UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

CARLOS BRUNO OLIVEIRA LOPES

Detecção Visual de Atividade de Voz com Base na Movimentação Labial

Dissertação apresentada como requisito parcial
para a obtenção do grau de
Mestre em Ciência da Computação

Prof. Dr. Jacob Scharcanski
Orientador

Porto Alegre, maio de 2013

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Lopes, Carlos Bruno Oliveira

Detecção Visual de Atividade de Voz com Base na Movimentação Labial / Carlos Bruno Oliveira Lopes. – Porto Alegre: PPGC da UFRGS, 2013.

68 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2013. Orientador: Jacob Scharcanski.

1. Método de Bayes, Segmentação de pele, Segmentação de lábios, Operadores Morfológicos, Cadeia de Markov Ocultas. I. Scharcanski, Jacob. II. Título.

“Se um dia tiver que escolher entre o mundo e o amor...
Lembre-se! Se escolher o mundo ficará sem o amor,
mas se escolher o amor com ele você conquistará o mundo.”
— ALBERT EINSTEIN

AGRADECIMENTOS

"Algumas pessoas marcam a nossa vida para sempre,
umas porque vão nos ajudando na construção do crescer,
outras porque nos apresentam projetos de sonhos
e outras ainda porque nos desafiam a construí-los".

Agradeço a conclusão deste trabalho às pessoas,
que sem o saberem, muito para ele contribuíram.

Agradeço,

A Deus, fonte de inspiração e protetor contra todos os males.

Aos meus Pais, que sempre me apoiaram, me deram forças para seguir em frente, e são meus exemplos de vida e orgulho.

A minha irmã, que sempre demonstra carinho, afeição e alegria para comigo fazendo-me lembrar que mesmo estando longe eu tenho uma família que não se esquece de mim e que me ama nas suas famosas frases: "ei mano, seu filho desnaturado, liga de vez em quando para dar sinal de vida, se a gente não ligar tu não liga."

A toda minha família pelo carinho, incentivo e compreensão.

A todos os meus Amigos, que foram e são companheiros de trabalhos, cúmplices de "presepadas", incentivadores, conselheiros, ou seja, AMIGOS.

E aos Professores, que contribuíram para minha formação acadêmica.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

SVM	*Support Vector Machine*
SMQT	*Successive Mean Quantization Transformation*
FCM	*Fuzzy C-Means*
ROI	*Region Of Interesting*
EM	*Expectation Maximisation*
P.D.F	Função de Densidade Probabilística (*Probability Density Function*)
VAD	*Voice Activity Detection*
CIELab	*Commission Internationale De l'Eclairage Lab*
XOR	*Exclusive Or*
HMM	*Hidden Markov Model*
ASR	*Automatic Speech Recognition*
VVAD	*Visual Voice Activity Detection*
HSV	(*Hue, Saturation, Value*)
RGB	(*Red, Green, Blue*)
AAM	*Active Appearance Models*
GMM	Modelo de Mistura de Gaussianas (*Gaussian Mixture Model*)
EBGM	*Elastic Bunch Graph Matching*
PCA	Análise das Principais Compentes *Principal Component Analysis*
AVSR	Reconhecimento Áudio-Visual da Fala (*Audio-Visual Speech Recognition*)

LISTA DE FIGURAS

RESUMO

O movimento dos lábios é um recurso visual relevante para a detecção da atividade de voz do locutor e para o reconhecimento da fala. Quando os lábios estão se movendo eles transmitem a idéia de ocorrências de diálogos (conversas ou períodos de fala) para o observador, enquanto que os períodos de silêncio podem ser representados pela ausência de movimentações dos lábios (boca fechada). Baseado nesta idéia, este trabalho foca esforços para detectar a movimentação de lábios e usá-la para realizar a detecção de atividade de voz. Primeiramente, é realizada a detecção de pele e a detecção de face para reduzir a área de extração dos lábios, sendo que as regiões mais prováveis de serem lábios são computadas usando a abordagem Bayesiana dentro da área delimitada. Então, a pré-segmentação dos lábios é obtida pela limiarização da região das probabilidades calculadas. A seguir, é localizada a região da boca pelo resultado obtido na pré-segmentação dos lábios, ou seja, alguns *pixels* que não são de lábios e foram detectados são eliminados, e em seguida são aplicados algumas operações morfológicas para incluir alguns *pixels* labiais e não labiais em torno da boca. Então, uma nova segmentação de lábios é realizada sobre a região da boca depois de aplicada uma transformação de cores para realçar a região a ser segmentada. Após a segmentação, é aplicado o fechamento das lacunas internas dos lábios segmentados. Finalmente, o movimento temporal dos lábios é explorado usando o modelo das cadeias ocultas de Markov (HMMs) para detectar as prováveis ocorrências de atividades de fala dentro de uma janela temporal.

Palavras-chave: Método de Bayes, Segmentação de pele, Segmentação de lábios, Operadores Morfológicos, Cadeia de Markov Ocultas.

Visual voice activity detection using as information the lips motion

ABSTRACT

Lips motion are relevant visual feature for detecting the voice active of speaker and speech recognition. When the lips are moving, they carries an idea of occurrence of dialogues (talk) or periods of speeches to the watcher, whereas the periods of silences may be represented by the absence of lips motion (mouth closed). Based on this idea, this work focus efforts to obtain the lips motion as features and to perform visual voice activity detection. First, the algorithm performs skin segmentation and face detection to reduce the search area for lip extraction, and the most likely lip regions are computed using a Bayesian approach within the delimited area. Then, the pre-segmentation of the lips is obtained by thresholding the calculated probability region. After, it is localized the mouth region by resulted obtained in pre-segmentation of the lips, i.e., some non-lips *pixels* detected are eliminated, and it are applied a simple morphological operators to include some lips *pixels* and non-lips around the mouth. Thus, a new segmentation of lips is performed over mouth region after transformation of color to enhance the region to be segmented. And, is applied the closing of gaps internal of lips segmented. Finally, the temporal motion of the lips is explored using Hidden Markov Models (HMMs) to detect the likely occurrence of active speech within a temporal window.

Keywords: Bayesian method, skin segmentation, lip segmentation, morphological operators, Hidden Markov Model.

1 INTRODUÇÃO

A fala é um sinal bimodal, tanto acústico quanto visual (SODOYER et al., 2009). Entretanto, muitos sistemas automáticos de reconhecimento de fala (ARS) usam somente a informação acústica. Tais sistemas geralmente são muitos sensivéis a questões de canais e ambientes (ruídos), o que tem motivado pesquisas em técnicas de pré-processamento de áudio e algoritmos de adaptações de ruídos (ROHANI et al., 2008). De fato, estudos tem demonstrado que a informação visual melhora a inteligibilidade da fala na presença de ruídos quando comutado somente da condição de áudio para áudio e visual. Mais especificamente, a informação visual ajuda na extração das feições acústicas, ou seja, "ver para ouvir melhor" (SODOYER et al., 2009). Por exemplo, as expressões faciais como espanto, surpresa, alegria, dentre outras que demonstram o sentimento do indivíduo na conversa; a movimentação corporal ao gesticular e tentar descrever o quê se está falando através do corpo; a movimentação dos lábios e língua que indicam ou passam a idéia de que a pessoa está falando alguma coisa. Dentre essa informações visuais, o movimento labial é uma das melhores características visuais para indicar o reconhecimento de quando uma pessoa está falando ou está em silêncio. Isto é justificável, uma vez que os lábios consigam se movimentar mais que 80% das vezes em que locutores humanos estão em atividade de discurso, ou seja, que eles estejam movimentando os seus lábios cerca de 80% das vezes ao falar (WANG; WANG; XU, 2010). Dentro desse contexto, o trabalho apresentado focaliza-se no esforço da detecção de atividade de voz (VAD) pela informação visual. Portanto, parte-se da idéia de obtenção das feições visuais da movimentação dos lábios para a sua análise e identificação de ocorrência de fala ou silêncio ao longo do tempo.

Iniciamos nosso trabalho minimizando a ação da variação de iluminação do ambiente sobre as cores da imagem que serão utilizadas como informações para localização dos *pixels* de lábios em nossa abordagem. Dessa forma, buscamos tornar as cores da imagem indiferentes à mudança do iluminante pela execução de sua correção através do vetor de ajuste que é obtido usando as informações de uma imagem tomada como referência e do quadro de vídeo de entrada, ou seja, realizamos uma correção das intensidades de cores (iluminação do ambiente) (HORDLEY et al., 2005) sobre o i-ésimo quadro de vídeo antes de aplicarmos a abordagem de detecção de lábios.

Concluída a fase de pré-processamento, concentramos-nos na detecção dos lábios visando o foco principal que é a obtenção das feições de movimentação labial para detecção de atividade de voz. Portanto, a abordagem proposta inicia-se pela restrição da área de busca dos *pixels* de lábios localizando-se primeiramente as regiões de intersecção dos *pixels* de face e dos *pixels* de pele após a aplicação de operações morfológicas. Sobre a região de intersecção é executada a detecção dos lábios que mensuram as regiões de alta

probabilidade de os *pixels* serem de lábios através de uma abordagem Bayesiana (WEBB, 2002) que utiliza as informação de cromaticidades como discriminantes. Então, os *pixels* de lábios são pré-segmentados pela limiarização das regiões mais prováveis de serem de lábios utilizando o método de Otsu (OTSU, 1979) (LIU; YU, 2009). Com o resultado obtido é realizada a localização da região onde se encontra a boca. Por fim, é realizada uma nova segmentação dos lábios sobre essa região seguida de um fechamento de lacunas entre as bordas superiores, inferiores e laterais dos lábios obtendo-se como resultado a inclusão de todos os *pixels* de lábios entre esses extremos e a inclusão dos *pixels* da região interna da boca (entre o lábio superior e inferior) quando o locutor estiver discursando (isto é, quando a boca estiver aberta).

Uma vez extraídas as feições de movimentação dos lábios, o problema da detecção visual de atividade de voz pode ser abordado. Logo, é explorada a evolução temporal das feições dos lábios usando o HMM como resolução para detectar as prováveis ocorrências de atividade de voz ou silêncio. Dessa forma, as taxas de aspectos dos lábios (área) dentro de uma janela temporal são usadas como medidas (ou informações) de abertura da boca que alimentam as cadeias ocultas de Markov (HMM) possibilitando a detecção visual da atividade de voz.

1.1 Objetivo Geral

O trabalho desenvolvido tem por objetivo realizar a detecção dos lábios e extraí-los como variáveis de observação que alimentarão dois HMMs que exploram a consistência temporal da movimentação dos lábios para realizar a detecção dos períodos de silêncio ou fala resultando na detecção da atividade de voz (VAD) pela informação visual.

1.2 Objetivos Específicos

1. Executar o método de correção de intensidade de cores no pré-processamento da imagem para minimizar a influência da iluminação do ambiente;
2. Realizar a detecção e segmentação dos lábios para obtenção das medidas observadas (variáveis observáveis);
3. Normalizar as variáveis de observação para poderem ser utilizadas pelo HMM;
4. Realizar a detecção de atividade de voz através dos HMMs;
5. Realizar a detecção de atividade de voz para múltiplos usuários identificando-os através do detector de faces;
6. Validar e analisar o método de detecção visual de atividade voz (VVAD) proposto;

1.3 Organização do Texto

O texto esta organizado em 6 capítulos. O capítulo 2 apresenta a revisão bibliográfica através do estado da arte e dos fundamentos téoricos (matemáticos) nos quais formaram o alicerce para o desenvolvimento do método de obtenção dos lábios e da realização da detecção de atividade de voz. O capítulo 3 expõe o método desenvolvido de detecção visual de atividade de voz utilizando como informação a movimentação labial. O capítulo

4 demonstra os experimentos e seus resultados. O capítulo 5 aduz a acerca das conclusões e dos trabalhos futuros a serem desenvolvidos. O capítulo 6 enuncia as contribuições e publicações aceitas e submetidas.

2 REVISÃO BIBLIOGRÁFICA

2.1 Estado da Arte

A detecção de atividade de voz é um problema abordado em muitas aplicações, tais como, vídeo-conferência para identificação de períodos de fala ou silêncio, melhoramento do som e foco no locutor ativo; sistemas de reconhecimento de fala para identificações e seleções dos quadros de áudio e vídeo a serem processados; e sistema de interação humano-computador (homem e máquina) que envolvam comando de voz e sua identificação (sistemas biométricos). Para sua realização, muitas pesquisas e técnicas ou soluções existentes geralmente são baseadas ou exploram apenas a informação sonora, ou seja, o áudio que é processado através de algoritmos que realizam a detecção de atividade de voz e tentam eliminar ruídos de ambiente e canal (ROHANI et al., 2008). Outras alternativas de abordagem de detecção ou reconhecimento de fala, exploram a informação visual através das expressões faciais ou corporais, e principalmente a movimentação de lábios (WANG; WANG; XU, 2010). Elas utilizam essas informações como feições que indicam a possível ocorrência de atividade de voz no período de tempo analisado.

A movimentação dos lábios destaca-se como feição a ser utilizada para a detecção visual de atividade de voz (VVAD) por apresentar um forte indicativo de ocorrência ou reconhecimento de fala, pois ao falar, uma pessoa move os seus lábios inúmeras vezes (WANG; WANG; XU, 2010). Ela é uma das feições visuais mais utilizada para indicar a presença de atividade de voz. Para sua exploração ou uso como informação visual para detecção de atividade de voz é necessário primeiramente conseguir extraí-la do resto da imagem por algum um metódo de detecção de lábios.

O estudo da detecção e obtenção dos lábios é uma área ativa de pesquisa em processamento de imagens, visão computacional e reconhecimentos de padrões. Elas têm suas aplicabilidades em detecção de faces, pois nos induz a conjectura da presença de faces no meio analisado; no reconhecimento do contexto em uma conversa através da leitura labial; na detecção de atividade de voz pela informação de movimentação labial; e na identificação do locutor ativo.

No trabalho realizado por Yao e Gao (YAO; GAO, 2001) é apresentada uma abordagem de detecção de lábios com intuito de detectar faces. Para isso, eles utilizaram as informações de lábios e pele tendo como objetivo investigar a relação entre as cromaticidades e seus componentes de cor focando-se no uso destas informações para detectar e localizar as regiões de interesse (ROI), no caso em específico a face. Em seu método estabelece-se uma transformação que melhora a discriminância entre as cromaticidades da pele e dos lábios sendo essas informações utilizadas para detectar a presença de faces em imagens. A abordagem inicia-se pela busca das regiões de pele através da aplicação das

transformações de crominâncias de pele seguida de uma limiarização. Por conseguinte, os mesmos processos são realizados para detecção dos lábios. Então, a face é detectada pela posição relativa da região de localização dos lábios e da pele seguindo um conjunto de restrições estabelecidas no trabalho de Yao e Gao (YAO; GAO, 2001).

Na pesquisa desenvolvida por Eveno, Caplier e Coulon (EVENO; CAPLIER; COULON, 2001) é proposta um nova transformação, chamada de mapa de curvas das cromaticidades, para aumentar a discriminância entre os lábios e a pele objetivando a segmentação dos lábios. O processo para segmentação inicia-se pela redução de dependência da luminância sobre os componentes de cores do espaço RGB computando-se novos componentes ($R_{cor}, G_{cor}, B_{cor}$). Então, é calculado o mapa de curvas das cromaticidades usando a informação dos novos componentes de cores. O mapa é dado através de três pontos $P_{1(x,y)}$, $P_{2(x,y)}$ e $P_{3(x,y)}$ que são associados aos *pixels* (x, y) da imagem e definem uma curva da parabóla. A segmentação dos lábios é dada através da computação dessa curva da parabóla pela seguinte regra: se o *pixel* avaliado obtiver um alto valor de curva ele é classificado como *pixel* de lábios e se obtiver um baixo valor é classificado como *pixel* de pele.

Dargham e Chekina (DARGHAM; CHEKIMA, 2006) propuseram em seu trabalho uma nova normalização dos componentes RGB para a realização da detecção de lábios. O método inicia-se pela aplicação da técnica proposta sobre a imagem em RGB, sendo chamada pelos autores de normalização das intensidades máximas que é dada através das seguintes equações:

$$\begin{aligned} r(x,y) &= \frac{R(x,y)}{Max(R+G+B)}, \\ g(x,y) &= \frac{G(x,y)}{Max(R+G+B)}, \\ b(x,y) &= \frac{B(x,y)}{Max(R+G+B)}. \end{aligned} \tag{2.1}$$

Em seguida, são gerados os histogramas usando as cromaticidades normalizadas r, g e b, ou suas combinações (i.e., $r \cdot g$, $r \cdot b$, r/g, r/b,$r - g$, $r - b$, $r + g - b \cdot 6$, $r + b - g \cdot 2$ e $r + b - g \cdot 6$). Então, limiares são aplicados para segmentar a imagem em regiões de pele e regiões de lábios, ou seja, para cada componente ou combinações de cromaticidades que são usadas para segmentação um valor de limiar é aplicado variando-se do menor para o maior valor. Para cada valor de limiar aplicado é calculado o erro de sua segmentação para os lábios e para a pele. Então, o melhor limiar é escolhido quando a porcentagem de erro total é mínima ou quando porcentagem de erro da detecção dos *pixels* de pele e lábios são iguais.

Na pesquisa de Salazar-Jiménez (SALAZAR-JIMÉNEZ et al., 2006) foi desenvolvido um método de detecção das condições dos lábios para o auxílio ao tratamento da fissura labial e da fenda palatina (*cleft lip and palate - CLP*) em crianças. Em sua abordagem a imagem é processada em duas etapas. Na primeira, é realizada a localização da boca dentro da região da face detectada. A boca é localizada pelos seguintes passos: restringe a imagem da face detectada à 1/3 de sua altura, aplicam-se algumas operações morfológicas sobre área da face, realiza-se uma transformação sobre as cores da imagem, binariza-se a imagem pelo limiar (*threshold*) e selecionam-se as regiões predominantes por meio de análise das conectividades. Na segunda etapa, é realizada a detecção dos contornos dos lábios e a estimava de variáveis geométricas. Os contornos são alcançados pela detecção

de um conjunto de 8 pontos poligonais que descrevem as bordas externas do formato dos lábios. Finalmente, os pontos são conectados formando um polígono que é usado como modelo de extração das variáveis que informam o conjunto de descritores para o reconhecimento das condições do lábios, ou seja, para a análise dos lábios com o objetivo de detectar uma provável fissura labial.

Os estudos de Rohani (ROHANI et al., 2008) focaram-se no desenvolvimento de uma abordagem de extração dos lábios baseado em "*fuzzy clustering*"como forma de contribuir para o melhoramento da acurácia na detecção da região dos lábios para aplicações de reconhecimento de fala que utilizem a informação visual. Neste método, inicialmente é delimitada a região de busca dos lábios executando-se o detector de faces. Esse detector utiliza como discriminante as variáveis locais SMQT (*Successive Mean Quantization Transformation*) e o classificador "*Split Up Snow*". Partindo da região que corresponde a face detectada, seleciona-se nesta face uma região menor onde localizam-se os lábios. E, aplica-se sobre a mesma região uma transformação nos dados chamado de falsa matiz (*pseudo-hue*). Então, executa-se o agrupamento dos dados em lábios e não-lábios sobre a região inferior da face pelo método de *Fuzzy C-Means* (FCM) que utiliza as informações de falsa-matiz e o canal 'b' do espaço de cores CIELab como dados discriminantes. Ao término do processo de agrupamento, é realizado um pós-processamento para eliminação de falsos-positivos através das operações morfológicas e a execução da melhor elipse correspondente sobre a região classificada. E por fim, é aplicado o filtro Gaussiano para suavizar a região dos lábios detectados.

Em seu trabalho, Wang (WANG; WANG; XU, 2010) procurou solucionar parte do problema de leitura labial, focando-se na detecção, localização e rastreamento dos lábios para reconhecimento de fala. Em sua abordagem, é proposta a combinação das variáveis de *Haar-Like* com a estimativa de variância local para construir um conjunto de dados que serão utilizados como discriminante no classificador de *Support Vector Machine* (SVM) para detecção dos lábios. O método inicia-se pela detecção de faces através do classificador de SVM utilizando os dados gerados pela combinação das variáveis de *Haar-Like* com a estimativa da variância local. Localizada a região da face, o classificador SVM é utilizado novamente para localizar a região da boca na parte inferior da face. Então, o rastreamento dos lábios é executado através do filtro de Kalman que estima o centro da boca nos quadros para o seu rastreamento em tempo real. Além disso, um pós-processamento é aplicado a cada quadro para melhorar a estimativa do filtro de Kalman, onde, aplica-se uma interpolação linear para preencher as falhas de segmentação dos *pixels* de lábios (i.e., *pixels* que não foram detectados) ao longo dos quadros; e o filtro da média para eliminar ruídos gerados.

Os trabalhos descritos acima apresentaram algumas formas para extração dos lábios tendo cada um focado no seu objetivo e na sua aplicabilidade específica. Trazendo para o nosso contexto, ao conseguirmos extrair as feições labiais podemos abordar o problema de detecção visual de atividade de voz (VVAD). Uma vez que a ocorrência de movimentação dos lábios através da informação visual nos infere a existência de períodos de fala enquanto que a sua ausência indica períodos de silêncio.

No trabalho de SODOYER (SODOYER et al., 2006), é proposto o uso da informação visual de fala, a movimentação dos lábios, como detecção de atividade de voz (VAD) para sinais de voz embutidos em ruídos não-estacionários. Em seu método são exploradas situações controladas da diferença temporal da relação interlabial da altura e largura, sendo estes parâmetros extraídos pelo uso do sistema de processamento da face. Ao final do processo a detecção visual de atividade de voz (VVAD) é realizada pela aplicação do

limiar para cada quadro do vídeo.

