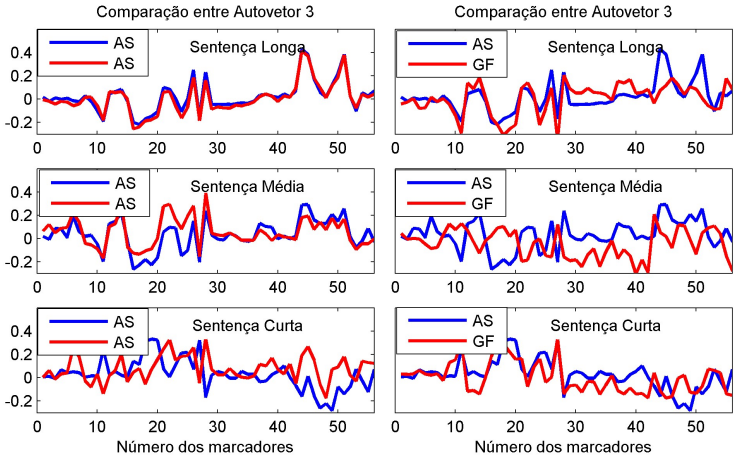


(a)

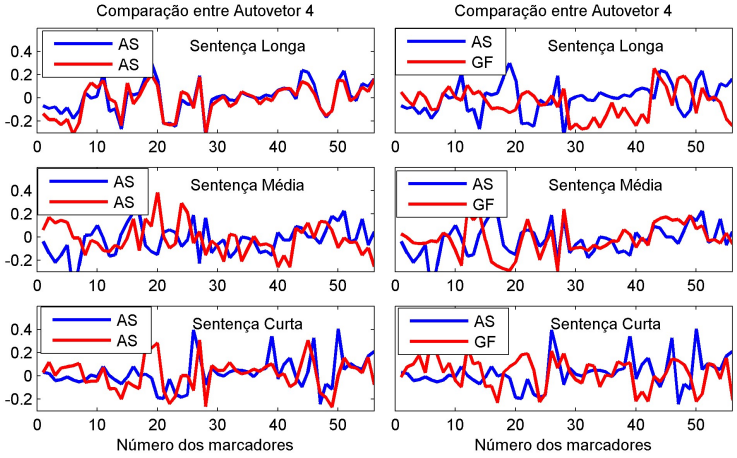


(b)

Figura 4.12: Autovetores do movimento facial para repetições de elocução de diferentes tamanhos de trechos para o mesmo e diferentes locutores (autovetores 1 e 2). (a) Primeiro autovetor para o mesmo locutor (AS e AS) e para locutores diferentes (AS e GF); (b) segundo autovetor para o mesmo locutor (AS e AS) e para locutores diferentes (AS e GF).



(a)



(b)

Figura 4.13: Autovetores do movimento facial para repetições de elocução de diferentes tamanhos de trechos para o mesmo e diferentes locutores (autovetores 3 e 4). (a) Terceiro autovetor para o mesmo locutor (AS e AS) e para locutores diferentes (AS e GF); (b) quarto autovetor para o mesmo locutor (AS e AS) e para locutores diferentes (AS e GF).

Tabela 4.12: Distância entre as componentes acusticamente alinhadas para repetições de elocuções de diferentes tamanhos e mesmo locutor. Em parênteses a correlação entre as componentes acusticamente alinhadas para diferentes tamanhos de trechos (locutor AS).

Autovetores	Trecho pequeno	Trecho médio	Trecho grande
1	0,47 (0,93)	0,29 (0,98)	0,15 (0,99)
2	0,67 (0,22)	1,00 (0,39)	0,42 (0,95)
3	0,87 (0,04)	0,95 (0,73)	0,70 (0,98)
4	1,03 (0,21)	1,04 (0,22)	0,60 (0,90)

Tabela 4.13: Distância entre os autovetores do movimento facial para repetições de elocuções de diferentes tamanhos de trecho e diferentes locutores. Em parênteses a correlação entre autovetores do movimento facial para diferentes tamanhos de trechos (locutor MC e locutor TS).

Autovetores	Trecho pequeno	Trecho médio	Trecho grande
1	0,75 (0,66)	0,33 (0,93)	0,46 (0,89)
2	0,80 (0,34)	0,72 (0,77)	0,61 (0,86)
3	1,02 (0,09)	1,05 (0,16)	0,64 (0,71)
4	1,14 (0,35)	1,13 (0,03)	0,92 (0,60)

duração de aproximadamente um minuto, porém com conteúdo acústico diferente. Este resultado reafirma que em um trecho grande os autovetores do movimento facial, especialmente o primeiro autovetor, adquire um padrão independente do conteúdo acústico. Aumentando o tamanho do trecho proferido, cria-se um autovetor de movimento padrão composto pelos possíveis movimentos de maior variância. Este vetor padrão contém também informações inerentes a cada locutor. No próximo capítulo, será vista uma análise mais aprofundada das características próprias a cada locutor contidas nos autovetores.

Tabela 4.14: Distância entre os autovetores do movimento facial para repetições de elocuções de diferentes tamanhos de trecho e diferentes locutores. Em parênteses a correlação entre autovetores do movimento facial para diferentes tamanhos de trechos (locutor KM e locutor AS).

Autovetores	Trecho pequeno	Trecho médio	Trecho grande
1	0,60 (0,85)	0,24 (0,97)	0,53 (0,89)
2	0,92(0,52)	0,57 (0,81)	0,69 (0,66)
3	0,93 (0,17)	0,87 (0,66)	1,02 (0,15)
4	0,95 (0,55)	0,82 (0,16)	0,95 (0,65)

Tabela 4.15: Distância dos autovetores do movimento facial para elocuções diferentes com conteúdo acústico diferente e mesmo locutor.

	Locutor		
Autovetores	AS _ AS	GF _ GF	BG_ BG
1	0,09	0,08	0,25
2	0,58	0,09	0,31
3	0,58	0,22	0,70
4	0,40	0,32	0,80

4.6 Relação entre a afiliação dos formantes à cavidade oral e o movimento facial

A geometria do trato vocal influencia o movimento facial e determina as freqüências dos formantes. Assim, existe um acoplamento entre as freqüências dos formantes e o movimento facial, uma vez que ambos estão ligados à geometria do trato vocal.

Os autovetores das componentes acusticamente alinhadas do movimento facial, ao longo do tempo, são menos constates do que os autovetores dos marcadores faciais, como mostra a Seção 4.3. Uma hipótese para esta maior variabilidade é de que a afiliação dos formantes à cavidade oral, na produção da fala, não ocorre seqüencialmente, i.e., o formante afiliado à cavidade oral é normalmente o segundo (ex. */a/*), mas pode também ser o terceiro (ex. */i/*). Este fato explica pelo menos parte do acoplamento variável que acontece no mapeamento entre a informação visual e a acústica da fala (Barbosa, 2004).

O movimento facial acopla-se de maneira desigual às freqüências dos formantes, sendo a freqüência mais fortemente acoplada aquela ligada à cavidade oral. Normalmente, na afiliação dos formantes, o primeiro formante afilia-se ao trato vocal como um todo, o segundo formante afilia-se à cavidade oral e o terceiro formante afilia-se à faringe (Apostol, Perrier, Baciu, Segebarth, e Badin, 2000; Silverman, 2006; Menard, Schwartz, Boe, e Aubin, 2007). Como citado anteriormente, na produção da vogal */a/*, por exemplo, o segundo formante está afiliado à maior cavidade do trato vocal que é a cavidade oral. Entretanto, na produção da vogal */i/*, é o terceiro formante que se afilia à cavidade oral, estando o segundo formante afiliado à cavidade da faringe que, no caso do */i/* torna-se maior que a cavidade oral. Deste modo, acredita-se que a relação entre as freqüências dos formantes e o movimento facial varie durante a fala. Esta mudança, entretanto, reflete o fato de o formante afiliado à cavidade oral variar com o conteúdo da fala, não havendo, portanto, uma real variação do acoplamento entre a acústica da fala e o movimento facial. O espectrograma de um locutor do sexo masculino produzindo o trecho */aio/* pode ser observado na Figura 4.14. As linhas sólidas indicam as freqüências dos três primeiros formantes. O formante $F1$ afilia-se ao trato vocal como um todo, o formante $F2$ afilia-se à maior cavidade no trato vocal que corresponde à boca para vogal */a/* e à faringe para vogal */i/*.

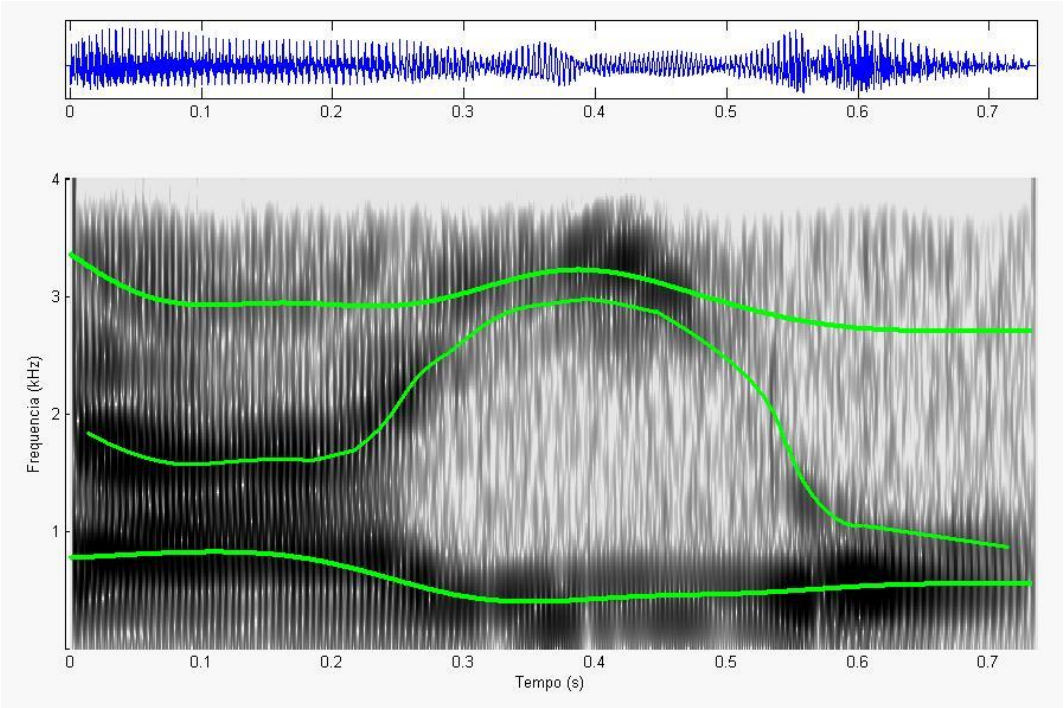


Figura 4.14: Espectrograma de */aio/* com as três primeiras freqüências de formantes indicadas pelas linhas sólidas.

Com o objetivo de analisar o acoplamento entre a acústica da fala e o movimento facial, levando-se em conta o formante afiliado à cavidade oral, realiza-se aqui um procedimento análogo ao da Seção 3.3, porém utilizando as freqüências dos três primeiros formantes em lugar dos parâmetros LSP para representar a acústica da fala.

De forma a verificar o efeito da variação do formante afiliado à cavidade oral, calculam-se os coeficientes de correlação entre as três componentes acusticamente alinhadas do movimento facial e estas mesmas três componentes estimadas (*i*) com base nos três primeiros formantes ordenados em ordem crescente de freqüência; (*ii*) com base nos formantes ordenados em função da cavidade de afiliação.

O trecho utilizado na análise é a palavra *sensacionais*, que tem para a vogal */i/* o formante $F2$ afiliado à laringe e o formante $F3$ afiliado à cavidade oral na estimação com base nos formantes ordenados em função da cavidade de afiliação. O espectrograma do trecho utilizado pode ser visto na Figura 4.15. A linha pontilhada indica os formantes $F1$ e $F2$ para a afiliação à cavidade oral.

