CENTRO UNIVERSITÁRIO FEI
MARCELO GONZAGA DE OLIVEIRA PARADA

BIOMETRIA MULTIMODAL BASEADA NOS SINAIS DE VOZ E FACIAL

São Bernardo do Campo
2018

MARCELO GONZAGA DE OLIVEIRA PARADA

BIOMETRIA MULTIMODAL BASEADA NOS SINAIS DE VOZ E FACIAL

Tese de Doutorado apresentada ao Centro Universitário FEI, como parte dos requisitos necessários para obtenção do título de Doutor em Engenharia Elétrica. Orientada pelo Prof. Dr. Ivandro Sanches.

São Bernardo do Campo

2018

Parada, Marcelo.
BIOMETRIA MULTIMODAL BASEADA NOS SINAIS DE VOZ E FACIAL / Marcelo Parada. São Bernardo de Campo, 2018.
110 f. : il.

Tese - Centro Universitário FEI.
Orientador: Prof. Dr. Ivandro Sanches.

1. Biometria. 2. Biometria multimodal. 3. Verificação automática de locutor. 4. Detecção visual de atividade de voz. 5. Identificação da região labial. I. Sanches, Ivandro, orient. II. Título.

Elaborada pelo sistema de geração automática de ficha catalográfica da FEI com os dados fornecidos pelo(a) autor(a).

centro universitário FEI

APRESENTAÇÃO DE TESE
ATA DA BANCA EXAMINADORA

Programa de Pós-Graduação Stricto Sensu em Engenharia Elétrica

Doutorado

PGE-10

Aluno: Marcelo Gonzaga de Oliveira Parada **Matrícula:** 514103-1

Título do Trabalho: Biometria multimodal baseada nos sinais de voz e facial.

Área de Concentração: Processamento de Sinais e Imagens

Orientador: Prof. Dr. Ivandro Sanches

Data da realização da defesa: 22/02/2018

ORIGINAL ASSINADA

Avaliação da Banca Examinadora

São Bernardo do Campo, 22 / 02 / 2018.

MEMBROS DA BANCA EXAMINADORA	
Prof. Dr. Ivandro Sanches	Ass.: _____
Prof. Dr. Carlos Eduardo Thomaz	Ass.: _____
Prof.ª Dr.ª Andréa Duarte Carvalho de Queiroz	Ass.: _____
Prof. Dr. Miguel Arjona Ramírez	Ass.: _____
Prof. Dr. Hae Yong Kim	Ass.: _____

A Banca Examinadora acima-assinada atribuiu ao aluno o seguinte:

APROVADO ☒ REPROVADO ☐

VERSÃO FINAL DA TESE

ENDOSSO DO ORIENTADOR APÓS A INCLUSÃO DAS RECOMENDAÇÕES DA BANCA EXAMINADORA

_____

Aprovação do Coordenador do Programa de Pós-graduação

_____

Prof. Dr. Carlos Eduardo Thomaz

Dedico este trabalho a minha esposa Flávia Gil que me acompanha em todos os meus caminhos

AGRADECIMENTOS

Ao orientador Prof. Dr. Ivandro Sanches por toda contribuição e suporte ao longo do desenvolvimento deste trabalho.

A minha esposa Flávia Gil pelo carinho e conforto nos momentos mais difíceis.

Aos meus pais Márcio e Sílvia por sempre me incentivarem e apoiarem a chegar cada vez mais longe.

Ao meu irmão Felipe pelas conversas e reflexões.

Aos amigos mais próximos que se envolveram indiretamente no desenvolvimento deste trabalho.

Ao centro Universitário FEI por acreditar no meu trabalho e por ter disponibilizado todo apoio necessário para o desenvolvimento das pesquisas.

A todos os demais que estiveram presentes de alguma forma durante esta etapa da minha vida.

RESUMO

Um sistema biométrico consiste no uso de informações biológicas ou comportamentais para reconhecimento de indivíduos, aplicadas em propósitos de segurança, acesso automático e ciência forense. Sua confiabilidade depende diretamente da qualidade da captura dos dados e precisão da etapa de processamento de sinal, seja ele um sinal de áudio, vídeo, imagem ou outras sequências temporais. Um dos principais desafios é a captura do sinal para ser utilizado na etapa de reconhecimento, já que algumas modalidades biométricas podem ser comprometidas dependendo da influência de fatores externos. Por exemplo, um sistema de identificação por imagem pode falhar se a luz ambiente não for adequada durante a captura e o desempenho de um sistema de reconhecimento por voz pode ser severamente degradado na presença de ruído ambiente. Até mesmo o simples incorreto posicionamento do usuário perante a localização do sensor biométrico pode ser um fator prejudicial para o processamento das informações e, por este motivo, o uso de modalidades biométricas baseadas em múltiplas características biológicas ou comportamentais, conhecidas como multimodais, vêm sendo aplicadas de forma a conferir maior robustez ao sistema. Esta tese propõe a combinação de características de movimento da região facial, especificamente da região labial, através da aplicação da Transformada Discreta dos Cossenos (DCT) aos vetores de movimento de um vídeo MPEG, em conjunto com características extraídas do sinal de voz, resultando em: um método para detecção de atividade de voz e remoção de silêncio; fusão de parâmetros extraídos do movimento e do áudio para finalidade de verificação automática de locutor; um método para extração da região labial baseado na média do movimento ao longo do tempo. A proposta faz uso de parâmetros já presentes em vídeo codificado em MPEG, eliminando a necessidade da etapa do cálculo dos parâmetros de movimento. Os testes biométricos foram realizados com uso da base de dados XM2VTS em diversas condições de relações de sinal-ruído no áudio e avaliados seguindo protocolo Lausanne. O desempenho do sistema foi comparado com diferentes propostas de biometria multimodal, obtendo resultados promissores para utilização em aplicações comerciais.

Palavras-chave: Biometria. Biometria multimodal. Verificação automática de locutor. Detecção visual de atividade de voz. Identificação da região labial.

ABSTRACT

A biometric system consists on the usage of biological or behavioural information for individual recognition being applied for security, automatic access and forensic science. Its reliability is directly related to the quality of the acquired data and precision of the signal processing, being the signal an audio, video, image or other time series. One of the major challenges is the acquisition of the signal to be used for recognition since some biometric modalities can be compromised depending on the influence of external factors. For example, an identification system based on image can fail if the ambience light is not adequate during the capture, the performance of a voice based recognition system can be severely degraded in the presence of background noise, or even the simple incorrect positioning of the user in relation to the location of the biometric sensor can be a harmful factor for the correct processing of the information. Therefore, biometric modalities based on multiple biological or behavioural information, known as multimodal biometrics, are being applied in order to provide greater robustness to the system. This thesis proposes the combination of motion features from the facial region, especially the lip region, with employment of the Discrete Cosine Transform (DCT) to the motion vectors of an MPEG video together with acoustic features, resulting in: a method for voice activity detection and silence removal; fused motion and audio features for automatic speaker verification; a method for lip region extraction based on the mean of the motion over time. The proposal makes use of parameters already present in MPEG encoded video, eliminating the need for the motion feature computation step. The biometric tests were performed with XM2VTS database under various signal-to-noise ratios in the audio and evaluated following the Lausanne protocol. The system performance was compared with different multimodal biometric proposals obtaining promising results for use in commercial applications.

Keywords: Biometrics. Multimodal biometrics. Automatic speaker verification. Visual voice activity detection. Lip-region detection.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AdaBoost	*Adaptive Boost*
ASI	*Automatic Speaker Identification*
ASV	*Automatic Speaker Verification*
AVC	*Advanced Video Coding*
AVI	*Audio Video Interleave*
CSRC	*Conversational Systems Research Center*
CVSSP	*Center for Vision Speech and Signal Processing*
DCT	*Discrete Cosine Transform*
DET	*Detection Error Tradeoff*
DV	*Digital Video*
EER	*Equal Error Rate*
EM	*Expectation Maximization*
FA	*Factor Analysis*
FAE	*False Acceptance Evaluation*
FAR	*False Acceptance Rate*
FFmpeg	*Fast Forward MPEG*
FRE	*False Rejection Evaluation*
FRR	*False Rejection Rate*
FTA	*Failure to Aquire*
FTE	*Failure to Enroll*
GMM	*Gaussian Mixture Model*
GOP	*Group of Pictures*
HEVC	*High Efficiency Video Coding*
HMM	*Hidden Markov Model*
HTER	*Half Total Error Rate*
IDCT	*Inverse Discrete Cosine Transform*
IEC	*International Electrotechnical Commission*
ISO	*International Organization for Standardization*
i-vector	*Identity-vector*
JFA	*Joint Factor Analysis*
JPEG	*Joint Photographic Experts Group*

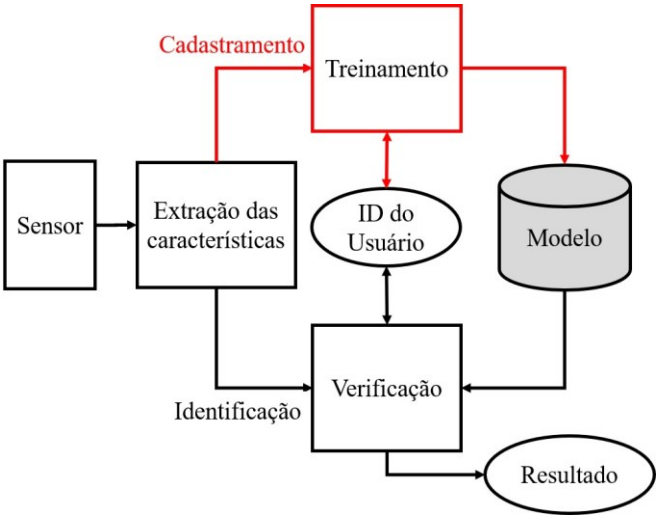
LLR	*Log-Likelihood Ratio*
MAP	*Maximum a posteriori*
MATLAB®	*MATrix LABoratory*
MFCC	*Mel-Frequency Cepstral Coefficients*
ML	*Maximum-likelihood*
MP3	*MPEG layer 3*
MPEG	*Moving Picture Experts Group*
MSR	*Microsoft Research*
PCM	*Pulse Coding Modulation*
PLDA	*Probabilistic Linear Discriminant Analysis*
Quadro B	*Bipredictive Frame*
Quadro I	*Intra Frame*
Quadro P	*Predictive Frame*
RLE	*Run Length Encoding*
RMSE	*Root Mean Squared Error*
ROC	*Receiver Operating Characteristic*
SNR	*Signal-to-noise Ratio*
TAR	*True Acceptance Rate*
TER	*Total Error Rate*
TRR	*True Rejection Rate*
UBM	*Universal Background Model*
VAD	*Voice Activity Detection*
VLC	*Variable Length Coding*
VQ	*Vector Quantization*
VR	*Verification Rate*
VVAD	*Visual Voice Activity Detection*
WAV	*WAVeform Audio format*
XM2VTS	*Extended Multi Modal Verification for Teleservices and Security applications*

SUMÁRIO

1 INTRODUÇÃO

Biometria consiste na medição de informações biológicas ou comportamentais que contém parâmetros suficientes para distinção de um indivíduo dentro de um grupo maior de indivíduos observados, sendo amplamente utilizado por sistemas computacionais para esta finalidade [1], [2], [3]. Há mais de um século, diferentes modalidades de biometria têm sido estudadas e aplicadas para propósitos de segurança e acesso automático, sendo as mais utilizadas comercialmente baseadas nas seguintes informações biométricas: impressão digital [4], [5]; face [6], [7]; íris [7]; voz [8], [9]; assinatura [10]; palma da mão [11]. Além de diferentes modalidades propostas nos últimos anos, baseadas em outras informações, por exemplo: padrão de digitação [12]; padrão de caminhada [13], [14]; orelha [15]; veias da mão [16]; vasos da retina [17]. Um sistema biométrico pode ser genericamente descrito conforme figura 1.

Figura 1 - Sistema biométrico genérico



Fonte: Autor
Legenda: fases de cadastramento e identificação de um sistema biométrico.

Inicialmente o sinal é capturado por um sensor que pode ser um microfone, uma câmera, leitor de impressão digital, entre outros. Posteriormente, o sinal capturado é convertido para um sinal digital e processado de modo a extrair os parâmetros ou características biométricas que serão analisadas, sendo esta etapa conhecida como *front-end* do sistema. No *back-end* da fase de cadastramento um modelo correspondente ao usuário é extraído e armazenado. Já na fase de verificação os parâmetros extraídos de um novo dado de entrada são comparados com o modelo previamente cadastrado verificando a correspondência com a identidade do usuário. As técnicas de análise diferem pelo tipo de sinal usado no processo, que pode ser originário de uma variedade

de fontes, tais como: imagem, vídeo, áudio ou outras sequências temporais que são capturadas pelo sensor biométrico.

1.1 MOTIVAÇÃO DA TESE

A confiabilidade do sistema biométrico depende diretamente da precisão da etapa de processamento de sinal, sendo um dos maiores desafios sua escolha e captura para ser utilizado na etapa de reconhecimento e, por esta razão, alguns tipos de biometria podem ser comprometidos dependendo de parâmetros externos, por exemplo: posicionamento do usuário não adequado perante a localização do sensor biométrico; luz ambiente não adequada para uma boa captura de imagem ou vídeo; presença de ruído ambiente que afeta o desempenho de um sistema de reconhecimento por voz. Desta forma, sistemas de biometria multimodais propõem a combinação de mais do que um tipo de sinal para superar as limitações de modalidades baseadas em apenas uma única informação [18], [19], [20]. Quando utilizadas em conjunto, uma modalidade pode compensar a outra, resultando em sistemas com maior robustez e confiabilidade.

Identificação ou verificação automática de locutor, que é o reconhecimento de indivíduos pelo sinal de voz é uma modalidade biométrica já estudada há muitos anos e utilizada em aplicações comerciais nas últimas décadas [8], [9], porém, ainda requer aprimoramento, principalmente relacionado à dificuldade da identificação quando o áudio é capturado sobre a influência de ruído ambiente, reverberação, entre outros fatores. Por este motivo, esta tese propõe o uso de novas técnicas para sistemas de verificação automática de locutor com a combinação de características do movimento da região facial, extraídos especialmente da região dos lábios, com características acústicas do sinal de voz, resultando em um reconhecimento mais confiável e garantindo um menor número de falsa aceitação ou falsa rejeição de usuários ao sistema. As principais contribuições resultantes desta tese são descritas na seção seguinte.

1.2 CONTRIBUIÇÕES DA TESE

As contribuições desta tese têm como finalidade propor novos métodos para aplicação na etapa de *front-end* de sistemas de verificação e identificação de locutor, sendo elas:

a) Nova técnica para detecção audiovisual de atividade vocal e remoção de trechos de silêncio de um áudio, baseada nos vetores de movimento de vídeo codificado em MPEG e treinamento baseado no sinal de áudio. Podendo ser utilizada para seleção dos

melhores instantes temporais para extração de parâmetros biométricos, resultando em uma melhoria do desempenho do sistema de reconhecimento pelo sinal de voz na presença de ruído ambiente. Apresentada no capítulo 7;

b) Combinação de sinais de movimento da região labial, utilizando vetores de movimento presentes em vídeo codificado em MPEG com parâmetros extraídos do sinal de voz, de forma a superar limitações destas modalidades quando utilizadas individualmente para finalidade de identificação e verificação de locutor. Esta técnica propõe ainda a redução da dimensionalidade dos parâmetros com aplicação da DCT-II, bidimensional, serialização zig-zag e seleção dos parâmetros extraídos das matrizes de movimento vertical e/ou horizontal. Apresentada no capítulo 8;

c) Nova técnica de uso dos vetores de movimento para detecção da região labial, baseada na média do movimento ao longo do tempo. Visa facilitar a identificação da região labial para aplicação das técnicas de extração das características de movimento labial utilizadas em sistemas de identificação e verificação de locutor, bem como em sistemas de reconhecimento de fala baseados em vídeo. Apresentada no capítulo 9.

1.3 ESTRUTURA DA TESE

Os capítulos 2 a 6 apresentam as ferramentas necessárias para o entendimento das propostas inovadoras desta tese que são descritas nos capítulos 7, 8 e 9. O capítulo 2 contém uma revisão bibliográfica para sistemas de reconhecimento de locutor, apresentando as técnicas de extração de parâmetros, modelagem e biometria multimodal. O Capítulo 3 apresenta técnicas para reconhecimento facial, reconhecimento labial e extração de parâmetros destas regiões para aplicações biométricas. O capítulo 4 descreve partes importantes dos codificares MPEG, como a estimativa e compensação de movimento que são utilizados nas propostas da tese. O capítulo 5 apresenta as métricas para avaliação de um sistema biométrico e o capítulo 6 a base de dados utilizada nos testes e seu protocolo de utilização. Os demais capítulos (7, 8 e 9) apresentam as contribuições desta tese, conforme descritas na seção anterior

1.4 NOTAÇÕES MATEMÁTICAS

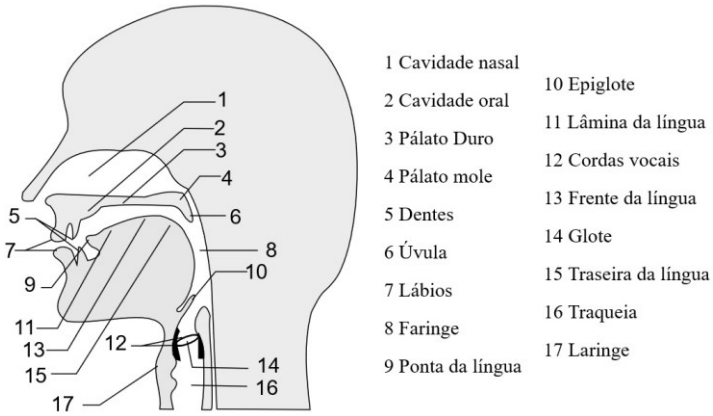
De forma a facilitar a leitura deste texto, todas as equações matemáticas foram padronizadas utilizando as seguintes notações: vetores são representados com letra minúscula, negrito; matrizes por letra maiúscula e negrito; constantes por letra maiúscula e itálico. Todos

os vetores são representados como vetores coluna e o correspondente vetor linha é o transposto do vetor, representado com índice sobrescrito T. Os elementos dos vetores são apresentados em letra minúscula e itálico e endereçados com índice subscrito, por exemplo, x_i é o i-ésimo elemento do vetor **x**. Um intervalo definido de elementos de um dado vetor é endereçado na forma x_i, $i = 1, \ldots, N$ com intervalo igual a um elemento quando não especificado de outra forma. Similarmente para matrizes os elementos são endereçados com dois índices subscritos, sendo o primeiro para linhas e o segundo para colunas, ou seja, um elemento da matriz **X** é representado como $x_{i,j}$. Matriz transposta é representada com o sobrescrito T, a matriz inversa com o sobrescrito -1 e matriz identidade por **I**.

2 RECONHECIMENTO AUTOMÁTICO DE LOCUTOR

A capacidade que nós, seres humanos, temos em reconhecer a identidade de vozes familiares decorre de peculiaridades nas características acústicas da fala de cada indivíduo, que estão diretamente relacionadas a formação particular do trato vocal, que compreende a estrutura desde a laringe até os lábios, incluindo a cavidade nasal e a cavidade oral, conforme figura 2. Este conjunto atua como um filtro, que modifica o espectro de frequências da onda de ar que vem dos pulmões durante a produção da fala e, embora tenha suas características constantemente modificadas durante a produção de diferentes fonemas (devido ao posicionamento da língua, abertura da boca, posicionamento das cordas vocais e da articulação da mandíbula [21]), algumas características permanecem inalteradas conferindo particularidades no som emitido por cada indivíduo, permitindo seu reconhecimento.

Figura 2 - Estrutura do trato vocal



Fonte: Tavin [22]
Legenda: estrutura da laringe aos lábios, incluindo as cavidades nasal e oral

Com base nessas características, sistemas computacionais, capazes de realizar a identificação ou verificação de indivíduos pelo sinal de voz, foram propostos inicialmente em 1976 por Atal et al. em [23], sendo atualmente uma técnica amplamente conhecida e explorada por diversos pesquisadores, como em [8], [9], [24], [25], [26]. Estes sistemas vêm sendo utilizados em aplicações biométricas comerciais nas últimas décadas, desde aplicações mais simples como o desbloqueio do smartphone com o sinal de voz ou até mesmo sistemas de segurança nacional como identificação de vozes em perícias criminais [27].

Esta modalidade biométrica, pode ser dividida em duas principais: verificação automática de locutor ou *Automatic Speaker Verification* (ASV), que consiste em verificar se a voz do usuário corresponde à uma identidade previamente cadastrada no sistema e alegada

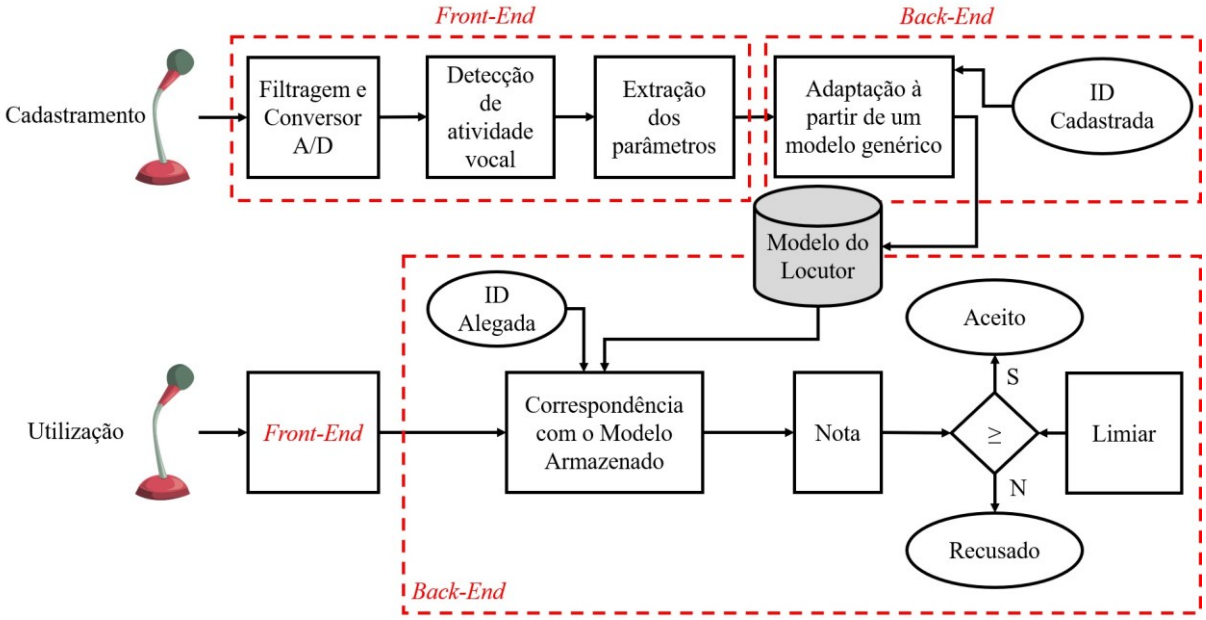
pelo usuário, resultando em uma comparação com um único modelo cadastrado; identificação automática de locutor ou *Automatic Speaker Identification* (ASI) onde o usuário é reconhecido automaticamente sem a necessidade de informar previamente ao sistema sua identidade, ou seja, a etapa de identificação deve buscar a melhor correspondência do sinal em análise em comparação aos modelos de usuários previamente cadastrados no sistema [9]. Embora as contribuições desta tese possam ser utilizadas para ambas modalidades, sistemas ASV foram utilizados para os testes e por este motivo serão abordados com maior detalhamento.

Os sistemas ASV ainda podem ser divididos em dois tipos principais: dependente de texto ou independente de texto [9]. No primeiro caso, as etapas de cadastramento e verificação são baseadas em um determinado texto, sendo muito utilizado em situações em que o usuário deve falar uma senha para autenticar o sistema, já na segunda, o locutor pode ser identificado falando qualquer texto, mesmo que diferente do cadastrado.

Os dois métodos, apesar de provados extremamente eficazes para variadas finalidades biométricas, encontram diversos desafios em sua implementação e uso, como: presença de ruído ambiente; separação de múltiplas vozes; reverberação; etc., degradando sua confiabilidade e robustez. Até mesmo o próprio locutor muda de voz ao longo do tempo por fatores fisiológicos que podem ser decorrentes do envelhecimento, de doenças que afetam o trato vocal ou mudanças de emoções [9], também afetando a qualidade do reconhecimento. Essas alterações são conhecidas como variabilidade intraespecífica, ou seja, do próprio locutor, enquanto a variabilidade interespecífica é decorrente das diferenças entre locutores, oriundas das próprias características vocais, mas, que também incluem variações nas condições da captura do áudio, como mudança no ruído ambiente e diferenças no canal de transmissão [26].

Um sistema genérico de verificação automática de locutor pode ser representado conforme a figura 3. Como pode ser observado nesta figura, duas fases distintas estão presentes, sendo elas: a fase de cadastramento e a fase de utilização do sistema. Em ambas as fases o sistema pode ser separado em duas principais etapas: *front-end* e *back-end*. A etapa de *front-end* é geralmente idêntica tanto no cadastramento quanto no uso do sistema e equivale ao conjunto das seguintes partes: conversão analógico/digital do sinal de fala; filtragem pré-ênfase; etapa opcional de detecção de atividade vocal ou *Voice Activity Detector* (VAD) [28], [29], que realiza a detecção dos trechos onde existe locução para remoção dos silêncios que ocorrem entre as falas; por último, a extração das características do sinal de voz, através de diferentes técnicas como, por exemplo, *Mel-frequency cepstral coefficients* (MFCC) [8], [27], conforme descrito na seção 2.1, para serem posteriormente utilizadas para modelagem do locutor.

Figura 3 - Diagrama em blocos do sistema ASV



Fonte: Autor
Legenda: etapas de *front-end* e *back-end* das fases de cadastramento e utilização do sistema ASV

A etapa de *back-end* da fase de cadastramento conta com a modelagem do locutor com base nas características extraídas no *front-end* e o armazenamento do modelo para uso posterior, através de técnicas de modelagem estatísticas, sendo as duas mais utilizadas atualmente para este propósito: *Gaussian Mixture Model* (GMM) [30], [31], [32], [33] e *i-vector* [34, 35, 36], que serão detalhadas nas seções 2.2 e 2.3, respectivamente. Já no *back-end* da fase de utilização do sistema, ocorre a extração dos parâmetros do locutor e em seguida a comparação com o modelo previamente armazenada da identidade alegada pelo usuário, com aplicação de medidas da correspondência dos modelos, como *Log-Likelihood Ratio* (LLR) descrito na seção 2.2.3. Esta medida é utilizada na etapa de decisão sendo que caso seja maior que um limiar adotado, este usuário é reconhecido conforme a identidade alegada, caso contrário é recusado.

Os tópicos seguintes descrevem com maior detalhamento os fundamentos teóricos necessários para implementação das etapas listadas acima.

2.1 COEFICIENTES MEL-CEPSTRAIS

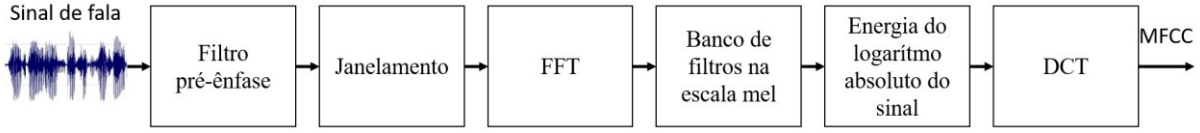
Coeficientes Mel-Cepstrais ou *Mel-frequency cepstral coefficients* (MFCC) são parâmetros fundamentais que descrevem um sinal de fala de forma compacta. O primeiro uso destes coeficientes foi proposto por Bridle e Brow em 1974 [37] e sugeridos para tarefas de

reconhecimento de fala por Mermelstein em 1976 [38], e Davis em 1980 [39]. Estes coeficientes continuam sendo utilizados amplamente até hoje por diversos pesquisadores tanto para tarefas de reconhecimento de fala como verificação e identificação de locutor, como em [40], [41], [42].