Na abordagem de AUBREY (AUBREY et al., 2007), são apresentados dois métodos para detecção visual de atividade de voz. No primeiro, usa-se os parâmetros que descrevem a aparência dos lábios, forma e textura, baseado no modelo de aparência ativa (AAM) como variáveis de amostragem (observáveis) ao longo do tempo. Então, essas variáveis amostrais são usadas pelo HMM para a detecção de atividade de voz (VAD). No segundo método é aplicado um filtro de retina para realçar os contornos dos lábios. A seguir, é realizado o cálculo da mudança de energia em cada quadro. E então, a classificação da existência ou ocorrência de atividade de voz é dada através da aplicação de um limiar sobre a informação de mundaças da energia ao longo dos quadros de vídeo.

No trabalho de Aoki (AOKI et al., 2007), procurou-se solucionar o problema de operações de comando por voz para motoristas quando as suas mãos estão ocupados dirigindo o carro no trânsito. Em seu trabalho, são prospotos uma detecção de atividade de voz integrando o resultado do processamento visual, movimento dos lábios, com o resultado do processamento acústico (som). Para a análise sonora, são usados dois modelos de misturas de gaussianas (GMM) para gerar uma taxa de log de verossimilhança ao longo do tempo onde é aplicado um limiar para identificar o que não é voz. Para análise visual, é empregada a *wavelet* de Gabor que extrai pontos característicos para construção do *Elastic Bunch Graph Matching* (EBGM) que corresponderá a um grafo genérico do rosto com suas características detectadas. Então, a taxa de aspecto dos lábios, altura sobre a largura, são utilizados para medir a abertura da boca. Por fim, as diferença de taxas de aspecto temporais são limiarizadas para produzir a detecção de atividade de voz (VAD) em vídeos sendo-o combinado com sinais de áudio para gerar o resultado final de detecção da voz do motorista.

Chin, Ang e Seng (CHIN; ANG; SENG, 2009) focaram o seu trabalho no reconhecimento áudio-visual da fala (AVSR). A informação visual é alcançada pela obtenção dos lábios obtida através da fronteira de interpolação do *cubic spline* que o agrupa as variáveis de cor no espaço YCbCr para identificar os lábios. Em seguida, o filtro de Sobel é utilizado para obtenção dos contornos dos lábios junto com o filtro de Kalman que é utilizado para o sistema de rastreamento dos lábios. Então, o reconhecimento visual da fala é dado através do método do HMM após a extração das variáveis visuais. Essas variáveis são obtidas pela análise das principais componentes (PCA). Já a informação sonora é extraída usando a *form of mel-frequency cepstral coefficients*. E seu reconhecimento também é dado através da utilização de modelos de HMM. O reconhecimento áudio-visual da fala é alcançado no final do processo pela combinação dos resultados isolados obtidos anteriormente com a informação de áudio e visual.

A pesquisa de Petsatodis e colegiado (PETSATODIS; PNEVMATIKAKIS; BOUKIS, 2009) apresentou um sistema que utiliza a informação áudio e visual para detecção de atividade de voz em aplicações com ambientes inteligentes que utilizem sensores *far-field* (FF). Seu objetivo foi obter um desempenho superior usando ambas as informações de cada componente unimodal em vez de apenas uma delas. O sistema proposto consiste em duas modalidades separadas. Uma de áudio VAD (A-VAD) e uma visual VAD (V-VAD), sendo ambas operadas por sensores áudio-visual *far-field* (FF). Em ambos os sub-sistemas são usados modelos das cadeias ocultas de Markov (HMM). Para A-VAD foi usado um par de HMMs para modelar a presença e a ausência de fala. Sobre os resultados obtidos com esses pares de HMMs são aplicados limiares adaptativos a cada quadro de vídeo que indicará se há presença de fala. Para V-VAD também foi usado dois HMMs para identificar o movimento vertical dos lábios. Um HMM é usado para modelar os movimentos

labiais que ocorrem durante a geração da fala enquanto que outro HMM é usado para identificar a ausência de movimentos labiais (i.e., ausência de fala). O reconhecimento da ocorrência de atividade de voz pela informação visual é dado pela comparação dos logs da verossimilhança produzida pelo par de HMMs, onde, a taxa com maior valor indicará se existe a ocorrência de fala ou não. Finalmente, a fusão das informações é dada pela combinação dos resultados. Para a combinação, são usados atribuições de pesos para duas modalidades. Estes pesos podem mudar o resultado de qualquer uma das modalidades dependendo da característica do sinal áudio-visul. Então, a regra final para o AV-VAD é demonstrada através da tabela verdade descrita no trabalho de pesquisa de Petsatodis e colegiado (PETSATODIS; PNEVMATIKAKIS; BOUKIS, 2009) que combina o resultado da detecção de face, A-VAD e o V-VAD.

No trabalho realizado em 2010 por Aubrey, Hicks e Chambers (AUBREY; HICKS; CHAMBERS, 2010) foi desenvolvido um método de detecção visual de atividade de voz (V-VAD) com *optical flow* visando sua aplicação na detecção de períodos de silêncio em vídeos. A abordagem utiliza o *optical flow* para estimar os vetores que descrevem a movimentação dos *pixels* entre quadros de vídeo consecutivos (i.e. estimar a movimentação dos lábios sobre a região da boca). A estimação do movimento é realizada pelo algoritmo de Magarey e Kingsbury que é baseado na aplicação da transformada complexa da *wavelet* discreta (CDWT) sobre o par de quadros de vídeo. Para cada par consecutivo de quadros de vídeo há um campo de movimento F cuja mudança dinâmica sobre o tempo é modelada usando HMMs, ou seja, os HMMs são usados para modelar a variação do vetor *optical flow* da região da boca do locutor. Cada observação O do vetor de movimento F que alimenta o HMM gera uma verossimilhança de valor P. Então, é aplicado um limiar de valor β sobre $P_f(j)$ (i.e., j-ésimo valor de verrossimilhança realçado pelo filtro da média) para classificar o quadro de vídeo como fala ou não-fala.

Os trabalhos de pesquisas acima descreveram alguns métodos desenvolvidos para detecção visual de atividade de voz. Eles demonstraram que é possível utilizar a informação visual da região da boca das pessoas para identificar a presença de atividade de voz. A comprovação deste fato é apresentado em 2009 por Sodoyer (SODOYER et al., 2009) que estudaram a relação existente entre o movimento dos lábios e a detecção de atividade de voz e concluíram que as formas dos lábios pode ser análogo a atividade de voz e de silêncio. Assim, os parâmetros dinâmicos podem fornecer separabilidade suficiente desde que uma janela temporal adequada seja utilizada no processo, ou seja, é possível realizar a detecção de atividade de voz usando a informação visual de movimentação dos lábios.

2.2 Fundamentos Teóricos

2.2.1 Visão Computacional

Os seres humanos em geral e outras espécies de animais confiam em seus sistemas visuais para planejar ou executar ações no mundo. Fótons de luz refletidos de objetos formam imagens que são detectadas e traduzidas em sinais multidimensionais. Estes fótons viajam ao longo das vias visuais graças a uma cadeia de processos químicos e elétricos no cérebro. Os sinais visuais não apenas passam de um neurônio para o outro, mas eles também sofrem vários processamentos de sinais para finalmente fornecer-nos a informação simplificada do meio ambiente visualizado por nós. Isto, possibilita-nos a tomada de decisões e o planejamento de ações acerca do ambiente à nossa volta (i.e., interação com o meio) (BIGUN, 2006).

A visão computacional é entendida como o conjunto de técnicas para adquirir, processar, analisar e compreender as complexidades das altas dimensões de dados do nosso meio ambiente para exploração científica que desenvolvem técnicas e sistemas com objetivo de reproduzir os processos realizados pelo nosso sistema visual (JAHNE; HAUSSECKER; GEISSLER, 1999).

Em visão computacional, o que tentamos fazer é descrever o mundo que vemos através dos nossos olhos em uma ou mais imagens e reconstruir suas propriedades, tais como, forma, iluminação, e distribuição de cor. Esse processo é facilmente realizados pelos humanos e os animais através dos seus sistemas de visão (cérebro e olhos), enquanto que os algoritmos de visão computacional são tão propensos a erros ou ineficientes (SZELISKI, 2010).

A boa notícia é que atualmente a visão computacional está sendo usada em uma ampla variedade de aplicações do mundo real, que incluem (SZELISKI, 2010):

- **Reconhecimento óptico de caracteres (OCR):** Leitura de código postal escrito a mão em cartas e reconhecimento automático de números de placa;

- **Inspeção por máquina:** parte das inspeções realizadas por máquinas para assegurar a qualidade usando a visão estéreo com iluminação especializada para medir as tolerâncias das asas de aviões, partes altas do corpo ou procurar por defeitos em peças de aço fundidas utilizando visão de raios X;

- **Varejo:** o reconhecimento automatizado de objetos para caixas expressos;

- **Construção de modelos 3D (fotogrametria):** construções automatizadas de modelos 3D a partir de fotografias aéreas;

- **Imagens médicas:** registro de imagem antes e durante a cirurgia ou a realização de estudos de longo prazo da morfologia do cérebro das pessoas à medida que envelhecem;

- **Segurança automatizada:** detecção de obstáculos inesperados, como pedestres nas ruas;

- **Correspondência de movimentos:** fundir imagens geradas por computador (CGI) com cenas de ação ao vivo através do rastreamento de pontos característicos no vídeo original para estimar o movimento 3D da câmera e formar o meio ambiente;

- **Captura de movimentos (mocap):** usando marcadores retro-refletivos e visualizados de múltiplas câmeras ou outras técnicas de baseadas em visão para capturar os atores da animação computadorizada;

- **Vigilância:** monitoramento de intrusos, análise de tráfego de estradas, e monitoramento de piscinas para evitar afogamentos;

- **Reconhecimento de digitais e biométricos:** para autenticação de acesso automático.

2.2.2 Reconhecimento de Padrões

Reconhecimento de padrões é a área de pesquisa que tem por objetivo a classificação de objetos (padrões) em um número de categorias ou classes através de um conjunto de propriedades ou características. O termo padrão pode ser usado para denotar as p-dimensões de dados do vetor $\mathbf{x} = (x_1, ..., x_p)^T$, cujas componentes x_i são medidas ou valores das características do objeto. Assim, as características são variáveis especificadas pelo investigador sendo importantes para a classificação. Na discriminação, assume-se que existam C grupos ou classes, denotados por $\omega_1, ..., \omega_C$, e associados a cada padrão $\mathbf{x}$ que é uma variável categórica z que indica o membro da classe ou grupo pertencente. Isto é, se $z = i$, então o padrão pertence a ω_i, $i \in \{1, ..., C\}$ (WEBB, 2002).

Exemplos de reconhecimentos de padrões são: medições de uma forma de onda acústica no problema de reconhecimento de fala; medições feitas em pacientes a fim de identificar uma doença (diagnóstico) e no sentido de prever um provável resultado (prognóstico); medidas meteorológicas para previsões de tempo; e uma imagem digitalizada para o reconhecimento de caracteres (WEBB, 2002).

Existem dois principais tipos de classificação: classificação supervisionada e classificação não-supervisionada. Na classificação supervisionada, temos um conjunto de dados amostrais (cada um consistindo de medições em um conjunto de variáveis) rotulados ao tipo de classe que são usados como exemplos para o projeto de classificador. Na classificação não-supervisionada, os dados não são rotulados e buscamos encontrar grupos de dados e características que distinguem um grupo do outro (WEBB, 2002).

2.2.2.1 Modelos Estatísticos para Reconhecimento de Padrões

Para o reconhecimento de padrão podemos usar a estatística para projetar classificadores e realizar a separação dos dados. Esses modelos estatísticos podem ser construídos a partir das informações de médias, covariâncias, processos estocásticos e distribuições probabilísticas dos padrões a serem reconhecidos a fim de inferir sobre qual classe i a variável observada pertence. Os modelos estatísticos usados neste trabalho para o reconhecimento de padrões são apresentados nas sub-secções a seguir.

2.2.2.1.1 Classificador de Bayes - Regra de Decisão de Bayes

O estimador de Bayes é um estimador ou regra de decisão que mapeia um dado observado para uma classe apropriada (WEBB, 2002).

Considere C classes, $\omega_1, ..., \omega_C$ com probabilidades a *priori* $p(\omega_1), ..., p(\omega_C)$ (probabilidade de ocorrência de cada classe) conhecidas (WEBB, 2002). Se desejarmos minimizar a probabilidade de erro de classificação entre as classes, e não temos outras informações acerca do objeto observado que não seja a distribuição probabilística de cada classe, então nós atribuiríamos um objeto a classe ω_j se

$$p(\omega_j) > p(\omega_k)\ para\ k = 1, ..., C; k \neq j. \tag{2.2}$$

classificando todos os objetos em umas das classes (WEBB, 2002). Para as classes com probabilidades iguais, os padrões são atribuídos arbitrariamente entre essas classes.

No entanto, se nós temos um vetor de observação $\boldsymbol{x}$ e desejamos atribuir $\boldsymbol{x}$ a umas das C classes. A regra de decisão (WEBB, 2002) baseada em probabilidades é atribuir $\boldsymbol{x}$ a classe ω_j se a probabilidade da classe ω_j dada a observação $\boldsymbol{x}$, $p(\omega_j|\boldsymbol{x})$, é maior sobre todas as classes $p(\omega_1), ..., p(\omega_C)$. Isto é, atribui-se $\boldsymbol{x}$ a classe $p(\omega_j)$ se

$$p(\omega_j|\boldsymbol{x}) > p(\omega_k|\boldsymbol{x})\ para\ k = 1, ..., C; k \neq j. \tag{2.3}$$

Esta regra de decisão (WEBB, 2002) divide o espaço calculado dentro de C regiões $\Omega_1, ..., \Omega_C$ tal que se $\boldsymbol{x} \in \Omega_j$ então $\boldsymbol{x}$ pertence a classe ω_j.

A probabilidade a *posteriori* (WEBB, 2002) $p(\omega_j|\boldsymbol{x})$ pode ser expressada em termos da probabilidade a *priori* e da função de densidade da classe-condicional $p(\boldsymbol{x}|\omega_j)$ usando o teorema de Bayes como

$$p(\boldsymbol{x}|\omega_j) = \frac{p(\omega_j|\boldsymbol{x})p(\omega_j)}{p(\boldsymbol{x})} \tag{2.4}$$

e assim a regra de decisão pode ser escrita como: atribui-se $\boldsymbol{x}$ a ω_j se

$$p(\boldsymbol{x}|\omega_j)p(\omega_j) > p(\boldsymbol{x}|\omega_k)p(\omega_k)\ para\ k = 1, ..., C; k \neq j \tag{2.5}$$

sendo esta conhecida como regra de Bayes para o erro mínimo.

2.2.2.1.2 Modelos de Mistura

Modelos de mistura têm sido explorados cada vez mais em aplicações, tendo suas aplicabilidades em modelos de distribuição onde as medições surgem para separação de grupos de dados cujos membros individuais são desconhecidos (WEBB, 2002). Como método de estimava de densidade, os modelos de mistura são mais flexíveis que os modelos baseados em normais simples, fornecendo uma melhor discriminância em alguns casos. As aplicações de modelos de mistura incluem detecção de falhas têxteis, classificação de forma de onda e classificação da ROI (WEBB, 2002).

Um modelo de mistura finita (YANG; AHUJA, 1999) é uma distribuição da forma

$$p(\boldsymbol{x}) = \sum_{j=1}^{g} \pi_j p(\boldsymbol{x}; \Theta_j) \tag{2.6}$$

onde g é o número de componentes de mistura; $\pi_j \geq 0$ são as proporções de mistura ($\sum_{j=1}^{g} \pi_j = 1$); e $p(\boldsymbol{x}; \Theta_j)$, $j = 1, ..., g$, são as funções de componente de densidade que dependem do vetor de parâmetros Θ_j. Existem três conjuntos de parâmetros à serem estimados: o valores de π_j, as componentes de Θ_j e o valor de g. As componentes de densidade podem ser de diferentes formas paramétricas e são especificadas usando conjunto de dados conhecidos, se disponível. No modelo de mistura normal (WEBB, 2002), $p(\boldsymbol{x}; \Theta_j)$ é uma distribuição normal multivariada, com $\Theta_j = \{\mu_j, \Sigma_j\}$.

Dado um conjunto de n observações $(x_1, ..., x_n)$, a função de verossimilhança (WEBB, 2002) é

$$L(\Psi) = \prod_{i=1}^{n} \sum_{j=1}^{g} \pi_j p(\boldsymbol{x}; \Theta_j) \tag{2.7}$$

onde Ψ denota o conjunto de parâmetros $\{\pi_1, ..., \pi_g; \Theta_j, ..., \Theta_g\}$ e denotamos a dependência das componentes de densidade sobre os seus parâmetros como $p(x|\Theta_j)$. Em geral, podemos solucionar o problema de encontrar os parâmetros do modelo de mistura finito empregando-se um método iterativo. Uma abordagem que maximiza a verossimilhança $L(\Psi)$ é usar uma classe geral de processo iterativo conhecido como algoritmo EM (*expectation maximisation*) (WEBB, 2002).

2.2.2.1.3 Método do Limiar de Otsu (*Threshold*)

O método de Otsu (OTSU, 1979) é usado para selecionar um limiar de forma autónoma em imagens em tons de cinza usando a informação do histograma. Ele seleciona um limiar (*threshold*) pela maximização da variância entre as classes. Esse limiar é usado para separar os *pixels* da imagem em *"foreground"* e *"background"* (LIU; YU, 2009).

Assumindo que uma imagem é representada em L níveis de cinza $[0, 1, ..., L-1]$. O número de *pixels* no nível i é denotado por $N = n_1 + n_2 + ... + n_L$. A probabilidade do nível de cinza i é obtido por

$$p_i = \frac{n_i}{N}, p_i \geq 0, \sum_{0}^{L-1} p_i = 1. \tag{2.8}$$

No método de limiarização bi-nível (LIU; YU, 2009), os *pixels* da imagem são divididos dentro de duas classes C_1 com níveis de cinza $[0, 1, ..., t]$ e C_2 com níveis de cinza $[t+1, ..., L-1]$ pelo limiar t. A distribuição probabilística do nível de cinza para as duas classes são

$$w_1 = Pr(C_1) = \sum_{i=0}^{t} p_i \tag{2.9}$$

e

$$w_2 = Pr(C_2) = \sum_{i=t+1}^{L-1} p_i. \tag{2.10}$$

As médias das classes C_1 e C_2 são

$$u_1 = \sum_{i=0}^{t} \frac{ip_i}{w_1} \tag{2.11}$$

e

$$u_2 = \sum_{i=t+1}^{L-1} \frac{ip_i}{w_2}. \tag{2.12}$$

A média total dos níveis de cinza é dada por u_T, onde

$$u_T = w_1 u_1 + w_2 u_2. \tag{2.13}$$

As variâncias das classes são

$$\sigma_1^2 = \sum_{i=0}^{t} (i - u_1)^2 \frac{p_i}{w_1} \tag{2.14}$$

e

$$\sigma_2^2 = \sum_{i=t+1}^{L-1} (i - u_2)^2 \frac{p_i}{w_2} \tag{2.15}$$

A variância dentro da classe é dada por

$$\sigma_W^2 = \sum_{k=1}^{M} w_k \sigma_k^2. \tag{2.16}$$

A variância entre as classes é dada por

$$\sigma_B^2 = w_1(u_1 - u_T)^2 + w_2(u_2 - u_T)^2. \tag{2.17}$$

A variância total dos níveis de cinzas é dada por

$$\sigma_T^2 = \sigma_W^2 + \sigma_B^2. \tag{2.18}$$

O método de Otsu (OTSU, 1979) escolhe o limiar (*threshold*) t pela maximização da variância entre as classes, que é equivalente a minimização da variância dentro das classes, desde que a variância total (a soma da variância dentro das classes e da variância entre as classes) seja constante para as diferentes partições (LIU; YU, 2009).

$$t = \arg\{\max_{0 \leq t \leq L-1}\{\sigma_B^2(t)\}\} = \arg\{\min_{0 \leq t \leq L-1}\{\sigma_W^2(t)\}\}. \tag{2.19}$$

2.2.2.1.4 Cadeias Ocultas de Markov

Um processo ou cadeia de Markov é uma sequência de eventos, usualmente chamados de *estados*, cuja probabilidade de ocorrência de cada um deles é dependente somente do evento imediatamente anterior a ele.

Considere um sistema que pode ser descrito em qualquer tempo como sendo um conjunto de N estados distintos, $S_1, S_2, \cdots, S_N$, como ilustrado na Figura 2.1 (onde $N = 5$). Denotamos esses instantes de tempo associados às mudanças de estado como

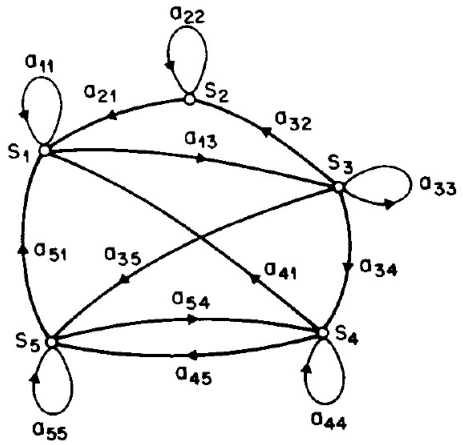


Figura 2.1: Ilustração do processo ou cadeia de Markov com 5 estados (rotulados como S_1 a S_5). Fonte: (RABINER, 1989)

.