Os resultados são mostrados na Tabela 4.16 e indicam, ainda que não conclusivamente, que a ordenação dos formantes em função da cavidade de afiliação conduz a uma melhor estimação das componentes acusticamente alinhadas do movimento facial.

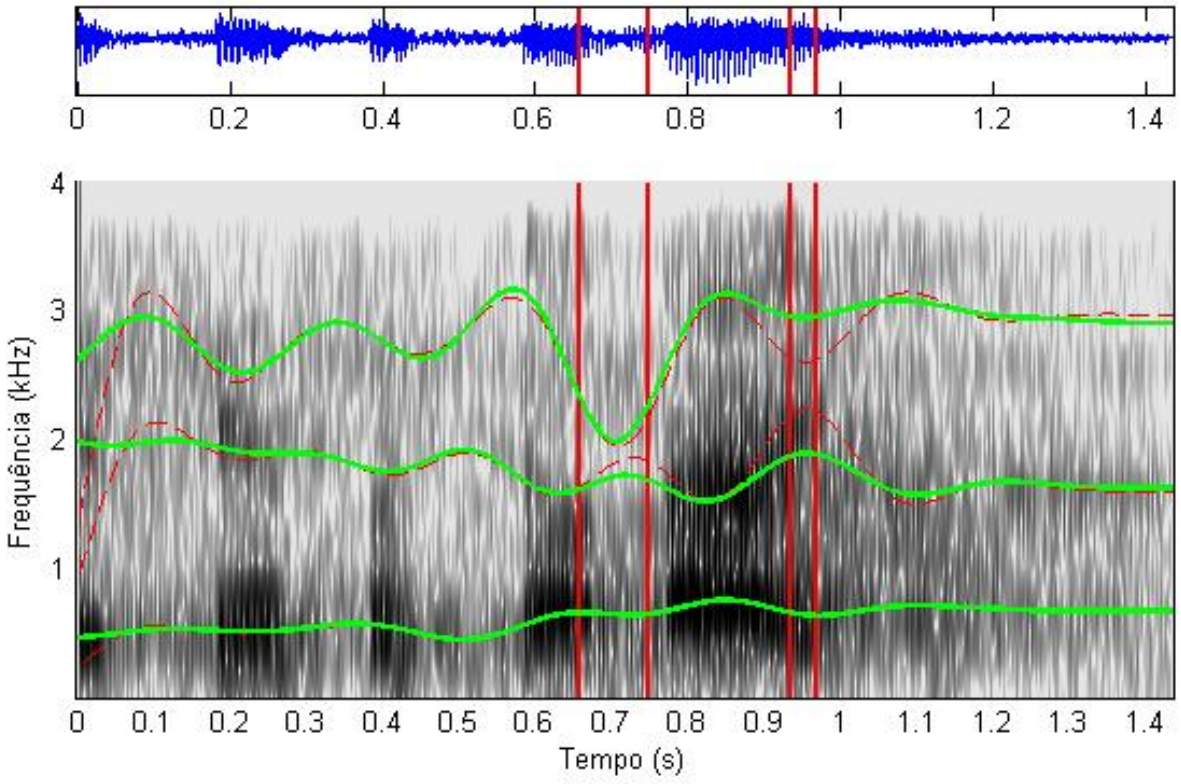


Figura 4.15: Espectrograma de *sensacionais* com as três primeiras freqüências de formantes indicadas pela linha sólida. A linha pontilhada indica a troca dos formantes, $F1$ e $F2$, para à afiliação a cavidade oral.

4.6.1 Relação entre a afiliação dos parâmetros LSP à cavidade oral e o movimento facial

A análise das componentes acusticamente alinhadas do movimento facial, descrito na Seção 4.3, utiliza parâmetros originados na localização de marcadores faciais e parâmetros LSP extraídos da acústica da fala (Yehia et al., 1998). O uso dos parâmetros LSP é justificado, pois eles são fortemente ligados aos formantes, que são determinados pela geometria do trato vocal, como visto na Seção 3.2.2 e ilustrado na Figura 4.16, em que cada par de parâmetros LSP está relacionado a um formante.

O objetivo desta seção é analisar o acoplamento entre a acústica da fala e o movimento facial, análogo ao realizado na Seção 4.6. Todavia, leva-se em conta o par de parâmetros LSP correspondente ao formante afiliado à cavidade oral. Para isto, utilizam-se os seis parâmetros LSP correspondentes aos três primeiros formantes. Na afiliação à cavidade oral, os parâmetros LSP, (w_2, θ_2), correspondentes ao formante $F2$, e os parâmetros LSP, (w_3, θ_3), correspondentes ao formante $F3$, têm seus pares de freqüências trocados para a estimação

Tabela 4.16: Coeficientes de correlação entre as componentes do movimento facial acusticamente alinhadas medidas (P_X) e estimadas com base nas freqüências dos três primeiros formantes ordenados (i) em ordem crescente ($P_{Xcrescente}$) e (ii) em função da cavidade de afiliação ($P_{Xafiliacao}$).

Número da componente	Coeficiente de Correlação	
	$\rho(P_X, P_{Xcrescente})$	$\rho(P_X, P_{Xafiliacao})$
1	0,68	0,71
2	0,77	0,79
3	0,70	0,81

com base nos formantes ordenados em função da cavidade de afiliação.

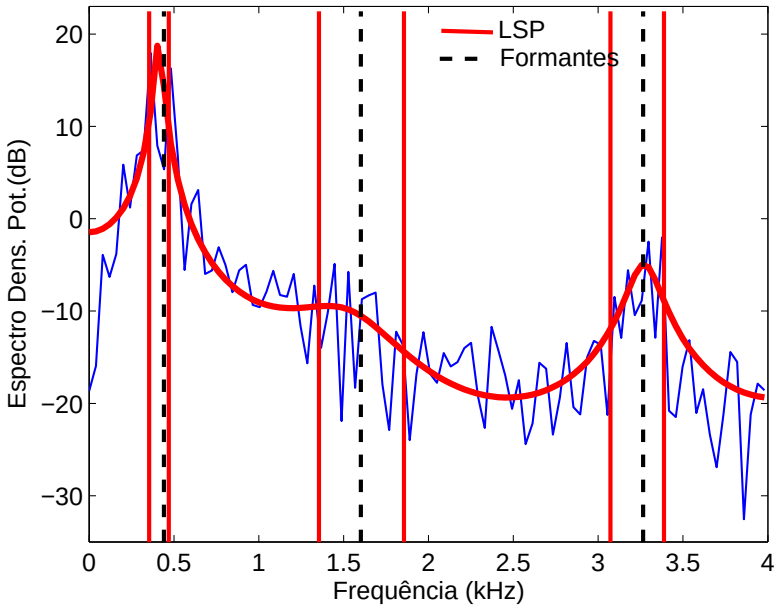


Figura 4.16: Parâmetros LSP fortemente ligados aos formantes. Cada par de parâmetros LSP representados pelas linhas verticais sólidas está associado a um formante representados pelas linhas verticais pontilhadas. A envoltória espectral é a resposta em freqüência do filtro LPC.

De forma a verificar o efeito da variação dos parâmetros LSP afiliados à cavidade oral, calculam-se os coeficientes de correlação entre as três componentes acusticamente alinhadas do movimento facial e estas mesmas três componentes estimadas (i) com base nos seis primeiros parâmetros LSP ordenados em ordem crescente de freqüência; e (ii) com base nos seis primeiros parâmetros LSP ordenados em função da cavidade de afiliação.

Os resultados são mostrados na Tabela 4.17 e indicam, similarmente à Seção 4.6, que a ordenação dos parâmetros LSP em função da cavidade de afiliação conduz a uma melhor estimação das componentes acusticamente alinhadas do movimento facial.

Uma outra forma de se verificar o efeito da ordenação dos parâmetros LSP em função da

Tabela 4.17: Coeficientes de correlação entre as componentes do movimento facial acusticamente alinhadas medidas (P_X) e estimadas com base nas freqüências dos seis primeiros parâmetros LSP ordenados (i) em ordem crescente ($P_{Xcrescente}$); e (ii) em função da cavidade de afiliação ($P_{Xafiliacao}$).

Número da componente	Coeficiente de Correlação	
	$\rho(P_X, P_{Xcrescente})$	$\rho(P_X, P_{Xafiliacao})$
1	0,90	0,92
2	0,84	0,87
3	0,82	0,73

cavidade de afiliação é pelo cálculo dos coeficientes de correlação entre as três componentes principais do movimento facial e as três componentes acusticamente alinhadas do movimento facial estimadas (i) com base nos seis primeiros parâmetros LSP ordenados em ordem crescente de freqüência; e (ii) com base nos seis primeiros parâmetros LSP ordenados em função da cavidade de afiliação.

Os resultados são mostrados na Tabela 4.18 e indicam também que a ordenação dos parâmetros LSP em função da cavidade de afiliação conduz a uma correlação maior entre as três componentes principais do movimento facial e as componentes acusticamente alinhadas do movimento facial estimadas.

Seguindo adiante com a análise, verifica-se agora o acoplamento entre os autovetores da matriz de covariância das posições dos marcadores e os autovetores obtidos a partir da matriz de correlação cruzada entre as posições dos marcadores e os parâmetros LSP calculados (i) com base nos seis primeiros parâmetros LSP ordenados em ordem crescente de freqüência; e (ii) com base nos seis primeiros parâmetros LSP ordenados em função da cavidade de afiliação. Na Figura 4.17, podem-se visualizar os três primeiros autovetores e, na Tabela 4.19, podem ser vistos os resultados dos coeficientes de correlação.

Analisando os resultados apresentados na Tabela 4.19 e na Figura 4.17, verifica-se que a ordenação dos parâmetros LSP em função da cavidade de afiliação conduz a uma maior correlação entre os autovetores dos marcadores e os autovetores originados a partir das componentes acusticamente alinhadas. Este resultado deve-se, possivelmente, a um melhor

Tabela 4.18: Coeficientes de correlação entre as componentes do movimento facial (P) e as componentes do movimento facial acusticamente alinhadas estimadas com base nas freqüências dos seis primeiros parâmetros LSP ordenados (i) em ordem crescente ($P_{Xcrescente}$); e (ii) em função da cavidade de afiliação ($P_{Xafiliacao}$).

Número da componente	Coeficiente de Correlação	
	$\rho(P, P_{Xcrescente})$	$\rho(P, P_{Xafiliacao})$
1	0,89	0,91
2	0,63	0,70
3	0,54	0,64

alinhamento entre o movimento facial e os parâmetros acústicos (LSP). Isto indica que os autovetores calculados a partir da matriz de correlação cruzada entre as posições dos marcadores e os parâmetros LSP ordenados em função da cavidade de afiliação são mais constates ao longo do tempo.

Um exemplo mais complexo pode ser visto na Figura 4.18 e nas tabelas 4.20, 4.21 e 4.22 que mostram os resultados para um experimento, em que a sentença *How are you?* é gravada por um locutor do sexo feminino. Nela, a afiliação dos formantes não ocorre seqüencialmente, sendo o formante $F1$ do primeiro fonema */u/* afiliado de alguma maneira à cavidade oral. No espectrograma visto na Figura 4.18, a linha pontilhada indica os formantes $F1$ e $F2$ para a afiliação à cavidade oral.

A correlação existente entre as componentes do movimento facial acusticamente alinhadas, medidas e estimadas com base nas freqüências dos três primeiros formantes ordenados (i) em ordem crescente; e (ii) em função da cavidade de afiliação é vista na Tabela 4.20. A Tabela 4.21 mostra os coeficientes de correlação entre as componentes do movimento facial acusticamente alinhadas, medidas e estimadas com base nas freqüências dos seis primeiros parâmetros LSP ordenados (i) em ordem crescente; e (ii) em função da cavidade de afiliação. Por fim, a Tabela 4.22 mostra os coeficientes de correlação entre os autovetores do movimento facial e os autovetores do movimento facial acusticamente alinhados, calculados com base nas freqüências dos seis primeiros parâmetros LSP ordenados (i) em ordem crescente; e (ii) em função da cavidade de afiliação. Por sua vez, a Figura 4.19 mostra os três primeiros autovetores. Um ponto importante a observar é que o primeiro autovetor do movimento facial acusticamente alinhado, calculado com base nas freqüências dos parâmetros LSP ordenados em função da cavidade de afiliação aproxima-se muito do autovetor oriundo dos marcadores.