Os MFCCs derivam da combinação de técnicas ainda mais antigas, que são fundamentais para o estudo do sinal de fala. Primeiramente, o termo *cepstral,* que deriva de *cepstrum*, foi primeiramente estudado em 1963 por Boegert em [43] e teve seu primeiro uso para aplicação no reconhecimento de fala sugerido por Noll em [44] em 1967. O termo teve origem na palavra *spectrum*, uma vez que que corresponde em termos genéricos a transformada do espectro do sinal [45], conforme descrito a seguir.

O processo da extração dos MFCCs pode ser representado conforme figura 4.

Figura 4 - Extração dos coeficientes Mel-cepstrais



Fonte: Autor
Legenda: sequência do algoritmo de extração do MFCC

A obtenção dos coeficientes cepstrais segue as seguintes etapas [45]:

a) Filtro pré-ênfase;
b) Janelamento do sinal;
c) Cálculo da Transformada de Fourier do sinal com uso da FFT;
d) Linearização conforme escala *Mel* com a aplicação de filtros triangulares conforme as bandas críticas da escala *Bark*;
e) Cálculo da energia do absoluto do sinal em escala logarítmica;
f) Aplicação da transformada dos cossenos ou *Discrete Cosine Transform* (DCT).

A primeira etapa consiste na aplicação de um filtro passa-altas pré-ênfase, que tem a função de enfatizar as altas frequências que costumam estar atenuadas na produção da voz. Sua aplicação é dada conforme equação:

$$y_n = x_n - \alpha x_{n-1}, \tag{1}$$

em que **y** é a saída do filtro e **x** o sinal de voz, sendo ambos discretos no tempo e indexados pelo índice de amostra $n = 1, \dots, n_{max}$, sendo n_{max} o total de amostras. O fator α é usualmente da ordem de 0.97 [46].

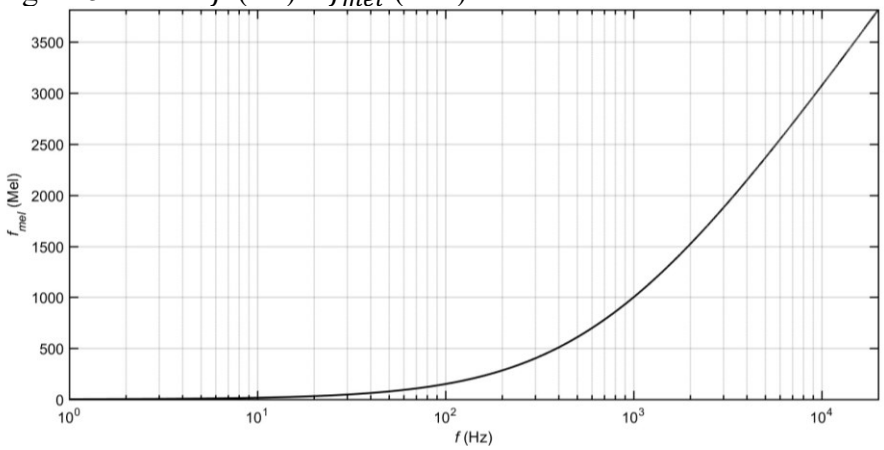
Como o sinal de voz é um sinal não-estacionário, ou seja, possui um espectro variável ao longo do tempo, os coeficientes MFCC são estimados durante janelas temporais do sinal, tipicamente de 10 a 30 ms, em que as características do espectro do sinal de fala amostrado permanecem próximas de uma representação estacionária [35]. As janelas temporais são ainda sobrepostas utilizando passos em torno da metade da duração da janela para uma melhor representação do sinal [9], sendo o janelamento mais comumente utilizado do tipo Hamming.

O termo *Mel* é uma abreviação da palavra *melody*, utilizado para denominar uma escala de *pitch* ou altura do som, que é definido como a percepção subjetiva do som relacionada à sua frequência [44]. Experiências conduzidas por Stevens, Volkmann E Newman na década de 30 e publicadas em 1937 [47] consistiram em testes psicofísicos de modo a estimar a relação da alteração de frequência de um som com a sensação desta alteração percebida por seres humanos, resultando nesta escala perceptual. Outras pesquisas foram conduzidas por Stevens e Volkmann em [48] e Beranek em [49], e posteriormente uma equação foi proposta por O' Shaghnessy em [50] resultando na relação da frequência em Hz (*f*) e a altura do som na escala *Mel* (f_{mel}), conforme segue:

$$f_{mel} = \frac{1000}{\ln\left(1+\frac{1000}{700}\right)}\ln\left(1+\frac{f}{700}\right) \cong 1127\ln\left(1+\frac{f}{700}\right). \quad (2)$$

Os resultados desta equação também podem ser representados graficamente, conforme figura 5. É possível verificar nesta curva que variações em Hz em frequências mais altas são percebidas como menores variações na altura do som do que intervalos em Hz em frequências mais baixas.

Figura 5 - curva f (Hz) x f_{mel} (Mel)



Fonte: Autor
Legenda: *Pitch* ou altura do som em escala Mel

Os MFCCs são, portanto, extraídos na linearização perceptual do espectro do áudio conforme esta escala da altura do som ou *pitch*. Para isso, filtros triangulares de bandas variáveis são aplicados ao espectro do sinal. Estes filtros têm origem na definição de bandas críticas de Zwicker, publicadas em [51] que propôs a utilização de uma escala, denominada *Bark*, em memória a Heinrich Barkhausen que definiu as primeiras medidas de *loudness*. Esta escala subdivide o espectro de frequências audíveis em 24 bandas, denominadas bandas críticas, de modo que cada banda esteja representada por um intervalo de 1 *Bark*, equivalente a 100 *Mels*. Desta forma, é possível verificar que variações de 100 *Mels* em alta frequência correspondem a um intervalo maior em Hz do que em baixa frequência. A tabela 1 apresenta o intervalo de cada banda crítica.

Tabela 1 - Bandas críticas

Banda Crítica	Frequência Central (Hz)	Banda (Hz)	Banda Crítica	Frequência Central (Hz)	Banda (Hz)	Banda Crítica	Frequência Central (Hz)	Banda (Hz)
1	60	80	9	1000	160	17	3400	550
2	150	100	10	1170	190	18	4000	700
3	250	100	11	1370	210	19	4800	900
4	350	100	12	1600	240	20	5800	1100
5	450	110	13	1850	280	21	7000	1300
6	570	120	14	2150	320	22	8500	1800
7	700	140	15	2500	380	23	10500	2500
8	840	150	16	2900	450	24	13500	3500

Fonte: Zwicker, 1961. [51].
Legenda: separação das bandas críticas em 24 bandas de intervalo 100 *Mels*

É importante notar que estas bandas apesar de possuírem larguras fixas conforme escala *Bark*, o ajuste da frequência central de cada banda pode ser alterado conforme sugerido por Zwicker em [51]. Desta forma, filtros triangulares podem ser ajustados conforme a divisão das bandas críticas de modo que os coeficientes cepstrais possam ser extraídos na linearização perceptual do espectro do áudio. A figura 6 representa uma possível configuração destes filtros, conforme sugerido por Davis e Merlmelstein em 1980 [39], sendo uma das configurações mais utilizadas.

O último passo consiste na aplicação da DCT para obtenção dos MFCCs finais. Como propriedade da própria transformada, os primeiros coeficientes são os que concentram a maior energia e, por este motivo, uma quantidade limitada destes pode ser suficiente para uma boa representação dos parâmetros, sendo uma configuração comum a utilização de doze coeficientes cepstrais, do segundo ao décimo terceiro [8, 9]. O primeiro carrega a informação da energia próximo a 0 Hz (nível médio) e sua inclusão muitas vezes também é considerada,

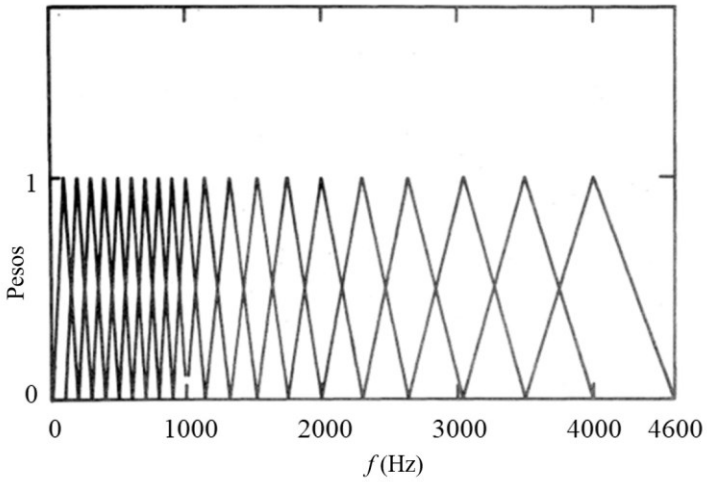
totalizando desta forma treze coeficientes.

É comum também a inclusão dos parâmetros resultantes da derivada primeira (velocidade) e derivada segunda (aceleração) dos MFCCs, conhecidos como delta-cepstrais e delta-delta cepstrais [9]. Esses coeficientes incluem informação da dinâmica temporal da fala e muitas vezes auxiliam sistemas de reconhecimento de fala ou até mesmo reconhecimento de locutor, calculados por [46]:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}, \tag{3}$$

sendo d_t um coeficiente delta cepstral e c_t um coeficiente Mel-cepstral em uma dada janela t. Os parâmetros delta-delta são encontrados da mesma forma, substituindo na equação 3 c_t por d_t. O valor de N determina o período da aplicação da derivada, sendo $N = 2$ geralmente utilizado [46].

Figura 6 - Filtros triangulares conforme escala Mel



Fonte: Autor “adaptado de” Davis e Mermelstein, 1980 [39]
Legenda: filtros triangulares para extração dos MFCCs

2.2 DETECÇÃO DE ATIVIDADE VOCAL

A detecção de atividade de voz, conhecida na literatura como VAD, desempenha um papel importante na etapa de *front-end* de sistemas de reconhecimento de fala e verificação ou identificação de locutor, sendo responsável por distinguir os trechos de locução e a sua ausência no sinal de áudio analisado [52] [53]. Esta etapa é fundamental para garantir que a extração das características seja realizada apenas durante os períodos em que ocorre locução,

removendo os trechos que são apenas silêncio, ruído ou outros sinais acústicos que não estão correlacionados com a produção de fala.

Muitos algoritmos para VAD baseados em parâmetros extraídos do sinal de áudio, já foram propostos para esta tarefa. Desde pesquisas mais antigas como o uso da energia espectral e cruzamento por zero [54] até pesquisas mais recentes baseadas no uso do fluxo espectral [55], frequências ultrassônicas [56], filtragem de uma única frequência [57], formantes espectrais do sinal de voz [58] ou até mesmo em análises de redes neurais [59], porém, todos eles podem ter os resultados de detecção degradados se o áudio for capturado na presença de ruído ambiente não estacionário.

Para tratar essa dificuldade, muitos autores propuseram uma detecção de atividade de voz que se baseia também em informações visuais, geralmente denominada na literatura como detecção visual de atividade vocal ou *Visual Voice Activity Detection* (VVAD) [60]. Autores em [61] propuseram a combinação de características extraídas do sinal de áudio, com o uso de MFCC, e parâmetros extraídos do movimento da região do lábio, com aplicação de mapa de difusão. Outros autores como em [62], [63] utilizaram algoritmos de fluxo óptico para extrair parâmetros de movimento de lábios, enquanto autores em [64] propuseram um algoritmo que permite a extração de movimento labial para detecção de atividade vocal mesmo sobre variação de iluminação ambiente. Embora todos esses métodos produzam resultados interessantes, nenhum deles utiliza informações que já estão disponíveis no vídeo codificado, conforme proposta desta tese.

2.3 MODELO DE MISTURAS GAUSSIANAS (*GAUSSIAN MIXTURE MODEL*)

Gaussian Mixture Model (GMM) é um modelo de misturas estatístico [65] baseado em curvas Gaussianas utilizado para modelagem e reconhecimento de sub-classes dentro de uma classe mais genérica. Foi proposto para utilização em sistemas biométricos por Reynolds no início dos anos 90 [30], [31], sendo amplamente utilizado desde então para finalidades de reconhecimento de locutor [66], [67], como uma forma de modelagem da distribuição probabilística de características acústicas do sinal de fala, por exemplo, as que podem ser obtidas através dos MFCCs descritos na seção anterior.

É descrito por uma Função Densidade de Probabilidade (FDP) paramétrica resultante da soma ponderada de C componentes de densidades Gaussianas, conforme equação:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{C} w_i \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\,, \tag{4}$$

sendo $p(\mathbf{x}|\lambda)$ a função densidade de probabilidade de um vetor $\mathbf{x}$ aleatório *D*-dimensional, por exemplo, composto de medidas ou até mesmo parâmetros Mel-cepstrais, dado um modelo λ. Os pesos das misturas definidos por $w_i, i = 1, \ldots, C$, sendo $w_i \geq 0$ e os componentes de densidade Gaussiana $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, representados da seguinte forma:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right\}, \tag{5}$$

sendo $\boldsymbol{\mu}_i$ o vetor das médias e $\boldsymbol{\Sigma}_i$ a matriz de covariância da *i*-ésima componente Gaussiana. Os pesos das misturas devem satisfazer a condição $\left(\sum_{i=1}^{C} w_i\right) = 1$.

O modelo completo λ é, portanto, parametrizado pelos vetores de média, matrizes de covariância e pesos de cada componente de densidade, sendo sua representação para um dado locutor escrita da seguinte forma [30, 31]:

$$\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}\,, i = 1, \ldots, C. \tag{6}$$

O uso da matriz de covariância completa geralmente não é necessário e apenas a matriz diagonal principal pode ser suficiente para uma boa modelagem conforme observado por Reynolds [31]. A matriz utilizada pode ainda ser única para cada componente ou compartilhada entre diferentes parâmetros observados. Essas escolhas são determinadas em função da disponibilidade de parâmetros, tempo de execução esperado e finalidade da aplicação, uma vez que impactam diretamente nos resultados biométricos. Portanto, uma investigação prévia da melhor parametrização se faz necessária.

2.3.1 Estimativa por máxima verossimilhança

Um dos maiores desafios no uso do algoritmo GMM é a estimativa dos parâmetros do modelo, sendo um dos métodos utilizados a estimativa por máxima verossimilhança ou *maximum-likelihood* (ML). A partir da observação de um conjunto de características de um dado locutor durante a fase de treinamento, por exemplo, uma sequência de *T* vetores MFCC $X = \{\mathbf{x}_t\}, t \in \{1, \ldots, T\}$, a estimativa por máxima verossimilhança tem a função de realizar a estimativa dos parâmetros do modelo estatístico, por exemplo um modelo de GMM

paramétrico λ conforme descrito na seção anterior. Isso é feito de forma a maximizar a verossimilhança dos parâmetros observados ao modelo, sendo calcula da seguinte forma [31]:

$$p(X|\lambda) = \prod_{t=1}^{T} p(\mathbf{x}_t|\lambda). \tag{7}$$

Ou calculada através da verossimilhança logarítmica, da seguinte forma [32]:

$$\log p(X|\lambda) = \sum_{t=1}^{T} \log p(\mathbf{x}_t|\lambda). \tag{8}$$

Para tornar a solução do problema tratável desta forma é preciso assumir a ausência de dependência entre os vetores, embora haja dependência na maioria dos casos, conforme observado por Reynolds em [31]. Devido a não-linearidade das equações em função de λ a estimativa é geralmente realizada através de um algoritmo de maximização da esperança matemática, conhecido como EM (*Expectation-Maximization*) que iterativamente encontra um novo modelo $\bar{\lambda} = \{\overline{w}_i, \overline{\boldsymbol{\mu}}_i, \overline{\boldsymbol{\Sigma}}_i\}, i = 1, \dots, C$ em que a probabilidade *a-posteriori* dos vetores de parâmetros em função do modelo é cada vez maior, até atingir um limiar pré-determinado de convergência, similarmente como na estimativa de parâmetros através do uso de HMM (*Hidden-Markov Models*) com algoritmo estatístico Baum-Welch [68].

A aplicação deste algoritmo para estimativa dos parâmetros do modelo foi proposta por Reynolds em [31] uma vez que os vetores não são rotulados e as classes acústicas são totalmente ocultas. O algoritmo é composto pelos seguintes passos:

a) Estimar primeiramente um modelo GMM paramétrico da forma $\lambda = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}, k = 1, \dots, C$, podendo ser um modelo genérico utilizando algoritmo VQ (*Vector Quantization*) [69] ou um modelo universal do tipo GMM-UBM (*Universal Background Model*) conforme será descrito na seção seguinte;

b) Calcular a verossimilhança logarítmica conforme Equação 8;

c) Calcular a probabilidade *a-posteriori* para um dado componente *i*, conforme equação:

$$P_r(i|\mathbf{x}_t, \lambda) = \frac{w_i \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{k=1}^{M} w_k \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}\ ; \tag{9}$$

d) Estimar novamente os parâmetros do novo modelo $\bar{\lambda}$ com base nesta probabilidade calculada, estimando as estatísticas de Baum-Welch de ordem zero (n_i), primeira ($\boldsymbol{f}_i$) e

segunda ordem ($\mathbf{S}_i$) respectivamente:

$$n_i = \sum_{t=1}^{T} P_r(i|\mathbf{x}_t, \lambda)\,, \tag{10}$$

$$\mathbf{f}_i = \sum_{t=1}^{T} P_r(i|\mathbf{x}_t, \lambda)\,\mathbf{x}_t, \tag{11}$$

$$\mathbf{S}_i = \sum_{t=1}^{T} P_r(i|\mathbf{x}_t, \lambda)\,\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}. \tag{12}$$

Obtendo os novos pesos $\overline{w}_i$:

$$\overline{w}_i = \frac{1}{T} n_i, \tag{13}$$

sendo $\left(\sum_{i=1}^{C} \overline{w}_i\right) = 1$, onde $\overline{w}_i \geq 0$. As novas médias $\overline{\boldsymbol{\mu}}_i$ e matrizes de covariância $\overline{\boldsymbol{\Sigma}}_i$ são obtidas através dos cálculos das esperanças, como sendo respectivamente:

$$\overline{\boldsymbol{\mu}}_i \; = E_i(\mathbf{x}_t|X) = \frac{\mathbf{f}_i}{n_i}, \tag{14}$$

$$\overline{\boldsymbol{\Sigma}}_i \; = E_i(\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}|X) = \frac{\mathbf{S}_i}{n_i}; \tag{15}$$

e) Verificar a convergência calculando novamente a verossimilhança conforme equação 8. Retornar para o passo c) caso a convergência não tenha sido atingida ou encerrar o algoritmo com o modelo resultante.

2.3.2 Adaptação MAP e UBM

Outro método amplamente utilizado para estimativa dos parâmetros do modelo GMM, principalmente em tarefas de reconhecimento de locutor, é conhecido como *Maximum a Posteriori* (MAP) [70]. Esta técnica consiste em encontrar um novo modelo a partir de um UBM, que é um GMM obtido a partir de parâmetros combinados extraídos de um grupo de

locutores, resultando em um único modelo genérico utilizado como base para a adaptação dos parâmetros específicos de cada locutor.

Dado um modelo UBM inicial (λ_{UBM}) e vetores de uma determinada classe representados por $X = \{\mathbf{x}_t\}, t \in \{1, \dots, T\}$ é possível encontrar o alinhamento probabilístico *a-posteriori* destes vetores ao modelo inicial, idêntico ao encontrado para o algoritmo EM, conforme a equação 9. A seguir são encontrados os valores dos pesos, médias e variâncias utilizando uma versão modificada do algoritmo EM descrito na seção anterior.

Primeiramente são encontrados os valores dos pesos $\overline{w}_i$, médias $\overline{\boldsymbol{\mu}}_i$ e covariâncias $\overline{\boldsymbol{\Sigma}}_i$ conforme equações 13, 14, 15, respectivamente, utilizando o modelo UBM como referência. Porém, diferentemente do algoritmo EM tradicional, as estatísticas calculadas do novo modelo são combinadas com o velho modelo utilizando um coeficiente de mistura dependente dos dados, de modo que possa ser ajustado para que na estimativa final dos parâmetros as misturas com alta contagem dos novos dados tenham mais dependência das novas estatísticas do que das estatísticas calculadas anteriormente [70]. O resultado é um modelo $\hat{\lambda} = \{\widehat{w}_i, \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i\}, i = 1, \dots, C$.

Os novos valores dos pesos podem ser encontrados conforme a equação 16, as médias conforme a equação 17 e as matrizes de covariância conforme a equação 18.

$$\widehat{w}_i = [\alpha_i^w \, \overline{w}_i + (1 - \alpha_i^w) w_i]\gamma, \tag{16}$$

$$\widehat{\boldsymbol{\mu}}_i = \alpha_i^m \overline{\boldsymbol{\mu}}_i + (1 - \alpha_i^m)\boldsymbol{\mu}_i, \tag{17}$$

$$\widehat{\boldsymbol{\Sigma}}_i = \alpha_i^v \overline{\boldsymbol{\Sigma}}_i \, + (1 - \alpha_i^v)\left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i^{\mathrm{T}}\right) - \widehat{\boldsymbol{\mu}}_i \widehat{\boldsymbol{\mu}}_i^{\mathrm{T}}, \tag{18}$$

sendo γ um fator para garantir que a soma de todos os pesos seja igual a 1 e os coeficientes $\alpha_i^w, \alpha_i^m \, e \, \alpha_i^v$ os fatores de adaptação que controlam o ajuste do novo modelo e estimados conforme a equação

$$\alpha_i^p = \frac{n_i}{n_i + r^p}, \tag{19}$$

sendo r^po fator de relevância do parâmetro $p \in \{w, m, v\}$ e n_i a estatística de Baum-Welch [68] de ordem zero definida na equação 10. É usual utilizar um único valor nos coeficientes de adaptação de modo que $\alpha_i^w = \alpha_i^m = \alpha_i^v$, ou seja, todos fatores de relevância com o mesmo

valor. Muitas vezes somente os vetores de média são adaptados, já que bons resultados podem ser obtidos sem adaptar os outros parâmetros, reduzindo assim a carga computacional [9], [71].

2.3.3 LOG LIKELIHOOD RATIO

A última etapa, após a criação dos modelos, é a realização do cálculo da métrica que faz a correspondência de parâmetros novos extraídos ($X = \{\mathbf{x}_t\}, t \in \{1, \dots, T\}$) com o modelo de um locutor armazenado (λ_{target}), compensado pelo modelo genérico UBM (λ_{UBM}). A métrica conhecida como *log likelihood ratio* (LLR) é normalmente utilizada para este propósito, calculada da seguinte forma [9]:

$$LLR_{avg}(X, \lambda_{target}, \lambda_{UBM}) = \frac{1}{T}\sum_{t=1}^{T}\left\{\log\left(p(\mathbf{x}_t|\lambda_{target})\right) - \log(p(\mathbf{x}_t|\lambda_{UBM}))\right\}, \quad (20)$$

sendo LLR_{avg} o LLR médio do conjunto de vetores de parâmetros analisados.

Quanto maior a função densidade de probabilidade de um vetor **x** para o modelo do locutor que se deseja verificar a identidade (λ_{target}), maior será o valor de LLR_{avg} obtido e, portanto, maior a correspondência do locutor avaliado em função do modelo cadastrado. Caso o resultado seja maior do que um limiar estipulado previamente o usuário é considerado aceito, caso contrário, recusado.

2.4 ANÁLISE FATORIAL

A Análise Fatorial ou *Factor Analysis* (FA) é uma técnica estatística de análise multivariada que busca representar a variabilidade de eventos observados em um número menor de variáveis ocultas. A aplicação deste método para verificação automática de locutor foi proposta inicialmente por Kenny [34], separando a análise do modelo GMM em duas componentes, sendo uma dependente do locutor e outra dependente do canal. Desta forma, um super-vetor GMM, que consiste na concatenação dos vetores de média de cada componente de mistura Gaussiana, pode ser escrito como a composição linear de dois super-vetores, da forma:

$$\mathbf{m} = \mathbf{s} + \mathbf{c}, \quad (21)$$

sendo $\mathbf{m}$ o super-vetor das componentes de média GMM, $\mathbf{s}$ o super-vetor dependente do locutor e $\mathbf{c}$ o super-vetor dependente do canal. Sendo as componentes $\mathbf{s}$ e $\mathbf{c}$ estatisticamente independentes e distribuídas conforme distribuição normal padrão $\sim\mathcal{N}(0,\mathbf{I})$.

2.5 JOINT-FACTOR ANALYSIS

Similarmente à técnica FA descrita na seção anterior, a análise fatorial combinada ou *Joint-Factor Analysis* (JFA) foi proposta por Kenny em [72] e trata-se da combinação das técnicas de auto-voz (*eigenvoice*) [73], [74], auto-canal (*eigenchannel*) [34] e adaptação MAP [31] para utilização em sistemas de reconhecimento de locutor.

Inicialmente, Kenny propõe em [72] a definição de um modelo de locutor, baseado no resultado do valor do parâmetro de média de um modelo GMM de C Gaussianas, obtidas a partir de vetores com F parâmetros, utilizando a técnica de adaptação MAP, conforme equação 17. O resultado é a obtenção de um super-vetor de dimensão CF x 1 obtido a partir da concatenação dos vetores de média dos GMM, obtido conforme equação:

$$\mathbf{s} = \mathbf{m}_{UBM} + \mathbf{Dz}, \tag{22}$$

sendo $\mathbf{m}_{UBM}$ um super-vetor independente do locutor e do canal de dimensão CF x 1 obtido a partir da concatenação dos vetores de média de um UBM, $\mathbf{D}$ uma matriz diagonal de dimensão CF x CF e $\mathbf{z}$ um vetor de dimensão CF x 1 de distribuição normal padrão $\sim\mathcal{N}(\mathbf{z}|0,\mathbf{I})$. Sendo a componente $\mathbf{Dz}$ dependente do locutor. No caso particular em que [9]:

$$\mathbf{D}^2 = \frac{1}{r}\mathbf{\Sigma}_{UBM}, \tag{23}$$

o problema passa a ser tratado identicamente como na definição da adaptação MAP descrita na seção 2.3.2. Sendo r o fator de relevância utilizado na equação 19 para estimar os fatores de adaptação $\alpha_i^w, \alpha_i^m \; e \; \alpha_i^v$ e $\mathbf{\Sigma}_{UBM}$ a matriz de bloco-diagonal onde os blocos correspondem às diagonais das matrizes de covariância do modelo UBM.

Similarmente, para a técnica conhecida como auto-voz (*eigenvoice*), Kenny propõe um super-vetor de um locutor também baseado na componente de média de um modelo GMM, sendo descrito como [72]:

$$\mathbf{s} = \mathbf{m}_{UBM} + \mathbf{Vy}, \tag{24}$$

sendo neste caso **V** uma matriz retangular de posto baixo, de dimensão CF x R onde $R \ll CF$ e representa o posto da matriz. Esta matriz abrange o subespaço definido pelo locutor e **y** são as variáveis ocultas de distribuição normal padrão $\sim\mathcal{N}(\mathbf{y}|0,\mathbf{I})$, dependentes do locutor e de dimensão R x 1.

Esta proposta é computacionalmente mais vantajosa para a aplicação das técnicas de EM na adaptação do modelo do que a utilização da técnica MAP, dado o tamanho reduzido da matriz **V** em comparação a matriz **D** definida na equação 22. Esta abordagem favorece o modelo adequado do locutor, porém, não garante boa modelagem do canal. Por outro lado, outra técnica conhecida como auto-canal (*eigenchannel*) favorece a modelagem do canal, sendo o super-vetor de um dado locutor definido como [72]:

$$\mathbf{s} = \mathbf{m}_{UBM} + \mathbf{Dz} + \mathbf{Ux}, \tag{25}$$

sendo que **x** representa os fatores dependentes do canal distribuídos na forma $\sim\mathcal{N}(\mathbf{x}|0,\mathbf{I})$ e **U** é uma matriz retangular de posto baixo, de dimensão CF x R_c, onde $R_c \ll CF$ e suas colunas representam os auto-vetores da matriz de covariância do canal.