$t = 1, 2, \cdots$, e o atual estado para o tempo t como q_t. Para o cadeia de Markov discreta, a probabilidade de transição entre o estado corrente e o antecessor é descrita por

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \cdots] = P[q_t = S_j | q_{t-1} = S_i]. \tag{2.20}$$

Dessa forma, o conjunto de probabilidades de transição de estado a_{ij} possui a seguinte forma

$$a_{ij} = P[q_t = S_j | q_{t-1} = S_i],\ 1 \leq i, j \leq N \tag{2.21}$$

com os coeficientes de transição de estado tendo as seguintes propriedades

$$a_{ij} \geq 0 \tag{2.22}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{2.23}$$

desde que eles obedeçam às restrições estocásticas padrões (RABINER, 1989).

O processo estocástico acima pode ser chamado de modelo Markov de observáveis desde que a saída do processo seja o conjunto de estados para cada instante de tempo onde cada estado corresponde ao evento físico (observações). Considere um simples modelo Markov de 3 estados de tempo. Assume-se que o tempo no dia é observado como apenas um dos seguintes estados:

Estado 1: chuvoso;
Estado 2: nublado;
Estado 3: ensolarado.

e que sua matrix A de probabilidades de transição de estado é

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \tag{2.24}$$

Dado que o tempo no dia 1 ($t = 1$) é ensolarado(estado 3), podemos responder a seguinte questão: Qual é a probabilidade de o tempo para os próximo 7 dias ser "ensolarado-ensolarado-chuvoso-chuvoso-ensolarado-nublado-ensolarado"?. Define-se a sequência de observação O como $O = \{S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3\}$ correspondendo a $t = 1, 2, \cdots, 8$. A probabilidade de O pode ser expressada como

$$\begin{aligned} P(O|Model) &= P[S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3] \\ &= P[S_3] \cdot P[S_3|S_3] \cdot P[S_3|S_3] \cdot P[S_1|S_3] \\ &\quad \cdot P[S_1|S_1] \cdot P[S_3|S_1] \cdot P[S_2|S_3] \cdot P[S_3|S_2] \\ &= \pi_3 \cdot a_{33} \cdot a_{33} \cdot a_{31} \cdot a_{11} \cdot a_{13} \cdot a_{32} \cdot a_{23} \\ &= 1 \cdot (0.8) \cdot (0.8) \cdot (0.1) \cdot (0.4) \cdot (0.3) \cdot (0.1) \cdot (0.2) \\ &= 1.536 \times 10^{-4} \end{aligned} \tag{2.25}$$

onde a notação

$$\pi_i = P[q_1 = S_i], \ 1 \leq i \leq N \tag{2.26}$$

é usada para denotar as probabilidades do estado inical (RABINER, 1989).

Para os modelos cujos os estados são desconhecidos a probabilidade pode ser avaliada como a probabilidade da sequência de observação

$$O = \{\underset{1}{S_i}, \underset{2}{S_i}, \underset{3}{S_i}, \cdots, \underset{d}{S_i}, \underset{d+1}{S_j} \neq S_i\}, \tag{2.27}$$

dado o modelo

$$P(O|Model, q_1 = S_i) = (a_{ij})^{d-1}(1 - a_{ij}) = p_1(d). \tag{2.28}$$

onde $p_1(d)$ é função de densidade de probabilidade discreta de duração d em estado i. Esta densidade de duração exponencial é característica do estado de duração na cadeia de Markov (RABINER, 1989).

Os modelos de Markov em que cada estado corresponde a um evento (físico) observável é muito restritivo para ser aplicável a muitos problemas de interesse. Dessa forma, o ampliação desse conceito é necessária para incluir os casos em que a observação é uma função probabilística do estado, ou seja, o modelo resultante chamado de modelo de Markov oculto é um processo estocástico incorporado duplamente com um processo estocástico subjacente que não é observável (é oculto), mas que pode somente ser observado através de um outro conjunto de processos estocásticos que produzem uma sequência de observações (RABINER, 1989).

Um HMM é caracterizado por:

1. N, o número de estados no modelo. Estes estados embora sejam ocultos, para muitas aplicações práticas há frequentemente algum significado físico ligado aos estados ou conjuntos de estados do modelo. Geralmente os estados estão interligados de tal forma que qualquer estado pode ser atingido a partir de qualquer outro estado. Denotamos os estados inviduais como $S = \{S_1, S_2, \cdots, S_N\}$, e o estado no instante t como q_t.

2. M, o número de observação distintas por estado, ou seja, o tamanho do alfabeto discreto. Os símbolos de observação corresponde à saída física do sistema a ser modelado. Denotamos os símbolos individuais como $V = \{v_1, v_2, \cdots, v_M\}$.

3. A distribuição de probabilidade dos estados de transição $A = \{a_{ij}\}$ onde

$$a_{ij} = P[q_{t+1} = S_i | q_t = S_i],\ 1 \leq i, j \leq N. \tag{2.29}$$

   Para o caso especial em que nenhum estado pode atingir a qualquer outro estado em uma única etapa, temos $a_{ij} > 0$ para todo i, j. Para outros tipos de HMM, teremos $a_{ij} = 0$ para um ou mais pares.

4. A distribuição de probabilidade da observação no estado j, $B = \{b_j(k)\}$, onde

$$b_j(k) = P[v_k \text{ para } t | q_t = S_j], \qquad \begin{matrix} 1 \leq j \leq N \\ 1 \leq k \leq M. \end{matrix} \tag{2.30}$$

5. A distribuição do estado inicial $\pi = \{\pi_i\}$ onde

$$\pi_i = P[q_1 = S_i],\ 1 \leq i \leq N. \tag{2.31}$$

Por conveniência, usamos a notação compacta

$$\lambda = (A, B, \pi). \tag{2.32}$$

para indicar que o conjunto de parâmetros completo do modelo (RABINER, 1989).

Dado a forma do HMM, há três problemas básicos de interesse que devem ser resolvidos para que o modelo possa ser usado em aplicações do mundo real. Estes problemas são os seguintes:

Problema 1: Dada a sequência de observação $O = O_1\ O_2\ \cdots\ O_T$, e um modelo $A = (A, B, \pi)$, como podemos calcular de forma eficiente $P(O|\lambda)$, a probabilidade da seqüência de observação, dado o modelo?

Problema 2: Dada a sequência de observação $O = O_1\ O_2\ \cdots\ O_T$, e o modelo λ, como podemos escolher uma correspondente seqüência de estado $Q = q_1\ q_2\ \cdots\ q_T$ que é ótima em algum sentido significativo (ou seja, que o melhor "explica" a observações)?

Problema 3: Como podemos ajustar os parâmetros do modelo $\lambda = (A, B, \pi)$ para maximizar $P(O|\lambda)$?

A solução do *problema 1* é dada através do algoritmo *Forward-Backward*. Considere a variável $\alpha_t(i)$ definida como

$$\alpha_t(i) = P(O = O_1\ O_2\ \cdots\ O_t, q_t = S_i|\lambda), \tag{2.33}$$

ou seja, a probabilidade da sequência de observação parcial, $O = O_1\ O_2\ \cdots\ O_t$, e o estado S_i no tempo t, dado o modelo λ. Assim, podemos resolver $\alpha_t(i)$ indutivamente, tal como se segue:

1. Inicialização:
$$\alpha_t(i) = \pi_i b_i(O_i),\ 1 \leq i \leq N. \tag{2.34}$$

2. Indução:
$$\alpha_{t+1}(j) = \left[\sum_{N}^{i=1} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \qquad 1 \leq t \leq T-1 \qquad 1 \leq j \leq N. \tag{2.35}$$

3. Finalização:
$$P(O|\lambda) = \sum_{N}^{i=1} \alpha_T i. \tag{2.36}$$

A solução do *problema 2* é baseada no método de programação dinâmica dado pelo algoritmo de Viterbi. Dessa forma, para encontrar a única melhor sequência de estados, $Q = \{q_1\ q_2\ \cdots\ q_T\}$, para uma dada sequência de observações $O = \{O_1\ O_2\ \cdots\ O_T\}$, é preciso definir a quantidade

$$\delta_t(i) = \underset{q_1, q_2, \cdots, q_{t-1}}{max} P[q_1\ q_2\ \cdots\ q_t = i,\ O_1\ O_2\ \cdots\ O_t|\lambda], \tag{2.37}$$

ou seja, $\delta_t(i)$ é a melhor pontuação (maior probabilidade) - *score* - ao longo de um caminho único, no tempo t, que representa as primeiras t observações e termina no estado S_i. Pela indução temos:

$$\delta_{t+1}(j) = [\max_i \delta_t(i)a_{ij}] \cdot b_j(O_{t+1}). \tag{2.38}$$

Para recuperarmos a sequência de estados, precisamos manter o controle do argumento que maximizou a Eq. 2.38, para cada t e j. Fazemos isso através do *array* $\psi_t(j)$. O processo completo para encontrar a melhor sequência de estados é dado como se segue:

1. Inicialização:

$$\begin{aligned} \delta_1(i) &= \pi_i b_i(O_1), \quad 1 \leq i \leq N. && (2.39) \\ \psi_1(i) &= 0. && (2.40) \end{aligned}$$

2. Recursão:

$$\begin{aligned} \delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i)a_{ij}]b_j(O_t), \qquad & 2 \leq t \leq T \\ & 1 \leq j \leq N. && (2.41) \\ \psi_t(j) = \operatorname*{argmax}_{1 \leq i \leq N} [\delta_{t-1}(i)a_{ij}], \qquad & 2 \leq t \leq T \\ & 1 \leq j \leq N. && (2.42) \end{aligned}$$

3. Finalização:

$$\begin{aligned} P^* &= \max_{1 \leq i \leq N} [\delta_T(i)] \\ q_T^* &= \operatorname*{argmax}_{1 \leq i \leq N} [\delta_T(i)]. && (2.43) \end{aligned}$$

4. Caminho (sequência de estados) pelo *backtracking*:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \cdots, 1. \tag{2.44}$$

Para o *problema 3* não há um caminho conhecido que ajuste os parâmetros (A, B, π) do modelo tal que a probabilidade da sequência de observações dado o modelo é maximizada. De fato, dado qualquer sequência finita de observações como treinamento de dados, não existe uma ótima maneira de estimar os parâmetros do modelo. No entanto, podemos encontrar um $\lambda = (A, B, \pi)$ tal que $P(O|\lambda)$ é maximizado localmente usando um processo iterativo através do método Baum-Welch.

No processo de reestimava dos parâmetros, primeiro é definido $\xi_t(i, j)$, a probabilidade de estar no estado S_i no tempo t e no estado S_j no tempo $t + 1$, dado o modelo e a sequência de observação, ou seja,

$$\xi_t(i, j) = P(q_t = S_i,\ q_{t+1} = S_j | O, \lambda). \tag{2.45}$$

Podemos reescrever $\xi_t(i,j)$ na forma

$$\begin{aligned}\xi_t(i,j) &= \frac{\alpha_t(i)\ a_{ij}b_j(O_{t+1})\ \beta_{t+1}(j)}{P(O|\lambda)} \\ &= \frac{\alpha_t(i)\ a_{ij}b_j(O_{t+1})\ \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)\ a_{ij}b_j(O_{t+1})\ \beta_{t+1}(j)}\end{aligned} \tag{2.46}$$

onde o termo do numerador é justamente o $P(q_t = S_i,\ q_{t+1} = S_j|O,\lambda)$ e o divisor $P(O|\lambda)$ é a medida de probabilidade desejada. Definindo-se $\gamma_t(i)$ como a probabilidade de estar no estado S_i no tempo t, dado a observação e o modelo, podemos relacionar $\gamma_t(i)$ para $\xi_t(i,j)$ pelo somatório sobre j, resultando em

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j). \tag{2.47}$$

Usando as fórmulas acima podemos reestimar os parâmetros de um HMM através da maximização da função de Baum's

$$Q(\lambda,\bar{\lambda}) = \sum_{Q} P(Q|O,\lambda)\ log[P(O,\ Q|\bar{\lambda})]. \tag{2.48}$$

sobre $\bar{\lambda}$. Onde a maximização de $Q(\lambda,\bar{\lambda})$ é dada por

$$\underset{\bar{\lambda}}{max}\ [Q(\lambda,\bar{\lambda})] \Rightarrow P(O|\bar{\lambda}) \geq P(O|\lambda). \tag{2.49}$$

2.2.3 Imagens a Cores

O processo seguido pelo cérebro humano na percepção de cores é um fenômeno físico-psicológico que ainda não é completamente compreendido, no entanto, a natureza física das cores pode ser expressa numa base formal suportada por resultados experimentais e teóricos. Basicamente, as cores que os seres humanos percebem num objeto são determinadas pela natureza da luz refletida do objeto. A luz visível é composta por uma banda de frequências relativamente estreita no espectro de energia eletromagnética. Um corpo que reflete a luz e é relativamente balanceado em todos os comprimentos de onda visíveis aparece como branco para o observador. Entretanto, um corpo que favoreça a reflectância em uma variação limitada do espectro visível exibe alguns tons de cores. Por exemplo, objetos verdes refletem a luz com comprimentos de onda primariamente no intervalo de 500 a 570nm (10^{-9}m), enquanto absorve a maior parte da energia de outros comprimentos de ondas (GONZALEZ; WOODS, 2003).

A caracterização da luz é essencial para a ciência das cores. Se a luz for acromática (sem cores), seu único atributo será sua intensidade, ou quantidade. Assim, o termo nível de cinza refere-se a uma medida escalar de intensidade que varia do preto aos cinzas, e finalmente ao branco (GONZALEZ; WOODS, 2003).

A luz cromática abarca o espectro de energia eletromagnética desde aproximadamente 400 até 700nm. Três valores são usados para descrever a qualidade de uma fonte de luz cromática: radiância, luminância e brilho. Radiância é a quantidade total de energia que

flui de uma fonte de luz. Luminância dá uma medida da quantidade de energia que um observador percebe de uma fonte de luz. Brilho é um descritor subjetivo, que é praticamente impossível de ser medido. Ele incorpora a noção acromática de intensidade, sendo um dos fatores chave na descrição da sensação de cores (GONZALEZ; WOODS, 2003).

Devido à estrutura do olho humano, todas as cores são vistas como combinações variáveis das três chamadas cores primárias: vermelho (R, do inglês *red*), verde (G, do inglês *green*) e azul (B, do inglês *blue*). As cores primárias podem ser adicionadas para produzir as cores secundárias da luz - magenta (vermelho e azul), ciano (verde e azul) e amarelo (vermelho e verde). A mistura das três cores primárias, ou uma secundária com sua cor primária oposta, em intensidades corretas produz a luz branca (GONZALEZ; WOODS, 2003).

As cores são representadas em dispositivos eletrônicos através de modelos que facilitam a especificação das cores em alguma forma de padrão e de aceite geral. Essencialmente, um modelo de cor é uma especificação de um sistema de coordenadas tridimensionais e um subespaço dentro deste sistema onde cada cor é representada por um único ponto (GONZALEZ; WOODS, 2003).

A maioria dos modelos de cores atualmente em uso são orientados ou em direção do *hardware* ou em direção a aplicações envolvendo manipulação de cores. Os modelos orientados para hardware mais comumente usados na prática são o RGB ("*red*, *green*, *blue*") para monitores e uma ampla classe de câmeras de vídeo a cores; o CMY ("*cyan*, *magenta*, *yellow*") para impressoras coloridas; e o YIQ, que é o padrão para transmissão de TV a cores. Entre os modelos frequentemente usados para a manipulação de imagens coloridas estão HSI ("matiz, saturação, intensidade") e o HSV ("matiz, saturação, valor"). E os modelos de cores mais frequentemente usados para processamento de imagens são o RGB, o YIQ, e o HSI (GONZALEZ; WOODS, 2003).

2.2.3.1 Cores Invariante a Intensidade do Iluminante

As Cores são descritores poderosos que frequentemente simplificam a identificação do objeto e a extração de uma cena. Elas podem pontencialmente fornecer informações úteis para uma variadade de aplicações em visão computacional e processamento de imagens, tais como, segmentação de imagens, recuperação de informação, reconhecimento de padrões e rastreamento (HORDLEY et al., 2005). Porém, para que ela seja de total ajuda na prática, as cores devem relacionar diretamente a propriedade intrínseca do objeto da imagem e ser independente de suas condições de captura tais como a iluminação da cena e tipo de dispositivo que foi obtida (HORDLEY et al., 2005). Para amenizar a ação de tais dependências na imagem é possível utilizar abordagens de invariância de cores que apresenta métodos de transformações sobre seus dados tais que a transformada dos dados sejam independente ao iluminante, ou abordagens de constância de cores que determina uma estimação da iluminação da cena (FINLAYSON; HORDLEY, 2001).

Adotando-se um modelo simples de formação de imagem em que a resposta de um dispositivo de captura depende de três fatores: a luz que o objeto é iluminado, a propriedade da superfície de reflectância do objeto, e as propriedades do sensor do dispositivo. Assume-se que a cena é iluminada pela uma única fonte de luz caracterizada pela sua distribuição do espectro de energia que denotamos por $E(\lambda)$ e que específica quanto de energia a origem erradia de cada comprimentos de onda (λ) do espectro electromagnético (HORDLEY et al., 2005). As propriedades de reflectância da superfície são caracterizadas pela função $S(\lambda)$ que define a proporção de luz incidente sobre ela. E, o sensor

é caracterizado por $Q_k(\lambda)$, ou seja, a função de sensibilidade espectral que específica a sensibilidade para energia de cada comprimento de onda do espectro. Onde k denota o *k*-ésimo sensor (HORDLEY et al., 2005). A resposta a um dispositivo de captura de imagem é definida como

$$q_k = \int_{\omega} E(\lambda)S(\lambda)Q_k(\lambda)d\lambda, \quad k = 1, ..., m, \tag{2.50}$$

onde na integral é assumido a faixa de comprimento de onda ω (a faixa para que o sensor tenha a sensibilidade diferente de zero). Segue-se assumindo que o dispositivo de captura possui três sensores $(m = 3)$ de modo que a resposta para um ponto na cena seja representado por um trio de valores: (q_1, q_2, q_3). Sendo comum chamarmos estes trios de valores como R, G e B ou apenas como RGBs.

A Eq. (2.50) é um modelo acurado do processo de formação da imagem para a superfície de Lambertian para que a luz incidente seja refletida igualmente para todas as direções de incidência (HORDLEY et al., 2005). A equação deixa claro que a resposta ao dispositivo depende de ambos, a propriedades do sensor ($Q_k(\lambda)$) e a iluminação predominante ($E(\lambda)$). Isto é, as repostas são dependentes do dispositivo e da iluminação. Logo, se não levarmos em consideração estas dependências, a representação de cores em RGB não pode ser considerada corretamente como uma propriedade intrínseca de um objeto (HORDLEY et al., 2005).

Um modelo simples para solucionar o problema de mudança de iluminação é o chamado modelo diagonal, sendo proposto que o sensor de respostas sobre o par de iluminantes são relacionados por uma transformação diagonal da matriz:

$$\begin{bmatrix} R^c \\ G^c \\ B^c \end{bmatrix} = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix} \begin{bmatrix} R^o \\ G^o \\ B^o \end{bmatrix} \tag{2.51}$$

onde os sobescrito o e c caracterizam o par de iluminantes. Adotando esse modelo simples de representação de iluminante invariante de uma imagem, podemos obtê-lo pela aplicação da seguinte transformada:

$$R' = \frac{R}{R_{ave}}, \quad G' = \frac{G}{G_{ave}}, \quad B' = \frac{B}{B_{ave}}, \tag{2.52}$$

onde o trio $(R_{ave}, G_{ave}, B_{ave})$ denotam as médias de R, G e B na imagem.

3 DETECÇÃO VISUAL DE ATIVIDADE DE VOZ COM BASE NA MOVIMENTAÇÃO LABIAL

A solução desenvolvida para detecção de atividade de voz é mostrada resumidamente na Figura 3.1. Ela apresenta todas as entradas de dados e saídas que são geradas durante todo o processo para detecção de atividade de voz que se estende desde a etapa de obtenção da informação visual da movimentação dos lábios nos quadros do vídeo até a detecção de atividade de voz pela classificação das observações em uma janela temporal utilizando HMM's como abordagem.

Para o processo de obtenção da informação visual realizamos primeiramente a etapa de correção das cores para obter a invariância de intensidade do iluminante onde procuramos amenizar o efeito que as mudanças do iluminante ocasionam sobre as cores dos objetos em cena do vídeo. Pelo diagrama da Figura 3.1 é demonstrado as principais etapas

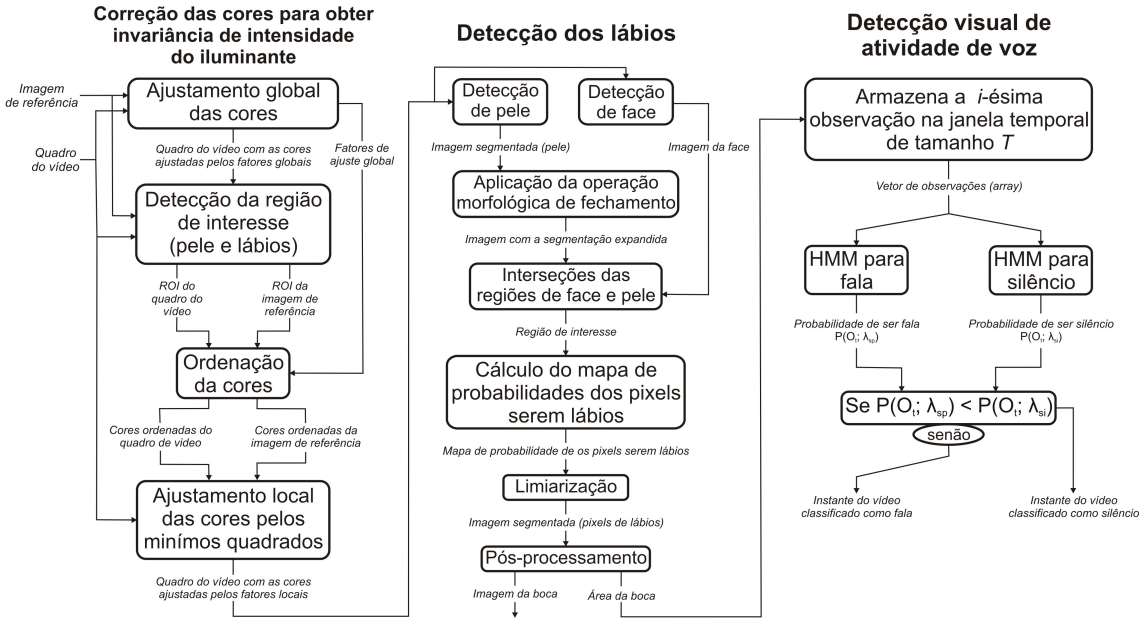


Figura 3.1: Diagrama de estrutura de dados do algoritmo de detecção de atividade de voz.

realizadas na correção das cores para obter a invariância de intensidade do iluminante. O processo de correção inicia-se pela computação do ajustamento global das cores sobre o quadro de vídeo. A seguir, é realizada a etapa de detecção da região de interesse, ou seja, é detectado a pele e os lábios. Então, as cores dos *pixels* encontrados dentro das regiões de interesse são ordenadas na etapa de ordenação. E na etapa de ajustamento local das

cores é realizada a correção de intensidade das cores dos quadros de vídeo que são usadas para o processo de detecção de lábios.