4.7 Conclusão

Neste capítulo, são apresentadas as análises da variabilidade dos autovetores do movimento facial e das componentes acusticamente alinhadas. Viu-se que os autovetores do movimento facial e as componentes acusticamente alinhadas (autovetores) variam com o tempo devido ao

Tabela 4.19: Coeficientes de correlação entre os autovetores do movimento facial (U) e os autovetores do movimento facial acusticamente alinhados calculados com base nas freqüências dos seis primeiros parâmetros LSP ordenados (i) em ordem crescente $(U_{XFcrescente})$; e (ii) em função da cavidade de afiliação $(U_{XFafiliacao})$.

	Coeficiente de Correlação	
Número do autovetor	$\rho(U, U_{XFcrescente})$	$\rho(U, U_{XFafiliacao})$
1	0,97	0,98
2	0,90	0,96
3	0,78	0,77

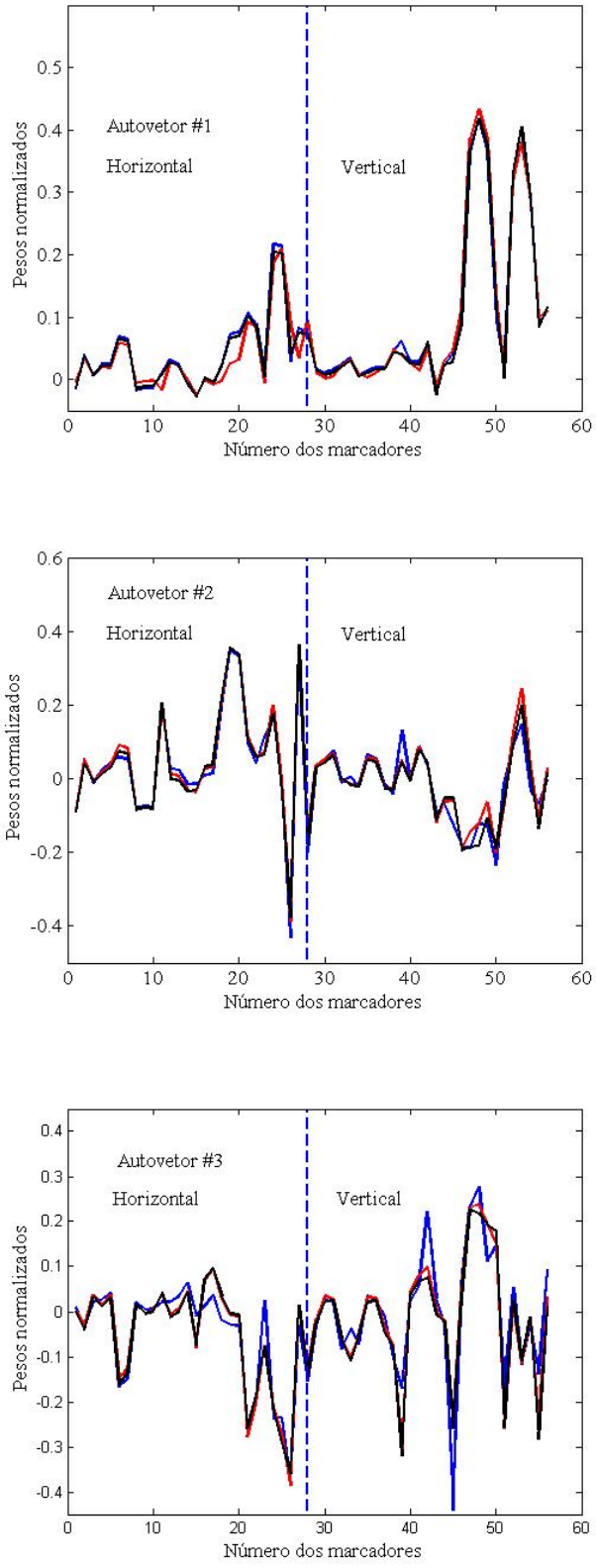


Figura 4.17: Três primeiros autovetores do movimento facial (em azul) e os autovetores do movimento facial acusticamente alinhados calculados (*i*) com base nas freqüências dos seis primeiros parâmetros LSP ordenados em ordem crescente (em vermelho); e (*ii*) com base nas freqüências dos seis primeiros parâmetros LSP ordenados em função da cavidade de afiliação (em preto).

Tabela 4.20: Coeficientes de correlação entre as componentes do movimento facial acusticamente alinhadas medidas (P_X) e estimadas com base nas freqüências dos três primeiros formantes ordenados (i) em ordem crescente ($P_{Xcrescente}$); e (ii) em função da cavidade de afiliação ($P_{Xafiliacao}$).

Número da componente	Coeficiente de Correlação	
	$\rho(P_X, P_{Xcrescente})$	$\rho(P_X, P_{Xafiliacao})$
1	0,91	0,95
2	0,82	0,90
3	0,62	0,60

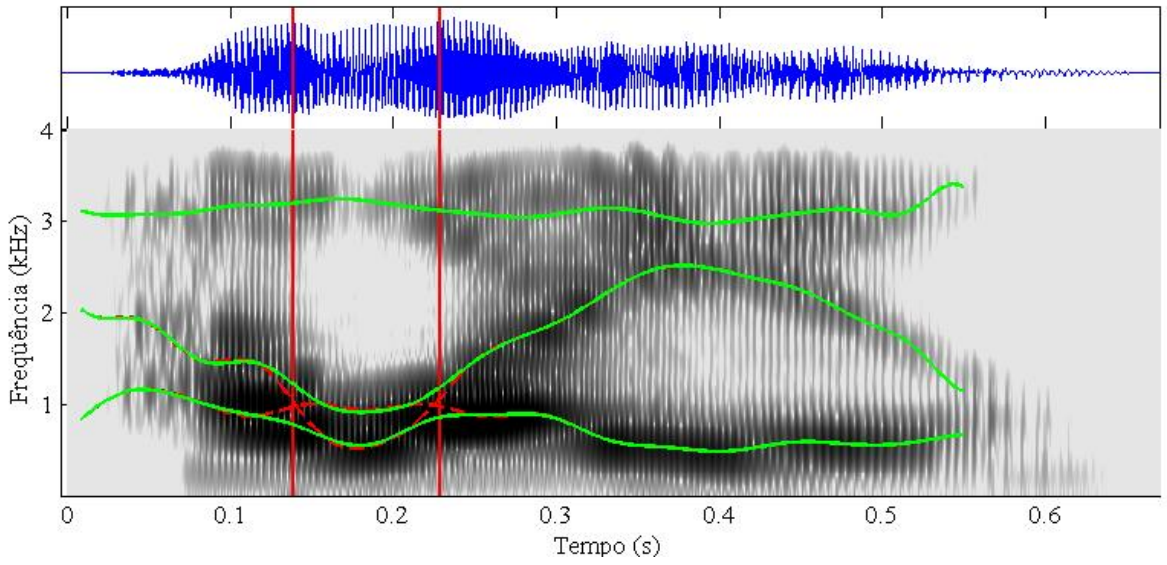


Figura 4.18: Espectrograma de *How are you?* com as três primeiras freqüências de formantes indicadas pela linha sólida. A linha pontilhada indica a troca dos parâmetros LSP para à afiliação a cavidade oral.

Tabela 4.21: Coeficientes de correlação entre as componentes do movimento facial acusticamente alinhadas medidas (P_X) e estimadas com base nas freqüências dos seis primeiros parâmetros LSP ordenados (i) em ordem crescente ($P_{Xcrescente}$); e (ii) em função da cavidade de afiliação ($P_{Xafiliacao}$).

Número da componente	Coeficiente de Correlação	
	$\rho(P_X, P_{Xcrescente})$	$\rho(P_X, P_{Xafiliacao})$
1	0,93	0,95
2	0,95	0,96
3	0,88	0,93

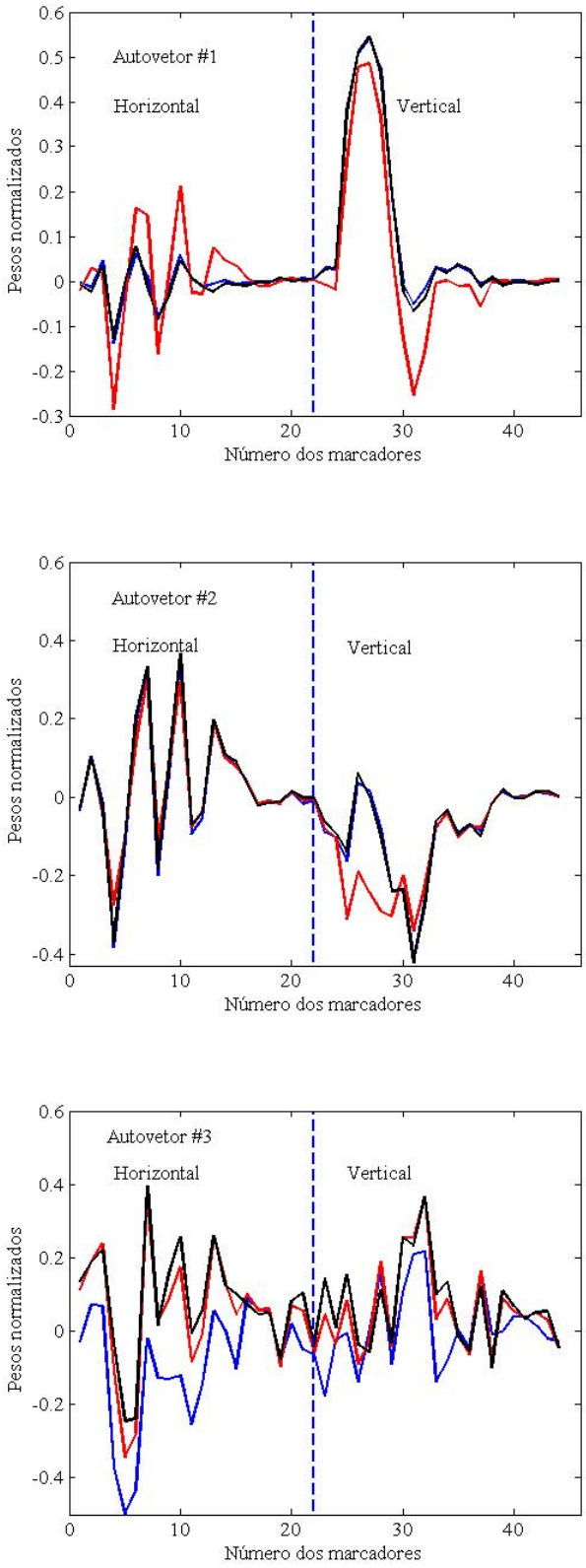


Figura 4.19: Três primeiros autovetores do movimento facial (em azul) e os autovetores do movimento facial acusticamente alinhados calculados (*i*) com base nas freqüências dos seis primeiros parâmetros LSP ordenados em ordem crescente (em vermelho); e (*ii*) com base nas freqüências dos seis primeiros parâmetros LSP ordenados em função da cavidade de afiliação (em preto).

Tabela 4.22: Coeficientes de correlação entre os autovetores do movimento facial (U) e os autovetores do movimento facial acusticamente alinhados calculados com base nas freqüências dos seis primeiros parâmetros LSP ordenados (i) em ordem crescente ($U_{XFcrescente}$); e (ii) em função da cavidade de afiliação ($U_{XFafiliacao}$).

Número do autovetor	Coeficiente de Correlação	
	$\rho(U, U_{XFcrescente})$	$\rho(U, U_{XFafiliacao})$
1	0,89	0,99
2	0,88	0,99
3	0,61	0,73

conteúdo acústico. Entretanto, o primeiro autovetor (1ª componente principal) é constante, variando muito pouco ao longo do tempo. Resultado similar é encontrado para o primeiro autovetor da componente do movimento facial acusticamente alinhada, porém, com uma variabilidade maior.