Combinando as diferentes técnicas apresentadas acima, a análise fatorial combinada resulta em um modelo do super-vetor para um dado locutor como:

$$\mathbf{s} = \mathbf{m}_{UBM} + \mathbf{Dz} + \mathbf{Ux} + \mathbf{Vy}. \tag{26}$$

2.6 I-VECTOR

A técnica de modelagem estatística *i-vector*, combina o uso de FA (*Factor Analysis*) para extração dos parâmetros como proposto por Kenny e Dehak em 2008 [34], com aplicação de diferentes técnicas de compensação de canal para diminuição dos efeitos decorrentes da variabilidade intraespecífica, ou seja, variabilidade entre os modelos obtidos para o próprio locutor em capturas de diferentes sessões. Este trabalho foi apresentado primeiramente na tese de doutorado de Dehak em 2009 [35] e nomeado como *i-vector* por Dehak em 2010 [36]. O termo é uma abreviação de *identity-vector* ou vetor-identidade, dado o propósito da aplicação e, desde então, tem sido utilizada como o estado da arte para reconhecimento de locutor [26], [75], [76], [77]. Sua utilização também é baseada na criação prévia de um modelo GMM através da adaptação MAP a partir de um modelo UBM, como descrito na seção 2.3.2.

Dado um UBM de C componentes Gaussianas e F características acústicas, o *i-vector* pode representar um trecho de locução por um vetor **m** especificado como [34], [35]:

$$\mathbf{m} = \mathbf{m}_{UBM} + \mathbf{Ti}, \tag{27}$$

onde $\mathbf{m}_{UBM}$ é o super-vetor da componente de média do modelo UBM conforme descrito anteriormente, **i** o *i-vector* assumindo uma distribuição normal padrão $\sim\mathcal{N}(\mathbf{i}|0,\mathbf{I})$ e **T** uma matriz retangular de posto baixo de dimensão CF x R, onde $R \ll CF$ é o posto da matriz e é obtida da mesma forma que a matriz **V** proposta por Kenny na modelagem eigenvoice [74], conforme equação 24.

Para uma sequência de T vetores MFCC $X = \{\mathbf{x}_t\}, t \in \{1, \ldots, T\}$ as estatísticas Baum-Welch são calculadas, sendo a de ordem zero (n_i) calculada conforme a equação 10 e primeira ordem centralizada ($\tilde{\mathbf{f}}_i$) similar a equação 11, porém, calculada subtraindo as médias da seguinte forma:

$$\tilde{\mathbf{f}}_i = \sum_{t=1}^{T} P_r(i|\mathbf{x}_t, \boldsymbol{\lambda})\,(\mathbf{x}_t - \boldsymbol{\mu}_i), \tag{28}$$

onde $\boldsymbol{\mu}_i$ é o vetor de média de um componente de mistura UBM, para cada Gaussiana $i = 1, \ldots, C$ e $P_r(i|\mathbf{x}_t, \boldsymbol{\lambda})$ a probabilidade *a-posteriori* da componente de mistura i, conforme equação 9. O vetor *i-vector* é representado para uma dada locução u, como sendo [34, 35]:

$$\boldsymbol{i} = \mathbf{B}^{-1}\mathbf{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}}_u, \tag{29}$$

sendo **B** a matriz de precisão, obtida como:

$$\mathbf{B} = \mathbf{I} + \mathbf{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{N}_u\mathbf{T}, \tag{30}$$

em que $\mathbf{N}_u$ é uma matriz diagonal de dimensão CF x CF onde os blocos diagonais são compostos por vetores $\mathbf{n}_i, i = 1, \ldots, C$ de dimensão F obtidos concatenando todos valores de n_i; $\tilde{\mathbf{f}}_u$ é um super-vetor de dimensão CF x 1 obtido concatenando todos os valores de $\tilde{\mathbf{f}}_i$ para uma dada locução u; $\boldsymbol{\Sigma}^{-1}$ é a matriz de covariância de dimensão CF x CF estimada durante o processo de treinamento baseado em análise fatorial e modela a variabilidade residual que não foi capturada na matriz retangular de posto baixo **T** de dimensão CF x R.

2.7 PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS

Análise Discriminante Linear Probabilística ou *Probabilistic Linear Discriminant Analysis* (PLDA) é uma técnica probabilística derivada da Análise Discriminante Linear ou *Linear Discriminant Analysis* (LDA), utilizada para separar informações desejadas de outras fontes indesejadas, similar a análise fatorial combinada (JFA) descrita na seção 2.5, que separa os fatores dependentes da classe (locutor) e aqueles dependentes da sessão (canal). A principal diferença é que esta técnica utiliza como fator dependente do locutor e independente do canal o *i-vector*, diferentemente do super-vetor GMM utilizado no JFA.

A técnica tem sido utilizada nos últimos anos para compensação de canal com o uso do *i-vector*, bem como para utilização na etapa de avaliação onde os modelos PLDA são comparados, conforme descrito em [78].

2.8 BIOMETRIA MULTIMODAL PARA RECONHECIMENTO DE LOCUTOR

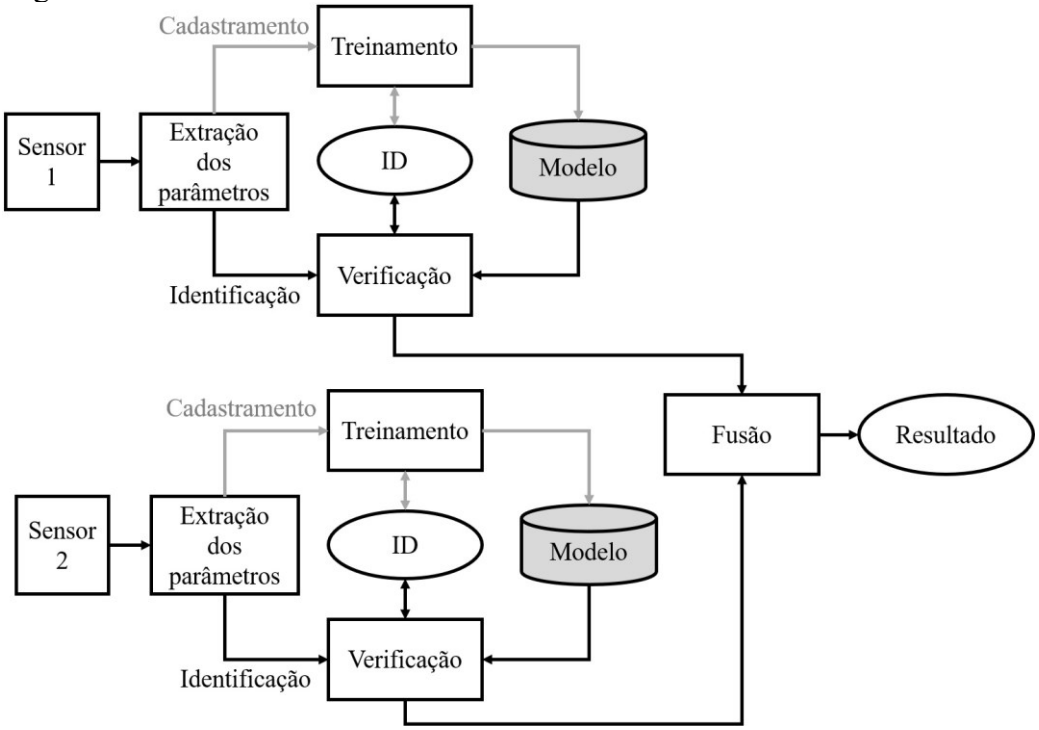
Conforme apontado anteriormente, sistemas baseados em uma única fonte podem ter seu desempenho facilmente degradado em função de falhas que ocorrem durante a captura do sinal, ruído ambiente no áudio, iluminação ruim para captura do vídeo, posicionamento errado do usuário perante o sensor biométrico, ou até mesmo variabilidade intraespecífica em que os parâmetros do indivíduo analisado podem variar em diferentes capturas, como no caso do sinal de voz. Por este motivo, sistemas multimodais que incorporam a combinação de múltiplos sinais para análise, têm sido foco de pesquisas e aplicações, como em [18], [19], [20]. Estes sistemas podem funcionar com maior precisão mesmo com qualidade ruim de uma das aquisições.

Para que a análise seja feita com a combinação dos múltiplos sinais, diferentes técnicas de fusão são utilizadas, como a técnica mais tradicional, que efetua a extração separada dos parâmetros de cada fonte e realiza a fusão somente no momento da verificação [79], conforme figura 7 ou com a combinação dos parâmetros extraídos anteriormente a modelagem, como em [80], [81], representado graficamente conforme figura 8.

No caso em que a fusão ocorre na etapa de decisão, conforme figura 7, ainda é possível ser separado em duas formas distintas: fusão no momento do cálculo da métrica de correspondência dos modelos, resultando em uma única medida para os diferentes parâmetros analisados; fusão no momento da decisão, sendo as métricas de correspondência dos modelos obtidas separadamente [82].

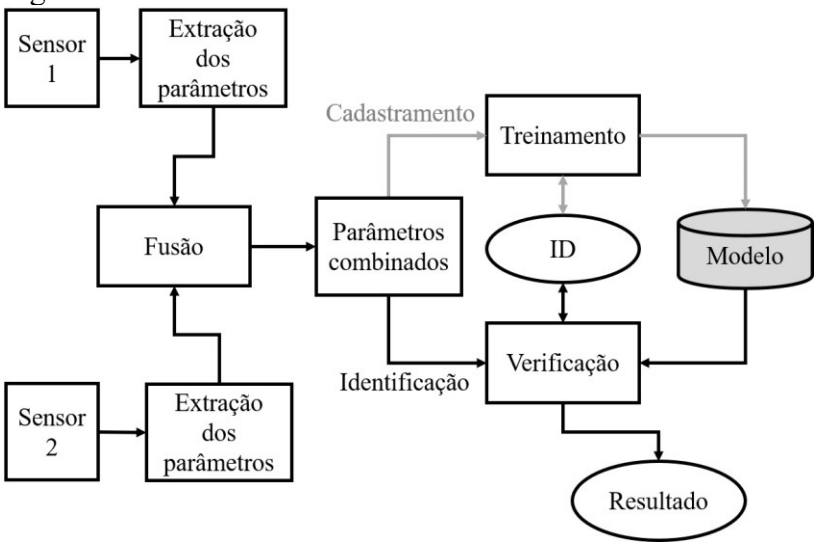
Esta tese apresenta no capítulo 8 uma fusão de parâmetros, conforme esquema de fusão da figura 8, com o uso de MFCC e DCT-II aplicada aos vetores de movimento labial. Testes são realizados utilizando modelagem *i-vector* e tendo o desempenho avaliado em diferentes condições de relação sinal-ruído do áudio.

Figura 7 - Sistema biométrico multimodal I



Fonte: Autor
Legenda: sistema multimodal com fusão na etapa de decisão

Figura 8 - Sistema biométrico multimodal II



Fonte: Autor
Legenda: sistema biométrico com fusão na extração dos parâmetros

2.9 DETECÇÃO LIVENESS

Um dos maiores desafios na área da biometria é garantir a segurança contra fraudes. Nos ataques, conhecidos como ataques diretos (*direct attacks*), o falsário utiliza algum tipo de artefato produzido sinteticamente, como dedos falsos, máscara, lente de contato com padrão da íris, foto do rosto e voz sintetizada/gravada, ou tenta imitar o comportamento de um usuário genuíno, como a forma de caminhar, a assinatura, etc. para acessar o sistema biométrico [83]. Como este tipo de ataque é realizado de forma que a interação com o sensor biométrico é feita seguindo o protocolo regular, os mecanismos usuais de proteção digitais, como encriptação, assinatura digital ou marca d'água não são efetivos. Além disso, estes ataques não requerem habilidades técnicas avançadas por parte do invasor e nem mesmo nenhum conhecimento aprofundado sobre o sistema biométrico. Uma amostra fraudulenta, caso não identificada, é processada pelo sistema e tratada como uma amostra biométrica genuína de um usuário válido cadastrado [84].

Detecção *liveness* é a habilidade de um sistema de distinguir o que é realmente um sinal proveniente de um ser humano vivo. Cada tipo de biometria necessita de uma diferente técnica para detecção *liveness* que tem sido estudada por muitos pesquisadores nos últimos anos, por exemplo, em reconhecimento facial [85], impressão digital [86], reconhecimento por íris [86], [87]. Estes trabalhos têm destacado a necessidade de métodos específicos de proteção contra estas ameaças, representando um problema desafiador, já que a detecção *liveness* deve satisfazer alguns requerimentos, como: não invasivos; amigável ao usuário; de rápida execução; de baixo custo; confiável [86].

Como os métodos de ataques continuam evoluindo e se tornando cada vez mais sofisticados, uma detecção *liveness* suportada por sistema biométrico multimodal, constitui um método promissor já que o invasor deve falsificar todas as fusões biométricas utilizadas [88]. Esta tese propõe a combinação de parâmetros do sinal de fala e movimento labial para reconhecimento de locutor, contribuindo para robustez do sistema contra fraudes uma vez que para falsificar a identidade seria preciso fraudar tanto a captura do sinal de áudio quanto vídeo. Uma técnica *liveness* baseada na relação dos movimentos labiais com o sinal de fala também pode ser derivada da aplicação deste *front-end* multimodal proposto.

2.10 MSR IDENTITY TOOLBOX

O MSR (*Microsoft Research*) *Identity Toolbox* consiste em um grupo de ferramentas para MATLAB®, desenvolvido pelo centro de pesquisa CSRC (*Conversational Systems Research Center*) da Microsoft, voltado para pesquisa e desenvolvimento de aplicações de reconhecimento de locutor [89]. As ferramentas disponíveis permitem a avaliação e desenvolvimento do *back-end* de um sistema de biometria que tem como base o sinal de fala, podendo também ser utilizado para aplicações voltadas para análise do sinal de fala.

No *front-end* apenas ferramentas de suporte são disponibilizadas, sendo necessário o uso de outro conjunto de funções ou a criação de novas funções para extração dos parâmetros e implementação de um sistema de VAD. Para tal, a documentação do MSR sugere o uso de dois pacotes: *The Auditory Toolbox* (Malcolm Slaney) [90] e *VOICEBOX* (Mike Brooks) disponível em [91]. O *front-end* implementando no capítulo 8 desta tese contou com a aplicação de funções do *VOICEBOX* [91] para cálculo dos parâmetros Mel-cepstrais.

Para o *back-end*, onde ocorre a criação dos modelos dos locutores na fase de cadastramento e identificação na fase de teste, o MSR disponibiliza implementações do GMM-UBM com adaptação do modelo UBM através da técnica MAP e uma segunda técnica utilizando *i-vector* com UBM como modelo universal e adaptação PLDA. O pacote inclui ainda funções para comparação dos modelos utilizando LLR bem como para verificação de métricas de avaliação do sistema biométrico. O *back-end* implementando no capítulo 8 desta tese utilizou funções deste pacote tanto para modelagem GMM-UBM quanto *i-vector* PLDA

3 BIOMETRIA FACIAL E LABIAL

Uma das modalidades biométricas mais explorada nas últimas décadas é o reconhecimento de indivíduos pela região facial, sendo utilizada desde aplicações mais comerciais como no reconhecimento de rostos em fotos publicadas na rede social virtual Facebook, com implementação do sistema DeepFace [92] ou até mesmo na ciência forense [93], [94]. Para a utilização desta modalidade, o sistema deve inicialmente realizar a detecção da região facial para extração dos parâmetros necessários para o reconhecimento, podendo utilizar diversos algoritmos como: algoritmo Viola-Jones [95]; segmentação *k*-means [96]; eigenfaces [97]; segmentação por misturas Gaussiana [98].

Após a identificação da região facial outras modalidades biométricas podem ser exploradas, como aquelas que dependem da extração do movimento labial. Dada a alta correlação do movimento labial com a produção de fala, alguns autores propuseram a fusão de parâmetros acústicos e do movimento dos lábios para tarefas de reconhecimento de fala, bem como reconhecimento automático de locutor dependente de texto como em [99], [100], [101], dada a singularidade do movimento que cada pessoa executa para falar um determinado texto. A detecção da região labial como em [102], [103] se faz necessária para aplicação destas modalidades, sendo aplicada após a detecção da região facial.

Os autores em [104] propuseram um algoritmo para rastreamento do contorno dos lábios para extração dos parâmetros e identificação de locutor. Os mesmos autores em [105] avaliaram os melhores parâmetros extraídos do movimento labial para verificação de locutor e reconhecimento de fala. Os autores de [106] propuseram a combinação de extração de movimento labial baseado em fluxo ótico com MFCC para reconhecimento de locutor e de fala. Outros pesquisadores como em [107], [108], [109] apresentaram pesquisas mais recentes do uso do movimento labial para reconhecimento de locutor, porém, há ainda uma carência de estudo aprofundado da melhor fusão e modelagem destes parâmetros.

Nesta tese, o reconhecimento labial e extração do movimento são utilizados para a aplicação de um detector de atividade labial, conforme proposto previamente pelo autor desta tese em [110], e para utilização no *front-end* de sistemas ASV. A tese propõe ainda, no capítulo 9, um algoritmo inovador para a extração da região labial que pode ser combinado com os demais métodos propostos.

3.1 ALGORITMO VIOLA-JONES

O algoritmo proposto por Viola e Jones em 2001 [95] representou uma mudança na forma como é feito o reconhecimento facial desde então, sendo uma das maiores referências no desenvolvimento de novos métodos para este propósito até hoje [111], [112]. Este algoritmo propõe: uso da imagem integral para extração dos parâmetros de base de funções Haar; seleção de parâmetros; classificador AdaBoost e um método em cascata para combinação dos classificadores.

A construção de uma imagem integral foi motivada pelo trabalho de Crow [113] que utiliza a técnica *mipmaps* para mapeamento de texturas da imagem. Esta imagem consiste no cálculo do somatório de todos os elementos a esquerda e acima de um ponto, incluindo o próprio elemento, conforme equações a seguir:

$$ii_{i,j} = \sum_{i' \leq i\ ,j' \leq j} f_{i',j'}\ , \tag{31}$$

sendo $ii_{i,j}$ cada pixel da nova imagem **II** e $f_{i',j'}$ cada pixel da imagem original **F**. A equação é solucionada com as seguintes recorrências:

$$s_{i,j} = s_{i,j-1} + f_{i,j}, \tag{32}$$

$$ii_{i,j} = ii_{i-1,j} + s_{i,j}, \tag{33}$$

sendo $s_{i,j}$ a soma acumulada das linhas da imagem, onde $s_{i,-1} = 0$ e $ii_{-1,j} = 0$.

Conforme apontado por Viola em [95], esta nova imagem é encontrada para facilitar o cálculo da soma de pixels dentro de uma região, conforme representado na figura 9. Nesta figura, os índices **i**, **j**, **k** e **l** indicados na matriz integral representam o valor dos pixels da direita inferior de cada quadrante (A, B, C, D) destacado na imagem. Estes valores são utilizados para encontrar a soma dos pixels de uma região definida na imagem original sendo:

a) O valor da soma dos pontos pertencentes ao retângulo A na imagem original é representado pelo valor da posição (**i**), que vale 8 neste exemplo;

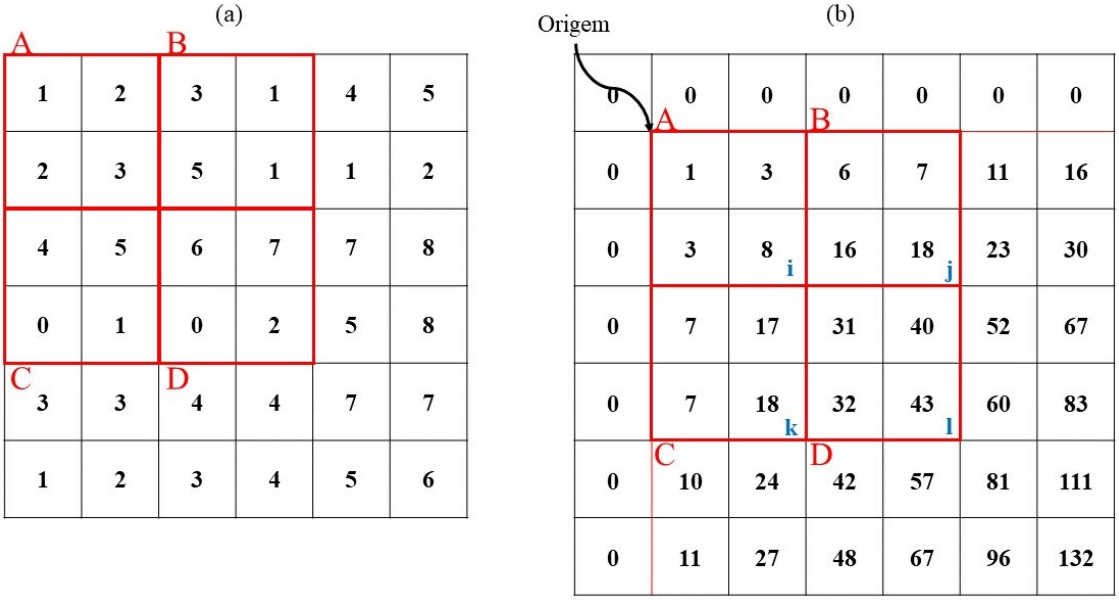
b) O valor da soma dos pontos pertencentes ao retângulo B é encontrado como o valor de ($\mathbf{j} - \mathbf{i}$), resultando em 10 neste exemplo;

c) O valor da soma dos pontos pertencentes ao retângulo C é encontrado como o valor de ($\mathbf{k} - \mathbf{i}$), resultando em 10 neste exemplo;

d) Por último o valor da região D é encontrado na forma $(\mathbf{l} + \mathbf{i}) - (\mathbf{j} + \mathbf{k})$ resultando em 15 neste exemplo.

Desta forma, através da operação de no máximo quatro valores, é possível encontrar a soma de pixels de qualquer região da imagem.

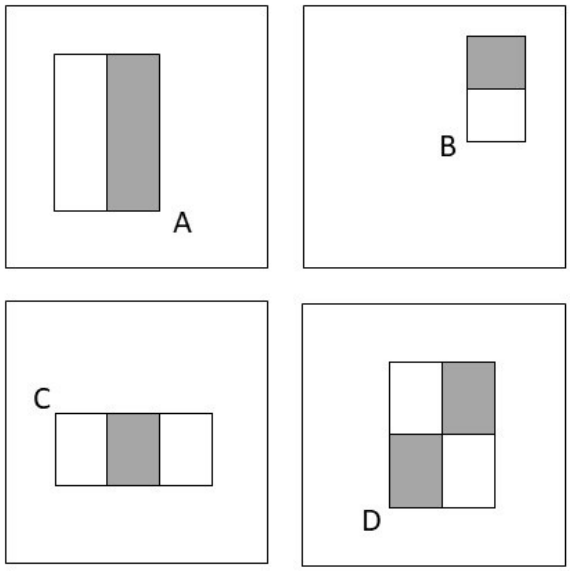
Figura 9 - Exemplo de imagem integral para algoritmo Viola-Jones



Fonte: Autor
Legenda: (a) imagem original, (b) imagem integral. Os quadrantes são indicados pelas letras A, B, C, D e os valores das posições i, j, k e l são utilizadas para o cálculo da soma nos quadrantes

Os parâmetros extraídos da segmentação da imagem em regiões retangulares são posteriormente classificados utilizando como base as transformadas de Haar, conforme figura 10 e proposto por Papageorgiou em [114]. Os blocos em cinza indicam somas negativas e em branco positivas. Desta forma, os parâmetros em A e B representam as diferenças entre dois blocos adjacentes da imagem, o parâmetro em C indica a soma de dois blocos menos a soma de um bloco central e por último em D representa a diferença entre pares diagonais. Estes parâmetros são úteis para extração de formatos na imagem como bordas, barras e outras importantes estruturas da imagem que podem servir de indicação para detecção do rosto.

Figura 10 - Parâmetros da transformada Haar



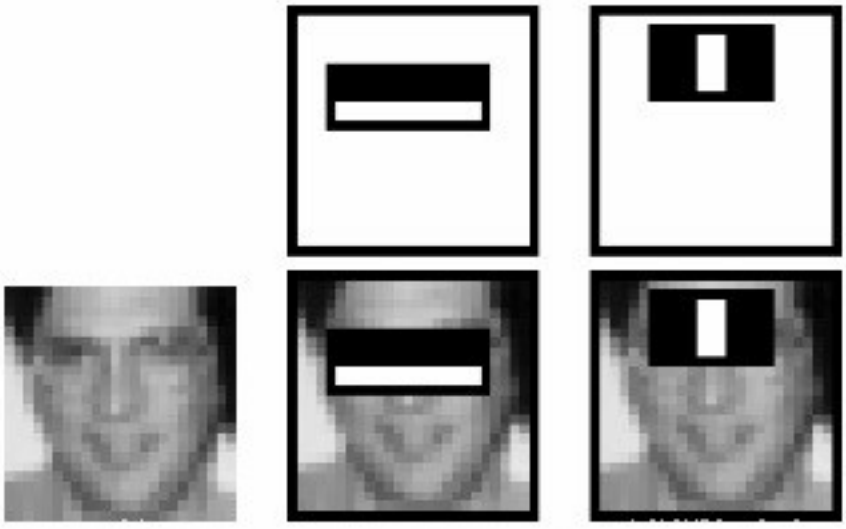
Fonte: Autor “adaptado de” Viola, 2001 [95]
Legenda: quadrantes em cinza indicam somas negativas e branco positivas. Parâmetros em A e B são resultados de blocos adjacentes, C parâmetro central e D diagonal

A classificação, bem como a seleção dos parâmetros são realizadas com algoritmo AdaBoost [115] em cascata e busca na imagem a região do rosto dado um determinado grupo de coeficientes extraídos de funções Haar. Para a escolha dos melhores parâmetros, um conjunto de imagens previamente sinalizadas ao sistema como exemplos positivos e negativos são utilizadas para treinar o classificador.

Inicialmente o algoritmo AdaBoost é utilizado para identificar um único classificador que seja mais discriminativo, ou seja, que melhor separe os exemplos positivos e negativos na entrada do sistema. Demais classificadores são posteriormente identificados e utilizados em cascata para obtenção de melhores resultados na identificação do rosto na imagem.

Os resultados de Viola-Jones apontam dois parâmetros como principais classificadores para tarefa de identificação da região facial, conforme figura 11. O primeiro parâmetro é selecionado devido a separação da região dos olhos e da parte inferior aos olhos, sendo a região dos olhos geralmente mais escura devido à sombra da órbita ocular. O segundo parâmetro ocorre devido ao mesmo fato, porém, identificado na separação da intensidade de luminância da região dos olhos e do nariz.

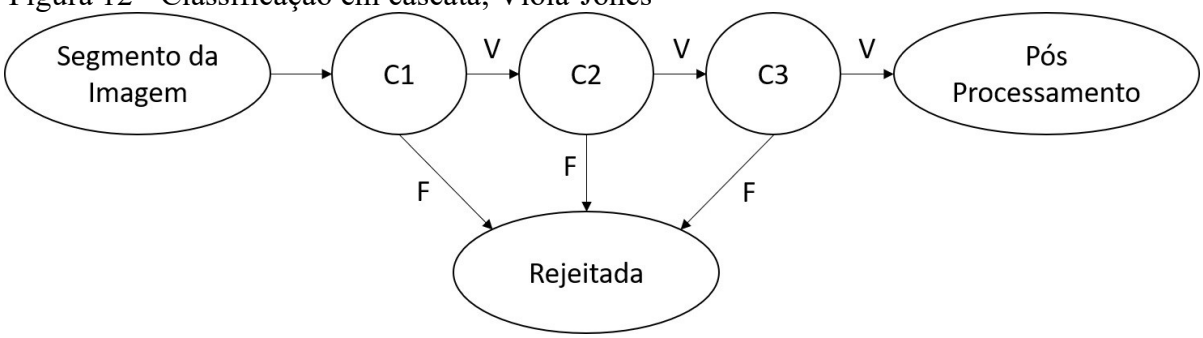
Figura 11 - Seleção dos parâmetros. Viola-Jones



Fonte: Viola, 2001 [95]
Legenda: da esquerda para a direita temos a imagem original, o primeiro e o segundo principal parâmetro selecionados

A classificação é realizada com aplicação do algoritmo AdaBoost em cascata, conforme figura 12, de forma que o sistema primeiramente segmenta a imagem em blocos e para cada bloco calcula as funções de Haar, caso os valores dessas funções sejam maiores ou iguais a um limiar pré-determinado para o classificador 1 (C1) ela passa para análise do classificador 2 (C2), caso contrário é automaticamente recusada e assim sucessivamente. A etapa de pós-processamento indicada na figura 12 pode consistir da aplicação de mais classificadores e/ou outros algoritmos que o sistema possa utilizar para melhorar a classificação. Deste modo, utilizando os classificadores ordenados em função do grau de discriminação, o sistema descarta rapidamente nos primeiros classificadores uma parte muito grande dos segmentos analisados, tornando o sistema mais eficiente.