No processo de detecção são obtidos os *pixels* de lábios utilizando a informação de intensidades de cores para discriminá-los. Inicializamos esse processo realizando primeiramente uma redução da área de busca (ROI) dentro do quadro de vídeo. Essa redução é dada através das etapas de detecção de pele, detecção de face, aplicação da operação morfológica de fechamento e da intersecção das regiões de face e pele como é demonstrado pelo diagrama da Figura 3.1. Sobre a região reduzida do quadro de vídeo ou simplesmente região de interesse é realizado a etapa de cálculo do mapa de probabilidades dos *pixels* serem de lábios. Então, os *pixels* de lábios são detectados aplicando-se um limiar na etapa de limiarização. O processo de obtenção dos lábios é finalizado com a aplicação da etapa de pós-processamento. Na etapa de pós-processamento, os *pixels* que foram detectados como falsos positivos são eliminados. Além disso, nessa etapa são inclusos todos os *pixels* da boca no resultado final da segmentação, ou seja, *pixels* de lábios que não foram detectados na etapa anterior são inclusos junto com os *pixels* que estiverem entre os lábios inferior e superior (i.e, *pixels* de dentes, língua, gengiva e céu da boca se a imagem a ser segmentada for de uma pessoa que esteja com a boca aberta).

Por fim, no processo de detecção visual de atividade de voz é computada a classificação da occorrência de atividade de voz (i.e., fala ou silêncio). A descrição detalhada de toda a solução encontrada para o método de detecção visual de atividade de voz baseado na movimentação labial é realizada nas subsecções que se seguem ao longo deste capítulo.

3.1 Correção das Cores para Obter Invariância de Intensidade do Iluminante

As cores são utilizadas como fonte de informação para extração dos nossos objetos de interesse, sendo-os, a pele e os lábios. Elas caracterizam sua essência individual discriminando-os quase de forma independente dos outros objetos que estejam em cena, ou seja, relacionam bem suas propriedades intrínsecas. No entanto, como as cores relacionadas aos objetos de interresse na imagem podem sofrer influência direta da intensidade do iluminante (iluminação em cena) durante sua captura, faz-se necessário minimizar a influência de tal dependência realizando-se uma abordagem que torne a cores invariantes à iluminação da cena (FINLAYSON; HORDLEY, 2001).

A invariância de cores ou sua constância é um fenômeno que mantém as propriedades e aparência das cores aproximadamente as mesmas sob diferentes iluminantes (FINLAYSON; HORDLEY, 2001), istó é, "enxergar a cores originais de um objeto independentes da sua iluminação", ou seja, uma bola vermelha é vermelha sobre o sol, sobre a sombra ou sobre lugares com baixa iluminação. Tal fenômeno ocorre no sistema nervoso humano que a partir da radiação detectada pela retina, extrai aquilo que é invariante sob diferentes mudanças de iluminação (penumbra). Embora a radiação sofra mudanças, o nosso cerébro reconhece certos padrões constantes nos estímulos percebidos. Ele agrupa e classifica os diferentes fenômenos como se fossem iguais. O que nós vemos não é exatamente o que está sendo perceptivo no exato momento através de nosso olhos, mas corresponde a um modelo simplificado da realidade compreendido pelo nosso sistema visual (HORDLEY et al., 2005). Baseado nesse contexto, buscamos minimizar os efeitos do iluminante sobre as regiões de interesse onde desejamos que as cores sob a intensidade do iluminante

desconhecido (iluminação da cena) mantenham a maior coerência possível com as cores sob a intensidade do iluminante conhecido das imagens usadas para geração do modelo na fase de treinamento. Portanto, procuramos solucionar o problema buscando uma transformação que ajuste as cores ao padrão em que a intensidade do iluminante é conhecido. Dessa forma, partimos do princípio da informação de um todo (toda imagem em si) para o específico de uma região, ou seja, a região de interesse (ROI), logo, não se está preocupado em manter coesa toda a informação de intensidade do iluminante sobre as cores de todos os objetos em cena na imagem a ser pré-processada, mas, das regiões dos objetos de interesse dividindo o processo de obtenção do ajuste das intensidades de cores para obter a invariância de intensidade do iluminante em duas etapas de ajuste, um ajuste global e um ajuste local.

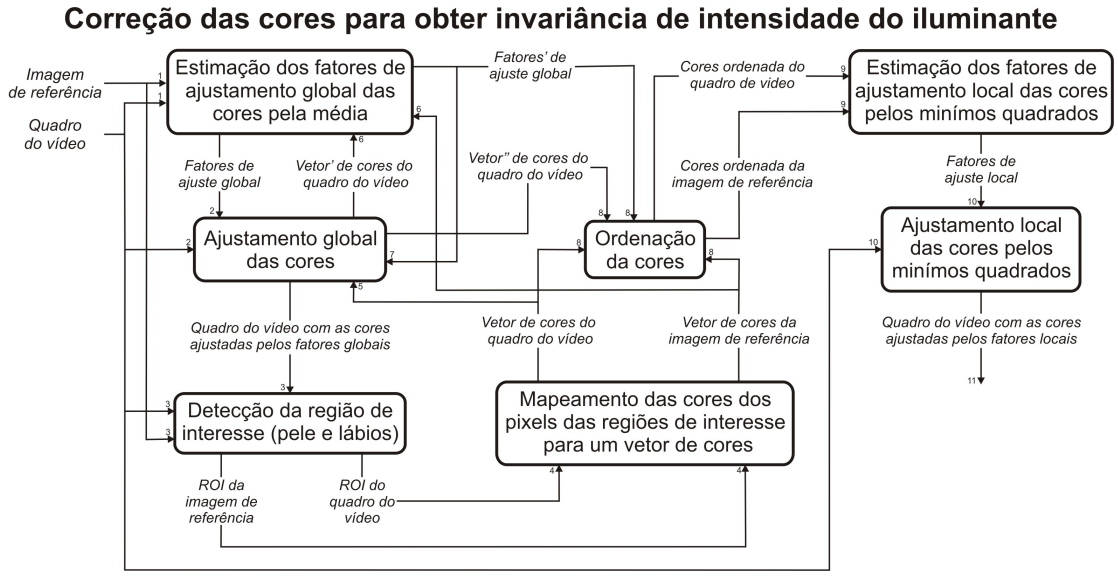


Figura 3.2: Diagrama de estrutura de dados do algoritmo de correção de cores para mudanças de intensidade do iluminante.

O diagrama na Figura 3.2 apresenta a visão geral do método de correção de cores para obter a invariância de intensidade do iluminante. Pelo diagrama podemos observar todos os passos tomados para realização da correção das cores no quadro do vídeo. O algoritmo começa pela estimação dos fatores de ajustamento global das cores através da média utilizando a informação dos canais de cores no espaço RGB do quadro do vídeo e da imagem usada como referência de iluminação. O cálculo dos fatores para cada canal é realizado como se segue

$$F_j = \frac{\mu'_j}{\mu_{j_n}}, \tag{3.1}$$

sendo μ'_j as médias de intensidades do canal *j* da imagem de referência e μ_{j_n} as médias de intensidades do canal *j* do n-esímo quadro do vídeo onde ambas as imagens estão representadas no espaço de cores RGB; e F_j é o fator de ajuste do j-esímo canal, onde $j = 1, 2\ e\ 3$ (ou R, G e B). A seguir, são aplicados os fatores encontrados sobre o quadro do vídeo para ajustar as cores a mudança global de iluminação. O ajuste da intensidade de cada canal é dado pela aplicação do fatores de ajuste sobre cada canal dos *pixels* do quadro do vídeo, como se segue

$$fr'_j(x, y) = fr_j(x, y) \cdot F_j, \tag{3.2}$$

onde $fr'_j(x, y)$ é o quadro do vídeo (*frame*) com a cores ajustadas; $fr_j(x, y)$ é o quadro original do vídeo (*frame*); e F_j é fator de ajuste, sendo $j = 1, 2\ e\ 3$ (corresponde aos canais do espaço de cores RGB). O ajuste de intensidade pelos fatores globais sobre o quadro do vídeo é realizado para podermos localizar a região de interesse, pele e lábios, através do detector de peles. Uma vez que as cores do quadro de vídeo sobre uma iluminação desconhecida é corrigida pelos fatores globais, podemos garantir que elas mantêm-se coerentes com as cores sobre a iluminação conhecida que são utilizadas na modelagem da solução.

(a) (b) (c)

Figura 3.3: Ilustração das altas probabilidades: (a) regiões de fundo (*background*); (b) regiões de cabelo; e (c) regiões de pele.

A detecção da região de interesse (ROI) é executada sobre o quadro de vídeo $fr'_j(x, y)$ (com as cores ajustadas) e sobre a imagem de referência para obter a localização dos *pixels* de interesse. Observe que aplicamos a detecção da ROI sobre o quadro do vídeo com as cores ajustadas apenas para obter a localização dos *pixels* de interesse do quadro original do vídeo, pois o quadro com as cores corrigidas preserva uma maior coerência com modelo de solução baseado em informação de cores. Dessa forma, realizamos através do detector de pele seguida da aplicação de operações morfológicas a detecção da ROI em ambas as imagens de entrada.

A detecção de pele é efetuada pelo classificador de Bayes descrito na Seção 2.2.2.1.1 que gera mapas probabilísticos utilizando um modelo de mistura de Gaussinas como descrito Seção 3.2, sendo o *pixel* avaliado pela regra de decisão de Bayes dada pela Eq.(2.5) para tomar suas decisões de classificação como demonstrado em resultados obtidos na Figura 3.3 e na Figura 3.6. Após a classificação, aplicam-se os operadores morfológicos de dilatação circular de 20 *pixels* e de erosão de 23 *pixels* para obter-se a ROI na imagem de referência I^{ref}_{ROI} e no quadro do vídeo I^{fr}_{ROI}. Um exemplo em resultado obtido experimentalmente da localização da região de interesse é ilustrado na Figura 3.4. Dando seguimento, realizamos o mapeamento das cores existentes dentro da região de interesse, ou seja, averiguamos quais são as cores existentes na ROI e armazenamos-as em um vetor para posterior uso na estimação dos fatores locais de ajuste. As cores armazenadas são triplas de intensidades que formam a representação das cores no espaço RGB, sendo o vetor representado por C_{ij} onde i é a i-esíma representação de cor e j representa a posição da tripla RGB (j = 1, 2 e 3 = R, G e B). O mapeamento das cores é realizado separadamente, um para o ROI da imagem de referência I^{ref}_{ROI} e um para o ROI do quadro de vídeo I^{fr}_{ROI}. Então, são gerado dois vetores, um vetor de cores do quadro de vídeo C^{*}_{ij} e um vetor de cores da imagem de referência C^{**}_{ij}, como ilustrado na Figura 3.2 pelo bloco de mapeamento das cores dos *pixels* das regiões de interesse para o vetor de cores.

As informações contidas nos vetores C_{ij}^{*} e C_{ij}^{**} são utilizados no cálculo dos fatores locais de ajuste pelos mínimos quadrados. No entanto, para podermos fazer uso dessas informações precisamos primeiramente relacionar as informações separadas e desordenadas das cores de cada vetor para que ambos obtenham uma coerência de proximidade entre si mantendo-se uma relação de procedência, isto é, suas posições i são ordenadas de maneira a preservar a proximidade das cores entre os dois vetores. Então, primeiramente aumentamos o grau de proximidades das cores entre os dois vetores aplicando duas transformações em C_{ij}^{*} . A primeira transformação realizada é a aplicação dos fatores globais de ajuste obtidos anteriormente sobre o vetor C_{ij}^{*} como se segue

$$C_{ij}^{\prime *} = C_{ij}^{*} \cdot F_j, \tag{3.3}$$

onde $C_{ij}^{\prime *}$ é o vetor linha de cores do quadro. Para a segunda transformação recalculamos os fatores de ajuste globais pela média usando as informações de cores dos vetores $C_{ij}^{\prime *}$ e C_{ij}^{**} conforme a Eq. 3.1 obtendo-se os fatores linhas de ajuste globais $F_j^{\prime}$. Então, aplicamos os fatores encontrados sobre $C_{ij}^{\prime *}$ da mesma forma como realizado na Eq. 3.3 obtendo-se o vetor duas linhas de cores do quadro $C_{ij}^{\prime\prime *}$. Após a execução das duas transformações sobre o vetor C_{ij}^{*} é relizado o cálculo da distância angular ou abertura angular para cada cor dos vetores C_{ij}^{*} e C_{ij}^{**} em relação a referência $F_j^{\prime}$.

A distância de cada cor do vetor de cores do quadro de vídeo C_{ij}^{*} é dada pelo seguinte cálculo

$$\theta_i^{*} = arccos\left(\frac{\vec{F^{\prime}} \cdot \vec{C_i^{\prime\prime *}}}{|\vec{F^{\prime}}| \cdot |\vec{C_i^{\prime\prime *}}|}\right), \tag{3.4}$$

onde θ_i^{*} é a i-ésima distância de cor referente ao vetor C_{ij}^{*}; já a distância de cada cor do vetor de cores da imagem de referência C_{ij}^{**} é obtida através do cálculo

$$\theta_i^{**} = arccos\left(\frac{\vec{F^{\prime}} \cdot \vec{C_i^{**}}}{|\vec{F^{\prime}}| \cdot |\vec{C_i^{**}}|}\right), \tag{3.5}$$

onde θ_i^{**} é a i-ésima distância de cor referente ao vetor C_{ij}^{**}. Depois de obtidas as distâncias θ_i^{*} e θ_i^{**} referentes aos vetores C_{ij}^{*} e C_{ij}^{**}, podemos realizar a ordenação das cores de cada vetor. A ordenação (*sorting*) dos vetores C_{ij}^{*} e C_{ij}^{**} é realizada de forma ascendente usando as distâncias das cores θ_i^{*} e θ_i^{**} para ordená-los, ou seja, a ordenação das cores do vetor C_{ij}^{*} é dada segundo as distâncias de suas cores dada pelo vetor θ_i^{*} e a ordenação das cores do vetor C_{ij}^{**} é dada segundo as distâncias de suas cores dada pelo vetor θ_i^{**} tendo em ambos os casos o indíce i como referência entre a cor do vetor e sua distância. A ordenação resulta em dois vetores ordenados, o vetor ordenado de cores do quadro de vídeo C_{ij}^{fr} e o vetor ordenado de cores da imagem de referência C_{ij}^{ref}, que podem ter tamanhos diferenciados impossibilitando o seu uso no método dos mínimos quadrados que exije que ambas informações tenha o mesmo tamanho. Por isso, após a ordenação do vetores, averiguamos qual deles tem o menor tamanho para elimarmos as últimas posições de cores do maior vetor até que ambos tenha o mesmo tamanho.

A estimava dos fatores de ajuste local das intensidades do iluminante sobre as cores é obtida através do critério dos mínimos quadrados utilizando como informação as intensidades dos vetores de cores ordenados C_{ij}^{fr} e C_{ij}^{ref} com o mesmo número de elementos, sendo o cálculo dado por

(a) (b)

Figura 3.4: Regiões de interesse: (a) ROI da imagem de referência I^{ref}_{ROI}; e (b) ROI do quadro do vídeo I^{fr}_{ROI}

$$\vec{A} = (X^t X)^{-1}(X^t \vec{Y}), \tag{3.6}$$

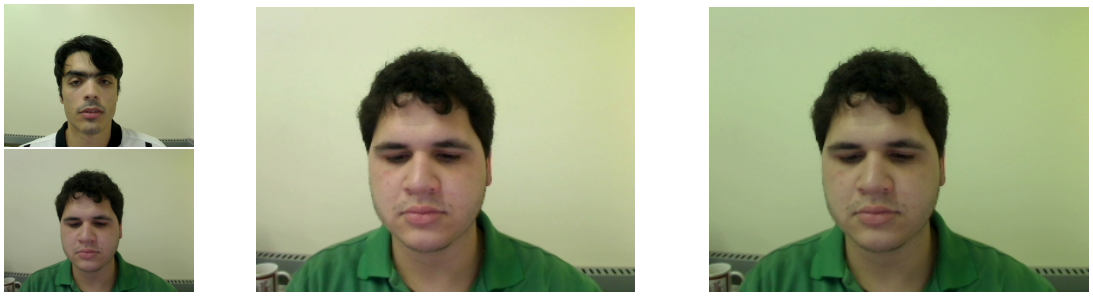
$$X = \begin{bmatrix} R_1^{fr} & 0 & 0 \\ 0 & G_1^{fr} & 0 \\ 0 & 0 & B_1^{fr} \\ \vdots & \vdots & \vdots \\ R_n^{fr} & 0 & 0 \\ 0 & G_n^{fr} & 0 \\ 0 & 0 & B_n^{fr} \end{bmatrix} \quad \vec{A} = \begin{bmatrix} Fl_R \\ Fl_G \\ Fl_B \end{bmatrix} \quad \vec{Y} = \begin{bmatrix} R_1^{ref} \\ G_1^{ref} \\ B_1^{ref} \\ \vdots \\ R_n^{ref} \\ G_n^{ref} \\ B_n^{ref} \end{bmatrix},$$

onde X é a matriz com as informações de intensidade do vetor de cores ordenado C^{fr}_{ij}; $\vec{Y}$ é o vetor com as informações de intensidade do vetor de cores ordenado C^{ref}_{ij}, sendo R^*_i, G^*_i e B^*_i a representação das intensidades da i-ésima cor em RGB com $i = 1, ..., n.$ e * = ref ou fr.; e $\vec{A}$ é o vetor com os fatores de ajuste de local, sendo Fl_R, Fl_G e Fl_B as transformações para os canais R, G e B do quadro do vídeo;

Finalmente o ajuste final das cores no quadro do vídeo para obter a invariância de intensidade do iluminante é dado pelas aplicações dos fatores locais de ajuste sobre as cores, como segue

$$\widetilde{\vec{Y}} = X\vec{A}, \quad X = \begin{bmatrix} R_1 & 0 & 0 \\ 0 & G_1 & 0 \\ 0 & 0 & B_1 \\ \vdots & \vdots & \vdots \\ R_n & 0 & 0 \\ 0 & G_n & 0 \\ 0 & 0 & B_n \end{bmatrix} \quad \vec{Y} = \begin{bmatrix} R'_1 \\ G'_1 \\ B'_1 \\ \vdots \\ R'_n \\ G'_n \\ B'_n \end{bmatrix}, \tag{3.7}$$

onde $\widetilde{\vec{Y}}$ é o vetor de *pixels* com as cores ajustados, sendo R'_i, G'_i e B'_i a representação do *i-ésimo pixels* cuja a cor em RGB foi ajustada com $i = 1, ..., n$; X é a matriz com as intensidades dos canais R, G e B que representam os *pixels* do quadro de vídeo; e $\vec{A}$ é o vetor de ajuste local dado pela Eq. (3.6). Um exemplo de resultado dos ajuste final de um quadro de vídeo é ilustrado na Figura 3.5 (c).



(a) (b) (c)

Figura 3.5: Exemplicação dos resultados das correções das cores para mudanças de intensidade do iluminante: (a) Imagem de referência e entrada; (b) Imagem após o ajuste inicial; e (c) Imagem após o ajuste final.