Também foram feitos experimentos para verificar o efeito Lombard (Junqua, 1993; Vatikiotis-Bateson, Chung, Lutz, Mirante, Otten, e Tan, 2006; Vatikiotis-Bateson, Barbosa, Chow, Oberg, Tan, e Yehia, 2007). Dois locutores proferiram trechos do texto *O Popular* (ver Seção 3.1) com diferentes intensidades entretanto não foi observado variações significativas nos resultados.

Outro resultado encontrado indica que os autovetores apresentam menores variabilidades para análises em trechos maiores. Assim, quando trechos com durações maiores são utilizados, perde-se na informação temporal instantânea e se ganha na informação espacial global do movimento. É observado também que, quando comparado o movimento da face entre locutores, estes apresentam diferenças maiores em relação às obtidas nas comparações para um mesmo locutor, indicando que, além do conteúdo, os autovetores exibem informações inerentes à anatomia de cada locutor. Este é o foco do Capítulo 5.

Por último, verifica-se que os autovetores das componentes do movimento facial acusticamente alinhadas são menos constates, ao longo do tempo, do que os autovetores dos marcadores faciais. Entretanto, uma menor variabilidade é observada quando os parâmetros LSP, usados na representação acústica da fala, são ordenados em função da sua cavidade de afiliação ao invés de simplesmente ordenados em ordem crescente.

Capítulo 5

Análise da caracterização de locutor em função das componentes principais do movimento facial

No Capítulo 4, analisa-se a variação dos autovetores do movimento facial para locutores e observa-se a existência de uma menor variabilidade para o primeiro autovetor. Para os demais autovetores observa-se uma variabilidade que ocorre principalmente devido à variação do conteúdo falado. Neste capítulo, é estudada a variabilidade dos autovetores do movimento facial entre diferentes locutores. O estudo verifica a possibilidade de identificação de locutor por meio das componentes principais do movimento facial de forma dependente e independente do conteúdo acústico. O objetivo é investigar se autovetores do movimento facial contêm características inerentes a cada locutor. Caso isto ocorra, podem-se utilizar tais autovetores para representar especificidades do movimento na tarefa de identificação de locutor.

5.1 Estudo de identificação de locutor

Locutores movimentam a face ao produzir informações acústicas (fala). A maioria dos movimentos da face está relacionada ao conteúdo lingüístico, à informação paralingüística (e.g. emoção) e à prosódia. Além disso, tais movimentos contêm informações sobre a especificidade do movimento de cada locutor, considerando que a forma do crânio, da mandíbula, a rigidez dos tecidos e o movimento dos músculos apresentam particularidades de cada locutor.

Na análise feita no capítulo anterior, observa-se que os primeiros autovetores do movimento facial capturam os principais movimentos originados no acoplamento entre as regiões faciais. Neste capítulo, a hipótese considerada é de que os autovetores da matriz de covariância entre os movimentos da face contêm informações sobre características específicas

do movimento facial de cada locutor. Se os autovetores do movimento facial forem capazes de capturar a individualidade do locutor, torna-se possível utilizar tal informação no processo de identificação de locutores.

5.1.1 Metodologia para caracterização de locutor

Seguindo um processamento similar ao descrito na Seção 3.2.1, oito locutores ($L = 8$) leram trechos da crônica *O Popular* (Tabela 3.1). Para cada trecho, foi determinado o primeiro autovetor da matriz de covariância do movimento facial[1]. Estes autovetores foram organizados na forma de uma matriz

$$\mathbf{U} = \left[\begin{array}{cccccccccc} \mathbf{u}_{1,1} & \cdots & \mathbf{u}_{1,Q} & \mathbf{u}_{2,1} & \cdots & \mathbf{u}_{2,Q} & & \cdots & \mathbf{u}_{L,1} & \cdots & \mathbf{u}_{L,Q} \end{array} \right].$$

Cada autovetor $\mathbf{u}_{lq}$, corresponde à q-ésima elocução do locutor l, é então decomposto em

$$\mathbf{u}_{lq} = \mu + \tau_l + \mathbf{e}_{lq}, \tag{5.1}$$

em que μ é o autovetor médio de todas as elocuções de todos os locutores, $\mu + \tau_l$ é o autovetor médio de cada locutor e $\mathbf{e}_{lq}$ é a componente de cada autovetor $\mathbf{u}_{lq}$ específica de cada elocução de cada locutor.

Foram feitos dois experimentos com oito meses de intervalo entre si. O número de locutores L é igual a 8 para o primeiro experimento, e igual a 3 para o segundo experimento. O número mínimo de elocuções é igual a 6 para cada locutor do primeiro experimento, e igual a 4 para cada locutor do segundo experimento.

Observando o primeiro autovetor dos marcadores faciais entre locutores diferentes, verifica-se a existência de um padrão de comportamento inerente à anatomia humana. Na produção da fala, todos os locutores movimentam, em maior ou menor intensidade, regiões específicas da face de forma semelhante. O primeiro autovetor do movimento facial para duas repetições de elocuções produzidas por quatro locutores é mostrado na Figura 5.1. Estes autovetores são obtidos em trechos com aproximadamente um minuto de duração. Esta duração relativamente longa realça informações específicas do locutor em contraste com durações menores que realçam características específicas da elocução.

Retirando da matriz $\mathbf{U}$ a média global μ, obtém-se a matriz

$$\mathbf{M} = \mathbf{U} - \mu, \tag{5.2}$$

a qual contém informações específicas de cada locutor:

$$\mathbf{M} = \left[\begin{array}{cccc} \mathbf{M}_1 & \mathbf{M}_2 & \cdots & \mathbf{M}_L \end{array} \right].$$

[1] Análises feitas com os demais autovetores não resultaram em informação que permitisse caracterização do locutor.

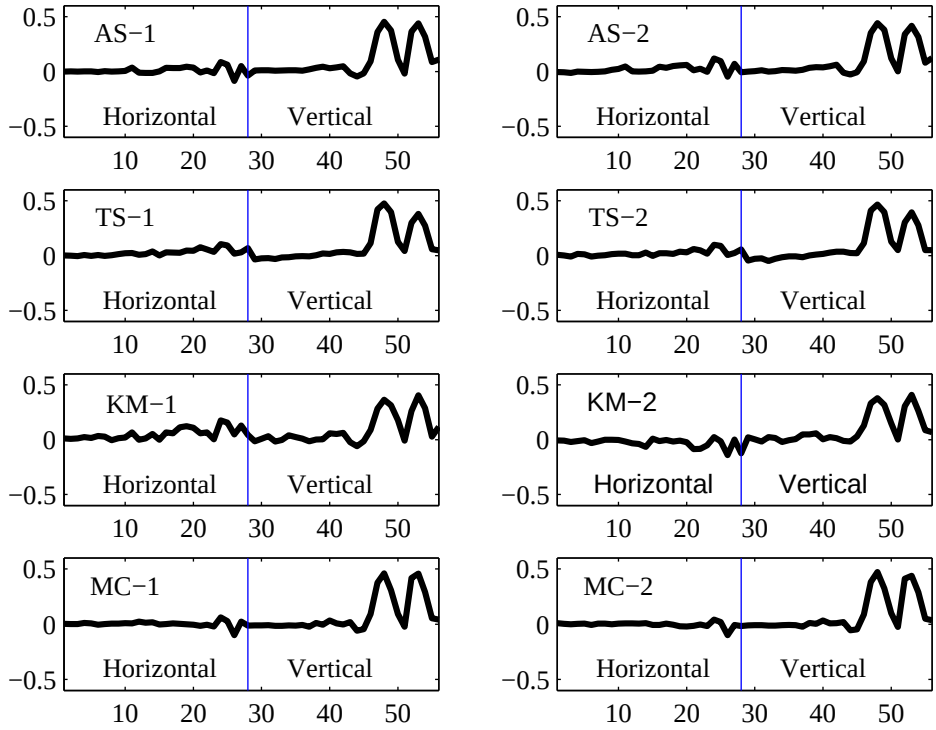
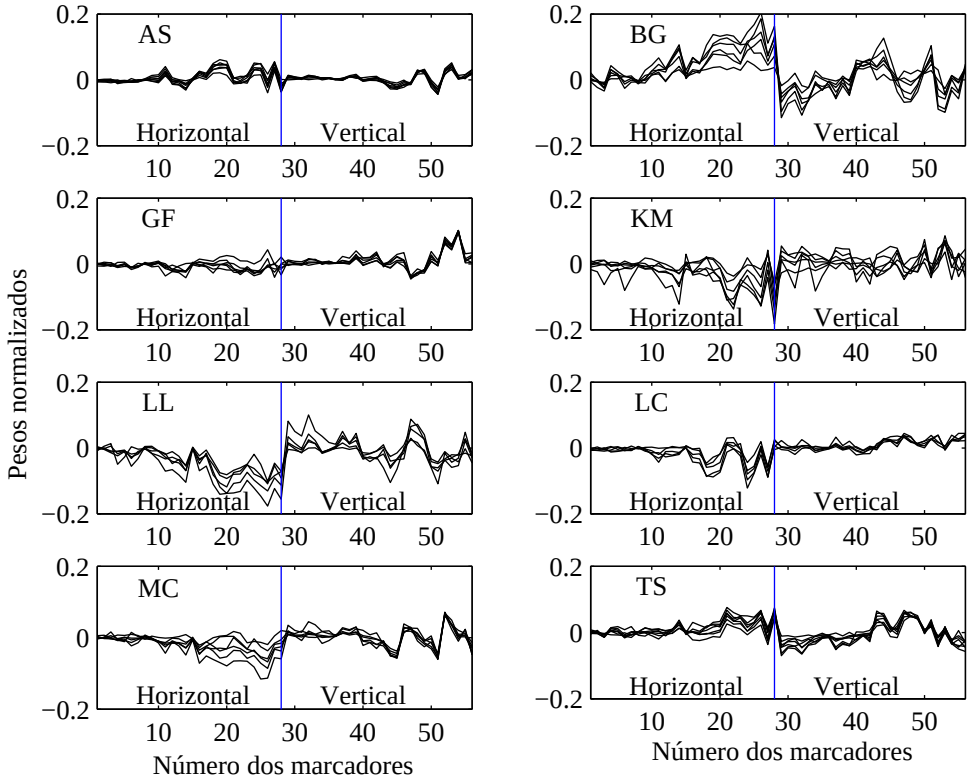


Figura 5.1: Comparação do primeiro autovetor entre locutores, demonstrando a existência de um padrão de comportamento semelhante. Locutores AS, TS, KM, e MC. Elocuções 1 (esquerda) e 2 (direita) do primeiro experimento.

Assim, para cada locutor, é formada uma matriz $\mathbf{M}_l$, composta por características próprias, τ_l, adicionada a informações específicas de cada elocução e_{lq}

$$\mathbf{M}_l = \tau_l + [\mathbf{e}_{l1} \quad \mathbf{e}_{l2} \quad \cdots \quad \mathbf{e}_{lQ}] \ .$$

A Figura 5.2 mostra os dados da matriz $\mathbf{M}_l$ para o primeiro autovetor. Nela, observa-se uma semelhança nos vetores originados dos autovetores de um mesmo locutor. Quantitativamente, o cálculo do coeficiente de correlação médio entre o vetor médio de cada matriz $\mathbf{M}_l$, $l = 1, 2, ..., L$ e os vetores da matriz $\mathbf{M}$ também indica a capacidade de identificação do locutor pelo movimento (ver Tabela 5.1). Os resultados do coeficiente de correlação são mostrados também na Figura 5.3 para o primeiro autovetor. Os dados utilizados para o cálculo da correlação contêm as observações de ambos os experimentos.



(a)

Figura 5.2: Matriz $\mathbf{M}_l$, composta por características próprias, τ_l, adicionada a informações específicas de cada elocução e_{lq}. Na comparação entre os vetores originados dos autovetores das observações de cada locutor retirando o autovetor médio global observa-se uma semelhança entre os autovetores originados de um mesmo locutor (Autovetor 1).