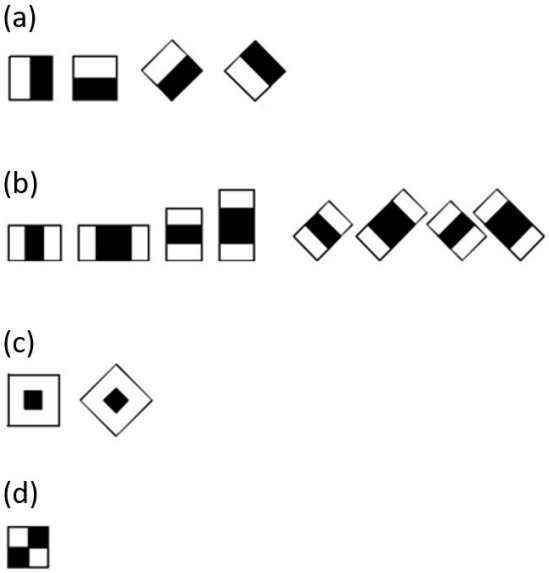
Figura 12 - Classificação em cascata, Viola-Jones



Fonte: Autor "adaptado de Viola, 2001 [95]
Legenda: C1, C2 e C3 são classificadores e a etapa de pós-processamento pode consistir mais classificadores ou outros algoritmos aplicados na sequência

Outros trabalhos posteriores à publicação de Viola-Jones, propuseram o uso de parâmetros da transformada Haar extraídos em outras direções resultando em uma melhoria no processo de classificação, como é o caso em [116], representado na figura 13.

Figura 13 - Novos parâmetros. Transformada Haar



Fonte: Autor "adaptado de" Lienhart [116]
Legenda: (a) parâmetros de borda, (b) parâmetros de linha, (c) parâmetros de centro e entorno e (d) parâmetro diagonal da transformada Haar

3.2 SEGMENTAÇÃO *K*-MEANS

Outra técnica muito comum em mineração de dados e problemas de aprendizado de máquina é a segmentação com aplicação do algoritmo k-means [96], [117] , que busca dividir um dado conjunto de dados (D) em k grupos $C_1, \dots, C_k$, de modo que a intersecção entre os novos conjuntos seja nula. A divisão é baseada no cálculo dos centroides $c_1, \dots, c_k$, referente a cada novo grupo, calculada como o ponto médio no espaço do conjunto. Para cada elemento do grupo D, sua distância Euclidiana para cada centroide é calculada, de modo que o elemento passa a pertencer ao centroide de menor distância.

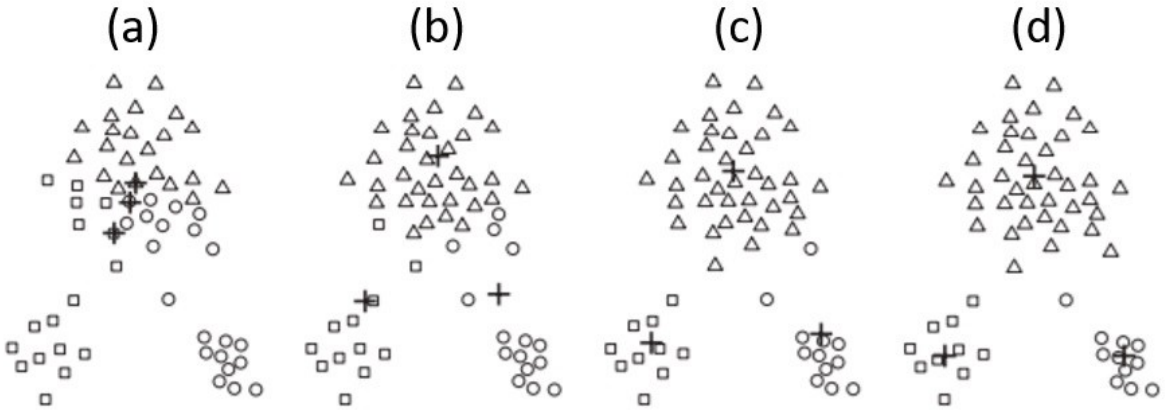
Após o primeiro agrupamento, os centroides são recalculados e as distâncias Euclidianas são novamente calculadas de modo que os pontos são atribuídos novamente ao grupo de centroide mais próximo, minimizando a distância quadrática Euclidiana de cada ponto para seu centroide. O algoritmo repete esses passos iterativamente até que os centroides não mudem de posição ou que a mudança seja menor que um fator pré-determinado para que

a convergência ocorra mais rapidamente.

Outras funções de proximidade além da distância Euclidiana podem alternativamente ser utilizadas como: distância de Manhattan; similaridade cossenoidal; divergência de Bregman. Para cada caso, o objetivo é sempre a minimização da medida de modo a agrupar melhor os elementos do conjunto, conforme descrito em [117].

A figura 14 mostra o funcionamento de uma segmentação *k*-means com $k = 3$, sendo a convergência atingida após 4 iterações.

Figura 14 - Segmentação *k*-means com 3 grupos



Fonte: Tan, 2005. [117]
Legenda: (a) iteração 1, (b) iteração 2, (c) iteração 3, (d) iteração 4

Este método pode ser aplicado para segmentar a distribuição da intensidade de cinza dos pixels em uma imagem, bem como a coloração dos pixels para identificar partes de uma imagem. No capítulo 9 desta tese, a intensidade média de movimento ao longo do tempo é segmentada, de forma a agrupar as regiões de movimento semelhante, sendo aplicada para encontrar a região labial, considerando a região facial como região de interesse.

4 COMPRESSÃO MPEG

Esse capítulo descreve a estrutura de um codificador de vídeo MPEG (*Moving Picture Experts Group*), seja ele MPEG-4 AVC (*Advanced Video Coding*) (H.264) [118] ou MPEG-H parte 2 HEVC (*High Efficiency Video Coding*) (H.265) [119], apontando suas particularidades quando necessário. A descrição do codec é necessária para entendimento dos algoritmos propostos nesta tese, que fazem uso dos vetores de movimento extraídos de vídeo codificado neste formato.

MPEG é um grupo de trabalho formado pela ISO (*International Organization for Standardization*)/IEC (*International Electrotechnical Commission*) que tem a função de definir e padronizar *codecs*. O termo *codec* é genericamente utilizado para descrever uma estrutura de um codificador/decodificador como um acrônimo das palavras *coder* e *decoder*, sendo sua função principal comprimir dados audiovisuais que podem ser posteriormente armazenados e/ou transmitidos.

Os *codecs* de áudio e vídeo definidos pelo grupo MPEG são os mais utilizados mundialmente, sendo os mais famosos: MP3 (MPEG-1/2 Camada de áudio 3) lançado em 1993 para compressão de áudio, muito utilizado até hoje em inúmeras aplicações comercias e o codec de vídeo MPEG-2 lançado em 1996 como um sucessor do MPEG-1 de 1993 para compressão de vídeo, tornando possível a aplicação em sistemas de transmissão de televisão digital, vídeo sob demanda, bem como armazenamento de vídeos digitais em mídias físicas como os próprios *smartphones*. Atualmente o codificador de vídeo MPEG-4 AVC (H.264) é amplamente utilizado para armazenamento de vídeos de alta resolução, sendo o seu sucessor o MPEG-H parte 2 HEVC (H.265) de 2013 o estado da arte para compressão de vídeo, e amplamente utilizado para vídeos de ultra resolução como 4K e 8K.

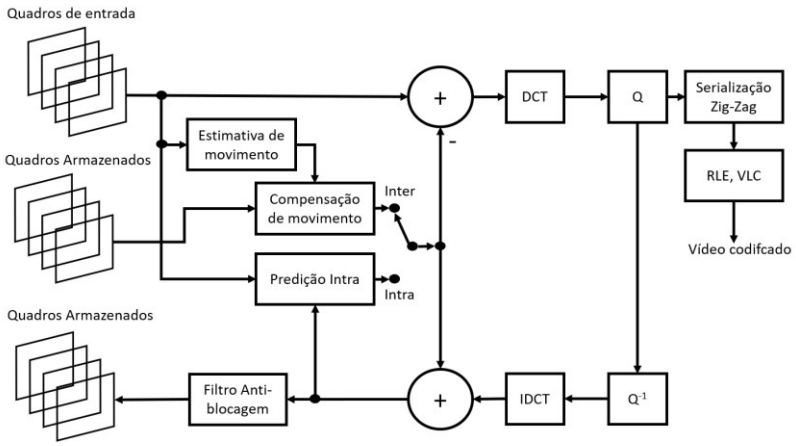
Para entender a estrutura básica de um codificador de vídeo é importante entender a estrutura de uma imagem digital, correspondente a um quadro do vídeo. A imagem é composta de 3 matrizes, usualmente armazenadas no formato **Y**, $\mathbf{P_B}$ e $\mathbf{P_R}$. A matriz **Y** corresponde ao canal de luminância da imagem e consiste na intensidade da cor em escala de cinza, sendo calculada em função das componentes RGB (*Red*, *Green*, *Blue*) e as matrizes $\mathbf{P_B}$ e $\mathbf{P_R}$ são as componentes de crominância, que representam conjuntamente as variações de tonalidade e saturação das cores, sendo $\mathbf{P_B}$ proporcional ao valor de azul menos a luminância e $\mathbf{P_R}$ proporcional ao vermelho menos a luminância [120], [121].

Essas matrizes podem ter todas as mesmas dimensões, correspondente a resolução de pixels da imagem, porém, é usual que as matrizes de $\mathbf{P_B}$ e $\mathbf{P_R}$ sejam subamostradas, ou seja,

tenham sua dimensão reduzida em comparação à matriz **Y** devido a características do olho humano, que é mais sensível a variações de intensidade luminosa. Esta compressão resultante da subamostragem de crominância pode ser ajustada pelo codec e a notação utilizada para indicação da subamostragem utilizada é $J:A:B$, sendo: J o número de pixels observados em um trecho de uma linha da matriz **Y**; A o número de amostras de crominância neste mesmo trecho e na mesma linha; B o número de amostras de crominância neste mesmo trecho, na linha seguinte. É padrão a utilização de $J = 4$ para esta notação. Seguem exemplos de valores padrões: 4:4:4 (sem subamostragem); 4:2:2 (metade do número de amostras na horizontal); 4:2:0 (metade do número de amostras na horizontal e metade na vertical); 4:1:1 (um quarto do número de amostras na horizontal) [120], [121].

A estrutura básica de um codificador MPEG-4 H.264 é descrita conforme figura 15. Dois tipos de compressão são utilizados: *Intra* e *Inter* [122]. A compressão do tipo *Intra* explora apenas a redundância espacial, resultando em estruturas conhecidas como quadro I (*Intra*). Já os quadros gerados com compressão do tipo Inter exploram a redundância temporal para extrair semelhanças entre quadros consecutivos. Os algoritmos de estimativa de movimento e de compensação de movimento são aplicados, resultando em uma representação de vetores de movimento do bloco de um quadro do vídeo, em comparação com um quadro de referência. Esses quadros podem ser do tipo P (*Predictive*) quando comparados apenas com quadros passados para estimativa de movimento ou B (*Bipredictive*) quando comparados com quadros passados ou futuros. Vale salientar que as técnicas aplicadas nos quadros *Intra* também são aplicadas as partes residuais dos quadros *Inter*, em que os movimentos dos blocos não foram detectados.

Figura 15 - Codificador MPEG-4 H.264



Fonte: Autor "adaptado de" Alencar, 2011 [120]
Legenda: diagrama em blocos do codificador MPEG-4 H.264, com representação das técnicas de compressão *Intra* e *Inter*

O vídeo codificado é armazenado como uma sequência de quadros comprimidos, sendo eles do tipo I, P ou B, resultando em uma sequência conhecida como GOP (*Group of Pictures*) que se repete dentro do codec, por exemplo, IBBPBBPBBPBBI, onde a periodicidade de quadros I é 12 e a distância entre quadros P e I ou P e P é 3. Essa estrutura é importante para periodicamente extrair um quadro I, que contém menor compressão e é usado como referência para os próximos quadros que utilizam compressão Inter [120].

O codec conta ainda com a possibilidade da aplicação de diferentes perfis que definem a qualidade do vídeo resultante, sendo os mais importantes dentro do codificador H.264 [120]:

a) *Baseline profile* - Apenas quadros I e P. Aplicado em videotelefonia, videoconferência, comunicação sem fio e radiodifusão televisiva como no sistema de televisão digital brasileiro para transmissão de vídeos de baixa resolução à dispositivos moveis;

b) *Main profile* - Suporta varredura de vídeo entrelaçada. Codificação utilizando também quadros do tipo B. Utilizado em radiodifusão televisiva;

c) *Extended profile* - Não suporta varredura de vídeo entrelaçada. Possui recuperação de erros aprimorada. Aplicado principalmente em aplicações de fluxo de mídia;

d) *High profile* - Suporte a representação de 8bits/símbolo com subamostragem de crominância 4:2:0, visa aplicações de alta resolução. HD-DVD, Blu-Ray, TV digital. Utilizado em radiodifusão televisiva, como no sistema de televisão digital brasileiro para transmissão de vídeos de alta resolução.

4.1 COMPRESSÃO INTRA

As técnicas utilizadas na compressão do tipo *Intra* do codificador MPEG exploram as semelhanças entre pixels vizinhos, ou seja, a redundância espacial de cada quadro do vídeo, resultando em quadros do tipo I. Essas técnicas são derivadas do codificador JPEG (*Joint Photographic Experts Group*) que é amplamente utilizado até hoje para compactação de arquivos de imagens estáticas. Todos os passos descritos a seguir são aplicados para cada um dos três canais da imagem separadamente: Luminância (**Y**) e crominância ($\mathbf{P_B}$ e $\mathbf{P_R}$) [121].

Inicialmente a matriz de cada canal da imagem é segmentada em blocos de tamanho fixo, usualmente de dimensão 8x8 pixels. Em cada bloco é aplicada uma DCT (*Discrete Cosine Transform*) bidimensional de forma que a matriz de pixels é transformada do domínio do espaço bidimensional para o domínio da frequência espacial bidimensional. A equação utilizada correspondente à DCT-II é definida como [123]:

$$\hat{s}_{u,v} = \frac{2}{\sqrt{NM}} C_u C_v \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} s_{x,y} \cos\left(\frac{(2x+1)u\pi}{2N}\right) \cos\left(\frac{(2y+1)v\pi}{2M}\right), \quad (34)$$

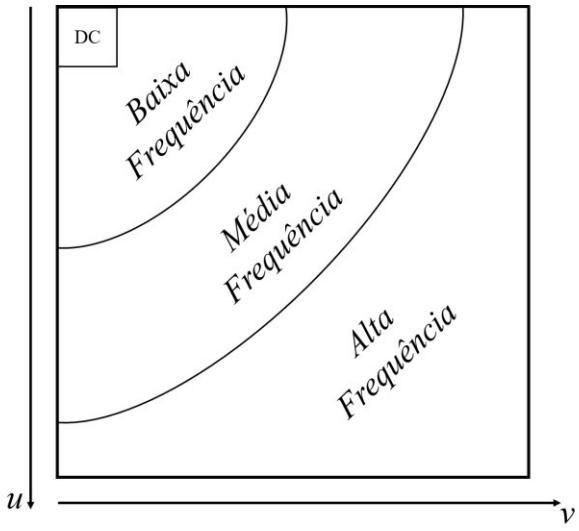
sendo $s_{x,y}$ correspondente a um pixel da matriz **S**, que representa um dos canais da imagem, $\hat{s}_{u,v}$ um coeficiente em frequência espacial armazenado na matriz $\hat{\mathbf{S}}$, para $u = 1, \ldots, M$ e $v = 1, \ldots, N$, sendo M e N o número de linhas e colunas do bloco da matriz **S** (usualmente 8 x 8 pixels), respectivamente. As constantes C_u e C_v são definidas da seguinte forma:

$$C_u = \begin{cases} \frac{1}{\sqrt{2}}, para\ u = 0 \\ 1\ caso\ contrário \end{cases}, \quad (35)$$

$$C_v = \begin{cases} \frac{1}{\sqrt{2}}, para\ v = 0 \\ 1\ caso\ contrário \end{cases}. \quad (36)$$

É possível observar a partir da equação que o primeiro coeficiente resultante da transformada, onde $u = v = 0$, corresponde ao nível médio (0 Hz ou nível DC) da matriz **S** e conforme os valores de u e v aumentam os coeficientes resultantes representam variações em maiores frequências da matriz, conforme figura 16.

Figura 16 - Variação em frequência da DCT bidimensional



Fonte:Autor
Legenda: aumento da frequência diretamente proporcional ao crescimento dos índices *u* e *v*

A matriz resultante (**Ŝ)** é posteriormente quantizada, de forma a comprimir o valor dos coeficientes. Como a alta frequência espacial é menos relevante para o olho humano [120], os coeficientes que sofrem maior compressão são os que possuem maiores índices u e v. Essa operação é realizada a partir do arredondamento da divisão ponto a ponto da matriz **Ŝ** pela matriz de quantização **Q**, resultando nos coeficientes da matriz **D**:

$$d_{u,v} = \text{round}\left(\frac{\hat{s}_{u,v}}{q_{u,v}}\right), \tag{37}$$

calculada para $u = 1, \ldots, M$ e $v = 1, \ldots, N$, sendo a matriz **Q** definida nas normas do codec [118], [119], conforme tabela 2 para compressão dos coeficientes do canal de Luminância (**Y**) e conforme tabela 3 para compressão dos coeficientes dos canais de crominância (**P_B** e **P_R**).

Tabela 2 - Matriz de quantização canal **Y**

Matriz Q – Canal de Luminância							
16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

Fonte: Autor “adaptado de” Robin, 2000 [121]
Legenda: matriz de quantização do canal de luminância definida na norma do codec H.264

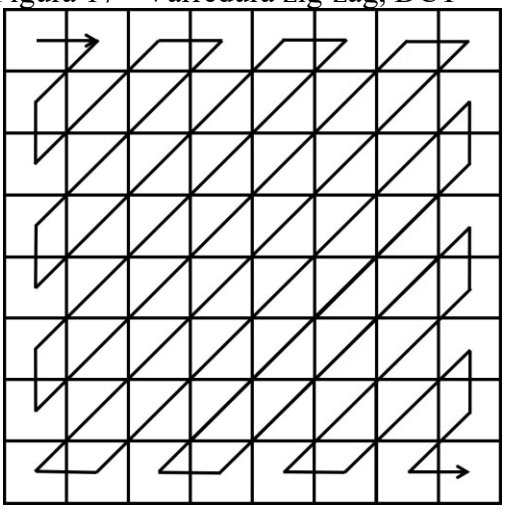
Tabela 3 - Matriz de quantização canal **P_B** e **P_R**

Matriz Q – Canal de Crominância							
17	18	24	47	99	99	99	99
18	21	26	66	99	99	99	99
24	26	56	99	99	99	99	99
47	66	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99
99	99	99	99	99	99	99	99

Fonte: Autor “adaptado de” Robin, 2000 [121]
Legenda: matriz de quantização dos canais de crominância definida na norma do codec H.264

A matriz **D** resultante de cada bloco transformado é posteriormente serializada aplicando uma técnica conhecida como varredura zig-zag, conforme figura 17, para posterior aplicação dos algoritmos VLC (*Variable Length Coding*) e RLE (*Run Length Encoding*) no vetor resultante [120], sendo o VLC uma codificação baseada na entropia dos dados, utilizando código de Huffman e RLE responsável por compactar sequências de dados repetidos [121], como os zeros que ocorrem devido ao arredondamento dos coeficientes durante a etapa de quantização.

Figura 17 - Varredura zig-zag, DCT



Fonte: Autor
Legenda: padrão de serialização da matriz transformada pela DCT

É possível observar dentro da estrutura do codificador, conforme figura 15, uma etapa de decodificação, com aplicação do processo conhecido como desquantização ou quantização inversa (Q^{-1}), seguido da transformada inversa dos cossenos ou *Inverse Discrete Cosine Transform* (IDCT). É importante observar que apesar de ser conhecido como quantização inversa essa etapa não retorna os coeficientes para os mesmos valores obtidos anteriormente à quantização, uma vez que este processo resulta em perdas que não podem ser recuperadas. Essa etapa de decodificação dentro do codificador é feita para que os quadros que são armazenados na memória interna e que são posteriormente utilizados como referência na compressão do tipo *Inter*, sejam idênticos aos quadros que serão recuperados pelo decodificador, diminuindo a propagação de erros.

4.2 COMPRESSÃO INTER

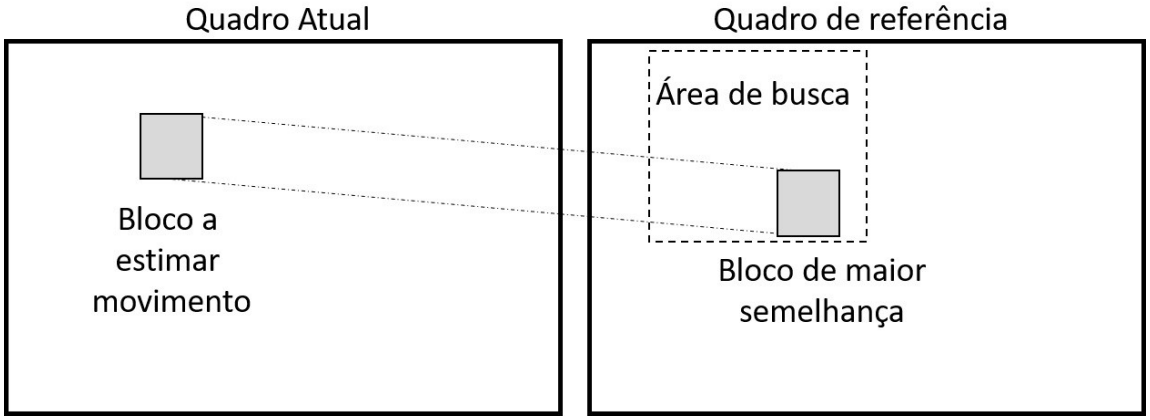
A etapa de compressão do tipo *Inter* explora a redundância temporal, ou seja, semelhança entre quadros sequenciais dentro de um vídeo, removendo partes repetidas e aplicando estimativa e compensação de movimento para maior compressão. Os quadros do tipo P utilizam outros quadros do tipo I ou P que foram comprimidos anteriormente como referência. Já os quadros do tipo B podem utilizar quadros I ou P passados e/ou futuros para melhor compensação do movimento. Esses algoritmos de implementados no codec MPEG são baseados em busca de blocos e resultam na representação dos quadros por vetores de movimento.

A estrutura de uma imagem é primeiramente segmentada em um macrobloco que corresponde a 16x16 amostras no canal de luminância e o equivalente no canal de crominância, dependendo da subamostragem utilizada. Por exemplo, para subamostragem 4:2:0 o tamanho do bloco no canal de crominância é a metade, ou seja, 8x8 amostras. Este macrobloco pode ser ainda subdividido em um bloco (8 x 8 amostras de luminância) ou até mesmo meio bloco (4 x 4 amostras de luminância). Para cada estrutura segmentada da imagem é realizada uma busca em um quadro vizinho de referência da melhor correspondência dessa estrutura ao redor de uma área de busca, conforme figura 18. Para o cálculo da melhor correspondência geralmente é utilizado o valor do erro quadrático médio dos pixels, conhecido como RMSE (*Root Mean Squared Error*), calculado como [109]:

$$RMSE = \sqrt{\frac{1}{NM}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(x_{i,j} - y_{i,j}\right)^{2}}, \tag{38}$$

sendo $x_{i,j}$ um elemento da matriz **X**, correspondente a um bloco de dimensão N x M a ser estimado o movimento, $y_{i,j}$ um elemento da matriz **Y** correspondente a um bloco de mesmo tamanho em um quadro de referência.

Figura 18 - Compensação de movimento MPEG



Fonte: Autor
Legenda: busca da melhor correspondência de um bloco em um quadro de referência

O resultado é uma representação de partes da imagem como vetores de movimento que são utilizados para comprimir a representação de um quadro do vídeo. Estes vetores são utilizados nesta tese para extração do movimento labial.

4.3 FFMPEG

Para a realização dos testes dos algoritmos propostos nos capítulos 7, 8 e 9, foi utilizada a biblioteca multiplataforma de código aberto FFmpeg (*Fast Forward MPEG*), disponível em (http://ffmpeg.org) [124]. Esta biblioteca foi utilizada para a manipulação dos vídeos codificados, como ferramenta para extração de parâmetros de movimento do vídeo codificado bem como para transcodificação do formato dos vídeos. Seu código permite as seguintes implementações:

a) codificar e decodificar em diversos formatos e padrões de codecs de áudio e vídeo, como o MPEG, com o uso da biblioteca libavcodec;

b) multiplexar e demultiplexar arquivos de áudio, vídeo e outras mídias com o uso da biblioteca libavformat;

c) transcodificar arquivos de áudio e vídeo com o uso da biblioteca FFmpeg;

d) filtrar arquivos de áudio e vídeo com o uso da biblioteca libavfilter;

e) reproduzir arquivos de áudio e vídeo com o uso da biblioteca ffplay;

f) realizar transmissões audiovisuais em tempo real com o uso do ffserver, entre outros.

4.4 MPEGFLOW

O mpegflow é um aplicativo desenvolvido por Kantorov e Laptev [125], disponível em (https://github.com/vadimkantorov/mpegflow), baseado no uso da biblioteca FFmpeg, que realiza a extração dos vetores de movimento de um vídeo codificado em MPEG. Nesta tese, o código foi modificado de forma que o arquivo resultante pudesse ser utilizado diretamente no MATLAB®, facilitando a integração com os outros algoritmos aplicados. A versão modificada pode ser acessada em <https://github.com/maparada/biometria-multimodal>.

5 MÉTRICAS PARA AVALIAÇÃO DO SISTEMA BIOMÉTRICO

Algumas métricas, padronizadas pelo comitê técnico ISO/IEC JTC 1/SC 37 [126], são utilizadas para avaliar um sistema biométrico e foram utilizadas na validação das propostas desta tese. Este comitê técnico da ISO/IEC é responsável por descrever padronizações no ramo da biometria, de forma a facilitar a interoperabilidade e intercâmbio de resultados entre diferentes sistemas e aplicações

Para entendimento das medições, suponha um sistema biométrico de verificação de identidade e um grupo de amostras de teste, que consiste tanto em amostras de clientes (usuários cadastrados que se identificam corretamente durante o uso) e impostores (usuários cadastrados que se identificam como outro usuário durante o uso ou usuários não cadastrados) previamente sinalizados ao sistema. A avaliação dos resultados pode ser realizada em função de algumas principais métricas, sendo medidas após etapa de decisão, como:

a) TAR (*True Acceptance Rate*) - número de usuários aceitos pelo sistema em função do número total de clientes avaliados;

b) TRR (*True Rejection Rate*) - número de usuários rejeitados pelo sistema em função do número total de impostores avaliados;

c) FAR (*False Acceptance Rate*) - número de usuários aceitos pelo sistema em função do número de impostores avaliados;

d) FRR (*False Rejection Rate*) - número de usuários rejeitados pelo sistema em função do número de clientes avaliados;

e) FTE (*Failure to enroll*) - número de tentativas falhas em função do número de tentativas totais de cadastramento no sistema;

f) FTA (*Failure to Aquire*) - número de tentativas falhas em função do número de tentativas totais de aquisição do sinal necessário para verificação biométrica;

g) EER (*Equal Error Rate*) - Como as taxas FRR e FAR variam de acordo com o limiar adotado na etapa de decisão, a taxa EER representa o valor em que FRR = FAR.