3.2 Detecção de Pele

A detecção de pele é obtida pelo classificador Bayes (WEBB, 2002). Esse classificador faz uso do critério estatístico para descriminar, avaliar e classificar os dados modelados e disponível na imagem. Portanto, a partir dos dados da imagem são construídos os mapas probabilísticos de cada uma das classes avaliadas, sendo-as, pele, cabelo e fundo (*background*). Esses mapas são obtidos *pixel* a *pixel* da imagem pela probabilidade a posteriori dada pela seguinte regra de Bayes (WEBB, 2002)

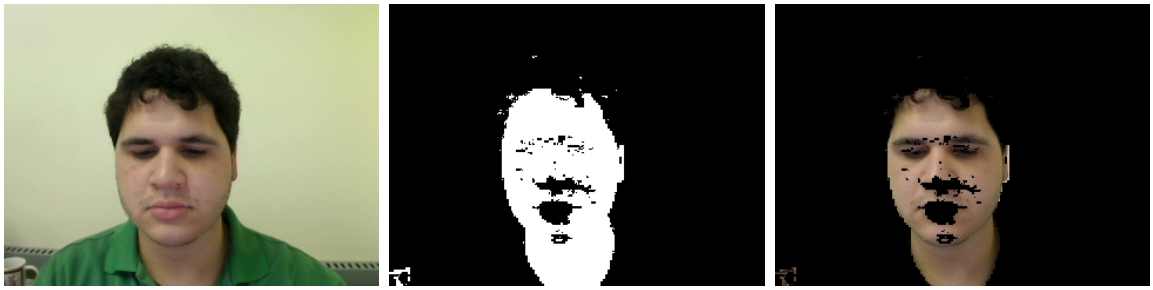
$$p(\omega_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_j)p(\omega_j)}{p(\boldsymbol{x})}, \tag{3.8}$$

onde ω_j é a j-ésima classe, sendo $j = 1, ..., n$; $p(\omega_j|\boldsymbol{x})$ é a probabilidade a posteriori; $p(\boldsymbol{x}|\omega_j)$ é a função de densidade probabilística (verossimilhança) modelada por uma mistura de Gaussianas; $p(\omega_j)$ é a probabilidade a priori da classe 'j' estimada do número de amostras usada para o treinamento do modelo; e $p(\boldsymbol{x})$ é a probabilidade a priori de 'x' ("*evidence probability*") dada pelo conjunto completo de treinamento dos dados (istó é, fator de normalização probabilística entre 0-1). As Figuras 3.3 (a), (b) e (c) ilustram exemplificações dos mapas probabilísticos da classe fundo, cabelo e pele obtidos pela Eq. 3.8. Esses mapas informam visualmente pelo seu nível de *brilho* (isto é, intensidade – "em branco") as regiões de alta probabilidade em *pixel* a *pixel* de um deles pertencerem a qualquer uma das classes avaliadas.

Para a construção das prioris do modelo de mistura Gaussiana da p.d.f. $p(\boldsymbol{x}|\omega_j)$ explorou-se a informação baseada em cores. Nos quais, utiliza-se a informação bivariada do espaço de cores CIELab para construção do modelo 2D da mistura de Gaussianas, ou seja, foram usados os canais de cromaticidades 'a' e 'b' do CIELab.

A função de densidade probabilística (P.D.F) modelada pela mistura de gaussiana bivariada é dada por

$$\begin{aligned} p(\boldsymbol{x};\Phi) &= \sum_{i=1}^{g} \pi_i p_i(\boldsymbol{x};\boldsymbol{\Theta}) \\ &= \sum_{i=1}^{g} \pi_i \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)}, \end{aligned} \tag{3.9}$$



(a) (b) (c)

Figura 3.6: Ilustração da detecção de pele: (a) imagem original; (b) região de segmentação binária; e (c) região de segmentação da pele.

onde g é o número de componentes de mistura; π_i é o parâmetro de mistura cuja a soma é igual a 1 e $\pi_i \geq 0$ (componente de proporcionalidade de contribuição de cada gaussiana); $p_i(\boldsymbol{x}; \boldsymbol{\Theta})$ é a p.d.f correspondente a Gaussiana i; $\boldsymbol{\Theta}$ denota o vetor de todos os parâmetros, sendo-os, o vetor de médias $\boldsymbol{\mu}_i$ e a matriz de covariâncias $\boldsymbol{\Sigma}_i$ para $i = 1, ..., g$; e x é o vetor das n variáveis de observação, ou seja, $x = [x_1, ..., x_n]$. Os parâmetros g, π_i e $\boldsymbol{\Theta}$ são obtidos através de um treinamento que é descrito na Secção 3.5.

O critério de decisão para classificação das classes é dado através da regra de decisão de Bayes, sendo o *pixel* x atribuído a uma classe avaliada se

$$p(\omega_j|\boldsymbol{x}) > p(\omega_k|\boldsymbol{x}) \;\; para \;\; k = 1, ..., C; \;\; k \neq j \tag{3.10}$$

no qual $p(\omega_j|\boldsymbol{x})$ e $p(\omega_k|\boldsymbol{x})$ é a probabilidade a posteriori das classes avaliadas. Portanto, a detecção de pele, como ilustrado na Figura 3.6 (c) é obtida pela regra decisão de Bayes após o cálculo dos mapas probabilísticos da imagem dado pela Eq. (3.8).

3.3 Detecção de Lábios

A detecção de lábios é executada após a correção das cores para obter a invariância de intensidade do iluminante, no qual, aplicou-se o ajuste da intensidade do iluminante sobre as cores da imagem de entrada como descrito na Secção 3.1. O processo de detecção é inicializado pela computação da restrição de área sobre a imagem pré-processada, ou seja, tomamos uma área menor da imagem. Dentro dessa pedaço de área de *pixels* da imagem será realizado a busca pelos *pixels* de lábios. Essa área de busca é obtida pela intersecção entre a região de *pixels* da face e dos *pixels* da pele expandidos.

A região de face é dada pelo detector de faces que é executado através do algoritmo de Bins (BINS et al., 2009). E a região de pele expandida é dada pela detecção de pele conforme descrita na Seção 3.2 seguida da aplicação das operações morfológicas de fechamento com um operador circular de dilatação de 25 *pixels* e um operador circular de erosão 23 *pixels*, sendo estes valores de *pixels* dos operadores morfológicos obtidos após a realização de inúmeros experimentos de ajuste desses parâmetros. Essa expansão ou inclusão de *pixels* a mais sobre os *pixels* de pele detectados é realizado para assegurar que todos *pixels* de lábios estejam incluídos dentro dessa região. Alguns resultados que exemplificam o processo de obtenção das regiões de *pixels* de face e de pele com os detectores, e a sua intersecção é ilustrado nas Figuras 3.6, 3.7 e 3.8 (c), sendo esta última utilizada para a detecção de lábios.

(a) (b)

Figura 3.7: Ilustração da detecção de face: (a) face detectada; e (b) região da face; .

Os *pixels* de lábios que estão dentro da região de busca na imagem são detectados pela discriminação das regiões de alta probabilidade de serem *pixels* de lábios ou não-lábios. Estas regiões probabilísticas dentro da imagem são construídas através da probabilidade a posteriori dada através da Eq. (3.8). Sendo, as seguintes probabilidades a prioris usadas na equação para o cálculo da probabilidade a posteriori: a priori $p(\omega_j)$ que é estimada do número de amostras usada para o treinamento do modelo; a priori de verrossimilhaça (p.d.f) $p(\boldsymbol{x}|\omega_j)$ que é modelada por uma mistura de Gaussiana conforme a Eq. (3.9); e a priori $p(\boldsymbol{x})$ ("*evidence probability*") que é dada pelo conjunto completo de treinamento dos dados.

(a) (b) (c)

Figura 3.8: Ilustração da região de busca: (a) face detectada; (b) pele detectada; e (c) região de busca.

A mistura de gaussianas foi modelada para utilizar as informações de cromaticidades da imagem como dados observáveis da distribuição. Portanto, modelou-se uma mistura de gaussianas bivariada a partir das informações de matiz do canal 'H' (*Heu*) do espaço de cores HSV, e dos canais 'R' e 'G' (*Red e Green*) do espaço de cores RGB. No canal 'H', tomamos o seu negativo e realizamos uma normalização discretizada entre a faixa de valores entre 0 e 255, como se segue

$$I'_{Hue}(x,y) = round\left(\frac{\max(I_{Hue}) - I_{Hue}(x,y)}{\max(I_{Hue}) - \min(I_{Hue})} \times 255\right), \tag{3.11}$$

onde $I'_{Hue}(x,y)$ é o canal 'H' (*Hue*) transformado como exemplificado através da Figura 3.9 (a); $I_{Hue}(x,y)$ são as magnitudes originais da matiz (canal 'H'); $\max(I_{Hue})$ é a magnitude máxima do canal; e $\min(I_{Hue})$ é a magnitude mínima do canal. Pelos canais

'R' (*Red*) e 'G' (*Green*) foi construído uma nova variável (*new feature*), como se segue

$$L_H(x,y) = round\left(\frac{G(x,y)}{R(x,y)} \times 255\right), \tag{3.12}$$

$$\text{Se} \quad L_H(x,y) > 255,$$

$$\text{Então} \quad L_H(x,y) = 255; \tag{3.13}$$

onde L_H é a nova variável como exemplificado através da Figura 3.9 (b).

(a) (b)

Figura 3.9: Ilustração das novas variáveis (*new feature*): (a) I'_{Hue}; and (b) L_H.

Similar ao caminho tomado na detecção de pele, nós calculamos o mapa de regiões de alta probabilidade (probabilidades a posteriori) dos *pixels* através da Eq. (3.8) para discriminar os lábios do resto da imagem como exemplificado através da Figura 3.10 (a). Em seguida, os mapas de altas probabilidades dos *pixels* de serem de lábios são reescalados e discretizados entre os valores de 0 a 255, como segue,

$$I_{gray}(x,y) = round(I_{pr}(x,y) \times 255), \tag{3.14}$$

onde $I_{gray}(x,y)$ é a imagem na nova faixa dinâmica de representação. Isto é realizado para determinarmos o limiar (*threshold*) de Otsu que será usado na classificação dos *pixels* em lábios e não-lábios. Portanto, as regiões de lábios detectadas na imagem são obtidas através da aplicação do limiar de Otsu encontrado, como se segue

$$I_{gray}(x,y) > limiar, \tag{3.15}$$

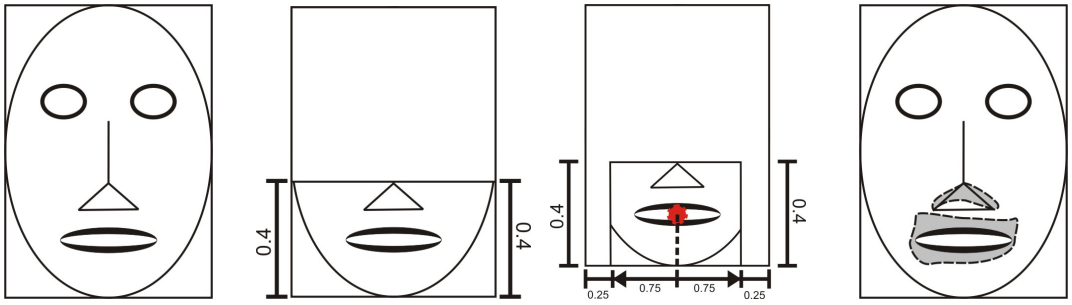
sendo '*x*' e '*y*' as posições dos *pixels* na imagem. Uma exemplificação de resultado obtido em experimento com a aplicação do limiar de Otsu é demonstrado através da Figura 3.10 (b) e (c). De fato, pode-se observar que a maior parte da região dos lábios são efetivamente detectados ao aplicar-se o limiar, porém alguns *pixels* em torno da região dos olhos e nariz são frequentemente detectados ocasionando os falsos positivos. Além disso, alguns *pixels* de lábios podem não ter sidos detectados durante o processo ocasionando os falsos negativos. Portanto, para refinar o resultado de detecção dos lábios é aplicada uma etapa de pós-processamento.

(a) (b) (c)

Figura 3.10: Ilustração da detecção de lábios: (a) regiões de alta probabilidade de serem lábios; (b) imagem binária dos *pixels* classificados com lábios; e (c) *pixels* detectados como lábios.

3.4 Pós-processamento

Na etapa de pós-processamento temos por objetivo refinar o resultado obtido com a detecção dos lábios visando a sua utilização como observáveis que alimentarão o HMM (cadeias ocultas de markov) para detecção visual de atividade de voz. Portanto, o resultado é refinado: eliminando-se alguns *pixels* que foram detectados erroneamente; incluindo-se alguns *pixels* de lábios que porventura não tenham sido detectados; e incluindo-se os *pixels* que estão entre os lábios toda a vez que o locutor abre a boca ao discursar.



(a) (b) (c) (d)

Figura 3.11: Ilustração da localização do ROI: (a) região da face; (b) eliminação da região superior da imagem da face; (c) localização da ROI; e (d) ROI em escala de cinza.

O processo de pós-processamento é iniciado pela localização da região de interesse (ROI) como ilustrado em torno da hachura em cinza da Figura 3.11 (d). Note que desejamos localizar uma região de *pixels* em que a boca do locutor esteja agregada junto a ela. Para isso, utilizamos a imagem binário do resultado anterior da detecção dos *pixels* de lábios sobre a imagem da face como ilustrado na Figura 3.10 (b). Sobre a imagem binária, tomamos 40% da região inferior correspondendo aproximadamente a metade da região de face, ou seja, eliminamos todos os *pixels* fora desta area como demonstrado na Figura 3.11 (b) e no exemplo de resultado na Figura 3.12 (b). Dando seguimento, é calculado o centróide sobre todos os *pixels* que foram detectados como lábios e que se encontram dentro da região delimitada na imagem, ou seja, utilizamos todos *pixels* '*1s*' (ligados) da imagem binário que não foram excluídos ou eliminados anteriormente (deslizados ou trocado seu valor para '*0*'). O centróide é dado por

$$\mu_C = \left(\frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n} \sum_{i=1}^{n} y_i \right) \tag{3.16}$$

onde μ_C é o centróide representado pelo par ordenado; n é o número total de *pixels* '*1s*' da imagem binária; e (x_i, y_i) são as i-ésimas coordenadas dos pixel '*1s*'. O resultado deste processo é bem ilustrado na Figura 3.11 (c) através do ponto vermelho representando o centróide. Após a localização do centróide, eliminamos na imagem 25% da região de

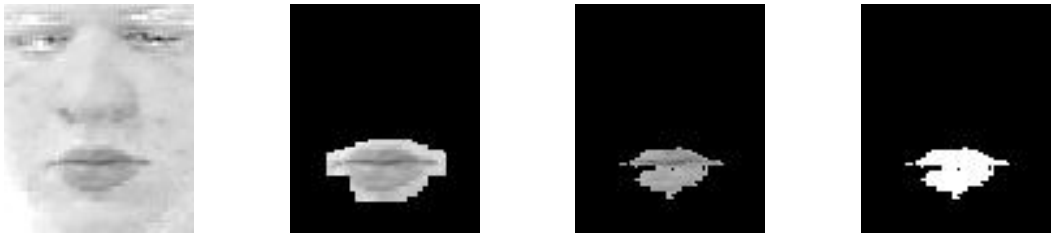
(a) (b) (c)

Figura 3.12: Ilustração da ROI: (a) *pixels* de lábios detectados; (b) eliminação de alguns *pixels*; e (c) região de interesse (região que agrega os *pixels* de lábios).

pixels que se encontram sobre a área do lado esquerdo e do lado direito do centróide, isto é, relacionando a distância do centróide ao seu lado esquerdo e ao direito excluimos ou "*desligamos*" uma percentagem dos *pixels* como desmonstrado através da Figura 3.11 (c). Finalmente, a região de interesse que agrega os *pixels* de lábios é obtido tomando-se os dois maiores componentes conexos dentro da região e aplicando-se a operação morfológica de dilatação usando um operador circular. Essa etapa é ilustrada na Figura 3.11 (d) e um resultado obtido em experimento é demonstrado na Figura 3.12 (c). O número de *pixels* usado pelo operador circular morfológico de dilatação é definido pela quantidade total de *pixels* de pele "$q_{skin} = \sum 1$" detectado sobre a imagem de face, como se segue

Se	$q_{skin} >= 30000,$	*Então*	*usar um operador com 12 pixels;*
Se	$q_{skin} >= 20000,$	*Então*	*usar um operador com 10 pixels;*
Se	$q_{skin} >= 10000,$	*Então*	*usar um operador com 7 pixels;*
Se	$q_{skin} >= 5000,$	*Então*	*usar um operador com 4 pixels;*
Senão	,	*Então*	*usar um operador com 3 pixels.*

Na etapa seguinte do pós-processamento, nós procuramos obter os *pixels* de lábios dentro da região de interesse localizada. Para este processo, aplicamos inicialmente a transformação sobre a imagem da face usando a Eq. (3.12) como demonstrado em resultado na Figura 3.13 (a). Em seguida, tomamos sobre a imagem transformada somente os *pixels* dentro da ROI para realizarmos a detecção dos lábios aplicando-se um limiar. O limiar (*threshold*) é calculado através do método de Otsu (OTSU, 1979) usando como informação apenas os *pixels* dentro da região de interesse da imagem transformada. Então, o limiar é usado para segmentar a imagem transformada segundo a Eq. (3.15). As Figuras 3.13 (a), (b), (c) e (d) demonstram alguns dos resultados experimentais obtidos com esse processo. Depois de alcançada a nova segmentação, nós excluímos todos os *pixels* da metade superior do resultado da imagem original obtida com detecção de lábios antes do pós-processamento e do resultado da nova segmentação obtida durante o pós-processamento. Então, realizamos a união dos resultados, como se segue

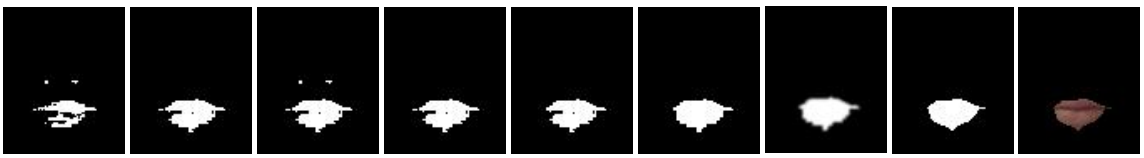


(a) (b) (c) (d)

Figura 3.13: Ilustração dos passos de segmentação dos lábios realizados na etapa de pós-processamento: (a) transformação sobre a imagem; (b) ROI à ser segmentada; (c) região segmentada; and (d) imagem binária da região segmentada.

$$I_U(x, y) = I_{seg}(x, y) \vee I'_{seg}(x, y), \tag{3.17}$$

onde $I_U(x, y)$ é o resultado da união; $I_{seg}(x, y)$ e $I'_{seg}(x, y)$ são os resultados da detecção de lábios original e nova (antes e depois do pós-processamento) representados pelas imagens binárias após a eliminação dos *pixels* descrita anteriormente. Dando seguimento, aplicamos a imagem binária da região de interesse como máscara sobre o resultado da união das imagens para obter-se somente os *pixels* que se encontram dentro daquela região. E finalmente, a região dos *pixels* de lábios é alcançada tomando-se o seu maior componente conexo. Alguns resultados obtidos em experimentos destas operações são demonstrados através da Figura 3.14 (c), (d) e (e).



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figura 3.14: Ilustração das etapas do pós-processamento: (a) segmentação original depois da eliminação da metade superior dos *pixels*; (b) nova segmentação depois da eliminação da metade superior dos *pixels*; (c) união das duas segmentações; (d) resultado da eliminação de todos os *pixels* que estavam fora da ROI; (e) maior componente conexo; (f) lacunas preenchidas; (g) aplicação do filtro gaussino; (h) resultado da suavização; e (i) resultado final.