Tabela 5.1: Coeficiente de correlação médio entre os autovetores de cada elocução de cada locutor e o autovetor médio de cada locutor.

	AS	BG	GF	KM	LL	LC	MC	TS
AS	0,93	0,03	0,11	0,21	0,22	0,21	0,10	0,32
BG	0,07	0,94	0,16	0,62	0,51	0,64	0,71	0,57
GF	0,11	0,17	0,95	0,18	0,36	0,12	0,18	0,42
KM	0,14	0,44	0,08	0,70	0,08	0,56	0,21	0,40
LL	0,16	0,47	0,19	0,15	0,65	0,29	0,33	0,22
LC	0,23	0,18	0,05	0,34	0,08	0,40	0,13	0,24
MC	0,09	0,62	0,14	0,26	0,33	0,47	0,84	0,32
TS	0,33	0,56	0,41	0,56	0,29	0,49	0,20	0,93

5.1.2 Variação dos autovetores interlocutores para o mesmo conteúdo acústico e variação dos autovetores intralocutor independentemente do conteúdo acústico

Com a finalidade de verificar o grau de variação nos autovetores para locutores diferentes, porém com o mesmo conteúdo acústico, calculam-se a distância euclidiana entre os autovetores de um trecho de um determinado locutor com duração de aproximadamente um minuto e os autovetores obtidos por meio de uma janela deslizante que percorre trechos com o mesmo conteúdo acústico, proferidos por locutores diferentes. O trecho padrão é utilizado na comparação entre os demais locutores para verificar as semelhanças existentes entre eles. A intenção é analisar as distâncias entre os autovetores do movimento facial para os locutores ao longo do tempo e verificar se a existência das características inerentes a cada locutor depende do conteúdo proferido.

O resultado percentual da distribuição das distâncias euclidianas ao longo do tempo é mostrado na Figura 5.4. A comparação acontece entre os autovetores do trecho padrão pertencente ao locutor A e os autovetores obtidos por meio da janela que percorre trechos pertencentes aos locutores A e B, proferindo o mesmo texto. Como pode ser observado, para o primeiro autovetor existe uma diferença significativa na localização das distribuições das distâncias, que são menores para o mesmo locutor do trecho padrão e maiores para o locutor diferente do que pronuncia o trecho padrão.

A análise com o primeiro autovetor demonstra que a não variação não depende do conteúdo falado, o que torna este autovetor eficaz para identificação de locutores. Porém, este resultado não é uniforme para todos os locutores, como mostra a matriz de confusão da Tabela 5.2. Esta matriz de confusão é usada na identificação de locutor com base na distância mínima entre o primeiro autovetor médio de cada locutor e os autovetores obtidos por meio da janela que percorre trechos pertencentes aos demais locutores.

De maneira semelhante à anterior, é analisada a variação dos autovetores intralocutores

Tabela 5.2: Matriz de confusão para identificação de locutor com base na distância mínima ao primeiro autovetor médio de cada locutor.

Ref \ Teste	AS	BG	GF	KM	LL	LC	MC	TS
AS	69,5	7,1	2,9	5	0,6	0	13,4	1,6
BG	1,2	49,4	0	25,1	0	20,5	3,8	0
GF	29,1	1	51,3	0,7	0	0	17,9	0
KM	0	0	0	100	0	0	0	0
LL	0	0	0	0	100	0	0	0
LC	0	6,1	33	0	0	60,9	0	0
MC	0	0	0	5,2	0	3,7	91,1	0
TS	0	1,8	0	0	2,1	0	0	96,1

independentemente do conteúdo acústico. A comparação acontece entre os autovetores do trecho padrão pertencente ao locutor A e os autovetores obtidos por meio da janela que percorre trechos pertencentes ao locutor A com conteúdos acústicos diferentes proferidos em experimentos diferentes. O resultado perceptivo da distribuição das distâncias euclidianas ao longo do tempo é mostrado na Figura 5.5. Nesta figura, observa-se que as distâncias do primeiro autovetor são pequenas para ambos os conjuntos de dados analisados. Estes resultados demonstram que, para um mesmo locutor, o primeiro autovetor mantém-se constante independentemente do conteúdo acústico.

5.2 Variabilidade do primeiro autovetor entre locutores

O primeiro autovetor está associado aos principais movimentos localizados na região da boca, os quais independem do conteúdo acústico. Este fato justifica a não variação do primeiro autovetor ao longo do tempo, podendo ser utilizado no processo de identificação de locutor por meio do movimento facial.

Para uma avaliação da capacidade de caracterização do locutor por meio do primeiro autovetor, verifica-se quantitativamente sua variabilidade total descrita pelo vetor da soma quadrática interlocutores dos marcadores:

$$\sigma_T^2 = \sum_{l=1}^{L}\sum_{q=1}^{Q}(\mathbf{u}_{lq} - \bar{\mathbf{u}})^2, \tag{5.3}$$

em que $\mathbf{u}_{lq}$ representa os vetores correspondentes ao locutor l do trecho proferido q do primeiro autovetor e $\bar{\mathbf{u}}$ é a média global. Este vetor da variabilidade (soma quadrática) é comparado com a variabilidade (soma quadrática) de intralocutor do autovetor 1:

$$\sigma_l^2 = \sum_{q=1}^{Q}(\mathbf{u}_{lq} - \bar{\mathbf{u}}_l)^2, \tag{5.4}$$

em que $\bar{\mathbf{u}}_l$ é a média global do locutor l.

A variabilidade do primeiro autovetor do movimento facial entre todos os locutores pode ser vista na Figura 5.6, Eq. 5.3. Comparada à variabilidade intrínseca de cada locutor, Eq. 5.4. Nesta figura observa-se também que a variação maior ocorre para o movimento horizontal na região da boca (ver Fig. 3.3).

5.3 Classificação de locutores por meio de Redes Neurais

O objetivo desta seção é verificar se os autovetores do movimento facial contêm informações inerentes a cada indivíduo. Neste sentido, os dados do primeiro autovetor são utilizados para reconhecimento dos locutores por meio de uma rede neural. Para isto são utilizados trechos de aproximadamente um minuto e trechos de aproximadamente 30 segundos compostos por conteúdos repetidos e diferentes. O que torna possível a classificação de locutores pelos movimentos faciais são as diferenças interlocutores oriundas de particularidades no acoplamento dos movimentos. Após a aquisição, uma rede neural artificial é utilizada para classificar o locutor. Cabe observar que, neste trabalho, utilizam-se marcadores para a aquisição do movimento. Entretanto, a utilização do reconhecedor de indivíduo por meio de movimentos faciais em uma aplicação real depende do desenvolvimento de um algoritmo que seja capaz de rastrear o movimento da face sem que seja necessário colocar marcadores sobre a face do locutor. Técnicas de fluxo ópticos podem ser usadas para esse fim (Horn e Rhunck, 1981).

A metodologia para classificação inicia-se com a normalização entre -1 e 1 dos vetores que compõem a matriz $\mathbf{M}$ (Eq. 5.2). Nesta seção, o grupo total de dados utilizado é composto por $Q = 243$ vetores pertencentes a oito locutores ($L = 8$) obtidos em dois experimentos diferentes. Apenas três dos locutores participaram do segundo experimento no qual se obteve um total de trinta e três trechos ($Q = 33$). Do grupo total de dados originaram-se dois grupos: um de treinamento e outro de validação. Foram escolhidos 60% dos dados para treinamento e 40% dos dados para validação.

No treinamento utiliza-se uma rede MLP (Perceptrons de Múltiplas Camadas) (Braga et al., 2000; Haykin, 2001). A arquitetura da rede consiste em um conjunto que constitui a entrada da rede, duas camadas ocultas de neurônios e uma camada de saída. O algoritmo de retropropagação utilizado para o treinamento é o algoritmo *Backpropagation - quasi-Newton method* (Haykin, 2001). Na camada oculta o número de neurônios é escolhido empiricamente com o objetivo de determinar o melhor resultado. A camada de saída apresenta um único neurônio cuja resposta varia de 1 a 8 classificando o locutor a ser reconhecido. Nas camadas ocultas utilizam-se funções de ativação do tipo tangente hiperbólica e, na camada de saída utiliza-se função linear para a ativação dos neurônios. A verificação do desempenho alcançado é obtida comparando os valores da saída da rede com o locutor.

O resultado da validação dos dados para a rede treinada é apresentado na Tabela 5.3. A rede treinada classificou $86,7\%$ dos locutores corretamente, demonstrando a existência de informações que são específicas de cada locutor.

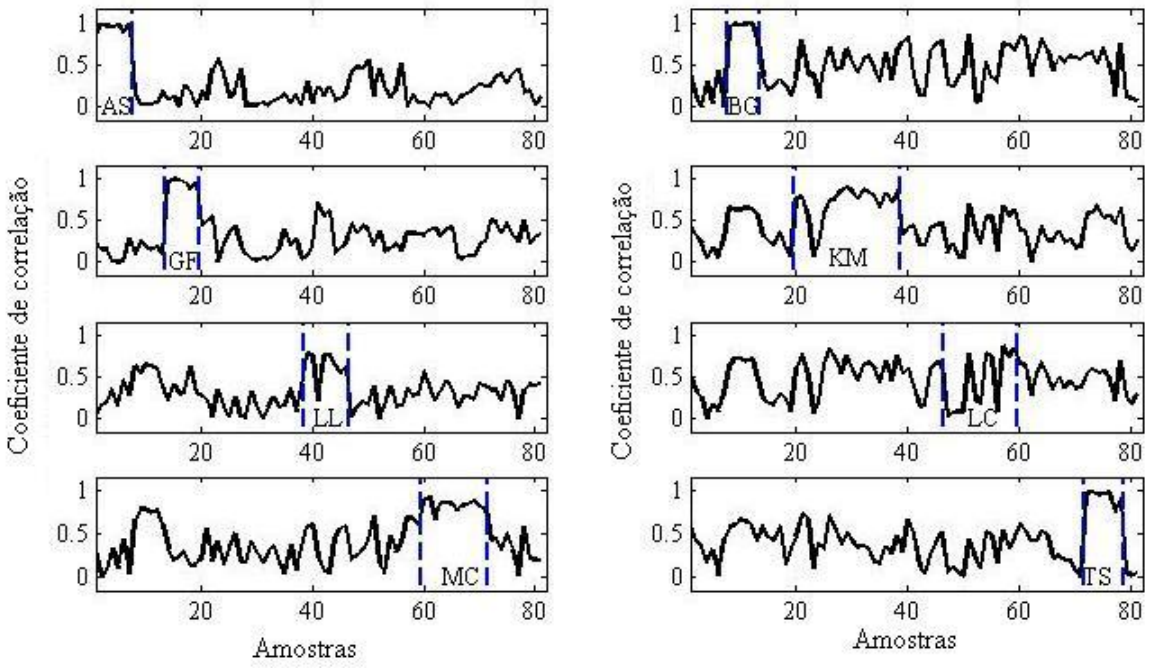
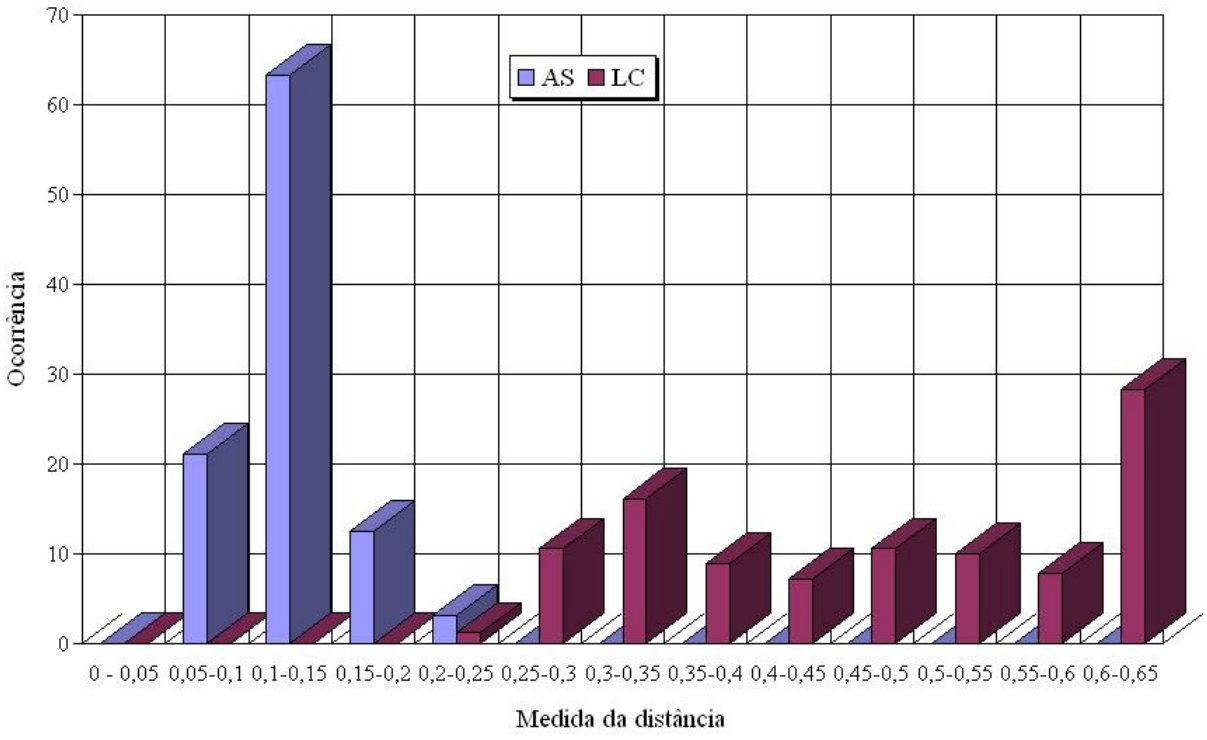