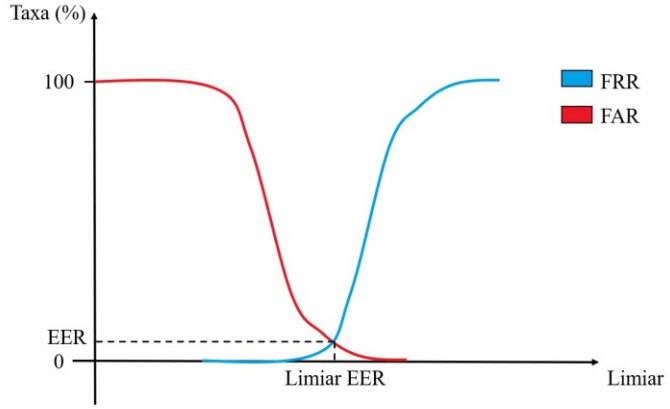
O EER é muitas vezes utilizado como o principal parâmetro para medir a qualidade de um sistema biométrico, sendo que quanto menor o valor obtido, maior a confiabilidade do sistema. Esta taxa é usualmente encontrada pelo resultado de uma das seguintes curvas: curva de FAR e FRR em função dos valores das medidas escolhidas como limiar na decisão; curva ROC (*Receiver Operating Characteristic*); curva DET (*Detection Error Tradeoff*), conforme descrito a seguir.

5.1 FRR E FAR EM FUNÇÃO DO LIMIAR.

As taxas de FRR e FAR analisadas em função da escolha de diferentes limiares utilizados na etapa de decisão, como o que pode ser obtido com o LLR descrito na seção 2.9, resulta na curva representada na figura 19. O ponto de cruzamento entre as duas curvas determina a taxa EER. Caso não ocorra cruzamento fora do eixo x, EER $= 0$ e o sistema de reconhecimento pode ser considerado ideal, pois existe um valor de limiar que ao ser utilizado resulta em 100 % de acerto na etapa de reconhecimento do sistema.

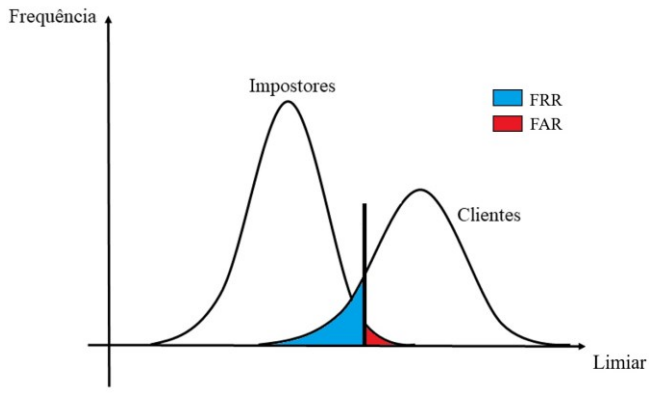
Este gráfico também pode ser obtido como a integral das áreas destacadas na figura 20, que representa a distribuição das métricas obtidas nos testes de correspondência dos parâmetros e modelos, tanto para clientes como impostores.

Figura 19 - Curva da distribuição de FRR e FAR



Fonte: Autor
Legenda: limiar x FRR e FAR. Cruzamento das curvas representa a taxa EER

Figura 20 - Curva da distribuição das métricas de clientes e impostores

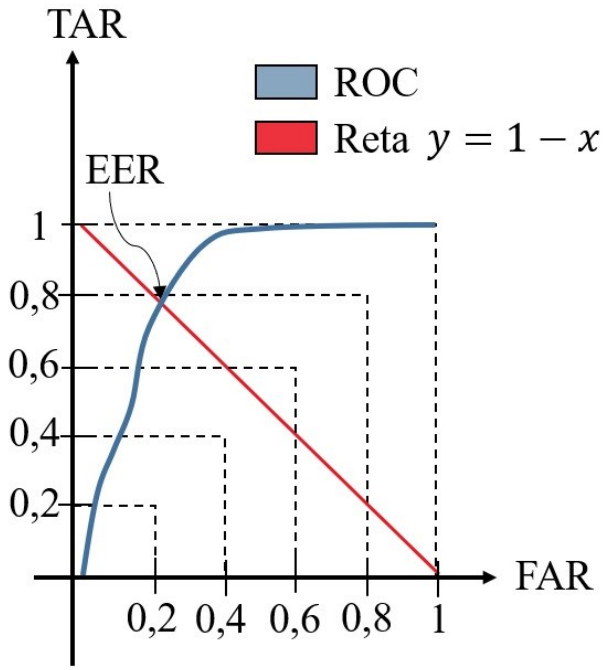


Fonte: Autor
Legenda: curva dos limiares de impostores e clientes. As áreas destacadas representam as taxas de FAR e FRR

5.2 CURVA ROC

Curva ROC é utilizada para representar graficamente a variação de TAR por FAR de um sistema de classificação binário, como por exemplo um sistema de verificação biométrico, ou outras aplicações como resultados de testes médicos e problemas de aprendizado de máquina. O cruzamento da curva ROC com a curva $y = 1 - x$ apresenta o valor de EER do sistema, conforme pode ser observado na figura 21.

Figura 21 - Curva ROC

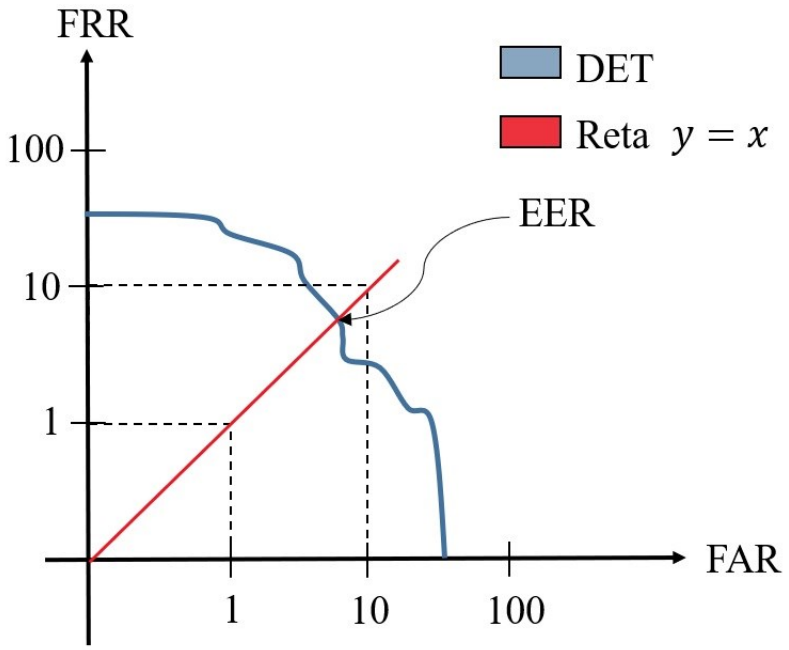


Fonte: Autor
Legenda: FAR x TAR. Cruzamento com a reta $y = 1 - x$ representa a taxa EER

5.3 CURVA DET

Detection Error Tradeoff (DET) é uma variação da curva ROC proposta em 1997 por Martin [127] onde é representado a variação de FRR em função de FAR, usualmente em escala logarítmica. A taxa EER pode ser encontrada pelo cruzamento desta curva com uma reta da função identidade, conforme a figura 22. Esta curva apresenta uma representação mais linear do que a curva ROC, ampliando a visualização.

Figura 22 - Curva DET



Fonte: Autor
Legenda: FAR x FRR. Cruzamento com a reta $y = x$ representa a taxa EER

6 BASE DE DADOS XM2VTS

XM2VTS é uma base dados audiovisual utilizada para testes e desenvolvimento de sistemas biométricos multimodais, resultado de um projeto desenvolvido no centro de pesquisa CVSSP (*Center for Vision Speech and Signal Processing*) da universidade de Surrey, Inglaterra, disponível em <http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/> [128].

6.1 CONTEÚDO DA BASE DE DADOS XM2VTS

Nesta base de dados, 371 voluntários foram gravados em quatro sessões, com diferença de um mês entre cada uma, sendo os dados audiovisuais de 295 destes locutores disponibilizados, em que cada indivíduo é referenciado na base de dados através de um número de três dígitos que segue a sequência em que foram capturados. Durante cada sessão foi realizada a captura audiovisual da locução de duas frases distintas:

a) "zero one two three four five six seven eight nine five zero six nine two eight one three seven four", nomeada nesta teste como frase 1.

b) "Joe took fathers green shoe bench out", nomeada nesta tese como frase 2.

Sendo a frase 1 capturada duas vezes em cada sessão e a segunda uma única vez. A base conta também com os arquivos de áudio disponíveis separados do vídeo. Neste caso, os áudios da frase 1 foram separados em dois arquivos, sendo um deles correspondente a "*zero one two three four five six seven eight nine*" e o outro a "*five zero six nine two eight one three seven four*", porém, como os testes realizados nesta tese combinam parâmetros do áudio e vídeo, os arquivos de áudio considerados foram extraídos diretamente do vídeo de modo que o áudio e vídeo correspondem a frase 1 completa.

Todos os vídeos da base de dados são codificados em DV AVI com resolução de 720 x 576 pixels e uma taxa de 25 quadros por segundo ou *frames per second* (*fps*) e os áudios em formato PCM WAV, 16 bits, 32 kHz. Para os testes realizados os vídeos foram transcodificados para MPEG-4 H.264, perfil *baseline* com subamostragem de crominância 4:2:0 e intervalo entre quadros I igual a 12, com o uso da biblioteca FFmpeg, mantendo a mesma resolução e quantidade de quadros por segundo, sendo seu áudio transcodificado para MPEG-4 HE-AAC v2. Esta transcodificação foi necessária para a extração dos vetores de movimento do vídeo codificado em MPEG que é utilizado como base para os algoritmos propostos.

6.2 PROTOCOLO LAUSANNE

O protocolo Lausanne [129] propõe uma padronização de utilização da base de dados XM2VTS para facilitar a possibilidade de comparação dos resultados com outros trabalhos, listando a quantidade de usuários e frases que devem ser utilizadas nas etapas de treinamento, avaliação e teste, bem como as métricas que devem ser utilizadas para medir o desempenho do sistema biométrico, de modo a padronizar os resultados.

Duas configurações possíveis (configuração I e II) podem ser utilizadas, diferindo entre elas os dados que são utilizados na fase de treinamento, avaliação e teste, conforme tabelas 4 e 5 respectivamente. Sendo a fase de treinamento utilizada para adaptação do modelo de cada cliente, a fase de avaliação utilizada para estimar o limiar a ser utilizado pelo sistema e a fase de teste utilizada para avaliação final do desempenho do sistema. Em ambas as configurações a frase 1 é utilizada e, portanto, suas duas capturas de cada sessão são indicadas como repetição 1 e 2.

Tabela 4 - Protocolo Lausanne, configuração I

Sessão	Repetição	Clientes	Impostores	
1	1	Treinamento	Avaliação	Teste
	2	Avaliação		
2	1	Treinamento		
	2	Avaliação		
3	1	Treinamento		
	2	Avaliação		
4	1	Teste		
	2	Teste		

Fonte: Luettin e Maitre, 1998 [129]
Legenda: treinamento, avaliação e testes com a utilização da frase 1

Tabela 5 - Protocolo Lausanne, configuração II

Sessão	Repetição	Clientes	Impostores	
1	1	Treinamento	Avaliação	Teste
	2	Treinamento		
2	1	Treinamento		
	2	Treinamento		
3	1	Avaliação		
	2	Avaliação		
4	1	Teste		
	2	Teste		

Fonte: Luettin e. Maitre, 1998 [129]
Legenda: treinamento, avaliação e testes com a utilização da frase 1

Na configuração I, dados de 3 diferentes sessões são utilizadas para treinamento, e dados destas mesmas sessões são utilizadas para fase de avaliação, sendo uma configuração boa para o treinamento por utilizar diferentes sessões nesta fase. Na configuração II, dados de apenas 2 sessões são utilizadas para treinamento e dados de uma outra sessão diferente são utilizados na fase de avaliação, tornando esta configuração ruim para o treinamento, porém boa por dissociar as sessões utilizadas no treinamento e na avaliação.

A configuração I foi utilizada nos testes realizados nesta tese, utilizando a frase 1 que foi capturada duas vezes em cada sessão. O protocolo sugere ainda o uso de 200 locutores como clientes, 25 impostores na fase de avaliação e 70 impostores na fase de teste. É importante ressaltar que o total de locutores da base de dados é 295, porém, como o áudio do locutor nomeado como 313 não foi capturado corretamente durante a sessão 2, este locutor foi desconsiderado dos testes. Desta forma, o número total de locutores foi reduzido para 294 e um total de 69 impostores foi utilizado na etapa de teste, em discordância aos 70 sugeridos pelo protocolo.

Como o protocolo não menciona os locutores que devem ser utilizados na criação do modelo UBM, todos os locutores foram utilizados para este propósito, de forma a criar o modelo mais genérico possível. A configuração final de locutores utilizada no treinamento, avaliação e testes é descrita pelas tabelas 6, 7 e 8 respectivamente.

Tabela 6 - Locutores na fase de treinamento

Clientes Cadastrados										
000	001	002	003	004	005	006	007	008	009	010
011	012	013	016	017	018	019	020	021	022	023
024	025	026	027	028	029	030	031	032	033	034
035	036	037	038	039	040	041	042	043	044	045
046	047	048	049	050	051	052	053	054	055	056
057	058	059	060	061	062	064	065	066	067	068
069	070	071	072	073	074	075	078	079	080	081
082	083	085	086	087	088	089	090	091	092	093
095	096	098	099	101	102	103	104	105	107	108
109	110	111	112	113	114	115	116	119	120	121
122	123	124	125	126	127	128	129	130	131	132
133	134	135	136	137	138	140	141	142	143	145
146	147	148	149	150	152	153	154	155	157	158
159	160	161	163	164	165	166	167	168	169	170
171	172	173	174	175	176	177	178	179	180	181
182	183	185	187	188	189	190	191	193	196	197
198	199	200	201	202	203	206	207	208	209	210
211	212	213	215	216	218	219	221	222	224	225
226	227									

Fonte: Autor
Legenda: locutores utilizados para treinamento dos modelos específicos

Tabela 7 - Impostores na fase de avaliação

Impostores na avaliação										
228	229	231	232	233	234	235	236	237	240	241
242	243	244	246	248	249	250	253	255	258	259
261	263	264								

Fonte: Autor

Legenda: locutores utilizados como impostores na fase de avaliação do sistema ASV

Tabela 8 - Impostores na fase de teste

Impostores no teste										
266	267	269	270	271	272	274	275	276	278	279
280	281	282	283	284	285	286	287	288	289	290
292	293	295	300	301	305	310	312	314	315	316
317	318	319	320	321	322	323	324	325	328	329
330	331	332	333	334	335	336	337	338	339	340
341	342	357	358	359	360	362	364	365	366	367
369	370	371								

Fonte: Autor

Legenda: locutores utilizados como impostores na fase de teste do sistema ASV. Locutor 313 removido devido a falhas na captura dos dados

Esta configuração resulta nas seguintes quantidades de combinações avaliadas:

a) 3 frases de treino por cliente, totalizando 600 combinações (200 clientes x 3 frases);

b) 3 frases de avaliação por cliente, totalizando 600 frases e 8 frases de avaliação para cada impostor, avaliadas para cada cliente, totalizando 40.000 combinações (25 impostores x 200 clientes x 8 frases);

c) 2 frases de teste por cliente, totalizando 400 combinações (200 clientes x 2 frases) e 8 frases de avaliação para cada impostor, sendo testadas para cada cliente, totalizando 110.400 combinações (69 impostores x 200 clientes x 8 frases).

Para verificação do desempenho do sistema, o protocolo sugere métricas para avaliação de forma a padronizar os resultados de teste utilizando esta mesma base de dados. As taxas de falsa aceitação FAR e falsa rejeição FRR, são encontradas tanto na etapa de avaliação como na etapa de teste em função do limiar (que pode ser resultado de uma métrica LLR) como sendo:

$$FAR = \frac{EI}{I} * 100\ \%, \tag{39}$$

$$FRR = \frac{EC}{C} * 100\ \%, \tag{40}$$

onde EI é o número de impostores aceitos, I é o número total de impostores, EC é o número de clientes recusados e C é o número total de clientes.

Embora a taxa de EER seja muito utilizada para verificação do desempenho de um sistema de biometria, esta métrica não representa um caso de uso real de um sistema uma vez que o limiar que leva a decisão de recusar ou aceitar um usuário é encontrada posteriormente, já em mãos da real identidade dos usuários. Desta forma, o protocolo sugere a escolha de três diferentes limiares dado um grupo de limiares observados (T) e conhecendo previamente a identidade dos locutores, sendo:

$$T_{FAR} = \underset{T}{\operatorname{argmin}}(FRR_E)\,;\; FAR_E = 0, \tag{41}$$

onde T_{FAR} representa o limiar mínimo na curva de falsa rejeição obtida na avaliação, FRR_E (*False Rejection Rate Evaluation*), para quando o valo de falsa aceitação na avaliação, FAR_E(*False Acceptance Rate Evaluation*), é igual a zero,

$$T_{EER} = (T); FRR_E = FAR_E, \tag{42}$$

onde T_{EER} representa o limiar no ponto de cruzamento da curva FRR_E e FAR_E,

$$T_{FRR} = \underset{T}{\operatorname{argmin}}(FAR_E)\,;\; FRR_E = 0, \tag{43}$$

onde T_{FRR} representa o ponto mínimo na curva FAR_E para quando FRR_E é igual a zero.

A etapa de teste é realizada utilizando separadamente cada um destes três limiares como critério de decisão, gerando um par de FAR e FRR para cada escolha. Para limiar igual a T_{FAR} é obtido:

$$FAR_{T_{FAR}}\ e\ FRR_{T_{FAR}}. \tag{44}$$

Para limiar igual a T_{EER}:

$$FAR_{T_{EER}}\ \text{e}\ FRR_{T_{EER}}. \tag{45}$$

Para limiar igual a T_{FRR}:

$$FAR_{T_{FRR}}\ \text{e}\ FRR_{T_{FRR}}. \tag{46}$$

Para cada limiar é encontrada a taxa *TER* (*Total Error Rate*), que representa soma de

FAR e FRR na etapa de teste, conforme segue:

$$TER_{T_{FAR}} = FAR_{T_{FAR}} + FRR_{T_{FAR}}, \quad (47)$$

$$TER_{T_{EER}} = FAR_{T_{EER}} + FRR_{T_{EER}}, \quad (48)$$

$$TER_{T_{FRR}} = FAR_{T_{FRR}} + FRR_{T_{FRR}} \quad (49)$$

e por último a taxa de *VR* (*Verification Rate*), que é utilizada para validação final do desempenho do sistema, é calculada como sendo:

$$VR = 100 - TER_{T_{EER}}. \quad (50)$$

Apesar de não considerada no protocolo, ainda é possível avaliar o desempenho em função da taxa de EER na etapa de teste, EER_T, podendo ser determinada através das curvas ROC ou DET apresentadas no capítulo 5.

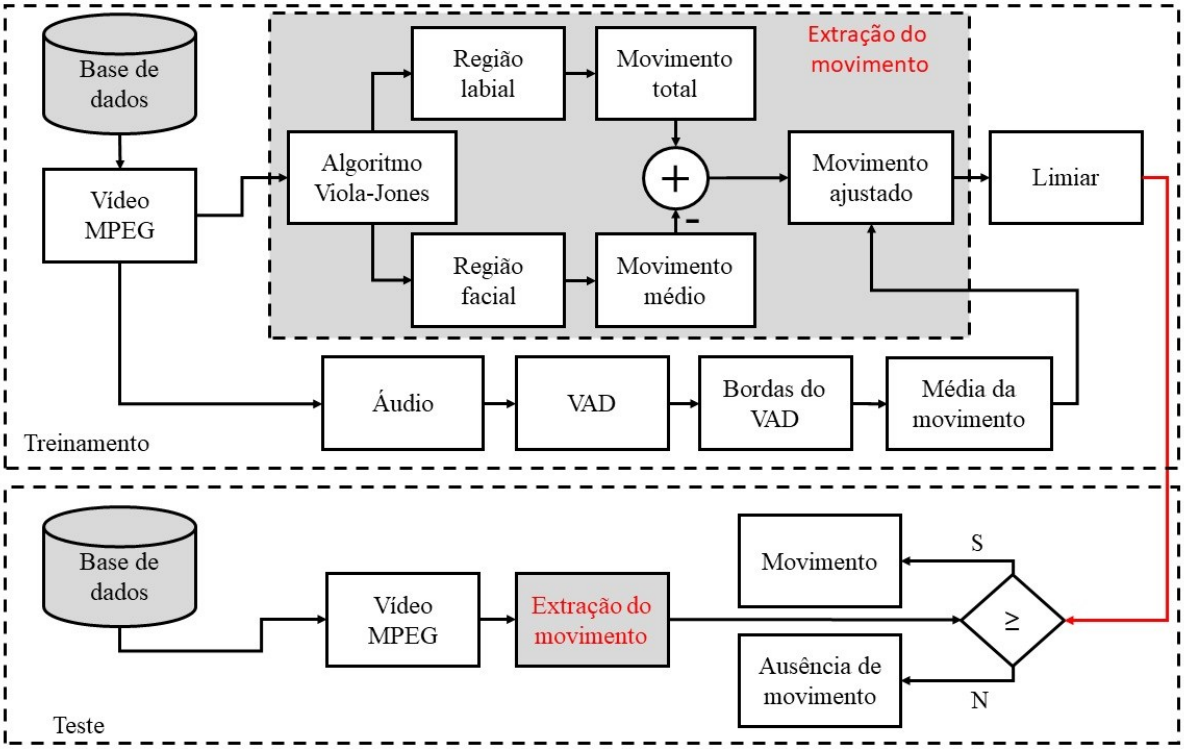
7 DETECÇÃO VISUAL DE ATIVIDADE VOCAL

O algoritmo de detecção de voz proposto pelo autor desta tese em [110], combina o uso de características do áudio e parâmetros extraídos do vídeo codificado em MPEG-4 H.264 de forma a tornar a detecção robusta mesmo com baixos valores de SNR no áudio. O uso de informações de movimento extraídas diretamente do vídeo codificado resulta em uma redução na carga computacional exigida para extração dos parâmetros do movimento do vídeo quando comparado com outras técnicas, como avaliado em [125].

7.1 ALGORITMO PROPOSTO

A metodologia do algoritmo proposto consiste de duas fases: fase de treinamento, onde o valor do limiar de movimento labial para um dado locutor durante a produção de fala é encontrado e fase de teste onde o VVAD é avaliado em função do limiar previamente encontrado. O processo completo é descrito pela figura 23 e descrito nas sessões seguintes.

Figura 23 - Diagrama em blocos do VVAD proposto



Fonte: Autor
Legenda: fases de treinamento e teste destacadas em pontilhado

7.1.1 FASE DE TREINAMENTO

Nesta fase os sinais de áudio e vídeo são analisados separadamente, sendo que para o vídeo a extração dos vetores de movimento do vídeo codificado é realizada, resultando para cada quadro do vídeo uma matriz $\mathbf{M} = \{m_{l,c}\}$, $l = 1, \dots, l_{max}$ e $c = 1, \dots, 4$, que armazena os valores dos movimentos horizontal e vertical para cada bloco ou macrobloco, sendo a posição inferior direita do bloco indicada nas duas primeiras colunas da matriz, conforme pode ser visto em um exemplo da matriz resultante da aplicação do mpegflow na tabela 9.

Esta matriz é resultado da aplicação do software mpegflow [125] descrito na seção 4.4, que é por sua vez baseado nas bibliotecas FFmpeg. Este aplicativo foi modificado para que os resultados das matrizes pudessem ser diretamente exportados para o MATLAB®, facilitando o restante da aplicação do algoritmo.

Tabela 9 - Matriz de movimento via mpegflow

Matrix M			
Posição Horizontal	***Posição Vertical***	***Movimento Horizontal***	***Movimento vertical***
8	8	-2	-1
16	8	2	1
24	8	1	2
32	8	0	2
⋮	⋮	⋮	⋮
8	16	0	4
16	16	-1	3

Fonte: Autor
Legenda: trecho da matriz **M,** resultante de uma execução do aplicativo mpegflow

O movimento é posteriormente separado em duas matrizes, sendo elas: **V**, que representa apenas as informações de movimento na direção vertical e **H** que representa as informações de movimento na direção horizontal, sendo descritas da seguinte forma:

$$I = \{m_{l,2} - 7, m_{l,2} - 6, m_{l,2} - 5, \dots, m_{l,2}\}, \tag{51}$$

$$J = \{m_{l,1} - 7, m_{l,1} - 6, m_{l,1} - 5, \dots, m_{l,1}\}, \tag{52}$$

$$v_{i,j} = m_{l,4};\ i \in I, \tag{53}$$

$$h_{i,j} = m_{l,3};\ j \in J. \tag{54}$$

São encontradas varrendo sequencialmente as linhas da matriz **M** indexadas por $l = 1, \ldots, l_{max}$, sendo l_{max} a quantidade de linhas da matriz **M**.

Utilizando a tabela 9 como exemplo, as matrizes de movimento vertical (**V**) e horizontal (**H**) são demonstradas na figura 24. A próxima etapa consiste em encontrar a média do movimento labial de um locutor durante a produção de fala. Para isso, duas regiões são detectadas na imagem (facial e labial). Os quadros do tipo *Intra* são utilizados para referência na compressão MPEG e não contam com a estimativa de movimento, sendo periódicos e de baixa compressão e, por este motivo, foram escolhidos para realizar a detecção das regiões periodicamente, uma vez que o usuário pode mover o rosto durante o passar do tempo do vídeo. As regiões estimadas nestes quadros são utilizadas para os seguintes quadros do tipo *Inter*. As matrizes de movimento nos quadros *Intra* são extraídas interpolando as matrizes de movimento de dois quadros *Inter* vizinhos ou, no caso do primeiro quadro do vídeo, consideradas nulas.

Figura 24 - Matrizes de movimento horizontal e vertical

(a)

Matriz **V**

linhas / colunas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1
2	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1
3	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1
4	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1
5	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1
6	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1
7	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1
8	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1
9	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	3
10	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	3
11	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	3
12	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	3
13	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	3
14	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	3
15	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	3
16	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	3

(b)

Matriz **H**

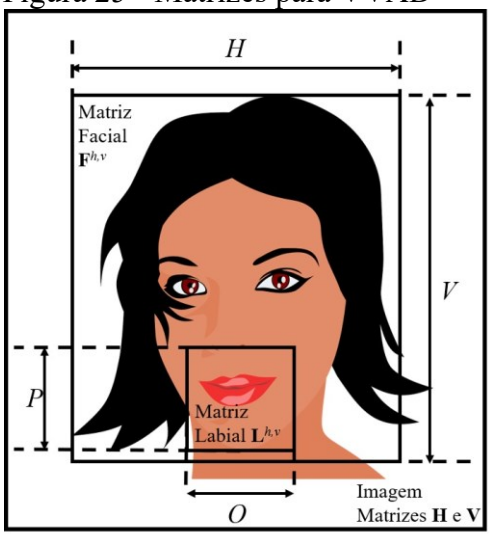
linhas / colunas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	-2	-2	-2	-2	-2	-2	-2	-2	2	2	2	2	2	2	2	2
2	-2	-2	-2	-2	-2	-2	-2	-2	2	2	2	2	2	2	2	2
3	-2	-2	-2	-2	-2	-2	-2	-2	2	2	2	2	2	2	2	2
4	-2	-2	-2	-2	-2	-2	-2	-2	2	2	2	2	2	2	2	2
5	-2	-2	-2	-2	-2	-2	-2	-2	2	2	2	2	2	2	2	2
6	-2	-2	-2	-2	-2	-2	-2	-2	2	2	2	2	2	2	2	2
7	-2	-2	-2	-2	-2	-2	-2	-2	2	2	2	2	2	2	2	2
8	-2	-2	-2	-2	-2	-2	-2	-2	2	2	2	2	2	2	2	2
9	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1
10	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1
11	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1
12	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1
13	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1
14	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1
15	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1
16	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1

Fonte: Autor
Legenda: (a) matriz de movimento vertical e (b) matriz de movimento horizontal obtidas conforme tabela 9

Para a tarefa de detecção das regiões faciais e labial, o algoritmo Viola-Jones é aplicado primeiramente para detecção da região facial e posteriormente a metade inferior desta região é utilizada como ROI para detecção labial utilizando o mesmo algoritmo, porém, treinado previamente para detecção de imagens do lábio. As matrizes de movimento **H** e **V** de dimensão

V x H, são segmentadas em partes correspondentes ao movimento de cada região, conforme figura 25.