As etapas anteriores do pós-processamentos são executadas para atingirmos o objetivo de eliminar alguns *pixels* que foram detectados erroneamente e incluir alguns *pixels* de lábios que porventura não foram detectados. Nas próximas etapas buscamos incluir os *pixels* que estão entre os lábios toda a vez que o locutor abre a boca ao discursar, e incluir os *pixels* das possíveis lacunas (buracos) de *pixels* que houverem nos lábios, ou seja, que não foram inclusas durante o processo anterior após a obtenção do resultado do maior componente conexo. A inclusão desses *pixels* é realizada pelo "preenchimento das lacunas" e sua execução é dada pelos seguintes passos:

1. Para cada coluna da imagem binária dos lábios, obter a coordenada 'y' máxima e mínima dos *pixels* de lábios, i.e., $y_{i_{max}}$ e $y_{i_{min}}$ onde $i = 1, ..., n$ colunas. Em

seguida, preencha as lacunas entre os $y_{i_{max}}$ e $y_{i_{min}}$ com demonstrado na Figura 3.15 (b);

2. Para cada linha da imagem binária dos lábios, obter a coordenada 'x' máxima e mínima dos *pixels* de lábios, i.e., $x_{i_{max}}$ e $x_{i_{min}}$ onde $i = 1, ..., n$ linhas. Em seguida, preencha as lacunas entre os $x_{i_{max}}$ e $x_{i_{min}}$ com demonstrado na Figura 3.15 (c);

3. Repita os passos no item 1 e 2 enquanto existir lacunas a ser preenchidas como ilustrado na Figura 3.15 (d).

Observe que no processo de preenchimento das lacunas só existirá inclusão dos *pixels* "*internos a boca*" (os *pixels* entre os lábios) quando a detecção de lábios ocorrer em quadros do vídeo em que o locutor esteja com sua boca aberta, ou seja, nos momentos de intervalo em que as imagens foram capturadas quando o locutor movimentou sua boca ao falar. A etapa final consiste na suavização das bordas após o resultado obtido com

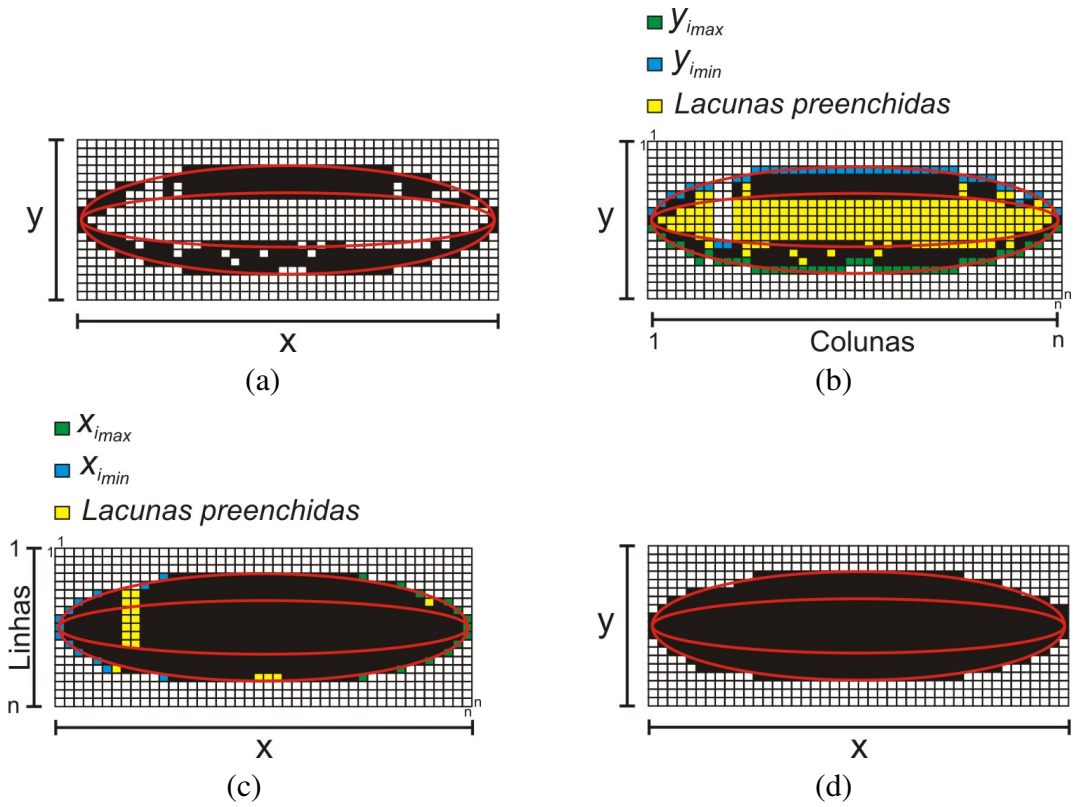


Figura 3.15: Ilustração do preenchimanto das lacunas: (a) *pixels* de lábios detectados; (b) preenchimento da lacunas na direção das colunas; (c) preenchimento da lacunas na direção das linhas; e (d) resultado alcançado depois de lacunas serem preenchidas.

preenchimento das lacunas. A suavização é alcançada pela aplicação de um pequeno borramento nas bordas usando um filtro gaussiano de tamanho 9×9 com variância = 2.0. Ele é aplicado sobre a imagem binária com a região de *pixels* da boca representado por '1s' e a região de *pixels* que não são da boca (complemento) representado por '0s'. Através da aplicação do filtro obtemos como resultado uma imagem com a região dos lábios borrada e cujos os seus valores de *pixels* estão distribuídos entre 0 e 1. Então, para eliminarmos partes das regiões borradas pelo filtro guassiano, e binarizarmos novamente

a imagem para se obter a região da boca suavizada, é aplicado um limiar de 0.5, como se segue

$$I_f = I_{gauss} > 0.5, \tag{3.18}$$

onde I_f é a imagem binária contendo região de *pixels* da boca suavizada; e I_{gauss} é a imagem borrada pelo filtro gaussiano. As Figuras 3.14 (a) até (i) ilustram alguns resultados obtidos experimentalmente nas etapas do pós-processamento até o seu resultado final para obtenção da segmentação da região de *pixels* da boca cujos os quais serão utilizados como variáveis observáveis que alimentarão o HMM na detecção visual de atividade de voz. A Figura 3.16 demonstra um resultado da localização na imagem original da região da boca circundada pela caixa retângula em vermelho (*bounder box*).

Figura 3.16: Localização da boca

3.5 Treinamento dos Parâmetros dos Modelos Estatístico de Cor de Pele e Lábios

Em nossa abordagem usamos o modelo de mistura gaussiana para construir as probabilidades a posteriori nos métodos de detecção de pele e lábios descrito nas Secções 3.2 e 3.3 respectivamente. Para podermos utilizar a mistura de guassianas foi necessário obter os seus parâmetros através de treinamentos. Os parâmetros encontrados nos treinamentos foram: número de componentes de mistura, as contribuições de proporcionalidade de cada gaussiana da mistura, os vetores de médias e a matrizes de covariância.

O treinamento dos parâmetros foi executado através do algoritmo do Figueiredo (FIGUEIREDO; JAIN, 2002) usando um conjunto verdade (*groundtruth*) de imagens rotuladas como ilustrados em exemplos através das Figuras 3.17 (a), (b), (c) e (d).

No modelo para detecção de pele foram usadas as informações de cromatacidades da imagem. Portanto, foram utilizados somente os canais '*a*' e '*b*' do espaço de cores CIELab. No treinamento utilizamos 10 conjunto verdades (*groundtruth*) de imagens para classe fundo (*background*), 100 para classe pele e 100 para classe cabelo.

No modelo para detecção de lábios foram usados as informações de cores baseadas na transformação 'L_H' obtido através da combinação dos canais '*R*'e '*G*' do espaço de cores RGB e as baseadas na transformação I'_{Hue} (negativo do canal) obtida sobre o canal '*H*' do espaço de cores HSV ambos descritos na Secção 3.3 e dados pelas Eq. 3.12 e 3.11 respectivamente. No treinamento utilizamos 100 conjunto verdades (*groundtruth*) de imagens para classe lábios e 100 para classe não-lábios.

(a) (b) (c) (d)

Figura 3.17: Ilustração do conjunto verdade de imagens (*groundtruth*): (a) fundo (*background*); (b) pele; (c) cabelo; e (d) lábios.

3.6 Detecção Visual de Atividade de Voz

A abordagem de detecção visual de atividade de voz explora a expectativa da ocorrência de movimentação dos lábios durante os períodos de fala para distingui-las dos períodos de silêncio. Istó é realizado por meio do uso de dois HMM concorrentes, um para o silêncio e o outro para fala, que utilizam como observáveis as variáveis obtidas com os resultados do metódo de detecção dos lábios descritos na Secção 3.3.

As cadeias ocultas de Markov (HMM) podem ser usadas para modelar sistemas dinâmicos que podem alterar seus estados ao longo do tempo. Um HMM com observáveis discretos é caracterizada através

$$\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}), \tag{3.19}$$

onde $\boldsymbol{A} = [a_{ij}]$ para $1 \leq i, j \leq N$ é a matriz de transição que contém as probabilidades de mudanças de estado; $\boldsymbol{B} = [b_i(O)]$ para $1 \leq i \leq N$ descreve as probabilidades de observação para cada estado; e $\boldsymbol{\pi} = [\pi_i]$ para $1 \leq i \leq N$ contém as probabilidades iniciais de cada estado.

Para a detecção de atividade de voz (VAD), foram usados dois HMM concorrentes. Um HMM para fala caracterizado por $\lambda_{\text{sp}} = (\boldsymbol{A}_{\text{sp}}, \boldsymbol{B}_{\text{sp}}, \boldsymbol{\pi}_{\text{sp}})$ e um HMM para silêncio caracterizado por $\lambda_{\text{si}} = (\boldsymbol{A}_{\text{si}}, \boldsymbol{B}_{\text{si}}, \boldsymbol{\pi}_{\text{si}})$. Para ambos os HMMs, existem dois estados ocultos relacionados com o estado atual da boca, sendo-os, aberta ou fechada.

As variáveis observáveis usadas para alimentar os HMMs são estimadas do tamanho da área da boca que foram obtidas através dos resultados de segmentação alcançados com o metódo de detecção dos lábios. Além disso, elas são normalizadas e discretizadas dentro de N valores para que possam ser usadas pela abordagem. A área da boca é dada por

$$Ar_{mouth} = \sum_{x=1}^{i} \sum_{y=1}^{j} I_{bin}(x, y), \tag{3.20}$$

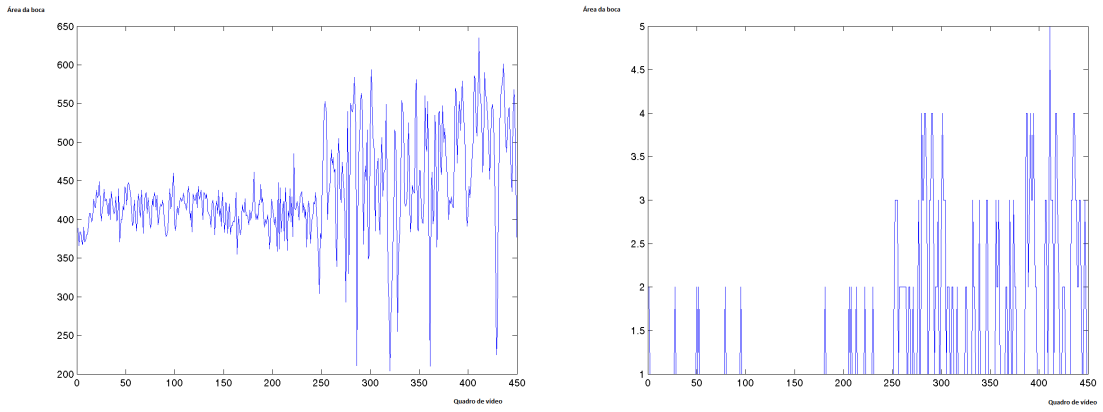
onde $I_{bin}(x, y)$ é a imagem binária do resultado obtido com detecção de lábios. E as variáveis observáveis são dadas através

$$n_{(i)} = \left(\frac{O_{(i)}}{\mu_{ArMouth}}\right) \times \left(\frac{\mu_{ArFace}}{A_{Face_{(i)}}}\right), \tag{3.21}$$

$$\begin{cases} \text{if} \quad n_{(i)} < mn, \text{ then } n_{(i)} = 1; \\ \text{if} \quad n_{(i)} > mx, \text{ then } n_{(i)} = 2; \end{cases}$$

$$O'_{(i)} = 1 + floor\left((N+1) \times \left(\frac{(n_{(i)} - mn)}{(mx - mn)}\right)\right). \tag{3.22}$$

onde $n_{(i)}$ é a i-ésima normalização da variável observável; $O_{(i)}$ é a i-ésima variável observável; $\mu_{ArMouth}$ e μ_{ArFace} são as médias de área da boca e face respectivamente; $A_{Face_{(i)}}$ é a i-ésima área da face; mx e mn é o valor máximo e mínimo que pode ser assumido na normalização (i.e., "o teto e o piso"); N é valor máximo que pode assumir a discretização; e $O'_{(i)}$ é a i-ésima variável observável normalizada e discretizada. Os parâmetros $\mu_{ArMouth}$ e μ_{ArFace} são obtidos calculando-se a média da área da boca e da face dos 15 primeiros quadros (*frames*) de vídeo tomados como informação de dimensão e escala iniciais da boca. O parâmetro N indica o número máximo de valores discretizados que as variáveis de observação podem assumir, sendo este parâmetro configurado para o valor 10 (obtido experimentalmente) neste trabalho. E os parâmetros mx e mn foram configurados para 2 e 1 também obtidos experimentalmente. As Figuras 3.18 (a) e (b) ilustram um resultado alcançado com as variáveis observáveis.



(a) (b)

Figura 3.18: Ilustração dos observáveis: (a) observáveis não normalizados e não discretizados; e (b) observáveis normalizados e discretizados.

Dado o HMM de fala λ_{sp} e o HMM de silêncio λ_{si}, e dada uma sequência de observações $\mathcal{O}_t = \{O(t-T), O(t-T+1), ..., O(t)\}$ com uma janela temporal de tamanho T, podemos calcular a probabilidade de $\mathcal{O}_t$ que cada HMM representa. Mas precisamente, é realizado pela computação de $P(\mathcal{O}_t; \lambda_{\text{sp}})$ e $P(\mathcal{O}_t; \lambda_{\text{si}})$ usando o algoritmo *Forward-Backward* (RABINER, 1989). Portanto, baseado em uma janela temporal de tamanho T, um dado t é classificado como silêncio se $P(\mathcal{O}_t; \lambda_{\text{sp}}) < P(\mathcal{O}_t; \lambda_{\text{si}})$, em caso contrário, é classificado como fala. Além disso, é importante ressaltar os efeitos da janela temporal T sobre o resultado. A aplicação de "altos" valores para T tendem a produzir resultados que são mais temporalmente coerente, mas também acarretam maiores atrasos (*delays*) para detectar a mudança de estado de silêncio para fala ou de fala para silêncio (desde que as transições apresentem as mesmas observações utilizadas tanto no HMM de fala e quanto

no de silêncio). Por outro lado, valores "menores" para T resultam em menores atrasos para detectar as transições, porém são mais suscetíveis ao ruído. Por isso, utilizamos um $T = 10$ que corresponde a aproximadamente um 1 segundo para sequências de vídeos que utilizem 10 ou 15 quadros por segundo.

3.6.1 Treinamento dos Parâmetros para a Detecção Visual de Atividade de Voz

Para obtenção dos paramêtros de ambos HMMs, silêncio e fala, um conjunto verdade (*ground truth*) de sequência de vídeos foram usadas, onde cada quadro (*frame*) foi manualmente rotulado como silêncio ou fala. Dentro desta sequência, cada conjunto de T quadros adjacentes marcados como fala ou silêncio foram usados para construir um conjunto de treinamento para cada HMM. O algoritmo Baum-Welch (RABINER, 1989) foi usado para obter as matrizes $\boldsymbol{A}$ e $\boldsymbol{B}$ para ambos os modelos, enquanto o vetor inicial de probabilidades $\boldsymbol{\pi}$ foi considerado uniforme. Mais especificamente, utilizamos 3 vídeos de 15 quadros por segundo para forma um sequência de aproximadamente 2 minutos de duração que foram usados para treinar os HMMs. Cada um dos vídeos continha apenas uma pessoa em cena que alternava entre os períodos de fala e silêncio durante a sua duração. Um exemplo de sequência de observáveis construída para o treinamento é demonstrado na Figura 3.19.

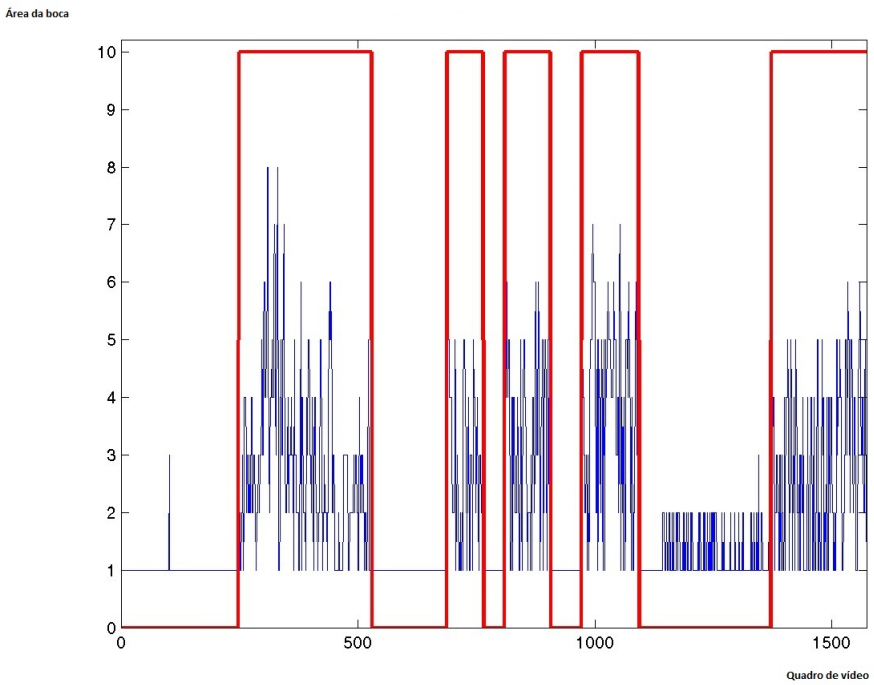


Figura 3.19: Ilustração da sequência de observáveis geradas para o treinamento: em azul a sequência de observações geradas e em vermelho o conjunto verdade (i.e, a rotulação em fala ou silêncio).

4 RESULTADOS EXPERIMENTAIS

Neste capítulo para validarmos os resultados obtidos em experimentos foram gerados para cada vídeo o seu conjunto verdade de sequência de vídeo onde cada quadro (*frame*) foi manualmente rotulado como silêncio ou fala. Por meio destes, foram obtidos em experimentos pela soma total de quadros a quantidade deles que eram Verdadeiros Positivos (TP), Verdadeiros Negativos (TN), Falsos Positivos (FP) e Falsos Negativos (FN) em comparação ao seu conjunto verdade. Por fim, foram calculados através das medidas anteriores a taxa de verdadeiros positivos (TPR), a taxa de verdadeiros negativos (TNR) e a Acurácia Total (OAcc), como se segue

$$TPR = 100 \times \left(\frac{TP}{(TP + FN)} \right), \tag{4.1}$$

$$TNR = 100 \times \left(\frac{TN}{(TN + FP)} \right), \tag{4.2}$$

$$OAcc = 100 \times \left(\frac{(TP + TN)}{(TP + TN + FP + FN)} \right). \tag{4.3}$$

Nos experimentos foram testados 3 tipos de casos: ideais, não ideais e com mais de um usuário em cena. Os casos ideais ou em condições normais são aqueles em que não há oclusões da boca, não há movimentão brusca da cabeça, não há movimentação dos lábios sem que não haja fala (som), não há risos, o usuário não fica de perfil ao falar e não há tentativas de enganar o sistema. Os casos não ideais ou em condições adversas são as situações opostas aos casos ideais. O caso em que há mais de um usuário em cena segue os mesmos padrões do caso ideal com a diferença de serem avaliados dois usuários em cena ao mesmo tempo.

Para os experimentos foram gerados 16 vídeos em laboratório, onde, 10 vídeos foram gravados para os testes de casos ideais com um usuário em cena, 5 vídeos para os casos não ideiais e 1 vídeo para o caso em há duas pessoas cenas. Em geral, o conjunto de vídeos gerados continha as seguintes características:

- Grupos de vídeo com a resoluções de 240x320 (4 vídeos gerados), 640x480 (7 vídeos gerados) e 960x720 (4 vídeos gerados);
- Duração de 30 segundos;
- E grupos com as taxas de 15 e 10 quadros por segundo.

Para os casos ideais ou em condições normais foram usados 10 vídeos para testes. Os resultados destes experimentos são apresentados através da Tabela 4.1. A taxa média de acurácia total foi de 88.65% em relação aos 10 vídeos usados. Os resultados individuais na Tabela 4.1 revelaram que:

- 5 dos vídeos obtiveram suas taxas de acurácia acima de 90.00% (vídeos 1, 2, 3, 4 e 10);
- 4 dos vídeos obtiveram suas taxas de acurácia entre 80.00% a 89.00% (vídeos 5, 6, 7 e 8);
- 1 vídeo obteve taxa inferior a 80.00% (vídeo 9).

Estas pequenas variações entre o resultados obtidos ocorrem devido os HMMs (silêncio e fala) utilizarem as variáveis de observação geradas a partir da segmentação da boca realizada no primeiro momento do método. Então, se a segmentação da boca não se mantiver estável, consistente e acurada isso acarretará influências diretas no resultado da detecção visual da atividade de voz gerando como consequência possíveis flutuações e erros na detecção. Baseado nesta observação, podemos afirmar que os vídeos simulados que alcançaram uma taxa de acurácia acima de 80.00% obtiveram uma sequência de segmentação consistente, estável e acurada. Exemplos em trecho de sequência de quadros de 2 vídeos testados em condições ideais e com uma segmentação de boca consistente, estável e acurada ao longo da duração da captura é demonstrado pelas imagens na Figura 4.1.

Tabela 4.1: Resultados obtidos com a simulação de vídeos em condições ideais.

	Resultados em condições ideais						
	TP	***TN***	***FP***	***FN***	***TPR (%)***	***TNR (%)***	***OAcc (%)***
Vídeo 1	192	248	0	10	95.05	100.00	97.78
Vídeo 2	176	248	0	26	87.13	100.00	92.22
Vídeo 3	136	270	22	22	86.08	92.47	90.22
Vídeo 4	293	143	9	5	98.32	94.08	96.89
Vídeo 5	166	226	44	14	92.22	83.70	87.11
Vídeo 6	115	222	15	48	70.55	93.67	84.25
Vídeo 7	188	194	4	64	74.60	97.98	84.89
Vídeo 8	224	150	6	70	76.19	96.15	83.11
Vídeo 9	193	152	103	2	98.97	59.61	76.67
Video 10	311	109	14	16	95.11	88.62	93.33
Média	*199.40*	*196.20*	*21.7*	*27.7*	*87.42*	*90.63*	*88.65*

Em contraste, o vídeo 9 obteve o pior resultado com uma taxa de acurácia 76.67% em decorrência de falhas de segmentação ao longo de algumas sequências de quadros de vídeo ocasionando sua baixa estabilidade, consistência e acurácia. Um exemplo de trecho da sequência de quadros com a falha de segmentação que acarretaram a baixa acurácia são demonstrado pelas imagens na Figura 4.2. Portanto, podemos afirmar que o aumento da taxa de acurácia depende da consistência de segmentação da boca ao longo da sequência de quadros do vídeo, uma vez que as variáveis de observação que alimentam os HMMs são oriundas do tamanho da área segmentada.