Figura 5.3: Correlação entre as características do locutor proveniente das observações menos a média global e a média correspondente à característica de cada locutor para o autovetor 1 (experimentos 1 e 2).

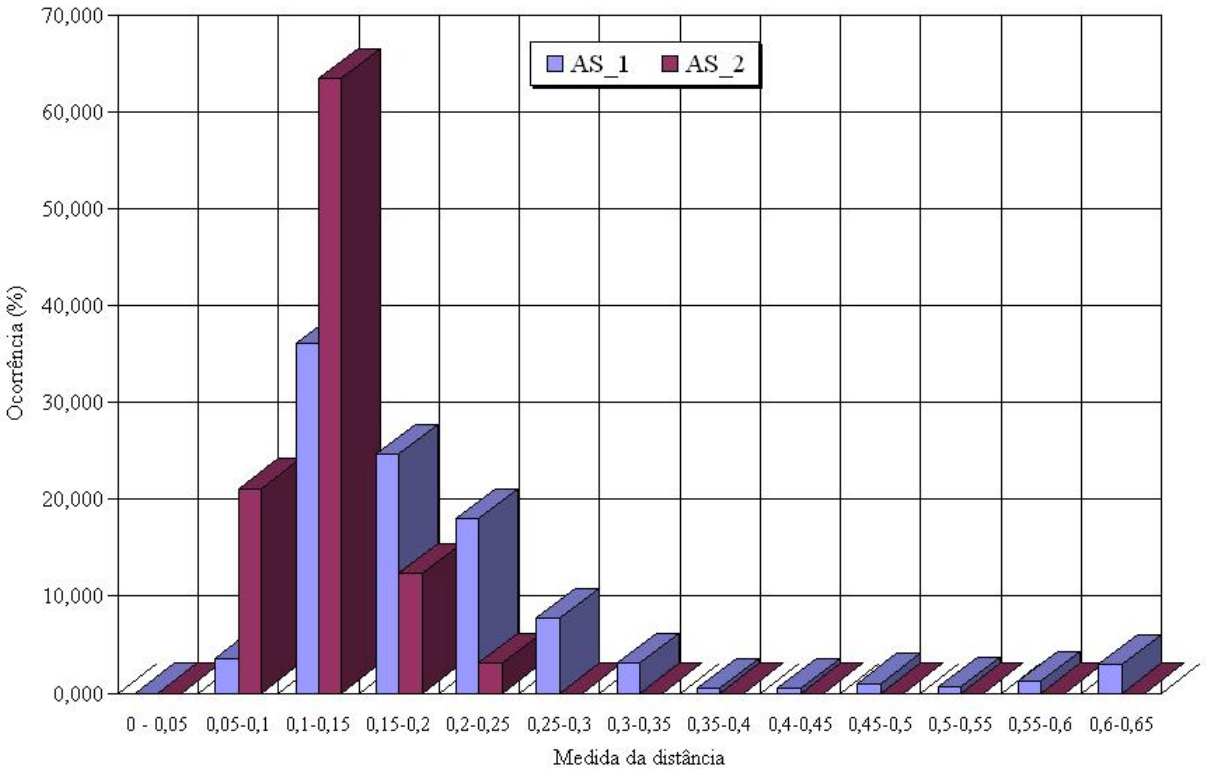
Tabela 5.3: Rede neural treinada pelo algoritmo *Backpropagation* para reconhecimento de locutores por meio do movimento facial. Cada linha apresenta o número de dados utilizados no treinamento e na validação para cada experimento. A última linha representa o reconhecimento obtido com os dados da validação.

Locutor		AS	BG	GF	KM	LL	LC	MC	TS
Experimento I e II	Treinamento	13	12	12	41	10	22	24	14
Experimento I e II	Validação	10	6	7	30	10	11	14	7
Resultado	Número de acertos na validação	9	2	7	29	9	8	14	4
Total de acertos na validação		86,7 %							



(a)

Figura 5.4: Distribuição percentual da distância euclidiana entre autovetores de um trecho padrão e autovetores originados de uma janela que percorre trechos com o mesmo conteúdo acústico para locutores diferentes (Autovetor 1).



(a)

Figura 5.5: Distribuição percentual da distância euclidiana entre autovetores de um trecho padrão e autovetores originados de uma janela que percorre trechos com diferentes conteúdos acústicos para o mesmo locutor (Autovetor 1).

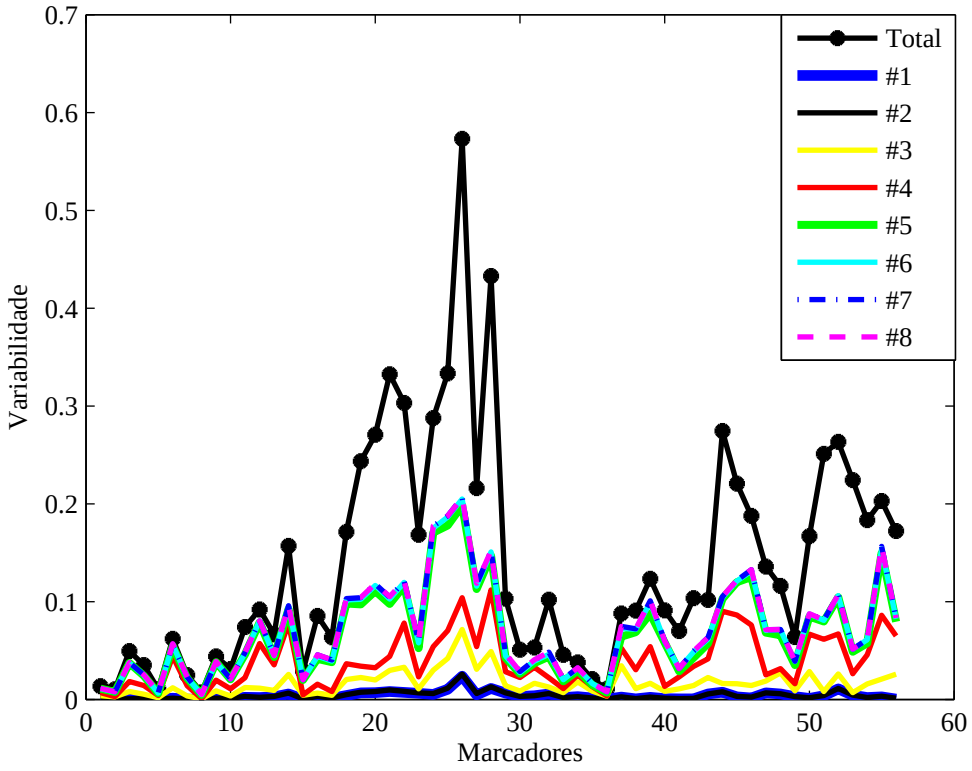


Figura 5.6: Variabilidade do primeiro autovetor do movimento facial entre todos os locutores comparada à variabilidade intrínseca de cada locutor. Observa-se que a variação maior ocorre para o movimento horizontal na região da boca.

5.4 Classificação de locutores por meio de imagens faciais com diferentes relações sinal/ruído adicionadas à informação do movimento

Classificação automática de locutores utilizando seqüências de imagens de faces é amplamente empregada no reconhecimento automático de pessoas. Contudo, o reconhecimento fica prejudicado quando a relação sinal/ruído é baixa, sendo muitas vezes necessário adicionar outras características do indivíduo para melhorar o desempenho dos classificadores. Com o objetivo de observar se o movimento contribui para um melhor reconhecimento, é feita uma comparação entre a classificação de locutores por meio de imagens faciais e a classificação de locutores por meio de imagens faciais acrescidas de informações de movimentos faciais. Nesta seção, a classificação é realizada por meio do cálculo da distância euclidiana, sem a utilização de redes neurais.

Como resultado da classificação de locutores por meio de imagens faciais com relação sinal/ruído elevada, obtêm-se 89% dos locutores classificados corretamente. Por outro lado,

Tabela 5.4: Caracterização de locutores utilizando a imagem da face do locutor e a imagem da face acrescida do movimento facial. Reconhecimento percentual para diferentes relação sinal/ruído.

Relação sinal/ruído (dB)		30	20	10	5	3	0	-3	-5	-10	-20
Face	Reconhecimento (%)	89	89	89	88	88	85	83	79	73	22
Face + Movimento	Reconhecimento (%)	89	89	89	87	87	86	85	86	75	30

como resultado da classificação de locutores por meio do movimento facial obtêm-se 83,3% dos locutores classificados corretamente. Agrupando as informações das imagens faciais às informações do movimento facial, obtêm-se como resultado da classificação 89% dos locutores classificados corretamente no caso da relação sinal/ruído elevada. Entretanto, ao reduzir a relação sinal/ruído da imagem, o resultado da classificação com a adição de informações de movimento facial torna-se superior à classificação com base apenas em informações de imagens faciais. Este fato é mostrado na Tabela 5.3, em que se observa o reconhecimento de indivíduos por meio de imagens faciais com diferentes relação sinal/ruído. Nesta mesma tabela verifica-se o reconhecimento de indivíduos por meio de imagens faciais com diferentes relações sinal/ruído acrescidas de informações de movimentos faciais. Observa-se que quando a relação sinal/ruído cai abaixo de 0 dB, o reconhecimento melhora com a utilização da informação do movimento.

5.5 Conclusão

Analisando os autovetores entre os diversos locutores, pode-se observar uma padronização destes autovetores. Este resultado é esperado já que locutores ao proferirem um trecho movimentam a face globalmente de forma semelhante. Entretanto, além das informações visuais devido ao conteúdo acústico, os autovetores carregam características inerentes a cada locutor. Porém, estas características estão sujeitas a ruídos causados pela imprecisão na localização do centro exato dos marcadores, informações lingüísticas e paralingüísticas e prosódia. Estes ruídos dificultam a análise dos dados quanto às características inerentes a cada locutor.

A rede neural utilizada para classificar os locutores apresentou $86,7\%$ de acerto. O número reduzido de locutores é um fator que deve ser considerado. Apesar das análises feitas indicarem presença de informações inerentes a cada locutor nos autovetores do movimento facial, mais estudos devem ser realizados para a classificação de indivíduos pelo movimento.