Figura 25 - Matrizes para VVAD



Fonte: Autor
Legenda: região facial e labial com suas matrizes e dimensões destacadas

Os passos abaixo descritos são realizados para cada quadro de um vídeo, de um dado locutor. Por este motivo, para facilitar a notação das equações, os índices referentes a quadros e vídeos serão omitidos até que se façam necessários. Para as equações seguintes, índices sobrescritos h indicam direção horizontal e sobrescrito v a direção vertical. Sendo $\mathbf{F}^h$a matriz de movimento horizontal da região facial e $\mathbf{F}^v$ a matriz de movimento vertical da região facial ambas de dimensão V x H, $\mathbf{L}^h$a matriz de movimento horizontal da região labial e $\mathbf{L}^v$ a matriz de movimento vertical da região labial, ambas de dimensão P x O, conforme indicado na figura 25.

Primeiramente os movimentos médios na região facial a^h e a^v são encontrados para serem utilizados na compensação do movimento labial, da forma:

$$a^h = \frac{1}{HV}\sum_{i=1}^{H}\sum_{j=1}^{V} f_{i,j}^h\,, \tag{55}$$

$$a^v = \frac{1}{HV}\sum_{i=1}^{H}\sum_{j=1}^{V} f_{i,j}^v\,. \tag{56}$$

sendo $f_{i,j}^{h}$ um elemento da matriz $\mathbf{F}^{h}$ e $f_{i,j}^{v}$ um elemento da matriz $\mathbf{F}^{v}$**.** Essa compensação de movimento é importante para que os movimentos da cabeça não influenciem falsamente na detecção do movimento labial.

Os movimentos ajustados totais da região labial (c^{h} e c^{v}) são calculados como:

$$c^{h} = \frac{1}{\sqrt{OP}} \sum_{i=1}^{O} \sum_{i=1}^{P} \left(\left| l_{i,j}^{h} - a^{h} \right| \right), \tag{57}$$

$$c^{v} = \frac{1}{\sqrt{OP}} \sum_{i=1}^{O} \sum_{i=1}^{P} \left(\left| l_{i,j}^{v} - a^{v} \right| \right). \tag{58}$$

sendo $l_{i,j}^{h}$ um elemento da matriz $\mathbf{L}^{h}$ e $l_{i,j}^{v}$ um elemento da matriz $\mathbf{L}^{v}$.

Estes valores são análogos ao valor médio do movimento ajustado, que compensa as diferentes distâncias que o locutor pode se encontrar diante da câmera. A divisão pelo valor da raiz quadrada da área no lugar da divisão pela área faz com que os valores desses coeficientes sejam iguais ao resultado do primeiro coeficiente resultante da aplicação de uma transformada discreta dos cossenos bidimensional nesta matriz de movimento compensada, sendo essa equivalência importante para compatibilidade com o algoritmo de extração de características de movimento relacionadas ao locutor proposto no capítulo 8.

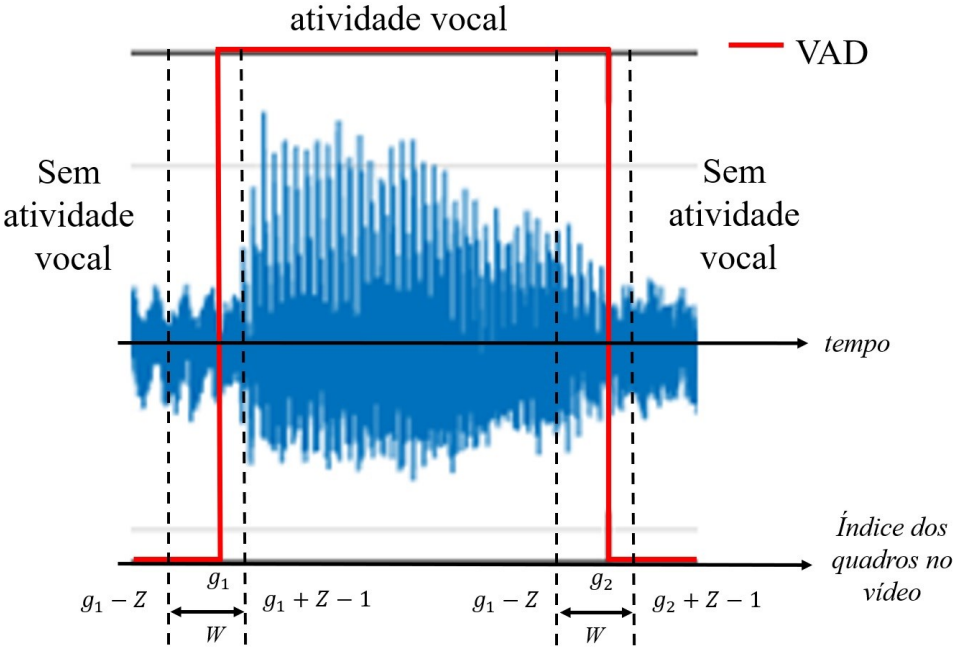
O áudio é processado separadamente, aplicando um detector de atividade de voz baseado no limiar da energia do sinal e cruzamentos por zero [54], sendo utilizado para detectar os instantes em que ocorrem o início e fim de uma locução. Próximo a esses instantes é possível assumir que há uma alta probabilidade de ocorrência de movimento labial, decorrente da produção da fala e por este motivo esses instantes são utilizados para extração do limiar de movimento de cada locutor. Para que o algoritmo seja corretamente aplicado é preciso garantir que os áudios da fase de cadastramento sejam capturados com valores elevados de relação sinal-ruído, de forma a garantir indicações confiáveis do VAD. Outras técnicas de VAD podem ser aplicadas nesta etapa, porém, esta foi escolhida devido a simplicidade de implementação e ótimos resultados para o caso de altos valores de SNR.

Os índices dos quadros do vídeo em que ocorrem as bordas do VAD, ou seja, o início e fim de atividade vocal são armazenados em um vetor $\mathbf{g} = \{g_i\}, i \in 1, \ldots, i_{máx}$, conforme figura 26, sendo considerada uma janela de W quadros em torno de cada índice deste vetor, para a extração do movimento, sendo Z calculado como:

$$Z = \text{round}\left(\frac{W}{2}\right). \tag{59}$$

O centro das janelas é deslocado para a esquerda dos instantes de detecção do VAD uma vez que foi avaliado que o instante dos picos de movimento labial geralmente precede as detecções de fala pelo sinal de áudio.

Figura 26 - Índices do VAD para obtenção dos limiares de movimento



Fonte: Autor
Legenda: instantes das bordas de detecção de atividade vocal armazenados no vetor **g**

A partir deste passo é considerado um conjunto de quadros de um vídeo, sendo f subscrito, utilizado como índice do número do quadro. Dado o vetor **g** com os índices do VAD encontrados e os valores dos movimentos ajustados totais da região labial c_f^h e c_f^v para cada quadro f do vídeo, os elementos dos vetores $\mathbf{a}^h$ e $\mathbf{a}^v$ que contêm os valores médios dos movimentos ajustados são encontrados para cada posição do vetor **g**, da seguinte forma:

$$a_i^h = \sum_{f=g_i-Z}^{g_i+Z-1} c_f^h\,, \tag{60}$$

$$a_i^v = \sum_{f=g_i-Z}^{g_i+Z-1} c_f^v \,, \tag{61}$$

onde $i = 1, \ldots, i_{máx}$ e $i_{máx}$ é o tamanho do vetor $\mathbf{g}$. Sendo que os movimentos médios para índices menores e iguais a zero são considerados da nulos:

$$c_f^h = c_f^v = 0, f \leq 0. \tag{62}$$

O valor ótimo da janela W será discutido na seção seguinte com a observação dos resultados dos testes. As médias dos vetores $\mathbf{a}^h$ e $\mathbf{a}^v$, nomeadas m^he m^v representam o movimento labial médio em ambas direções de um locutor em um dado vídeo e são calculadas da seguinte forma:

$$m^h = \frac{1}{i_{máx}} \sum_{i=1}^{i_{máx}} a_i^h \,, \tag{63}$$

$$m^v = \frac{1}{i_{máx}} \sum_{i=1}^{i_{máx}} a_i^v \,, \tag{64}$$

onde $i_{máx}$ é o tamanho dos vetores $\mathbf{a}^h$ ou $\mathbf{a}^v$, de mesmo tamanho que o vetor $\mathbf{g}$.

Todos os passos até agora foram descritos para apenas um único vídeo de um locutor. Considerando agora um conjunto de *S* vídeos de treinamento para um dado locutor, os valores dos movimentos médios podem ser encontrados como a média dos valores de m^h e m^v, obtidos para cada vídeo p (indexados abaixo pelo índice sobrescrito):

$$\mu^h = \frac{1}{S} \sum_{p=1}^{S} m_p^h \,, \tag{65}$$

$$\mu^v = \frac{1}{S} \sum_{p=1}^{S} m_p^v \,, \tag{66}$$

sendo os desvios padrões corrigidos pelo fator de Bessel encontrados da seguinte forma:

$$\sigma^h = \sqrt{\frac{1}{S-1}\sum_{p=1}^{S}(m_p^h - \mu^h)^2}\,, \tag{67}$$

$$\sigma^v = \sqrt{\frac{1}{S-1}\sum_{p=1}^{S}(m_p^v - \mu^v)^2}\,. \tag{68}$$

Os limiares podem ser finalmente computados como a média do movimento nos vídeos decrescidos de uma parcela do desvio padrão, conforme segue:

$$t^h = \mu^h - k\sigma^h, \tag{69}$$

$$t^v = \mu^v - k\sigma^v, \tag{70}$$

sendo k um fator que pode ser utilizado para ajuste fino do algoritmo. Melhores resultados foram obtidos para $k = 1{,}5$, sendo este valor utilizado nas seções seguintes.

7.1.2 FASE DE TESTE

A fase de teste conta com o uso dos limiares de movimento t^h e t^v encontrados na fase de treinamento para verificar os trechos do vídeo em que ocorre movimento suficiente para ser considerado movimento correspondente a fala.

Para um dado vídeo, os passos da fase de teste são iguais aos da fase de treinamento até encontrar os valores de c^h e c^vpara cada quadro do vídeo $f = 1, \ldots, f_{max}$ onde f_{max} é o número total de quadros do vídeo. Os valores obtidos são então comparados com os limiares t^h e t^v encontrados na fase de treinamento, resultando no vetor **m** que contém as informações de movimento detectado, sinalizado com valor 1 ou movimento não detectado com valor 0, para cada quadro *f* do vídeo:

$$m_f = \begin{cases} 1, c_f^h \geq t^h \text{ ou } c_f^v \geq t^v \\ 0, c_f^h < t^h \text{ e } c_f^v < t^v \end{cases}. \tag{71}$$

De forma a suavizar a detecção, ou seja, eliminar instantes esporádicos de detecção que

podem ocorrer em torno de instantes de não-detecção ou vice-versa, um filtro de média móvel é aplicado no vetor $\mathbf{m}$, resultando no vetor $\widehat{\mathbf{m}}$ definido da seguinte forma:

$$\widehat{m}_f = \frac{1}{N}\sum_{j=1}^{N} m_{f+j}\,, \tag{72}$$

sendo N a janela do filtro, ajustada para $N = 3$ nos exemplos avaliados. Deste modo, de forma a resultar novamente em um vetor de zeros e uns que indicam os instantes de detecção, a seguinte comparação é realizada, resultando no vetor $\widetilde{\mathbf{m}}$:

$$\widetilde{m}_{f-1} = \begin{cases} 1, \widehat{m}_f > 1/N \\ 0, \widehat{m}_f \leq 1/N \end{cases}. \tag{73}$$

Devido ao atraso causado pelo filtro é necessária a compensação da posição do quadro, conforme pode ser observado no índice f da equação acima. Como resultado final, os instantes correspondentes a um único quadro isolado de movimento ou de ausência de movimento são removidos.

7.2 AVALIAÇÃO DO ALGORITMO

Para avaliar o algoritmo, uma sequência de testes foi realizada com a base de dados XM2VTSdb. Para todos os locutores, os vídeos das sessões 2, 3 e 4 foram utilizadas na fase de treinamento para estimativa do limiar, enquanto a sessão 1 foi utilizada para teste. Os testes foram realizados tanto com os vídeos da frase 1 ("*zero one two three four five six seven eight nine five zero six nine two eight one three seven four*"), como da Frase 2 ("*Joe took father's green shoe bench out*"), sendo que o treinamento e o teste foram realizados com três combinações distintas, sendo elas:

a) configuração I: treinamento com a frase 2 e teste com a mesma frase, supondo uma forma de dependência do texto do algoritmo dado a relação do movimento labial com a frase;

b) configuração II: treinamento com a frase 1 e teste com a frase 2, verificando a possibilidade de uso do algoritmo independente do texto;

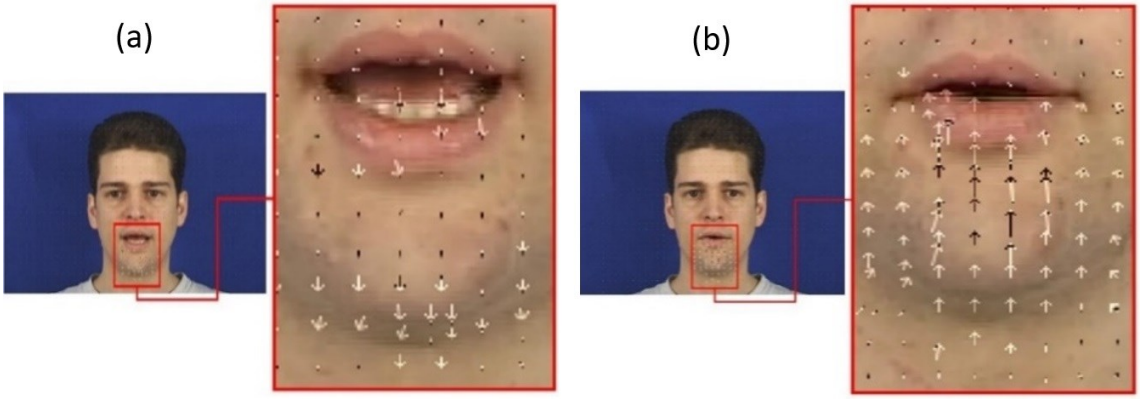
c) configuração III: treinamento combinado com vídeos da frase 1 e 2 e teste com a frase 2.

Todos os testes foram realizados utilizando a biblioteca FFmpeg para Linux em

combinação com a versão modifica do aplicativo mpegflow [125] para extração dos vetores de movimento, processados via MATLAB® com uso do grupo de funções do *Computer Vision System*™ para aplicação do algoritmo Viola-Jones e utilizando outras funções criadas pelo autor conforme algoritmo proposto, sendo os códigos disponíveis em <https://github.com/maparada/biometria-multimodal>.

Os resultados apresentados a seguir foram obtidos através da aplicação do algoritmo para o locutor 000 da base de dados. Os vetores de movimento extraídos podem ser observados graficamente na figura 27.

Figura 27 - Vetores de movimento

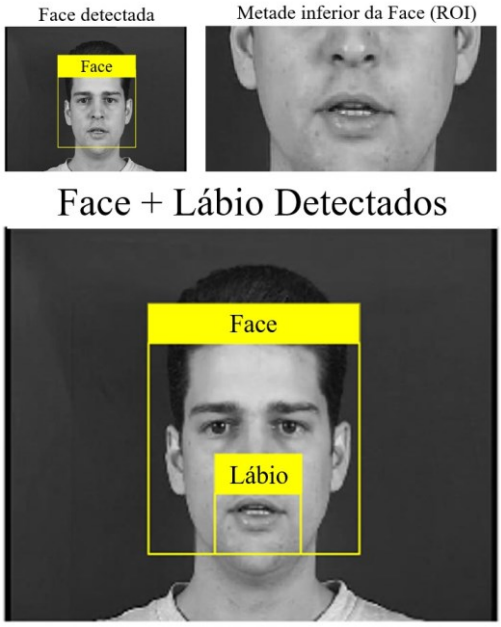


Fonte: Autor "adaptado de" Messer, 1999 [128]
Legenda: (a) abertura labial (b) fechamento labial. Locutor 000 XM2VTS

A região facial é encontrada com o uso do algoritmo Viola-Jones, e a metade inferior desta região é utilizada como ROI para detecção da região labial com treinamento prévio do modelo correspondente. Os resultados são apresentados graficamente na figura 28. Uma nova técnica proposta para extração da região labial e seus resultados são descritos no capítulo 9.

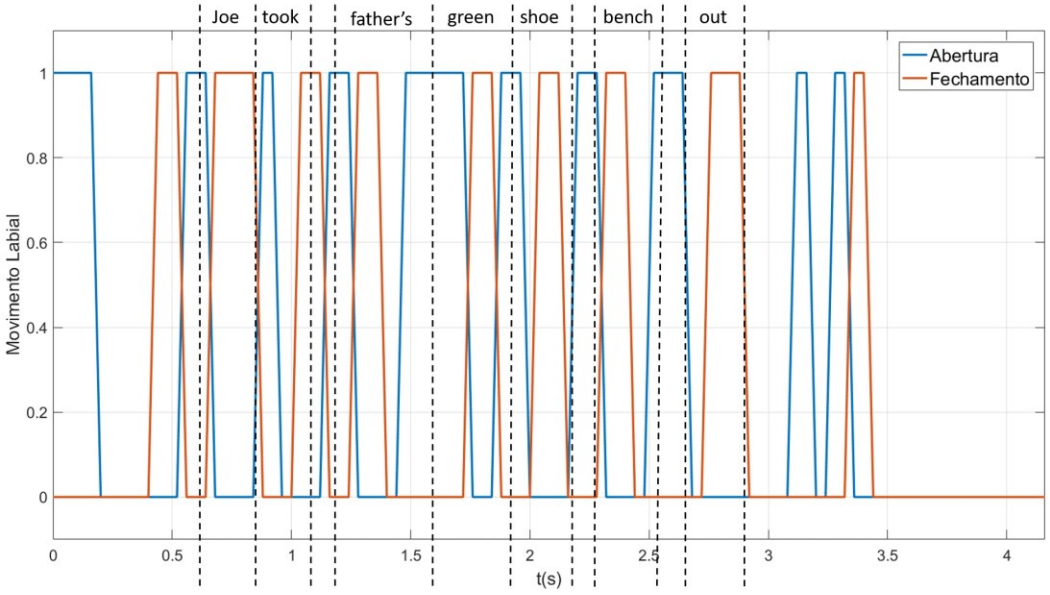
Antes da extração final dos instantes de movimento labial, uma análise por observação do vídeo foi realizada, verificando os instantes em que ocorrem movimento de abertura e fechamento labial para efeitos comparativos com os resultados obtidos de forma automática. Essa análise também auxiliou na determinação de uma estimativa do valor de W, correspondente a janela de observação de movimento em torno dos instantes de detecção do VAD, a ser utilizado na fase de teste. O resultado pode ser observado graficamente na figura 29, para o vídeo do locutor 000, sessão 1 e frase 1 utilizada no teste. Definiu-se o valor de amplitude igual a 1 nos instantes em que ocorre movimento e zero, caso contrário.

Figura 28 - Detecção facial e labial



Fonte: Autor "adaptado de" Messer, 1999 [128]
Legenda: detecção facial, seleção da ROI para detecção labial e segmentação final. Locutor 000, XM2VTS

Figura 29 - Detecção de movimento por observação.



Fonte: Autor
Legenda: os picos das curvas de abertura e fechamento representam os instantes de movimento observados, enquanto os vales representam ausência de movimento. As legendas superiores indicam os instantes de cada locução. Locutor 000, sessão 1, frase2, XM2VTS

Foi possível verificar que para este e outros vídeos observados desta base de dados, um movimento completo de abertura ou fechamento do lábio tem duração aproximada de 3 a 5 quadros do vídeo, por este motivo, o valor médio de $W = 4$ foi utilizado, o que equivale a

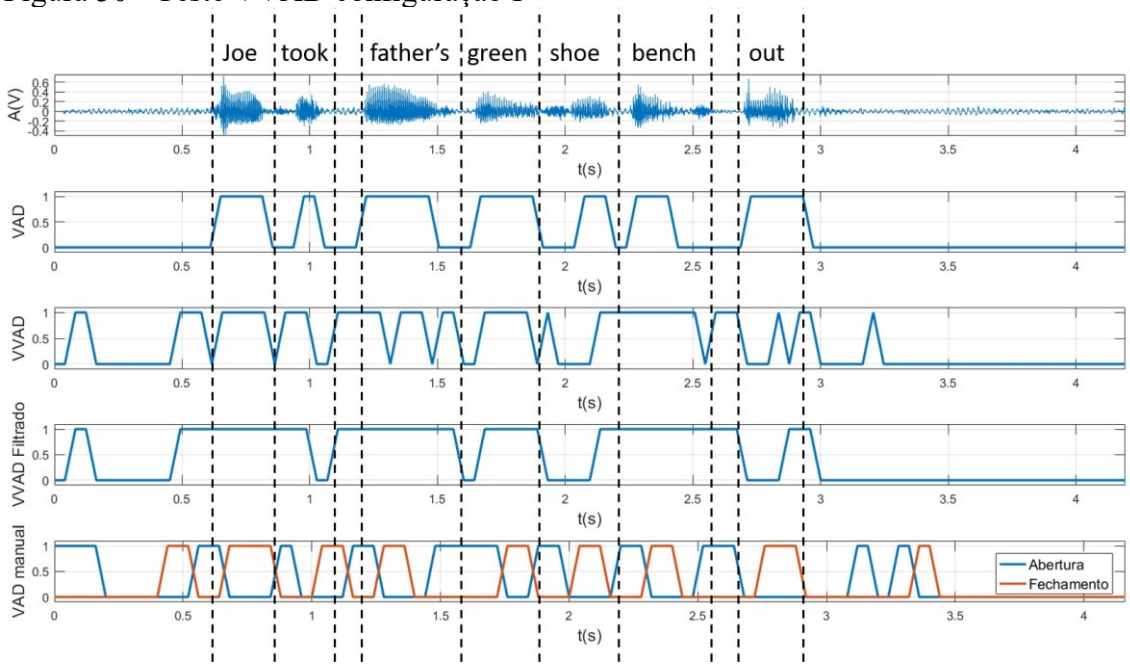
160 ms na escala temporal uma vez que os vídeos dessa base de dados foram capturados a uma taxa de 25 *fps*.

A aplicação do VVAD foi avaliada em comparação com os resultados obtidos pela aplicação do VAD, baseado no limiar da energia do sinal de áudio e em função do resultado da observação dos quadros do vídeo que resultaram na detecção dos instantes de movimento de abertura e fechamento dos lábios de forma manual, conforme figura 29. Os resultados são apresentados de acordo com as três diferentes combinações de treinamento/teste descritas no início desta seção.

A figura 30 exibe o resultado quando o treinamento foi feito conforme configuração I, com a frase 2 das sessões 2, 3 e 4 e o teste com a mesma frase da sessão 1. Nesta figura é possível verificar que a detecção do VVAD corresponde na maioria dos instantes de tempo com as detecções VAD baseado somente no sinal de áudio, sendo que os longos instantes de silêncio que ocorrem no início e final do áudio foram corretamente detectados, porém, algumas detecções de movimento labial ocorreram dentro deste período de silêncio, como, por exemplo, logo no início do vídeo e em torno de 3.2 s. Como pode ser visto na curva de movimento observado, obtida através da visualização quadro a quadro do vídeo, da mesma forma que na figura 29, os instantes detectados como ocorrência de movimento labial com o algoritmo de VVAD, durante os períodos de silêncio, correspondem a movimentos labiais mesmo que na ausência de produção de fala. Alguns desses movimentos como são esporádicos puderam ser removidos com o filtro de média móvel, como o que pode ser observado em 3.2 s na curva do VVAD filtrado, porém, caso ocorra um movimento labial de longa duração sem a produção de fala esse artifício não é suficiente para a correção. O impacto dessas falsas detecções depende da aplicação, sendo avaliado para Verificação Automática de Locutor, no capítulo seguinte.

Na figura 30 também é possível observar comparando a curva do áudio e a curva dos movimentos observados, que a ocorrência de movimentos de abertura labial precede os instantes de ocorrência de fala, sugerindo que há uma pequena defasagem entre o início do movimento labial e a produção sonora da fala que deve ser, confirmando a hipótese sugerida na fase de treinamento do VVAD. A figura 31 exibe os resultados conforme configuração II, para quando o treinamento foi feito com a frase 1 das sessões 2, 3 e 4 e o teste com a frase 2 da sessão 1. Neste caso, os resultados do VVAD apresentaram diversas detecções esporádicas que acabaram sendo removidas após aplicação do filtro de média móvel, resultando em trechos em que a detecção de voz não foi realizada corretamente, mostrando que os valores dos limiares não puderam ser calibrados corretamente.

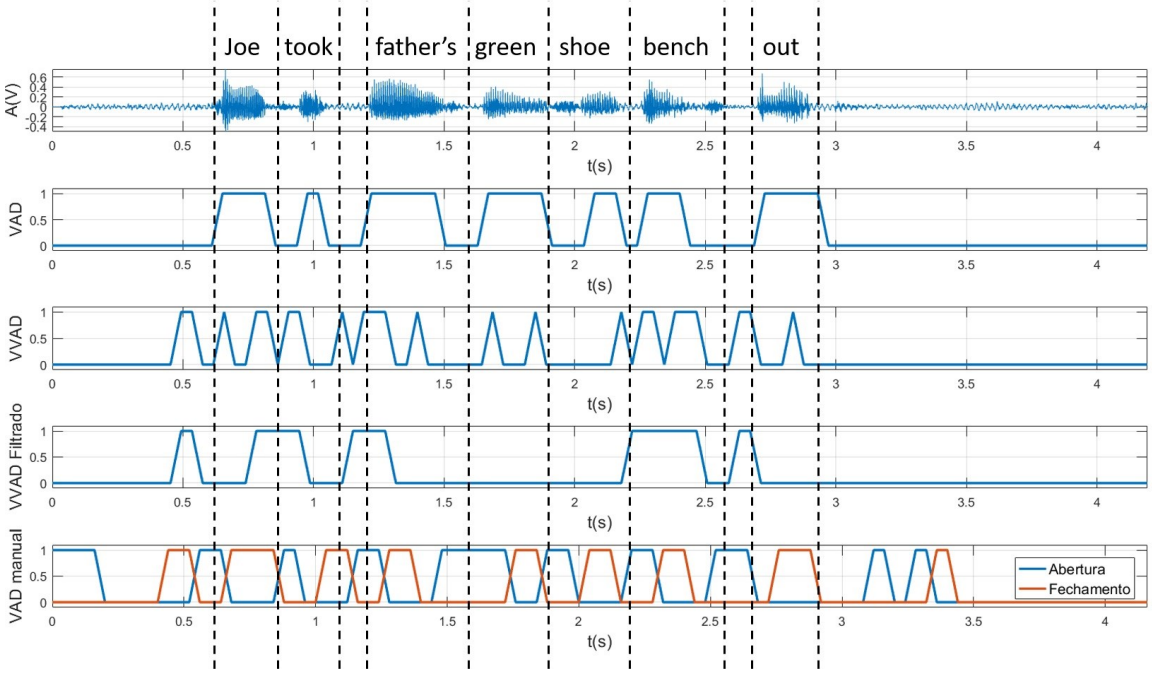
Figura 30 - Teste VVAD configuração I



Fonte: Autor

Legenda: Os picos das curvas: VAD, VVAD, VVAD filtrado e movimento observado representam os instantes de detecção de atividade vocal para cada situação, enquanto os vales representam os períodos de silêncio ou ausência de movimento labial. Os rótulos superiores indicam os instantes de cada locução. Teste realizado para o Locutor 000 da base de dados XM2VTS

Figura 31 - Teste VVAD configuração II

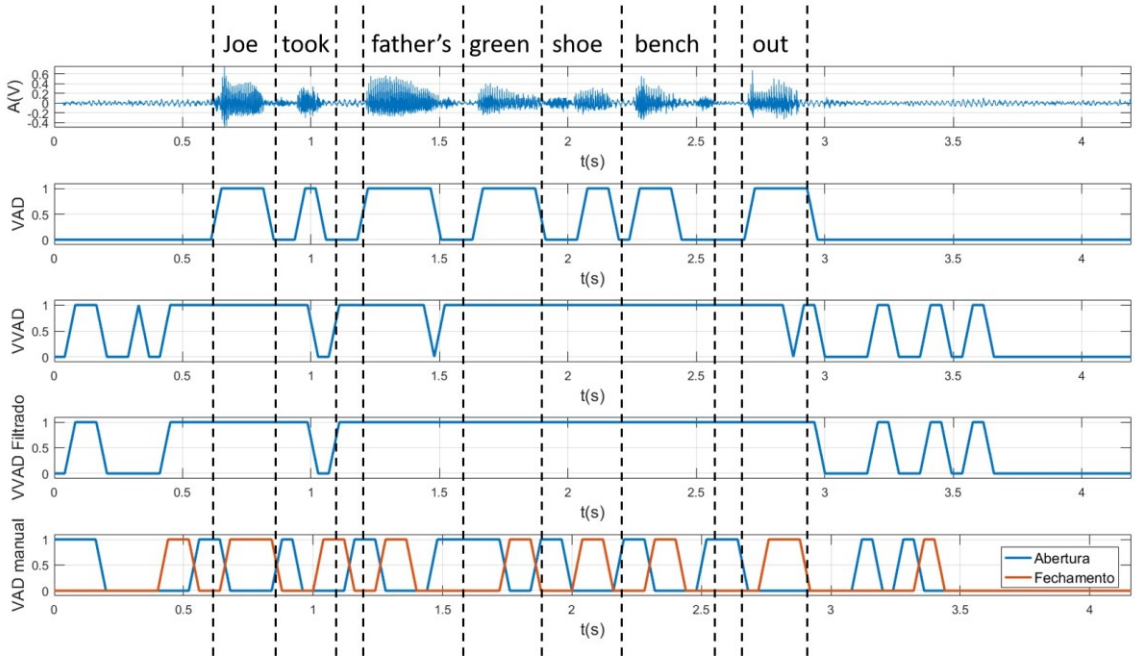


Fonte: Autor

Legenda: Os picos das curvas: VAD, VVAD, VVAD filtrado e movimento observado representam os instantes de detecção de atividade vocal para cada situação, enquanto os vales representam os períodos de silêncio ou ausência de movimento labial. Os rótulos superiores indicam os instantes de cada locução. Teste realizado para o Locutor 000 da base de dados XM2VTS

Já no caso da configuração III em que o treinamento foi realizado com a combinação das frases 1 e 2 e o teste realizado com a frase 2, como observado na figura 32, os resultados obtidos mostraram que as detecções foram realizadas dentro do período correto em que ocorre alta taxa de movimento labial, correspondendo com a detecção por observação, porém, havendo uma maior ocorrência de movimentos fora do trecho de locução do que no primeiro caso avaliado. Esse último caso, porém, mostra que é possível a utilização deste algoritmo independentemente do texto, sendo que no treinamento seja utilizada uma mistura de diferentes frases de um dado locutor. Neste caso foi utilizada uma mistura de apenas duas frases, conforme disponibilidade desta base de dados. Porém, trabalhos futuros pretendem avaliar o algoritmo para um número maior de sentenças utilizadas no treinamento e teste.