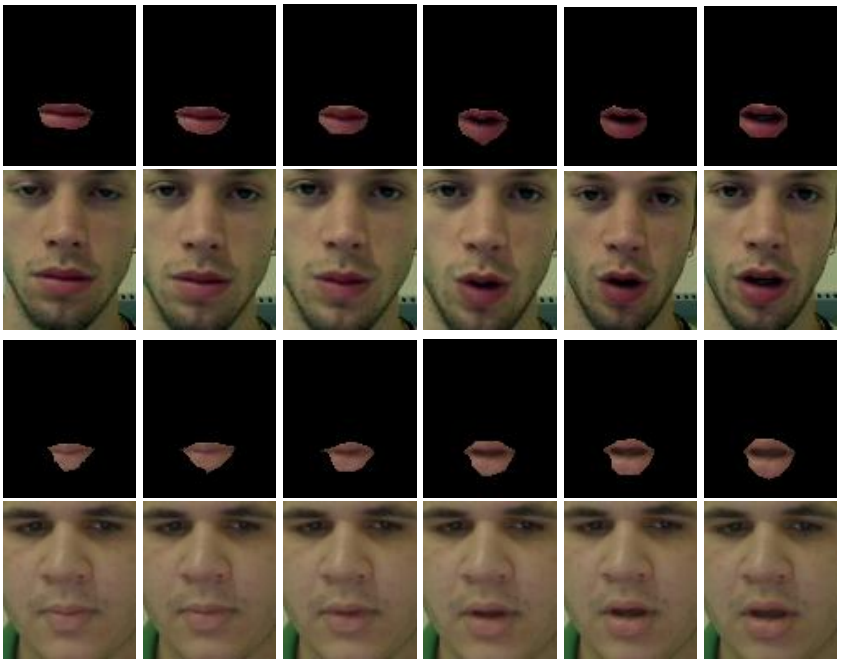


Figura 4.1: Ilustração de 2 trechos de sequência de quadros da simulações de vídeos em casos ideais e suas respectivas segmentações da boca.

Para os casos não ideais ou em condições adversas foram usados 5 vídeos para testes. Os resultados destes experimentos são apresentados através da Tabela 4.2. A taxa média de acurácia total foi de 52.71% em relação aos 5 vídeos usados. Os resultados individuais na Tabela 4.2 revelaram que:

- 2 dos vídeos obtiveram suas taxas de acurácia acima de 60.00% (vídeos 3a e 4a);
- 3 vídeos obtiveram suas taxas inferiores a 60.00% (vídeos 1a, 2a e 5a).

Estes resultados demonstram um baixo desempenho já esperado porque a solução abordada pelo método não trata oclusões, não soluciona o problema de falsas detecções de atividade de voz quando o locutor move os seus lábios sem dizer nada (i.e., enganando a solução) e não distingui a fala de risos. É importante salientar que no vídeo 3a obtivemos

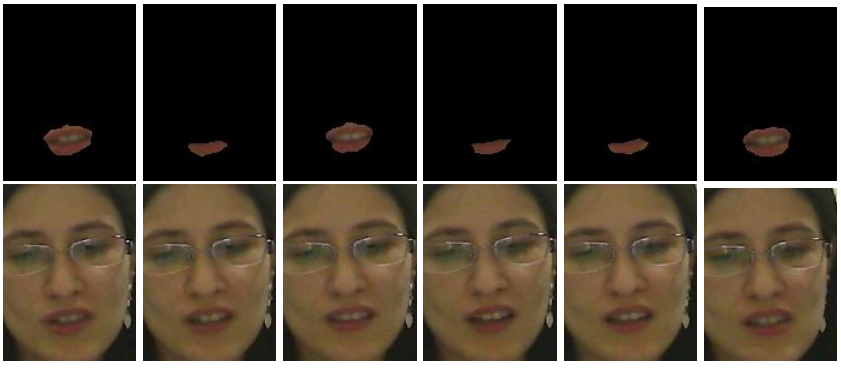


Figura 4.2: Ilustração de um trecho de uma sequência de quadros com falha de segmentação.

uma boa taxa de acurácia com 89.78% de acerto, isso decorreu porque nessa simulação pedimos para o voluntário (usuário) apenas alternar entre silêncio e risos durante a

captura. Já nos outros vídeos, pedimos para voluntários alternarem entre fala, silêncio, oclusões, "falsas falas" (i.e., enganar o sistema) e risos o que resultou na baixa acurácia apresentada nos vídeos 1a, 2a, 4a e 5a pela Tabela 4.2. Dessa forma, nosso objetivo foi averiguar o comportamento do método de VVAD quando introduzido às situações ou restrições que não foram tratadas na solução. Uma vez que, os HMMs que são alimentados pelas variáveis de observação que apenas informam as mudanças de estados de silêncio para fala e vice-versa e são baseadas na mudanças de picos de área da boca que foi segmentada sem quaisquer informação sobre o efeito de oclusões e risos. Além disso, se movimentarmos os lábios da boca simulando uma fala e não dizermos nada (sem som) as variáveis de observação geradas a partir desta situação se assemelharão às observações geradas nos casos em que a locutor de fato está falando ocasionando uma confusão para o sistema, i.e., o usuário estará enganando o método que não possui a informação de áudio que poderia ajudar a confirmar a ocorrência de atividade de voz. Essas observações se tornam mais claras através das Figuras 4.5 (a), (b) e (c) que demonstram o resultado da classificação dos vídeos 1, 2 e 3 para os casos ideais onde observamos claramente que a atividade de voz é detectada quando há "fortes oscilações" de picos de área da boca.

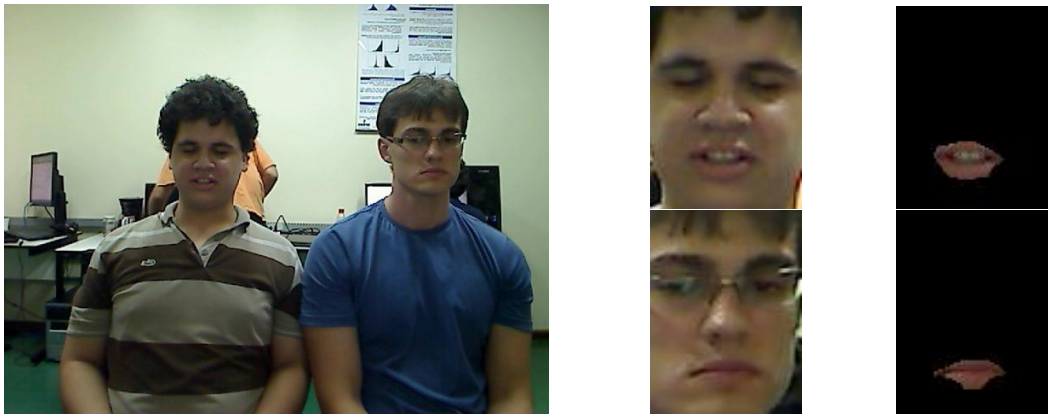
Tabela 4.2: Resultados obtidos com a simulação de vídeos em condições não ideais (adversas).

	Resultados em condições não ideiais (adversas)						
	TP	***TN***	***FP***	***FN***	***TPR (%)***	***TNR (%)***	***OAcc (%)***
Vídeo 1a	27	144	64	215	11.16	69.23	38.00
Vídeo 2a	128	89	193	40	76.19	31.56	48.22
Vídeo 3a	0	404	46	0	0.00	89.78	89.78
Vídeo 4a	133	148	82	87	60.45	64.35	62.44
Vídeo 5a	76	37	194	143	34.70	16.02	25.11
Média	*72.80*	*164.40*	*115.80*	*97.00*	*36.50*	*54.19*	*52.71*

Para os casos com mais de um usuário em cena foi usado um vídeo com 2 usuários para os testes. Neste experimento, o detector de faces é usado para distinguir um usuário do outro sendo processado a VVDA sobre cada face detectada independemente uma da outra. A Figura 4.3 (a) demonstra um quadro da cena do vídeo testado no experimento e as Figura 4.3 (b) e (c) retratam suas faces detectadas. Os resultados deste experimento são apresentados através da Tabela 4.3. A taxa média de acurácia total foi de 93.11% em relação aos 2 usuários em cena. Os resultados da Tabela 4.1 de cada usuário em cena revelaram que:

- Ambos os usuários obtiveram uma taxa de acurácia acima de 90.00%;
- O 1° usuário obteve uma taxa de acurácia de 93.56%;
- O 2° usuário obteve uma taxa de acurácia de 92.67%.

Através dos resultados desses experimentos procuramos demonstrar que o método VVAD pode ser aplicado quando há a ocorrência de mais de um usuários em cena. O que possibilita através do detecção de atividade de voz identificar o locutor ativo da conversa e manter o foco sobre ele.



(a) (b) (c)

Figura 4.3: Ilustração da segmentação da boca para mais de uma pessoa em cena: (a) quadro do vídeo; (b) faces detectadas; e (c) bocas segmentadas.

Tabela 4.3: Resultados obtidos com a simulação do vídeo em condições ideais e com mais de um usuários em cena

	Resultados alcançados						
	TP	***TN***	***FP***	***FN***	***TPR (%)***	***TNR (%)***	***OAcc (%)***
Video 1b - 1° person	190	231	17	12	94.06	93.15	93.56
Video 1b - 2° person	164	254	26	7	95.88	90.71	92.67
Média	*177*	*242.50*	*21.50*	*9.50*	*94.97*	*91.93*	*93.11*

As situações testadas em experimentos nos permitiram observar o comportamento da detecção visual de atividade voz sobre diferentes situações. Além disso, podemos validar o método através das taxas de verdadeiros positivos, verdadeiros negativos e principalmente pela acurácia total obtidos com cada vídeo testado. Através desse resultados confirmamos que a idéia de explorar a informação visual da movimentação dos lábios esperada durante fala é justificável. No entanto, a detecção visual de atividade de voz ("fala") somente pode ser garantida sem o uso da informação de áudio quando a incidência de movimentação dos lábios sem que a pessoa esteja realmente falando (i.e, falsos discursos - Enganar) é baixa. Portanto, a detecção de atividade de voz utilizando a informação visual somente é válida se assumirmos que todas as vezes que uma pessoa "abre sua boca", "ela está realmente falando alguma coisa". A Figura 4.4 ilustra bem o fato onde a informação visual pode ser efetiva para VAD, desde que a movimentação dos lábios seja altamente correlacionada com atividade de voz (i.e., a fala usualmente é acompanhada de movimentação labial) e que os períodos de fala coincidam com maior movimentação em torno da área da boca.

4.1 Discussão dos Resultados Experimentais

Tradicionalmente, o processo de detecção de atividade de voz (VAD) envolve a identificação dos períodos de quando um pessoa está em silêncio ou falando usando somente

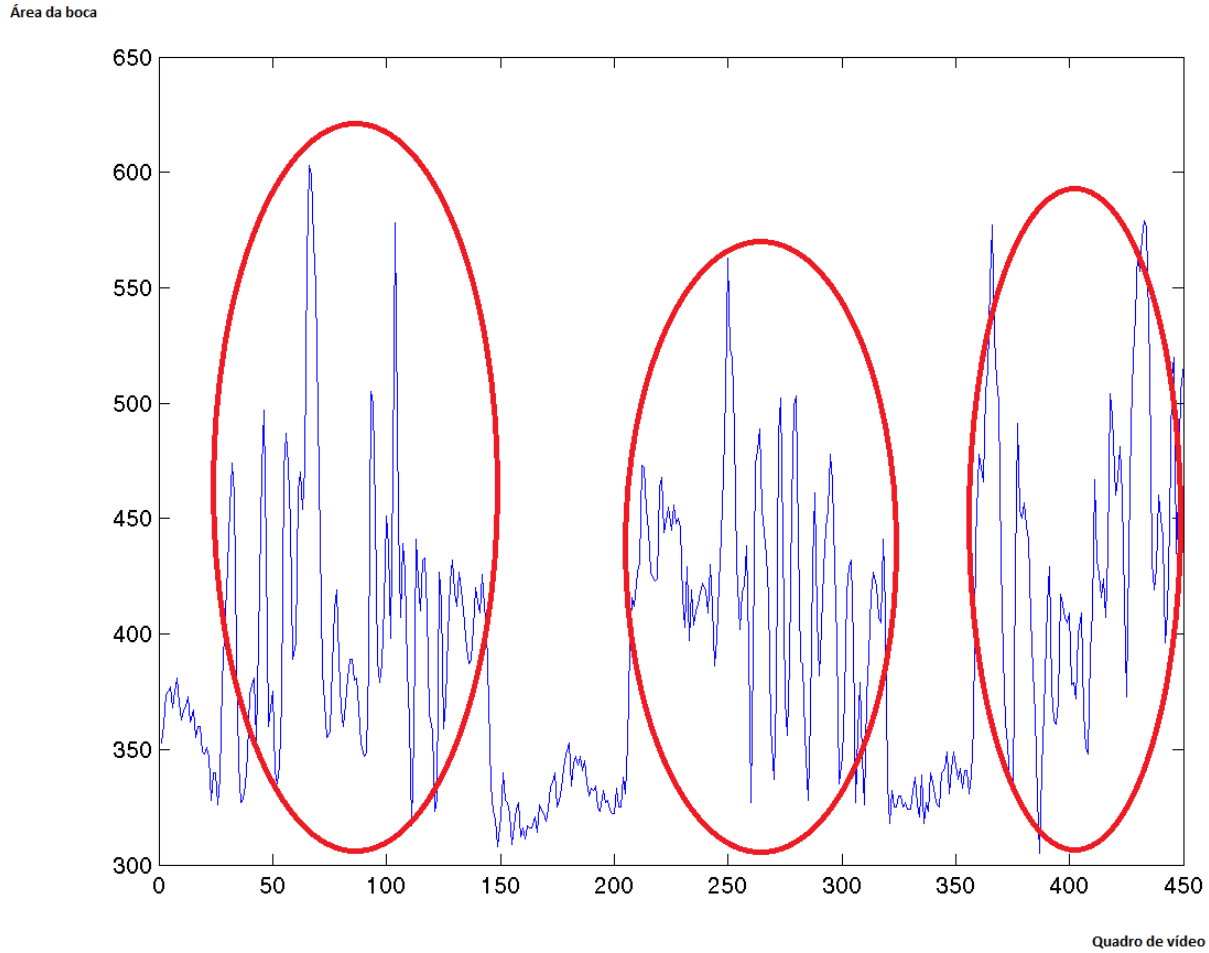


Figura 4.4: Ilustração do gráfico de observações da área da boca ao longo de um vídeo: a área dentro das elipses vermelhas indicam os períodos de fala e área fora indicam os períodos de silêncio.

a informação acústica (sonora) (AUBREY; HICKS; CHAMBERS, 2010). A acurácia de tais métodos para muitos casos é dependente das origens acústicas do ambiente em que elas são aplicadas. Por exemplo, a competividade de origens acústicas ou ruídos não estacionários como sons de fundo providas da circulação de carros, pessoas caminhando ou movimentando-se, sons de dispositivos eletrônicos, dentre outros (AUBREY; HICKS; CHAMBERS, 2010) (SODOYER et al., 2006). Neste trabalho foi desenvolvido um método de detecção visual de atividade de voz (VVAD) que não é suscetível aos problemas associados com os VADs baseados na informação de áudio. Isso possibilita o seu uso em ambiente acusticamente ruídosos (AUBREY; HICKS; CHAMBERS, 2010).

O trabalho abordado explora a informação visual de movimentação dos lábios para identificar a ocorrência de atividade de voz. Ele detecta os períodos de fala ou silêncio pela ausência ou ocorrência de movimentação dos lábios usando dois HMM's, um para silêncio e outro para fala. A movimentação dos lábios são dadas pelas oscilações de área da boca obtida após a segmentação como demonstram os gráficos das Figuras 4.5, onde, os gráficos representados pela cor em preto indicam as variáveis de observação (área da boca); os gráficos representados pela cor em azul indicam os conjuntos verdades (i.e., *groundtruth*); e os gráficos em vermelho indicam os resultados da classificação.

Os experimentos realizados afirmaram que a detecção de atividade de voz (VAD) baseada na informação visual pode ser alcançada pela utilização e exploração da informação de movimentação labial baseado na mudança de área da segmentação da boca. Este fato confirma a correlacionalidade da movimentanção dos lábios com a produção de som ou fala do ser humano. Uma vez que os lábios estão envolvidos com o complexo processo de produção sonora pela vibração do ar que é realizado pelas nossas cordas vocais que conseguem produzir diferentes sons através da movimentação da língua atrelado a abertura da boca (movimentação dos lábios) que permite a passagem do ar para a formação das ondas sonoras (ou acústicas).

No primeiro conjunto de experimentos testamos alguns vídeos que obedeciam as condições de restrição do algoritmo de detecção de atividade de voz. Esse primeiro conjunto de testes tiveram por objetivo averiguar o comportamento, a eficácia e a acurácia da abordagem proposta. Através dos resultados alcançados nos testes pudemos constatar que:

- O método proposto possui uma boa acurácia apresentando uma taxa média acima de 85,00 %;
- A detecção de atividade de voz apresentada classifica o período avaliado como fala ou silêncio pela competividade probabilística dado através da maior probabilidade resultante dos HMM's de fala e silêncio, ou seja, ela não utiliza limiares para distinguir o sinal avaliado como fala ou silêncio;
- A abordagem não é suscetível aos problemas associados com VADs baseados na informação de áudio, pois utiliza informação visual;
- O algoritmo não necessita de marcações manuais do contorno dos lábios ou da boca para sua inicialização;
- A simplicidade do método proposto pode permite que sua implementação seja voltada para aplicativos em tempo real;
- As segmentações obtidas ao longo do vídeo tem influência direta no resultado da classificação do VAD, pois a área da segmentação é utilizada como variável de observação;
- O uso de um pequeno intervalo de tempo no início do vídeo em que o usuário precisa ficar em silêncio e com boca fechada para a tomada autonôma das medidas iniciais da boca (i.e., área da boca e tamanho de face iniciais do usuário) se mostrou válido para realização da normalização das variáveis de observação usadas nos HMM's, pois ajustam o tamanho e a escala inicial da boca do usuário.

No segundo conjunto de experimentos testamos alguns vídeos que não obdeciam as condições de restrição e que tentavam enganar o algoritmo. Esse segundo conjunto de testes tiveram por objetivo averiguar o comportamento do sistema em situações que não foram abordada pelo método. Através desses resultados testados pudemos constatar que:

- A aplicação do método em condições imprevistas acarreta uma baixa acurácia com vários erros de detecção de atividade de voz, principalmente em relação a classificação de fala, como esperado;
- Oclusões afetam a performance do sistema de forma imediata acarretando erros, pois é perdida a informação visual da movimentação dos lábios;

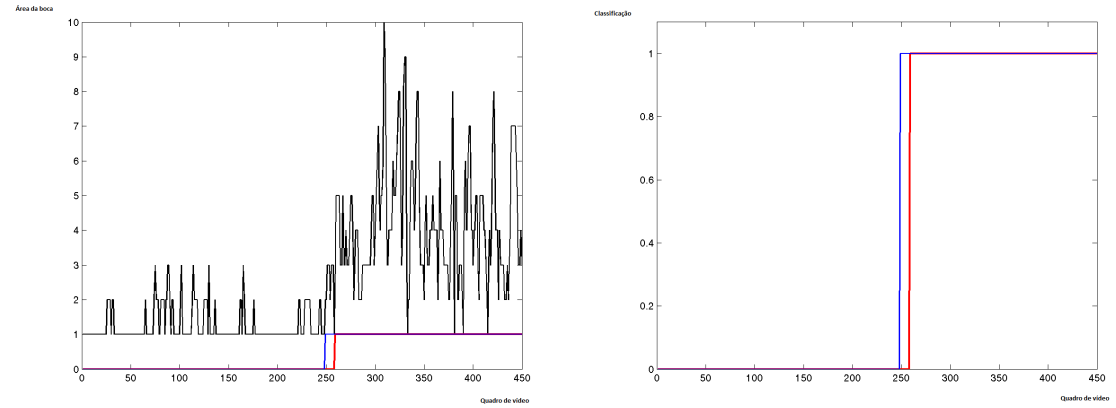
- Mudanças de posicionamento do estado frontal para perfil acerratam erros ou classificações não previstas, pois a abordagem não trata a situação em que o usuário está de perfil para câmera;

- A tomada errada das medidas de tamanho da face e da boca no início do processo que são usadas para a normalização das variáveis de observação ocasionam erros de classificação no sistema;

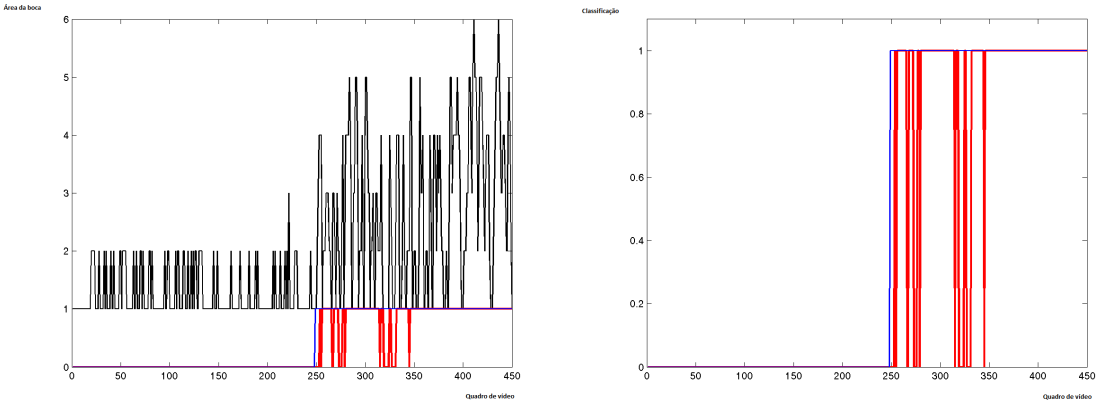
- Os movimentos bruscos da cabeça acarretam em erros de segmentação da boca tendo como consequência erros de classificação do VAD nesses intervalo de tempo em que a ação está sendo executada.

No terceiro e último conjunto de experimentos testamos a possibilidade de detecção de atividade de voz para múltiplos usuários em cena. Nesse teste averiguamos que o método pode ser aplicado isoladamente a cada usuário em cena utilizando o detector de faces para distiguir cada usuário em cena obtendo-se resultados similares ao primeiro conjunto de experimentos para cada usuário detectado.