A importância deste estudo é demonstrar a existência de informações específicas de cada locutor no primeiro autovetor do movimento facial, além de demonstrar que estes movimentos podem ser utilizados em conjunto com outras características para melhorar o desempenho de um sistema reconhecimento.

Capítulo 6

Conclusão

As deformações faciais são partes integrantes da comunicação controladas pelo sistema nervoso e são em grande parte resultantes do movimento do trato vocal. Desse modo, movimentando principalmente a língua e a mandíbula, modificam-se a forma e o comprimento da cavidade oral produzindo os diversos sons da linguagem. Na face, o movimento ocorre em maior proporção na região dos lábios, em que a maior amplitude acontece na vertical para o lábio inferior e queixo. Os movimentos da cabeça e sobrancelhas acontecem esporadicamente e neste estudo foram compensados e desconsiderados, respectivamente.

Ao longo deste trabalho, a análise que avalia quantitativamente a variabilidade de parâmetros obtidos no acoplamento existente entre partes da face, utiliza-se dos autovetores da matriz de covariância entre a posição dos marcadores da face. Do mesmo modo, a análise que avalia quantitativamente a variabilidade de parâmetros obtidos no acoplamento entre a acústica da fala e a informação facial utiliza-se dos autovetores da matriz de correlação cruzada entre os parâmetros que representam a informação visual e a informação acústica. Esses autovetores obtidos da relação entre partes faciais e entre o acoplamento acústico e facial demonstram a direção da variabilidade do movimento dos marcadores e do acoplamento acústico apresentado em ordem decrescente de variância. No estudo do movimento facial e do acoplamento acústico foram utilizadas seis componentes principais, suficientes para representar 95% da variância do movimento.

Analisando os resultados dos experimentos, conclui-se que o primeiro autovetor indica, principalmente, o movimento do lábio inferior e da mandíbula e representa aproximadamente 55% da variância do movimento facial. O restante da variabilidade do movimento está nos demais autovetores que representam as relações entre as diversas partes da face. Os autovetores 2 e 3 indicam os movimentos nas regiões do lábio superior, lábio inferior na horizontal e cantos dos lábios e representam aproximadamente 30% dos movimentos da face. O autovetor 4 indica os movimentos nas regiões das bochechas e área da boca que se movimenta menos e representa aproximadamente 7% dos movimentos da face. Finalmente, os demais autovetores indicam os movimentos do lábio inferior no eixo horizontal, além de outros

movimentos de menor importância, representando aproximadamente 3% dos movimentos da face.

Dando seqüência ao processo de caracterização do modelo de produção acústica, obteve-se como resultado a variabilidade dos autovetores do movimento facial. O primeiro autovetor apresenta-se extremamente constante, e esta não variação independe do conteúdo acústico. Assim, existe na face uma parcela de movimento que, independentemente do conteúdo acústico e do contexto, não se altera. Outro fato verificado é que quanto maior a variância representada pelo primeiro autovetor, menor é a sua variabilidade. Por outro lado, os demais autovetores demonstram dependência do conteúdo acústico resultando em vetores variáveis ao longo do tempo.

Similarmente, o primeiro autovetor da componente acusticamente alinhada do movimento facial também apresenta mais constante em relação aos demais autovetores. Entretanto, fazendo uma comparação por meio da distância euclidiana entre os autovetores do movimento facial e os autovetores da componente acusticamente alinhada obtêm-se valores maiores, (i.e. maior variabilidade), no alinhamento acústico. Este fato, em parte, é conseqüência do modelo linear utilizado para representar o mapeamento entre as componentes acústicas e as deformações faciais. Outra justificativa para variabilidade maior no alinhamento acústico é o acoplamento variável que acontece no mapeamento entre a informação visual e a acústica da fala ao longo do tempo. Se esse mapeamento acontece utilizando os parâmetros LSP seqüencialmente tem-se como resultado uma variabilidade maior dos dados. Porém, a variabilidade dos autovetores das componentes do movimento facial acusticamente alinhadas diminui quando se leva em conta a afiliação dos parâmetros LSP à cavidade oral.

Na análise da variabilidade dos autovetores de diferentes tamanhos de sentenças para aplicá-los no processo de caracterização de locutor, observa-se uma padronização do movimento facial para trechos mais longos (60 segundos). Nestas sentenças, a forma de falar do locutor e os movimentos que acontecem na face se sobressaem às pequenas variações, resultando em pouca dependência dos autovetores ao conteúdo acústico. Entretanto, em trechos curtos (2 segundos), não há uma padronização do movimento, sendo que qualquer variação pequena na forma de falar representa uma diferença grande na variabilidade do autovetor. Isto acontece mesmo para o primeiro autovetor, porém com uma intensidade menor. Outro fato observado é que, em algumas vezes, os autovetores 3 e 4 apresentam valores menores na variabilidade dos dados para trechos curtos em relação a trechos de média duração (5 segundos). Isto se deve ao fato de os autovetores exibirem variações ao longo do tempo causadas pelo conteúdo acústico. Assim, em trechos curtos, as informações acústicas são preponderantes apresentando maiores semelhanças nestes vetores. Com base nestes resultados conclui-se que em sentenças longas os autovetores não indicam informações acústicas, e sim informações comportamentais do movimento facial do locutor. O quanto a informação

visual obtida na produção da fala varia de locutor para locutor é o que propicia a identificação de locutores pelo movimento facial. Deve-se ressaltar que uma parcela da informação comportamental é semelhante para cada locutor devido à anatomia humana, mas diferenças na forma do crânio, na textura da pele, no tamanho, forma e grau de ativação dos músculos da face entre locutores resultam em movimentos correlacionados diferentemente.

No processo de identificação do indivíduo por meio do movimento, a variabilidade de cada autovetor em relação aos locutores indica uma maior dependência do autovetor ao locutor e não ao conteúdo falado. Os experimentos realizados com sentenças de diferentes tamanhos demonstram que o primeiro autovetor apresenta menor dependência ao conteúdo acústico. Neste trabalho, este conteúdo é usado como informação que caracteriza os indivíduos.

Em um sistema de reconhecimento de locutor, utiliza-se o primeiro autovetor de diversos locutores como entrada para uma rede MLP, obtendo-se um resultado de aproximadamente $86,7\%$ de acerto na tarefa de identificação de locutor. Isto é uma evidência de que os autovetores do movimento facial são inerentes à anatomia humana, porém também apresentam características que são um facilitador no reconhecimento de pessoas. Neste contexto, a caracterização do movimento do indivíduo somada a características morfológicas pode melhorar sistemas de identificação de pessoas.

O estudo até o momento abre caminhos para a realização de novas tarefas, que devem dar continuidade ao trabalho. Como proposta de atividades visando a alcançar uma maior integração do sistema como um todo, bem como a melhoria da qualidade final dos resultados do trabalho, são sugeridas as seguintes propostas de continuidade:

- estudar uma maneira para determinar a afiliação dos parâmetros LSP à cavidade oral durante a fala, de forma a encontrar um modelo mais realístico do acoplamento entre a acústica da fala e o movimento da face.

- comparar a influência do movimento no reconhecimento para um sistema bimodal que conjugue características estáticas e dinâmicas no reconhecimento de locutores por meio de visão computacional.

Referências Bibliográficas

Jake K. Aggarwal e Qin Cai. Human motion analysis: A review. *Computer Vision and Image Understanding: CVIU*, 73(3):428–440, 1999.

Lian Apostol, Pascal Perrier, Monica Baciu, Christoph Segebarth, e Pierre Badin. Using the formant cavity affiliation to study the inter-speaker variability: assessment from mri data. *Proceed. 5 th Speech Production Seminar*, pages 213–216, May 2000.

Ognjen Arandjelovi e Roberto Cipolla. Face recognition from face motion manifolds using robust kernel resistor-average distance. *Conference on Computer Vision and Pattern Recognition Workshop*, 5, 2004.

B. S. Atal e Suzanne L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, April 1971.

Adriano Vilela Barbosa. Codificação audiovisual integrada da fala. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil, 2000.

Adriano Vilela Barbosa. *A study on the relations between audible and visible speech.* PhD thesis, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil, 2004.

Adriano Vilela Barbosa e Hani Camille Yehia. Measuring the relation between speech acoustics and 2d facial motion. *26th International Conference on Acoustics, Speech, and Signal Processing - ICASSP'2001*, 1:181–184, 2001.

Juliano J. Bazzo e Marcus V. Lamar. Recognizing facial actions using gabor wavelets with neutral face average difference. *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 505–510, May 2004.

Chiraz BenAbdelkader e Ross Cutler. Motion-based recognition of people in eigengait space. *IEEE International Conference on Automatic Face and Gesture Recognition*, 02:1–4, 2002.

Michael J. Black e Yaser Yacoob. Recognizing facial expressions under rigid and non-rigid facial motions using local parametric models of image motion. *International Journal of Computer Vision*, 25(1):23–48, July 1997.

Volker Blanz, Curzio Basso, Tomaso Poggio, e Thomas Vetter. Reanimating faces in images and video. *Proceedings of EUROGRAPHICS 2003*, 2003.

Antônio de Pádua Braga, André P. L. F. Carvalho, e Teresa Bernarda Ludermir. *Redes Neurais Artificiais Teoria e Aplicações*. LTC - Livros Técnicos e Científicos Editora S.A., 2000.

Christoph Bregler e Yochai Konig. Eigenlipsfor robust speech recognition. *Acoustics, Speech, and Signal Processing (ICASSP-94)IEEE International Conference on*, 2:669–672, Apr 1994.

Roberto Brunelli e Tomaso Poggio. Face recognition: features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1042–1052, October 1991.

Claudette Cédras e Mubarak Shah. Motion based recognition: A survey. *IEEE Proceedings Image and Vision Computing*, 13(2):129–155, March 1995.

Li-Fen Chen, Hong-Yuan Mark Liao, e Ja-Chen Lin. Person identification using facial motion. *IEEE - International Conference on Image Processing*, 02:677–680, 2001.

Claude C. Chibelushi, John S. D. Mason, e Farzin Deravi. Feature-level data fusion for bimodal person recognition. *Image Processing and Its Applications, Sixth International Conference on*, 1:399–403, Jul 1997.

Claude C. Chibelushi, Farzin Deravi, e John S. D. Mason. A review of speech-based bimodal recognition. *Multimedia, IEEE Transactions on*, 4(1):23–37, 2002.

Greg I. Chiou e Jenq-Neng Hwang. Lipreading from color video. *IEEE Trans. on Image Processing*, 6(8):1192–1195, Aug 1997.

Michael M. Cohem e Dominic W. Massaro. Synthesis of visible speech. *Behavior Research Methods: Instruments & Computers*, pages 260–263, 1990.

John R. Cowell e Aladdin Ayesh. Extracting subtle facial expression for emotional analysis. *IEEE International Conference on Systems*, 2004.

Ingemar J. Cox, Joumana Ghosn, e Peter N. Yianilos. Feature-based recognition using mixture-distance. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'96)*, pages 209–216, June 1996.

Juliana Paulino da Silva. Parametrização facial para codificação audiovisual integrada da fala. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil, 2001.

Bruce A. Draper, Kyungim Baek, e Marian Stewart Bartlett. Recognizing faces with pca and ica. *Computer Vision and Image Understanding*, 91:115–137, 2003.

Meng Joo Er, Shiqian Wu, Juwei Lu, e Hock Lye Toh. Face recognition with radial basis function (rbf) neural networks. *IEEE Trans. Neural Networks*, 13:697–710, May 2002.

Mohamad Hani Ahmad Fadzil e Abu Bakar H. Human face recognition using neural networks. *IEEE International Conference Image Processing*, 3:936–939, Nov. 1994.

James L. Flanagan. *Speech analysis, synthesis, and perception.* Springer, 1972.

Sadaoki Furui e M. Mohan Sondhi. *Advances in Speech Signal Processing.* Marcel Dekker, New York-Besel-Hong Kong, 1991. ISBN 0824785401.

Dariu M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999. URL `citeseer.ist.psu.edu/gavrila99visual.html`.