Figura 32 - Teste VVAD configuração III



Fonte: Autor

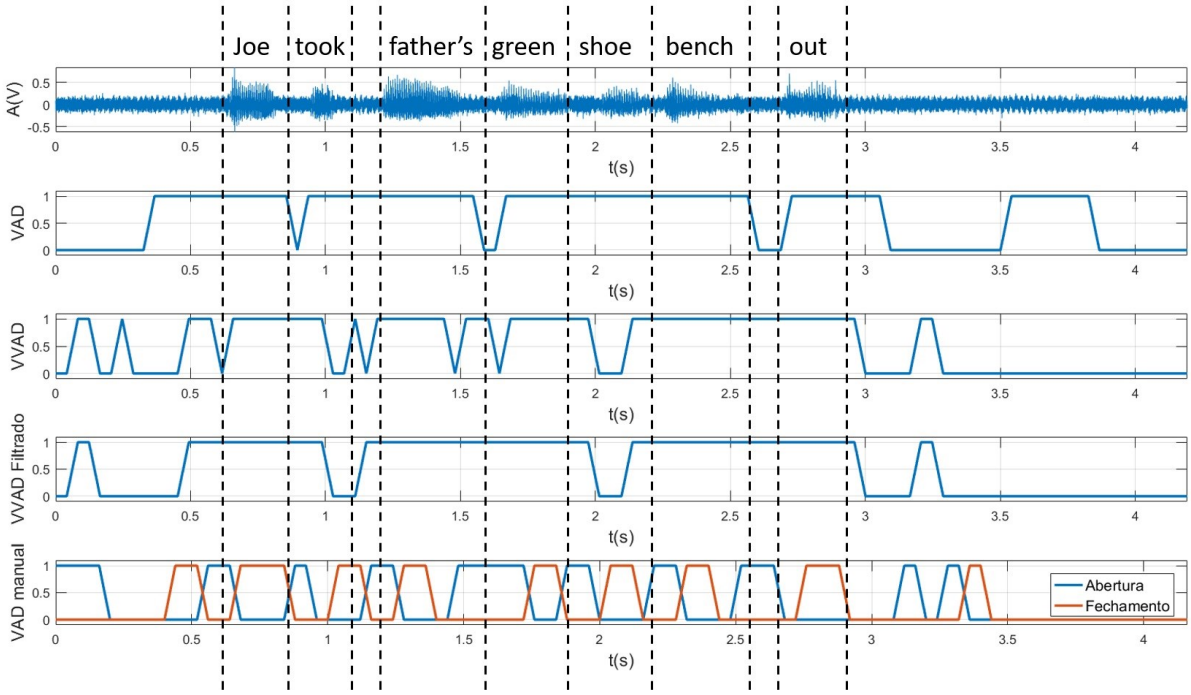
Legenda: Os picos das curvas: VAD, VVAD, VVAD filtrado e movimento observado representam os instantes de detecção de atividade vocal para cada situação, enquanto os vales representam os períodos de silêncio ou ausência de movimento labial. Os rótulos superiores indicam os instantes de cada locução. Teste realizado para o Locutor 000 da base de dados XM2VTS

Como o algoritmo utiliza informações do vídeo para detecção dos instantes de atividade de voz, seu resultado não é afetado por ruído no áudio, como no caso dos algoritmos tradicionais de VAD. A figura 33 ilustra o desempenho do algoritmo proposto, conforme configuração I, quando ruído do tipo branco é adicionado ao sinal de áudio, resultando em relação sinal-ruído SNR = 5 dB. Neste caso é possível observar a degradação da qualidade do VAD com a presença de falsas detecções que não ocorreram na mesma configuração, porém

sem ruído adicionado, conforme demonstrado na figura 30. Já o resultado do VVAD permanece inalterado, mostrando ser uma melhor opção para aplicação no *front-end* de sistemas de reconhecimento de locutor, quando há forte presença de ruído ambiente no sinal de áudio.

O algoritmo pode também ser aplicado para sistemas de reconhecimento que são baseados em características do movimento labial extraídos do vídeo, uma vez que remove os instantes em que o movimento não é detectado, conforme aplicado no sistema multimodal apresentado no capítulo 8.

Figura 33 - Teste VVAD, configuração I, SNR = 5 dB



Fonte: Autor

Legenda: Os picos das curvas: VAD, VVAD, VVAD filtrado e movimento observado representam os instantes de detecção de atividade vocal para cada situação, enquanto os vales representam os períodos de silêncio ou ausência de movimento labial. Os rótulos superiores indicam os instantes de cada locução. Teste realizado para o Locutor 000 da base de dados XM2VTS

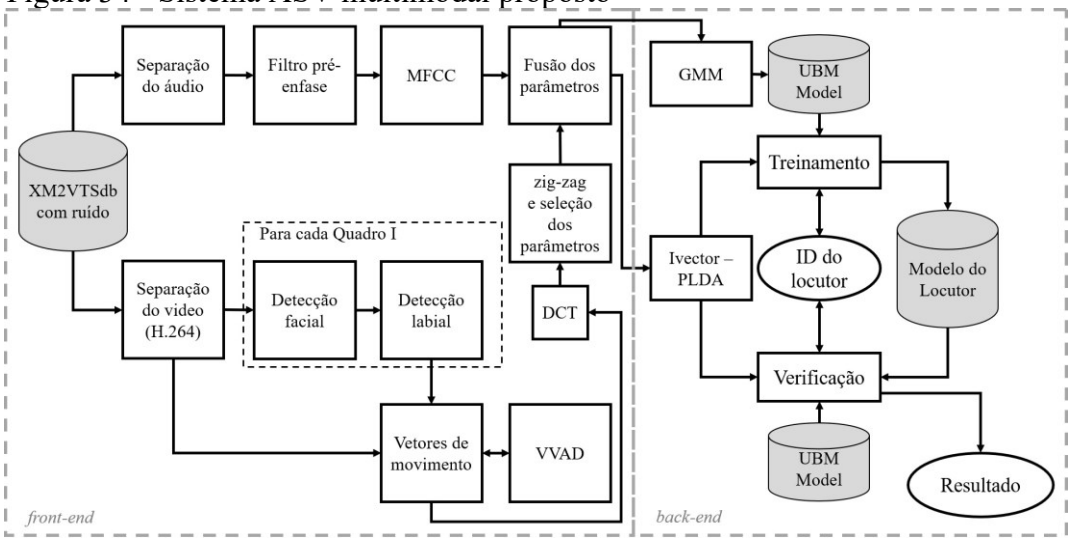
8 FUSÃO DE PARÂMETROS MULTIMODAL PARA ASV

A proposta aqui apresentada consiste na fusão dos coeficientes MFCC com parâmetros extraídos da transformada das matrizes de movimento vertical e/ou horizontal extraídas do vídeo codificado em MPEG-4 H.264 para verificação automática de locutor. Nesta proposta, uma matriz única de parâmetros é combinada, diferindo das técnicas de fusão mais tradicionais para biometria multimodal, onde usualmente os coeficientes de áudio e vídeo são analisados separadamente e a medida combinada das duas avaliações são unificadas no momento da decisão, conforme descrito na seção 2.8. O uso destes coeficientes combinados para tarefa de reconhecimento de locutor foi avaliado com o uso de modelagem *i-vector* PLDA, testado em diferentes situações de ruído ambiente no áudio, contando ainda com a aplicação do VVAD proposto no capítulo anterior. Os resultados foram avaliados de modo a obter a melhor combinação dos parâmetros, bem como melhor configuração do número de misturas no modelo UBM e dimensão da matriz de variabilidade **T** utilizada na modelagem *i-vector*.

8.1 ALGORITMO PROPOSTO

A estrutura do sistema ASV implementado é apresentada pela figura 34.

Figura 34 - Sistema ASV multimodal proposto



Fonte: Autor
Legenda: *back-end* e *front-end* (destacados em pontilhado) do sistema ASV multimodal proposto

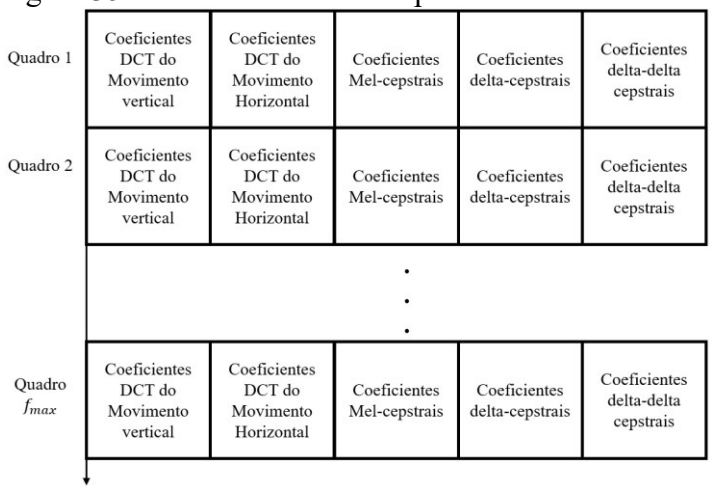
Inicialmente diversas versões da base de dados XM2VTS foram criadas adicionando dois tipos de ruído. No total 13 versões da base de dados foram utilizadas nos testes, sendo:

uma a base original sem ruído adicionado; 6 versões em que ruído branco foi adicionado de forma a resultar em diferentes taxas de SNR de 0 dB a 25 dB com passo de 5 em 5 dB; 6 versões em que ruído do tipo *babble* foi adicionado, nas mesmas taxas de SNR anteriores. O ruído do tipo *babble* equivale a um ruído gerado por diversas vozes de fundo, como ocorre em um ambiente movimentado, mostrando-se uma aproximação mais realista de um caso de uso do sistema ASV. Esse ruído foi extraído do material complementar de [21].

Primeiramente a extração dos coeficientes MFCC é realizada no sinal de áudio. Para extração dos parâmetros de movimento labial os quadros do tipo I são utilizados para detecção da região facial e labial com o algoritmo Viola-Jones. A posição da região labial detectada é considerada para os próximos quadros do tipo P ou B para extração das matrizes de movimento.

As matrizes de movimento são então transformadas com o uso da DCT-II e serializadas com varredura do tipo zig-zag, sendo os primeiros *N* coeficientes (valor a ser definido) de cada matriz utilizados na composição da matriz de parâmetros. Cada linha da matriz resultante corresponde a um instante temporal de um quadro e conta com a concatenação de *N* coeficientes resultantes da aplicação da DCT na matriz de movimento vertical, seguido da mesma quantidade de *N* coeficientes resultantes da DCT aplicada à matriz de movimento horizontal (quando utilizado) e na sequência os 39 coeficientes MFCC (13 coeficientes Mel-cepstrais, 13 delta-cepstrais e 13 delta-delta cepstrais). O valor de *N*, que representa o número de coeficientes utilizados após aplicação da DCT nas matrizes de movimento, será discutido na sessão 8.3. O formato da matriz resultante é exibido na figura 35.

Figura 35 - Matriz da fusão dos parâmetros



Fonte: Autor.
Legenda: Parâmetros combinados com DCT do movimento e MFCC, sendo f_{max} o número total de quadros do vídeo

Para que seja possível a extração desta matriz, é necessário que os coeficientes MFCC sejam obtidos em janelas temporais de duração equivalente ao tempo de um quadro do vídeo. Como na base de dados a taxa de quadros por segundo é de 25 *fps*, esta janela tem duração de 40 ms, sendo um intervalo bastante razoável para a extração dos parâmetros MFCC. Porém, resultados melhores poderiam ser obtidos reduzindo a dimensão desta janela. Como observado em [35], janelas entre 10 e 30 ms são ideais para melhor parametrização do sinal de voz.

Quando o VVAD é aplicado, o primeiro coeficiente resultante da DCT aplicada a matriz de movimento horizontal e o primeiro coeficiente resultante da DCT aplicada a matriz de movimento vertical são comparados com os correspondentes limiares de movimento encontrados na fase de treinamento, de modo que a extração dos parâmetros referente ao quadro só é efetuada caso um dos dois coeficientes seja superior ao limiar, conforme descrito na seção 7.2 na equação 71 ou podendo ainda ser suavizado com o filtro de média móvel conforme descrito pelas equações 72 e 73.

8.2 AVALIAÇÃO DO *BACK-END*

O *back-end* foi implementado com *i-vector* PLDA, avaliando previamente a melhor configuração do número de misturas Gaussianas (C) utilizadas para criação do modelo UBM, e dimensão do subespaço de variabilidade (R), correspondente ao posto da matriz de variabilidade **T**. Para isso, diferentes configurações foram avaliadas para uma mesma versão da base de dados contendo ruído *babble* a 10 dB conforme tabela 10

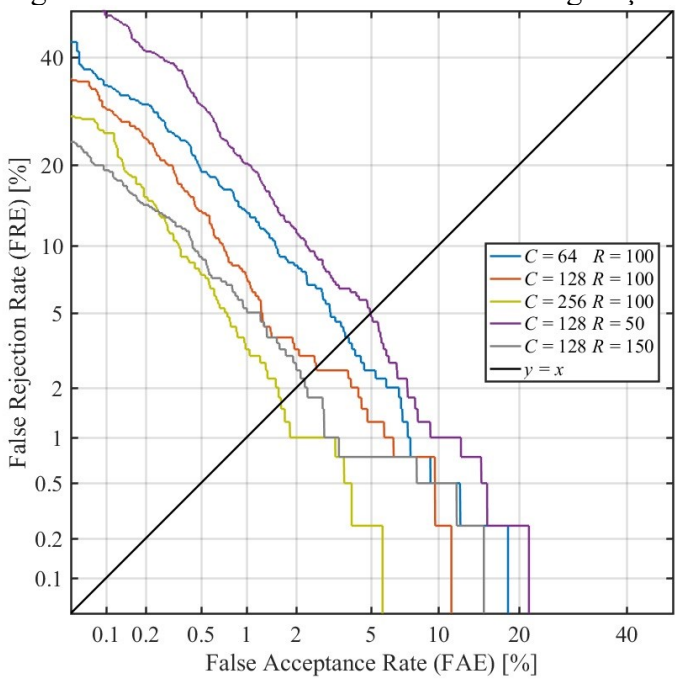
Tabela 10 - Configurações do *back-end*

C	R
64	100
128	100
256	100
128	50
128	150

Fonte: Autor
Legenda: variação de C e R para avaliação do *back-end*

Estes testes foram realizados conforme protocolo Lausanne configuração I, conforme descrito na seção 5.1 e *front-end* configurado somente com obtenção dos MFCCs. Os resultados foram avaliados em função da taxa de EER na etapa de teste, EER_T, obtida através da curva DET para as diferentes configurações, conforme figura 36.

Figura 36 - Curva DET das diferentes configurações



Fonte: Autor
Legenda: curvas DET para os diferentes valores de C e R avaliados. Cruzamento com a curva $y = x$ correspondem aos valores de EER_T

É possível observar na figura e apresentado numericamente na tabela 11, que conforme aumenta o número de misturas, o valor de EER_T (que corresponde ao cruzamento da curva DET com a curva $y = x$) diminui, ou seja, melhores resultados biométricos são obtidos. O mesmo ocorre aumentando a dimensão da matriz **T**. Para as configurações avaliadas, melhores resultados foram obtidos com $C = 256$ e $R = 100$. Outras configurações, com valores superiores de C e R foram avaliadas, porém, resultaram em tempos de execução muito longos não justificando o pequeno aumento obtido no desempenho do sistema e, portanto, não foram consideradas. Esta configuração foi ajustada para os demais testes do *front-end*, privilegiando os resultados biométricos sem considerar o aumento no tempo de execução.

Tabela 11 - Resultado das configurações avaliadas

C	R	EER_T
64	100	3,75 %
128	100	2,57 %
256	100	1,62 %
128	50	4,95 %
128	150	2,22 %

Fonte: Autor
Legenda: EER_T obtido para cada configuração do *back-end* avaliado

8.3 AVALIAÇÃO DO *FRONT-END*

Os testes para avaliação do *front-end* seguiram a configuração I do protocolo Lausanne, conforme descrito na seção 5.1 e foram realizados em três configurações diferentes de parâmetros biométricos extraídos, sendo elas:

a) somente MFCC;
b) somente coeficientes da DCT do movimento;
c) MFCC combinado com os coeficientes da DCT do movimento.

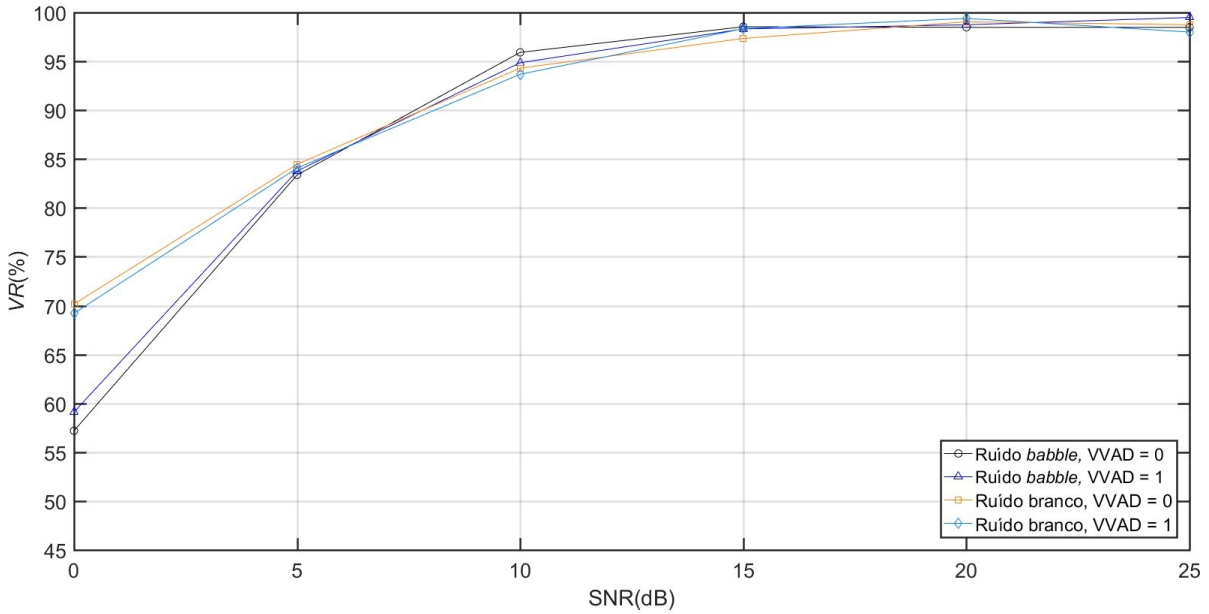
Cada uma dessas configurações contou ainda com diferentes variações da relação sinal-ruído, número de coeficientes utilizados e inclusão ou não de parâmetros do movimento horizontal, de forma a avaliar o desempenho do sistema em função destas variáveis, possibilitando a observação da melhor combinação de parâmetros a ser utilizada neste sistema de fusão proposto.

8.3.1 Somente MFCC

Os testes com a extração de parâmetros baseados somente na matriz de coeficientes MFCC foram realizados para as diferentes versões de SNR da base de dados para os dois tipos de ruído adicionado, conforme previamente descrito. Os testes foram ainda realizados incluindo ou não o VVAD para verificação do impacto desta etapa no desempenho geral do sistema. A figura 37 mostra o resultado dos diferentes testes em função da taxa VR obtida na etapa de teste, conforme sugerida pelo protocolo, enquanto a figura 38 exibe os resultados em função da taxa de EER_T. É possível notar pelas curvas que enquanto o valor de EER_T decresce em função do crescimento do SNR, o valor de VR cresce mostrando um melhor desempenho do sistema, como esperado.

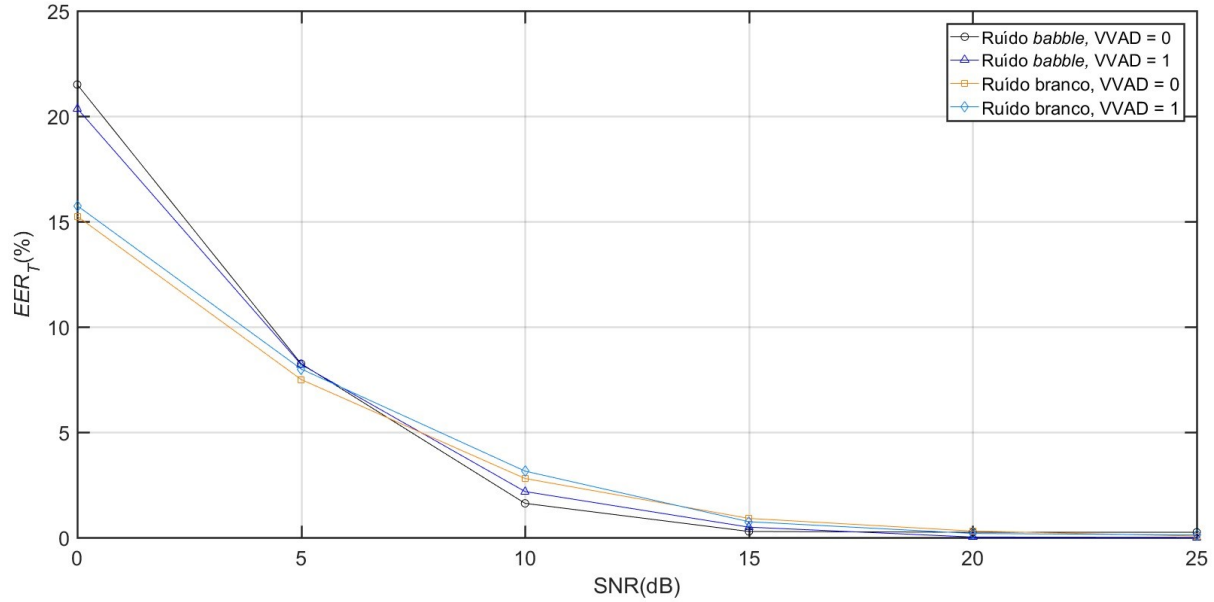
Os resultados obtidos com ruído *babble* mostram um pior desempenho do sistema, uma vez que os áudios são afetados por vozes de fundo, aumentando a variabilidade interespecífica dos dados entre as fases de avaliação e teste, prejudicando a modelagem, o que resulta em piores taxas de VR do que EER_T uma vez que na primeira o limiar deve ser escolhido previamente na fase de avaliação. É possível notar que apesar de não causar grande impacto no desempenho do sistema, o VVAD proposto melhora, na maioria dos casos, os valores de VR e EER_T obtidos e auxilia ainda a reduzir a quantidade de quadros em que os parâmetros são extraídos, resultando em uma demanda menor de carga computacional do sistema nas etapas de treinamento e verificação.

Figura 37 - SNR (dB) x VR (%). Somente MFCC



Fonte: Autor
Legenda: Resultado da métrica VR (%) para os valores de SNR (dB) avaliados com e sem a inclusão do VVAD

Figura 38 – SNR (dB) x EER_T (%). Somente MFCC



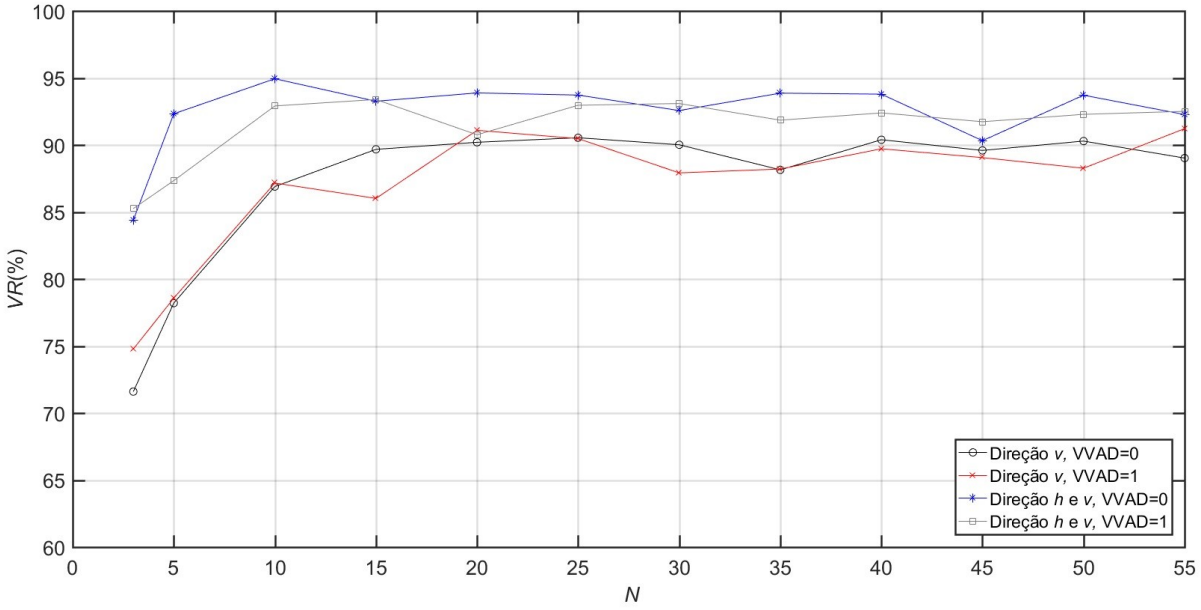
Fonte: Autor
Legenda: Resultado da métrica EER_T (%) para os valores de SNR (dB) avaliados com e sem a inclusão do VVAD

8.3.2 **Somente parâmetros dos vetores de movimento**

O segundo teste avaliado foi realizado com a matriz de parâmetros composta somente dos parâmetros provenientes da transformada DCT das matrizes de movimento. Os testes avaliaram o desempenho com a inclusão apenas de N coeficientes provenientes do movimento vertical e posteriormente com a combinação destes com N coeficientes provenientes do movimento horizontal. Desta forma, foi possível avaliar a necessidade da inclusão de parâmetros do movimento na direção horizontal uma vez que grande parte do movimento labial durante a produção de fala é proveniente da direção vertical.