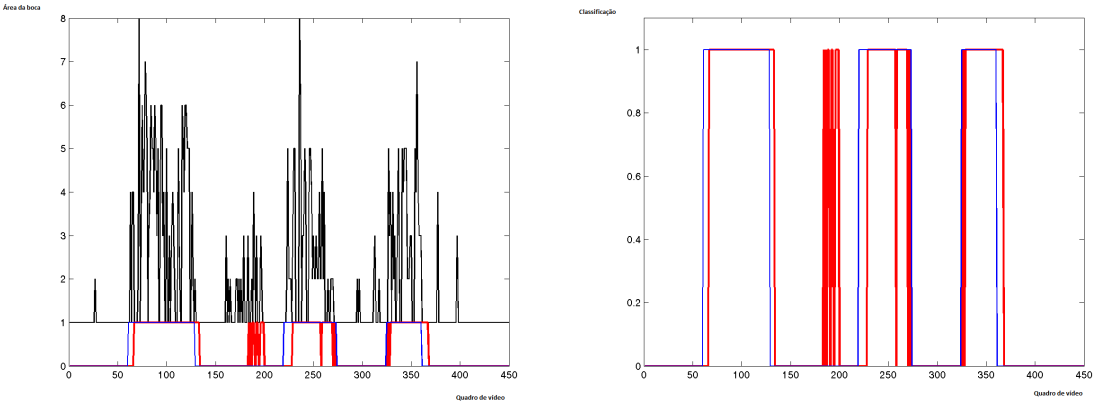
Todos esses conjuntos de experimentos citados acima exploraram diferentes cenários e situações do cotidiano que possibilitaram realizar a avaliação da solução proposta e encontrar possíveis falhas ou pontos fracos na estratégia tomada. Os experimentos, em geral, confirmaram que o proposta abordada baseada na informação visual é eficaz e pode ser usada para detecção de atividade de voz com a vantagem de ser imune aos problemas acarretados em abordagem baseada na informação sonora.



(a)



(b)



(c)

Figura 4.5: Ilustração do resultado da VVAD: em *azul* o conjunto verdade, em *vermelho* o resultado da classificação, e em *preto* as observações - (a) vídeo 1; (b) vídeo 2; e (c) vídeo 3.

5 CONCLUSÕES

Este trabalho apresentou uma abordagem de extração dos movimentos labiais baseado em cores objetivando a detecção de atividade de voz que se utiliza da segmentação da boca para gerar as variáveis de observação usada no HMM para a realização da detecção. Na extração da boca, o estimador Bayesiano foi empregado para obter-se uma pré-segmentação dos lábios. A detecção da boca alcançada por uma nova segmentação sobre a região de interesse seguida de um pós-processamento. Já na detecção visual de atividade de voz, o resultado das sequências de quadros dos vídeos processados, segmentações da região da boca sobre as imagens da face humana detectadas, foram usados como variáveis de observação para alimentar dois estados do HMM que exploraram as consistências temporais da detecção dos lábios a fim de detectar os períodos de silêncio ou fala (VAD). Por este meio, o objetivo de explorar a informação visual como abordagem eficaz de detecção de atividade de voz foi alcançado. Uma vez que, a informação carregada pelo movimento dos lábios é altamente correlacionada com o "processo de produção de voz do locutor", ou seja, quando uma pessoa fala se faz necessário articular a boca para proferir um som (i.e., os lábios se movimentam à permitir a entrada e saída do ar que produzirá o som pela sua vibração através das cordas vocais localizada na garganta e pela língua). Este fato pode ser observado pelo gráfico da Figura 4.4 que ilustra o aumento significativo de picos e oscilação da área da boca quando há períodos de ocorrência de fala.

Os resultados experimentais demostraram que a abordagem proposta por meio da idéia de usar a informação visual da movimentação dos lábios como discriminante da fala ou silêncio é plausível e eficaz, como ilustrado pelas taxas de acurácia apresentadas no Capítulo 4. Sobre essa constatação, a abordagem de detecção visual de atividade de voz permitiria-nos usá-la junto com o método clássico de VAD que utiliza somente o som como informação discriminante. Assim, os métodos unificados poderiam se ajudar onde qualquer um deles pudessem falhar durante o processo de detecção de atividade de voz ao longo do tempo de duração da captura de vídeo e som. Além disso, o método proposto pode ser aplicado com múltiplos usuários em cena permitindo-se executar múltiplas detecções de atividade de voz através das faces detectadas que são usadas para localizar as pessoas em cena como exemplificado em experimento no Capítulo 4. Este atributo também abre-nos a possibilidade de executar o reconhecimento e a localização em 2-dimensões do locutor ativo ou locutores ativos em cena. E é importante salientar que o método só funcionará adequadamente se obtivermos as características individuais de cada usuário, ou seja, temos que obter as informação iniciais de tamanho e escala da boca no inicio da captura do video para que possamos realizar a normalização das váriaveis de observação. Essa restrição gera uma certa incomodação ao(s) usuário(s), pois o(s) mesmo(s) terão de ficar parados e em silêncio por uns poucos segundos para que suas características

individuais sejam tomadas. Assim como não permitirá a entrada de novos usuários em cena sem que o passo inicial seja reiniciado. Então, obtemos no início do método a medida de tamanho que é obtida da média da dimensão de áreas iniciais da boca no estado de silêncio e a medida de escala que é obtida da média das áreas da face sendo ambas as medidas utilizadas como parâmetros de normalização como descrito na Secção 3.6 do Capítulo 3. Observe que a etapa inicial de obtenção desses parâmetros é muito importante por influenciar diretamente o cálculo da normalização das variávies de observação obtidas ao longo da captura dos quadros do vídeo. E, se esses parâmetros obtidos não condizem com as características do usuário, i.e., contém erros, esses erros serão propagados para as observações normalizadas, uma vez que são elas que alimentam os HMMs que estimam os períodos de silêncio ou fala.

Os problemas detectados que afetam diretamente os resultados acarretando uma baixa taxa de precisão (erros de detecção) são as falhas de segmentação, oclusões da boca e movimentos bruscos da cabeça, como demonstraram os resultados de alguns experimentos realizados no Cap. 4. As falhas de segmentação sucessivas e aleatórias ao longo de diferentes sequências de quadros da duração do vídeo podem prejudicar o desempenho da solução abordada porque usamos a informação do tamanho da área da boca que foi segmentado como variáveis observáveis pelo HMM. As oclusões prejudicam a solução proposta porque o método perde a sua referência ou observação, ou seja, não obtem mais as variáveis observáveis que alimentam o HMM. E como não são tratados as oclusões no método de VVAD, isso irá acarretar falhar de detecção nos períodos em que elas surgirem. Os movimentos bruscos de posição da cabeça provocam falhas de segmentação porque introduzem ruídos e borramento em alguns quadros no momento do movimento prejudicando a imagem a ser segmentada. Por fim, é importante salientar que o método foi desenvolvido para uma perspectiva frontal da face (i.e., o usuário está voltada de frente para câmera) não tendo uma abordagem ou tratamento para os casos em que o usuário esteja de perfil durante ou em algum momento da captura do vídeo. Como consequência possíveis erros devem ser gerados no VVAD nos momentos que o usuário esteja de perfil durante o vídeo.

5.1 Trabalhos Futuros

Como trabalhos futuros temos que encontrar um caminho que possamos realizar a normalização das variáveis de observação sem que seja preciso impor ao usuário alguns minutos iniciais do vídeo para que se tome os parâmetros de normalização. Essa questão poderia ser resolvida se mudássemos o tipo de medida da abertura de boca que utiliza o seu tamanho de área, sendo-o uma medida que pode variar de 0 a $+\infty$, para uma medida cuja a região de valores é delimitada como o espaço de medidas de ângulos. Outras melhorias que podem ser realizadas no método seriam o aumento da precisão de estimação da movimentação dos lábios através do uso de contornos ativos que se ajustassem às bordas internas e externas dos lábios usando como informação os resultados obtidos com a segmentação e detecção da boca. Uma outra seria a inserção da consistência temporal por métodos probabilísticos que predizem os resultados futuros objetivando aumentar a confiança pelas medidas tomadas da movimentação dos lábios podendo-se detectar as possíveis falhas de estimação de medidas em períodos de quadros computados. Além disso, poderia-se solucionar parte do problema das oclusões pela consistência temporal para os casos em que o usuário por algum motivo coloca o mão sobre a boca ou movimenta o braço passando pela frente da cabeça em curtos períodos de tempo na cena. E,

modelar uma proposta de solução para os casos em que os usuários mudam a sua posição que está de frente em relação a câmera para a posição de perfil.

Para questões de aplicabilidade como método final de uso para o mercado poderíamos realizar uma forma de unificar as técnicas de detecção de atividade de voz, som e visual (imagem), objetivando uma maior confiança nos seus resultados e um melhor desempenho. Além disso, faria-se justificável desenvolver através do método unificado de detecção de atividade de voz um localizador em 2D do locutor ou dos locutores ativos em cena com intuito de focalizar a câmera autonomamente sobre ele(s).

6 PUBLICAÇÕES E CONTRIBUIÇÕES

ARTIGOS PUBLICADOS:

1. ***Shading Attenuation in Human Skin Color Images***. Com: CAVALCANTI, P.; SCHARCANSKI, J.; LOPES, C.. Em: BEBIS, G. e colegiado (Ed.). ***Advances in Visual Computing***. Springer Berlin Heidelberg, 2010. p.190-198.

   (CAVALCANTI; SCHARCANSKI; LOPES, 2010) O artigo publicado apresenta um novo método automático para atenuar a degradação de cor devido a sombras em imagens coloridas de pele humana. As sombras são ocasionadas pela variação de iluminação através da cena devido as mudanças na orientação local da superfície, condições de iluminação, dentre outros fatores. Na abordagem é estimado a variação de iluminação modelando-a através de uma função quadrática que altera a iluminação em torno dos *pixels* de pele com uma simples operação. Então, o resultado obtido é aplicado na solução de dois problemas típicos, na segmentação de lesão de pele pigmentada e na detecção de faces, para ajudar na redução de complexidade do problema de análise de imagens coloridas para aquelas aplicações.

2. ***Color-based lips extraction applied to voice activity detection***. Com: LOPES, C.; GONÇALVES, A. L.; SCHARCANSKI, J.; JUNG, C. R.. Em: 18º ***International Conference on Image Processing*** - IEEE, 2011. p.1057-1060.

   (LOPES et al., 2011) O artigo publicado apresenta uma nova abordagem de detecção de visual de atividade de voz baseado em cores para extração da informação de movimentação labial. Para extração dos lábios é realizado primeiramente a segmentação de pele para reduzir a área de busca para então ser detectado as regiões mais prováveis de serem lábios pela abordagem Bayesiana. Então, a extração dos lábios é obtida pela aplicação da limiarização das regiões de probabilidades e pela aplicação de simples operadores morfológicos. Finalmente, o movimento temporal dos lábios é dado pelas suas dimensões de altura sendo usado os modelos ocultos de Markov (HMM) para detectar as prováveis ocorrências de atividade de fala sobre uma determinada janela temporal.

3. ***Audiovisual Voice Activity Detection Based on Microphone Arrays and Color Information***. Com: MINOTTO, Vicente P.; LOPES, Carlos B. O.; SCHARCANSKI, J.; JUNG, C.; LEE, BOWON. Em: Journal of Selected Topics in Signal Processing (J-STSP), 2012.

(MINOTTO et al., 2012) O artigo publicado aborda a detecção de atividade de voz pela informação audio e visual. A detecção áudio-visual de atividade de voz é uma etapa necessária para solução de vários problemas, tais como teleconferência, reconhecimento de voz e interação humano-computador. O movimento dos lábios e análise de áudio fornecem uma grande quantidade de informações que podem ser integrados para produzir sistemas mais robusto de detecção áudio-visual de atividade de voz (VAD). A movimentação dos lábios é muito útil para detectar o orador, sendo essa característica explorada na nova abordagem apresentada no artigo para detecção visual de atividade de voz (VVAD). Primeiro, é realizado uma ajuste sobre as cores para diminuir a complexidade e torná-las mais próximas dos modelos utilizados. Em seguida, o algoritmo executa a segmentação de pele para reduzir a área de busca de extração dos lábios e as regiões mais prováveis de serem lábios e não-lábios são computadas usando uma abordagem Bayesiana sobre a região delimitada. Extraídos os lábios, os seu movimento é detectado pelo uso dos modelos ocultos de Markov (HMM) que estimam a probabilidade de ocorrência de atividade de fala dentro de uma janela temporal. A informação de áudio é capturado por um conjunto de microfones sendo-a usada para detecção de atividade de voz (VAD) baseado no som que está relacionada com a busca de fontes espaço-temporalmente coerentes de som através de um outro conjunto de HMMs. Então, para aumentar a robustez do sistema proposto de detecção de atividade de voz, uma fusão final das estratégias é utilizada para combinar o resultado de cada modalidade (áudio e vídeo) obtendo-se resultados experimentais que indicaram que a abordagem proposta audio-visual apresenta melhores resultados quando comparado com os algoritmos individuais existentes para VAD.

ARTIGOS SUBMETIDOS:

1. ***A New Approach for Recognizing Lip Motion in Visual Voice Activity Detection***. Com: LOPES, Carlos; SCHARCANSKI, Jacob; JUNG, Claudio. Em: ***Speech Communication***, 2013.

   (LOPES; SCHARCANSKI; JUNG, 2012) O artigo submetido apresenta uma nova abordagem para extração dos lábios baseada na informação de cores sendo-o aplicado a detecção visual de atividade de voz. O movimento dos lábios é uma informação visual relevante para detectar o locutor ativo e para o reconhecimento de voz. Ao movimentar os lábios passamos a idéia de ocorrência de uma possível fala ou períodos de discursos para um observador, enquanto os períodos de silêncios podem ser representados pela ausência de seu movimento. Na abordagem aplicamos para cada quadro de vídeo as correções de cores para as mudanças locais de iluminação em relação ao modelo de cores utilizado no treinamento sobre a iluminução ambiente conhecida. Em seguida, a detecção dos lábios é realizada sobre o quadro com as cores ajustadas. No algoritmo de detecção executamos a segmentação da pele e a detecção de face para reduzir a área de busca para a extracção dos lábios. E através da abordagem Bayesiana obtemos dentro da área delimitada as regiões mais prováveis de serem lábios. Então, é realizado a pré-segmentação dos lábios pela limiarização das regiões de probabilidade calculada. Em seguida, localiza-se a região da boca pelo resultado obtido, isto é, alguns *pixels* de não-lábios são

eliminados e é aplicado uma simples operações morfológicas para incluir *pixels* com uma margem de segurança que garantam a inclusão dos *pixels* da boca. Então, uma nova segmentação dos lábios é realizada sobre a região da boca depois de ser aplicado uma transformação de cor para realçar a região de lábios a ser segmentada. E, é aplicado o preenchimento das lacunas internas da segmentação dos lábios e a suaviazação das bordas obtendo-se como resultado final a boca. Finalmente, a detecção visual de atividade de voz é obtida pela detecção da movimento temporal dos lábios explorando-se os modelos ocultos de Markov (HMM) para detectar as prováveis ocorrência de atividade de voz dentro de uma janela temporal.

REFERÊNCIAS

AOKI, M. et al. Voice activity detection by lip shape tracking using EBGM. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 15., New York, NY, USA. **Proceedings...** ACM, 2007. p.561–564.

AUBREY, A. et al. Two Novel Visual Voice Activity Detectors based on Appearance Models and Retinal Filtering. In: EUROPEAN SIGNAL PROCESSING CONFENRENCE. **Proceedings...** [S.l.: s.n.], 2007. v.1, p.2409–2413.

AUBREY, A.; HICKS, Y. A.; CHAMBERS, J. Visual voice activity detection with optical flow. **IET image processing**, [S.l.], v.4, n.6, p.463–472, December 2010.

BIGUN, J. **Vision with Direction**. [S.l.]: Springer, 2006.

BINS, J. et al. Feature-Based Face Tracking for Videoconferencing Applications. In: IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA, 2009., Washington, DC, USA. **Proceedings...** IEEE Computer Society, 2009. p.227–234. (ISM '09).

CAVALCANTI, P. G.; SCHARCANSKI, J.; LOPES, C. B. O. Shading attenuation in human skin color images. In: ADVANCES IN VISUAL COMPUTING - VOLUME PART I, 6., Berlin, Heidelberg. **Proceedings...** Springer-Verlag, 2010. p.190–198. (ISVC'10).

CHIN, S. W.; ANG, L.-M.; SENG, K. P. Lips detection for audio-visual speech recognition system. In: INTELLIGENT SIGNAL PROCESSING AND COMMUNICATIONS SYSTEMS, 2008. ISPACS 2008. INTERNATIONAL SYMPOSIUM ON. **Anais...** [S.l.: s.n.], 2009. p.1 –4.

DARGHAM, J.; CHEKIMA, A. Lips Detection in the Normalised RGB Colour Scheme. In: INFORMATION AND COMMUNICATION TECHNOLOGIES, 2006. ICTTA '06. 2ND. **Anais...** [S.l.: s.n.], 2006. v.1, p.1546 –1551.

EVENO, N.; CAPLIER, A.; COULON, P.-Y. New color transformation for lips segmentation. In: MULTIMEDIA SIGNAL PROCESSING, 2001 IEEE FOURTH WORKSHOP ON. **Anais...** [S.l.: s.n.], 2001. p.3 –8.

FIGUEIREDO, M.; JAIN, A. Unsupervised learning of finite mixture models. **Pattern Analysis and Machine Intelligence, IEEE Transactions on**, [S.l.], v.24, n.3, p.381 – 396, mar 2002.

FINLAYSON, G. D.; HORDLEY, S. D. Color constancy at a pixel. **Journal of the Optical Society of America A**, [S.l.], v.18, n.2, p.253–264, 2001.

GONZALEZ, R. C.; WOODS, R. E. **Processamento de Imagens Digitais**. [S.l.]: Edgard Blücher, 2003.

HORDLEY, S. D. et al. Illuminant and device invariant colour using histogram equalisation. **Pattern Recognition**, [S.l.], v.38, p.2005, 2005.

JAHNE, B.; HAUSSECKER, H.; GEISSLER, P. **Handbook of Computer Vision and Applications. Volume 1. Sensors and Imaging**. [S.l.]: Academic Press, 1999.

LIU, D.; YU, J. Otsu Method and K-means. In: HYBRID INTELLIGENT SYSTEMS, 2009. HIS '09. NINTH INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2009. v.1, p.344 –349.

LOPES, C. B. O.; SCHARCANSKI, J.; JUNG, C. R. A New Approach for Recognizing Lip Motion in Visual Voice Activity Detection. **Speech Communication**, [S.l.], December 2012.

LOPES, C. et al. Color-based lips extraction applied to voice activity detection. In: IMAGE PROCESSING (ICIP), 2011 18TH IEEE INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2011. p.1057 –1060.

MINOTTO, V. P. et al. Audiovisual Voice Activity Detection Based on Microphone Arrays and Color Information. **IEEE Journal of Selected Topics in Signal Processing**, [S.l.], December 2012.

OTSU, N. A Threshold Selection Method from Gray-Level Histograms. **Systems, Man and Cybernetics, IEEE Transactions on**, [S.l.], v.9, n.1, p.62 –66, jan. 1979.

PETSATODIS, T.; PNEVMATIKAKIS, A.; BOUKIS, C. Voice activity detection using audio-visual information. In: DIGITAL SIGNAL PROCESSING, 16., Piscataway, NJ, USA. **Proceedings...** IEEE Press, 2009. p.216–220. (DSP'09).

RABINER, L. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. **Proceedings of the IEEE**, [S.l.], v.77, n.2, p.257–286, February 1989.

ROHANI, R. et al. Lip segmentation in color images. In: INNOVATIONS IN INFORMATION TECHNOLOGY, 2008. IIT 2008. INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2008. p.747 –750.

SALAZAR-JIMÉNEZ, A. et al. Feature extraction and lips posture detection oriented to the treatment of CLP children. In: IEEE EMBS ANNUAL INTERNATIONAL CONFERENCE, 28., New York, NY, USA. **Anais...** [S.l.: s.n.], 2006. p.5747–5750.

SODOYER, D. et al. An Analysis of Visual Speech Information Applied to Voice Activity Detection. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING. **Proceedings...** [S.l.: s.n.], 2006. v.1, p.I–I.

SODOYER, D. et al. A study of lip movements during spontaneous dialog and its application to voice activity detection. **The Journal of the Acoustical Society of America**, [S.l.], v.125, n.2, p.1184–1196, 2009.

SZELISKI, R. **Computer Vision**: algorithms and applications. 1st.ed. New York, NY, USA: Springer-Verlag New York, Inc., 2010.

WANG, L.; WANG, X.; XU, J. Lip Detection and Tracking Using Variance Based Haar-Like Features and Kalman filter. In: FRONTIER OF COMPUTER SCIENCE AND TECHNOLOGY (FCST), 2010 FIFTH INTERNATIONAL CONFERENCE ON. **Anais...** [S.l.: s.n.], 2010. p.608 –612.

WEBB, A. R. **Statistical Pattern Recognition, 2nd Edition**. [S.l.]: John Wiley & Sons, 2002.

YANG, M.-H.; AHUJA, N. Gaussian mixture model for human skin color and its application in image and video databases. In: ITS APPLICATION IN IMAGE AND VIDEO DATABASES. PROCEEDINGS OF SPIE '99 (SAN JOSE CA. **Anais...** [S.l.: s.n.], 1999. p.458–466.

YAO, H.; GAO, W. Face detection and location based on skin chrominance and lip chrominance transformation from color images. **Pattern Recognition**, [S.l.], v.34, n.8, p.1555 – 1564, 2001.

APÊNDICE A ARTIGOS PUBLICADOS