Alan J. Goldschen, Oscar N. Garcia, e Eric Petajan. Continuos optical automatic speech recognition by lipreading. *28th Annual Asilomar Conference on Signals, Systems, and Computers*, 1994.

Hans Peter Graf, Eric Cosatto, Volker Strom, e Fu Jie Huang. Visual prosody: Facial movements accompanying speech. *IEEE Computer Society*, 00:0396, 2002.

Simon Haykin. *Redes Neurais Príncipios e Prática.* Bookman, 2 edition, 2001.

Berthold K.P. Horn e Brian G. Rhunck. Determining optical flow. *Artificial Intelligence*, 17: 185–203, 1981.

Roger A. Horn e Charles R. Johnson. *Matrix Analysis.* Cambridge Univ. Press, New York-NY USA, 1985.

Ping S. Huang, Chris J. Harris, e Mark S. Nixon. A statistical approach for recognizing humans by gait using spatial-temporal templates. *IEEE*, 01:178–182, 1998.

Ping S. Huang, Chris J. Harris, e Mark S. Nixon. Human gait recognition in canonical space using temporal templates. *IEEE*, 146-2, April 1999.

Xuedong Huang, Alex Acero, e Hsiao-Wuen Hon. *Spoken Language Processing.* Prentice Hall PTR, 2001.

Koji Iwano, Satoshi Tamura, e Sadaoki Furui. Bimodal speech recognition using lip movement measured by optical-flow analysis. *International Workshop on Hands-Free Speech Communication (HSC 2001)*, pages 187–190, 2001.

Jintao Jiang, Abeer Alwan, Lynne Bernstein, Patricia Keating, e Ed Auer. On the correlation between facial movements, tongue movements and speech acoustics, 2000.

Jintao Jiang, Abeer Alwan, Patricia A. Keating, Edward T. Auer, e Lynne E. Bernstein. On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP JOURNAL ON APPLIED SIGNAL PROCESSING*, 11, 2002.

Jean-Claude Junqua. The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 1993.

Ramanujan Kashi, Jianying Hu, Winston Nelson, e William Turin. A hidden markov model aproach to online handwritten signature verification. *International Journal on Document Analysis and Recognition*, 1:102–109, June 1998.

Barbara Knight e Alan Johnston. The role of movement in face recognition. *Visual Cognition*, pages 265–273, 1997.

Christian Kroos, Takaaki Kuratate, e Eric Vatikiotis-Bateson. Video-based face motion measurement. *Journal of Phonetics*, 30:569–590, 2002.

Munhall Jeffery Callan Kuratate, Jeff Jones, Daniel Callan, Takaaki Kuratate, e Eric Vatikiotis-Bateson. Visual prosody and speech intelligibility. head movement improves auditory speech perception. *Psychological Science*, 15(2):133–137, 2004. doi: 10.1111/j.0963-7214.2004.01502010.x. URL `http://www.blackwell-synergy.com/doi/abs/10.1111/j.0963-7214.2004.01502010.x`.

Takaaki Kuratate e Eric Vatikiotis-Bateson. Estimating 3d face expression postures for animation from photographs using a 3d face database. In *Symposium on Computer Animation (SCA2004)*, pages 22–23. Poster & demo session, 2004.

Takaaki Kuratate, Kevin G. Munhall, Philip E. Rubin, Eric Vatikiotis-Bateson, e Hani Camille Yehia. Audio-visual synthesis of talking faces from speech production correlates. *Proc. 6th European Conference on Speech Communication and Technology (EuroSpeech'99)*, 3:1279–1282, September 1999.

Marcus V. Lamar, Md. Shoaib Bhuiyan, e Akira Iwata. Hand alphabet recognition using morphological pca and neural networks. *Proc. of International Joint Conference on Neural Networks*, 4:2839–2844, July 1999a.

Marcus V. Lamar, Md. Shoaib Bhuiyan, e Akira Iwata. Hand gesture recognition using morphological principal component analysis and an improved combnet-ii. *Proc. of IEEE International Conference on System*, IV:57–62, Octobe 1999b.

Juho Lee e Hyun Seung Yang. A model based estimation method of rigid and non-rigid face. *Institute of Electronics, Information and Communication Engineers - IEICE TRANS. INF. & SYST.*, 89(1):20–27, January 2006.

Jorge C. Lucero e Kevin G. Munhallb. A model of facial biomechanics for speech production. *Acoustical Society of America*, 5(106):2834–2842, November 1999.

Lain Manhews, J. Andrew Bangham, e Stephen Cox. Audiovisual speech recognition using multiscale nonlinear imagedecomposition. *Spoken Language (ICSLP 96), Fourth International Conference on*, 1:38–41, Oct 1996.

Glenn A. Martin. Lipreading by optical flow correlation. *Proceedings of the National Conference on Undergraduate Research*, 1992.

Uwe Meiel, Wolfgang Hurst, e Paul Duchnowski. Adaptive bimodal sensor fusion for automatic speechreading. *Acoustics, Speech, and Signal Processing (ICASSP-96), IEEE International Conference on*, 2:833–836, May 1996.

Lucie Menard, Jean-Luc Schwartz, Louis-Jean Boe, e Jérôme Aubin. Articulatory acoustic relationships during vocal tract growth for french vowels: Analysis of real data and simulations with an articulatory model. *Journal of Phonetics*, 35:1–19, 2007.

Kétia Soares Moreira e Hani Camille Yehia. Analysis of the variability of the coupling between facial motion and speech acoustics. *7th International Seminar on Speech Production - ISSP*, pages 109–116, 2006.

Mark S. Nixon, John N. Carter, J. M. Nash, P. S. Huang, D. Cunado, e S. V. Stevenage. Automatic gait recognition. *BIOMETRICS-Personal Identification in Networked Society*, 1999.

Magnus Nordstrand, Gunilla Svanfeldt, Björn Granström, e David House. Measurements of articulatory variation in expressive speech for a set of swedish vowels. *Speech Communication*, 44:187–196, September 2004.

Alice J. O'Toole, Dana A. Roark, e Hervé Abdi. Recognizing moving faces: A psychological and neural synthesis. *Journal of Vision*, 2(7), March 2002. The University of Texas at Dallas.

Majla Pantic e Leon JM Rothkrantz. An expert system for multiple emotional classification of facial expressions. *IEEE Computer Society - 11th IEEE International Conference on*, pages 113–120, 1999.

Guy Perelmuter, Enrique Vinicio Carrera E., Marley Vellasco, e Marco Aurélio Pacheco. Reconhecimento de imagens bidimensionais utilizando redes neurais artificiais. *VIII SIBGRAPI*, pages 197–203, outubro 1995.

Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, e Iain Matthews. Audio-visual automatic speech recognition: An overview. 2004. Issues in Visual and Audio-Visual Speech Processing.

Lawrence Rabiner e Biing-Hwang Juang. *Fundamentals of Speech Recognition*. PTR Prentice Hall, 1993a.

Lawrence Rabiner e Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall PTR, 1993b. ISBN 0130151572.

Lawrence Rabiner e Ronald W. Shafer. *Digital Processing of Speech Signals*. Prentice-Hall Series in Signal Processing, New Jersey, 1978.

Krishnan Rangarajan, William Allen, e Mubarak Shah. Recognition using motion and shape. *11'th Intern. Conf. on Pattern Recognition*, 1, Aug.30-Sept.3 1992.

Ioannis M. Rekleitis. Optical flow recognition from the power spectrum of a single blurred image. *IEEE Signal Processing Society*, Sept. 1996.

Abdenour Sehad, Abdenour Hadid, H. Hocini, M. Djeddi, e S. Ameur. Face recognition using neural networks and eigenfaces. *Vision Interface Conference*, 2000.

Jamie D. Shutler, Mark S. Nixon, e Chris J. Harris. Statistical gait description via temporal moments. *Proc. SSIAI 2000*, pages 291–295, April 2000.

Daniel Silverman. *A critical introduction to phonology: of sound, mind, and body*. Continuum Intl Pub Group, continuum intl pub group edition, 2006. ISBN 0826486614.

Noboru Sugamura e Fumitada Itakura. Speech analysis and synthesis methods developed at ecl in ntt-from lpc to lsp-. *Speech Commun.*, 5(2):199–215, 1986. ISSN 0167-6393. doi: http://dx.doi.org/10.1016/0167-6393(86)90008-7.

Eric Vatikiotis-Bateson e Hani Camille Yehia. Estimation and generalization of multimodal speech production. *IIEEE Signal Processing Society Workshop*, 1:23–32, December 2000.

Eric Vatikiotis-Bateson e Hani Camille Yehia. Unified physiological model of audible-visible speech production. *V EUROSPEECH*, pages 22–25, 1997.

Eric Vatikiotis-Bateson, Kevin G. Munhall, Makoto Hirayama, Yuenchang Lee, e Demetri Terzopoulos. The dynamics of audiovisual behavior in speech. *Speechreading by humans and machines (NATO-ASI Series F)*, 150:221–232, 1996a.

Eric Vatikiotis-Bateson, Kevin G. Munhall, Y. Kasahara, F. Garcia, e Hani Camille Yehia. Characterizing audiovisual information during speech. *Spoken Language - ICSLP 96*, 3: 1485–1488, 1996b.

Eric Vatikiotis-Bateson, Inge-Marie Eigsti, Sumio Yano, e Kevin G. Munhall. Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, 60(6): 926–940, 1998a.

Eric Vatikiotis-Bateson, Takaaki Kuratate, e Hani Camille Yehia. Kinematics-based synthesis of realistic talking faces. *International Conference on Auditory-Visual Speech Processing (AVSP'98)*, pages 185–190, 1998b.

Eric Vatikiotis-Bateson, Hani Yehia, e Takaaki Kuratate. Speaking mode variability im multimodal speech production. *IEEE Transaction on Neural Networks*, 13(4):894–899, July 2002.

Eric Vatikiotis-Bateson, Victor Chung, Kevin Lutz, Nicole Mirante, Jolien Otten, e Johanna Tan. Auditory, but perhaps not visual, processing of lombard speech. *The Journal of the Acoustical Society of America*, 2006.

Eric Vatikiotis-Bateson, Adriano Vilela Barbosa, Cheuk Yi Chow, Martin Oberg, Johanna Tan, e Hani Camille Yehia. Audiovisual lombard speech: Reconciling production and perception. *International Conference on Auditory-Visual Speech Processing 2007*, 2007.

Luis Fernando Veríssimo. *O popular: crônicas, ou coisa parecida*. L&PM, 1984.

Yunhong Wang, Tieniu Tan, e Yong Zhu. Face verification based on singular value decomposition and radial basis function neural network. *Chinese Academy of Sciences.*, 1991.

Satosi Watanabe. *Pattern recognition: human and mechanical*. John Wiley & Sons, Inc., New York, USA, 1985. ISBN 0-471-80815-6.

Chew Yean Yam, Mark S. Nixon, e John N. Carter. Performance analysis on new biometric gait motion model. *Fifth IEEE Southwest Symposium on Image Analysis and Interpretation*, 02, 1-4 2002.

Junji Yamato, Jun Ohya, e Kenichiro Ishii. Recognizing human action in time-sequential images using hidden markov model. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–385, 1992.

Hani Camille Yehia, Philip Rubin, e Eric Vatikiotis-Bateson. Quantitative association of orofacial and vocal-tract shapes. *Audio-Visual Speech Processing*, pages 41–44, 1997.

Hani Camille Yehia, Philip Rubin, e Eric Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26:23–43, October 1998.

Hani Camille Yehia, Takaaki Kuratate, e Eric Vatikiotis-Bateson. Using speech acoustics to drive facial motion. volume 1, pages 631–634. 14th International Congress of Phonetic Sciences - ICPHS'99, August 1999.

Hani Camille Yehia, Takaaki Kuratate, e Eric Vatikiotis-Bateson. Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, 30:555–568, 2002.

Willard R. Zemlin. *Princípios de anatomia e fisiologia em fonoaudiologia.* Artes Médicas Sul, 4ª edition, 2000. ISBN 857307700x.