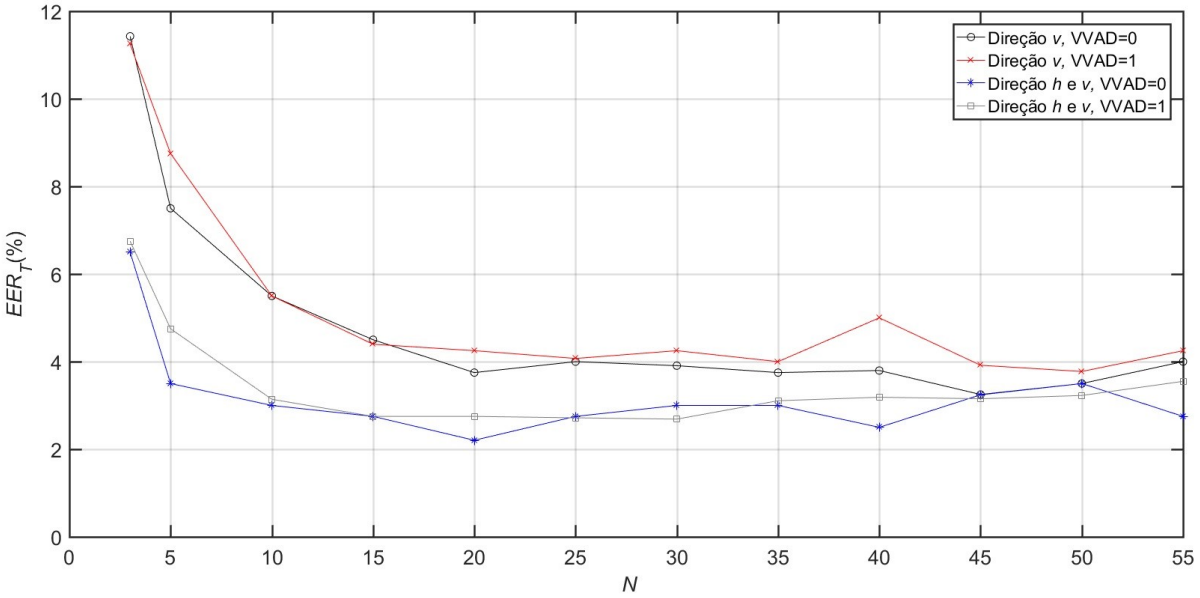
Os resultados foram ainda avaliados variando o valor de N em cada matriz, sendo indicado nos gráficos com o índice h para direção horizontal v e para direção vertical. A inclusão ou não do VVAD também foi considerada. Os resultados em função do valor de VR podem ser observados na figura 39 e os resultados em função do valor de EER_T na figura 40.

Figura 39 - N x VR (%). Somente coeficientes da DCT do movimento



Fonte: Autor
Legenda: Resultado da métrica VR (%) para os diferentes valores de N avaliados com e sem a inclusão do VVAD e com e sem a inclusão do movimento horizontal

Figura 40 - N x EER_T (%). Somente coeficientes da DCT do movimento



Fonte: Autor
Legenda: Resultado da métrica EER_T (%) para os diferentes valores de N avaliados com e sem a inclusão do VVAD e com e sem a inclusão do movimento horizontal

É possível observar que para valores de N maiores do que 15 pouca variação ocorre tanto na taxa de VR como de EER_T, ou seja, este valor é uma escolha ótima da quantidade de coeficientes a ser utilizada na matriz de parâmetros combinados sugeridos na próxima seção. Este método de avaliação proposto faz com que a quantidade de parâmetros utilizada seja a menor possível em função do desempenho do sistema biométrico, sendo uma boa ferramenta para a escolha dos parâmetros a serem utilizados.

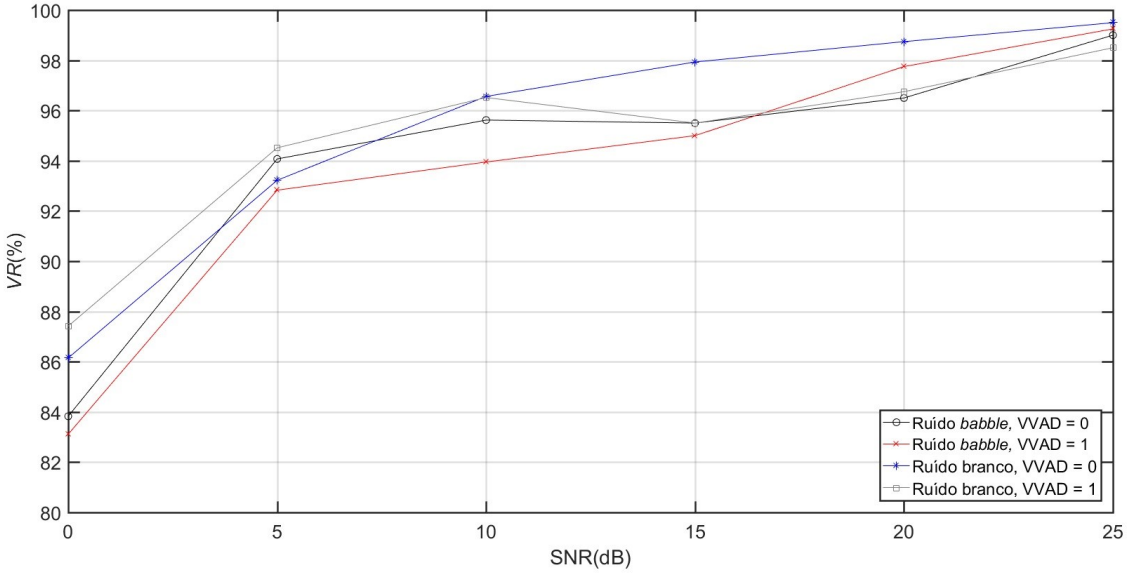
A inclusão dos parâmetros de movimento horizontal levou a resultados ligeiramente superiores, que podem ser observados tanto no aumento da taxa VR como no decréscimo da taxa EER_T, porém, sua inclusão deve ser considerada dependendo da aplicação uma vez que o tamanho da matriz de parâmetros aumenta e, para valores de N variando entre 15 e 55, pouca diferença na taxa de VR pôde ser observada. Na seção seguinte, porém, a matriz de parâmetros combinada inclui o movimento horizontal priorizando os melhores resultados possíveis do sistema proposto.

8.3.3 Combinação de MFCC e coeficientes DCT do movimento

A última configuração conta com a combinação dos parâmetros acústicos e do movimento labial para verificação do desempenho do sistema final proposto. Desta forma, a matriz final de parâmetros inclui em cada linha, a concatenação de: 15 coeficientes resultantes

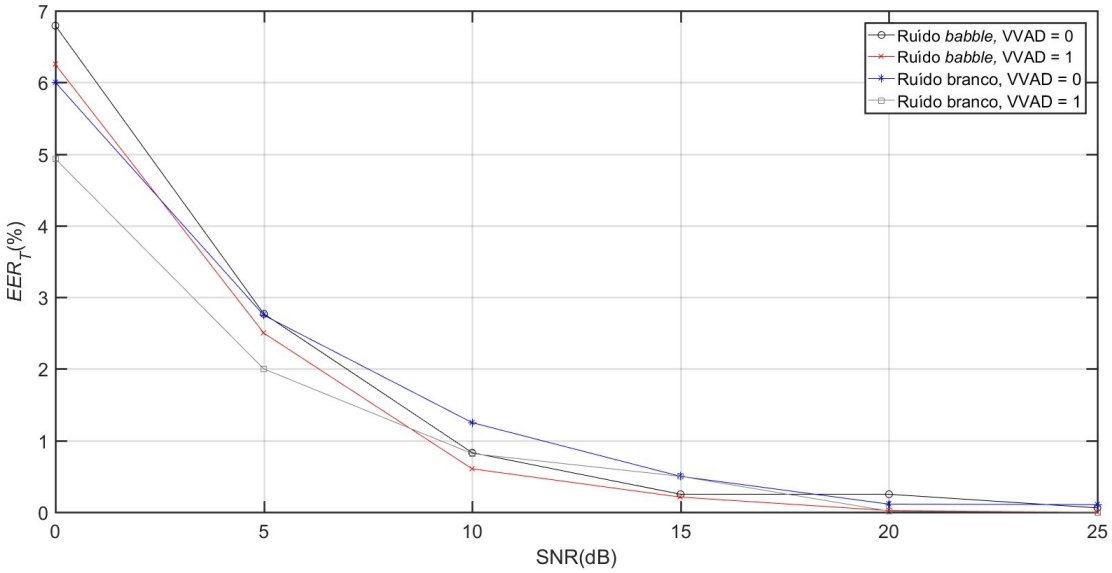
da DCT da matriz de movimento vertical; 15 coeficientes resultantes da DCT da matriz de movimento horizontal; 39 coeficientes MFCC (13 Mel-cepstrais, 13 delta-cepstrais e 13 delta-delta cepstrais). Os resultados foram avaliados novamente em função do SNR do áudio de forma a comparar os resultados obtidos com aqueles onde apenas os coeficientes MFCC foram considerados, sendo apresentados em função do valor de VR, na figura 41 e os resultados em função do valor de EER_T na figura 42.

Figura 41 - SNR (dB) x VR (%). Parâmetros combinados



Fonte: Autor
Legenda: Resultado da métrica VR (%) para os valores de SNR (dB), com e sem a inclusão do VVAD

Figura 42 - SNR (dB) x EER_T (%). Parâmetros combinados



Fonte: Autor
Legenda: Resultado da métrica EER_T (%) para os valores de SNR (dB), com e sem a inclusão do VVAD

É possível observar no sistema multimodal proposto uma melhoria significativa do desempenho do sistema em especial quando da ocorrência de baixos valores de SNR, sendo que o valor de *VR* no pior dos casos resultou em 83% e com o uso apenas de coeficientes MFCC o pior caso resultou em 57% quando avaliado no pior caso em que há presença de ruído do tipo *babble* com relação sinal-ruído de 0dB. O mesmo pode ser observado no decréscimo da taxa de EER_T, provando que a utilização dos parâmetros combinados levou a desempenhos mais estáveis do sistema, mesmo quando o áudio está sobre influência de ruído ambiente elevado. Os resultados observados com e sem a inclusão do algoritmo VVAD mostraram que o impacto positivo no desempenho final não foi muito grande. Isso ocorre devido ao fato de que a base de dados foi gravada de forma a permitir a ocorrência de poucos instantes de silêncio. Porém, é importante notar que a remoção de quadros em que o movimento labial não é significativo pode ser utilizado como uma ferramenta importante para seleção dos melhores quadros para extração dos parâmetros e seu impacto positivo mais percebido em situações com mais intervalos de ausência de fala.

Para valores de relação sinal-ruído acima de 5dB é possível notar que as taxas VR ficaram superiores a 93% e as taxas de EER_T abaixo de 3%. O sistema também foi avaliado sem presença de ruído aditivo no áudio com uso do VVAD, resultando em EER_T de 0,005% e taxa VR de 99,49%. Estes valores foram utilizados para comparação com trabalhos anteriores, que também propuseram sistemas multimodais baseados na combinação de parâmetros extraídos do sinal de áudio e parâmetros extraídos do lábio de forma estática (parâmetros geométricos) ou dinâmica (movimento labial), sendo avaliados com a mesma base de dados e utilizando o mesmo protocolo de avaliação (protocolo Lausanne), porém, sem presença de ruído aditivo no áudio. A tabela mostra a comparação do desempenho do sistema biométrico multimodal proposto em diferentes trabalhos, conforme listados em [130].

Os resultados obtidos em diferentes trabalhos foram avaliados em função da taxa EER na etapa de teste ou em função da taxa HTER (Half Total Error Rate), que corresponde ao valor médio entre a taxa de falsa aceitação e falsa rejeição na etapa de teste, utilizando como limiar o valor obtido previamente na fase de avaliação. Os valores foram obtidos com uso da base de dados XM2VTS original, sem adição de ruído. Nesta tese, os valores foram avaliados em função das taxas EER_T e VR, porém, para facilitar a comparação do desempenho do sistema com outros trabalhos, a taxa HTER foi calculada para o caso em que não há ruído adicionado no áudio, com aplicação do VVAD e fusão dos parâmetros propostos, resultando em 0,25%.

É possível observar através das taxas EER_T e HTER, que o desempenho obtido é superior aos demais trabalhos observados, que também combinam dois conjuntos de

parâmetros, sendo um deles extraídos do lábio e o outro extraído do sinal de áudio.

Tabela 12 - Comparação do desempenho do sistema multimodal proposto

Proposta	Fusão dos parâmetros	Locutores	Desempenho
Broun [131]	Labial (geométricos) + áudio	261	HTER 6,3%
Faraj [132]	Labial (Dinâmico e Indep. do texto) + áudio	295	EER 2%
Sanchez [133]	Labial (Dinâmico e Dep. do texto) + facial	295	HTER 2,62%
Sanchez [133]	Labial (Dinâmico e Dep. do texto) + áudio	295	HTER 0,7%
Parada	Dinâmico (parâmetros do MPEG) + áudio	294	EER_T 0,005%; HTER 0,25%

Fonte: Autor “adaptado de” Chan [130]
Legenda: Valores comparativos do desempenho de diferentes sistemas propostos, sendo a última linha correspondente ao sistema multimodal proposto nesta tese, avaliado sem presença de ruído no áudio.

9 DETECÇÃO DA REGIÃO LABIAL

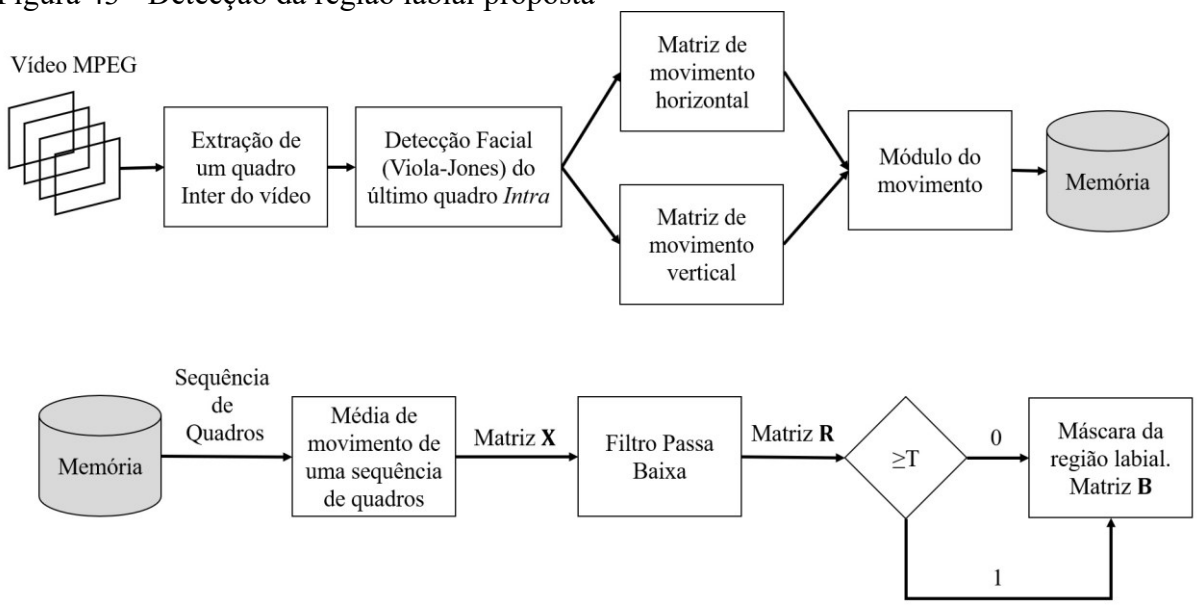
O algoritmo proposto para detecção da região labial é baseado na média do movimento da região facial ao longo do tempo, agrupando as regiões próximas em que ocorre maior quantidade de movimento. O algoritmo pode ser utilizado para as diferentes modalidades biométricas que dependem da segmentação da região labial, como reconhecimento de fala, leitura labial e reconhecimento de locutor. Além de auxiliar na obtenção da região onde são extraídos os parâmetros de movimento, também pode ser utilizado como uma indicação do formato dos lábios.

Assim como os demais algoritmos propostos nesta tese, a detecção da região labial conta com a análise das matrizes de movimento já extraídas no vídeo codificado em MPEG.

9.1 ALGORITMO PROPOSTO

O diagrama em blocos da figura 43 mostra a sequência de aplicação do algoritmo

Figura 43 - Detecção da região labial proposta



Fonte: Autor
Legenda: Sequência da aplicação do algoritmo de detecção da região labial baseado no movimento

Incialmente o algoritmo realiza a separação dos quadros de um vídeo MPEG em quadros, sendo os quadros do tipo *Intra* utilizados para aplicação do algoritmo Viola-Jones para detecção da região facial e os quadros do tipo *Inter* para extração do movimento similarmente aos outros algoritmos apresentados nos Capítulos 7 e 8. As matrizes de

movimento extraídas em cada quadro são utilizadas para o cálculo do módulo do movimento, de forma a abstrair a direção e mantendo apenas as informações de intensidade do movimento.

Para as equações seguintes, índices sobrescritos h indicam direção horizontal e sobrescrito v direção vertical. Sendo $\mathbf{F}^h$ a matriz de movimento horizontal da região facial e $\mathbf{F}^v$ a matriz de movimento vertical da região facial ambas de dimensão V x H, conforme previamente descritas no Capítulo 7. Os elementos $d_{i,j}$ da matriz do módulo do movimento, **D**, são calculados da seguinte forma, para cada quadro do vídeo:

$$d_{i,j} = \sqrt{\left(f_{i,j}^h\right)^2 + \left(f_{i,j}^v\right)^2}, i = 1, \dots, V \text{ e } j = 1, \dots, H. \tag{74}$$

Cada matriz de módulo é calculada e armazenada em uma memória para uso posterior. A quantidade de quadros armazenada depende da profundidade de análise temporal ajustada no algoritmo e tem impacto direto na quantidade de memória utilizada e tempo de execução do algoritmo. As matrizes armazenadas são posteriormente utilizadas para a obtenção de uma única matriz de movimento médio dentro do período analisado. Para isso, considerando $f \in \{1, \dots, F\}$ um conjunto de F quadros de um dado vídeo, sendo f utilizado como indexador sobrescrito nas equações abaixo, a matriz de movimento médio (**X**) é calculada da seguinte forma:

$$\mathbf{X} = \frac{1}{F}\sum_{f=1}^{F} \mathbf{D}^f . \tag{75}$$

A matriz resultante pode conter trechos de alta frequência espacial, decorrentes de ruído, que podem ocorrer devido à variação de iluminação, sombra ou movimentação esporádica de partes da imagem que não correspondem ao movimento contínuo do lábio durante a produção de fala. Para suavizar as altas frequências espaciais da matriz e melhorar a segmentação da região, um filtro passa baixa Gaussiano é aplicado à matriz **X**, da seguinte forma:

$$\mathbf{R} = \mathbf{X} ** \mathbf{G}, \tag{76}$$

sendo **R** o resultado da convolução bidimensional da matriz **X** pela matriz **G** que corresponde a uma matriz Gaussiana bidimensional de média zero e dimensão I x J, sendo utilizada comumente como *kernel* padrão sendo seus elementos $g_{i,j}$ definidos da seguinte forma:

$$g_{i,j} = \frac{1}{2\pi\sigma} e^{\frac{-i^2+j^2}{2\sigma^2}} \; ; \; i = 1, \dots, I \text{ e } j = 1, \dots, J. \tag{77}$$

O valor do desvio padrão (σ) pode ser ajustado para definição da intensidade do filtro, sendo que valores maiores resultam em uma imagem com maior suavização. Nos experimentos realizados, um valor de $\sigma = 10$ foi utilizado e a dimensão da matriz do filtro ajustada para 41 x 41. A aplicação do filtro Gaussiano auxilia também para que a detecção não resulte em uma região com bordas serrilhadas uma vez que o movimento é estimado pelo codec em blocos da imagem. Desta forma, a imagem fica com um aspecto mais uniforme e as regiões detectadas com as bordas mais arredondadas.

O último passo para extração da região labial consiste na obtenção de uma matriz binária a partir da análise dos valores dos pontos da matriz **R** e comparação com um limiar, resultando na matriz **B**, sendo seus elementos $b_{i,j}$ encontrados da seguinte forma

$$b_{i,j} = \begin{cases} 1 \; p/ r_{i,j} \geq T \\ 0 \; p/ r_{i,j} < T \end{cases}, i = 1, \dots, V \text{ e } j = 1, \dots, H, \tag{78}$$

sendo $r_{i,j}$ um elemento da matriz **R** de dimensão V x H resultante da aplicação do filtro Gaussiano na matriz média e T um limiar encontrado como:

$$T = \mu + k\sigma, \tag{79}$$

onde k é uma variável que realiza o ajuste fino do algoritmo, sendo o valor utilizado nos testes apresentado na seção seguinte, μ é o valor médio da matriz **R** encontrado como:

$$\mu = \frac{1}{VH} \sum_{i=1}^{V} \sum_{j=1}^{H} r_{i,j} \tag{80}$$

e σ o desvio padrão da matriz **R** encontrado como:

$$\sigma = \sqrt{\frac{1}{VH} \sum_{i=1}^{V} \sum_{j=1}^{H} \left(r_{i,j} - \mu\right)^2}. \tag{81}$$

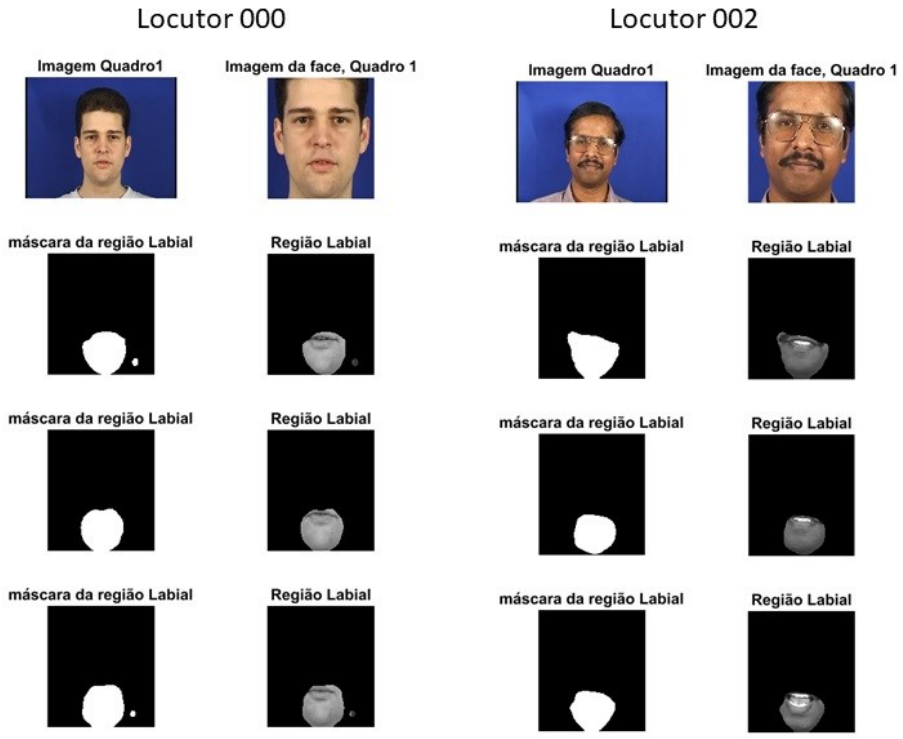
Similarmente ao que é realizado para treinar o limiar de detecção de atividade de voz no capítulo 7 nas equações 69 e 70, porém, neste caso há a soma da média com o desvio padrão como forma de restringir ainda mais a separação da região em que há maior movimento.

Diferentemente do que é feito para a detecção de atividade de voz, este limiar é adaptativo e dependente apenas da imagem sendo utilizada no momento do uso do algoritmo, não necessitando de um treinamento prévio.

9.2 AVALIAÇÃO DO ALGORITMO

Para avaliação do algoritmo, foi aplicado para todos os locutores da base de dados XM2VTS. A profundidade temporal ajustada nos testes foi de 50 quadros, ou seja, a média de movimento é estimada dentro desta janela temporal, resultando em uma detecção da região labial a cada dois segundos, uma vez que os vídeos foram gravados a 25 *fps*. O valor de k da equação 79 foi ajustado para 1. Sua representação gráfica para os locutores 000 e 002, sessão 1, frase 1 pode ser observada na figura 44. É possível notar que para estes locutores a detecção foi realizada de forma a entregar boas estimativas da região do lábio e do queixo que é onde concentra maior parte do movimento decorrente da produção da fala.

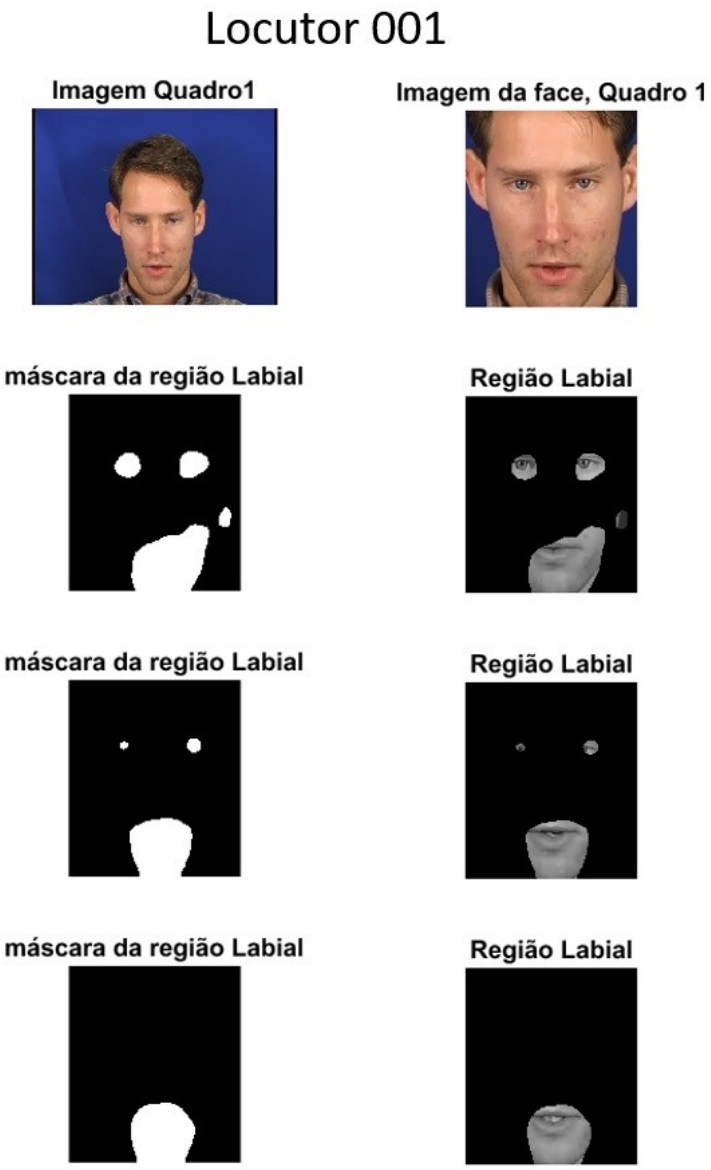
Figura 44 - Detecção da região labial. Locutores 000 e 002



Fonte: Autor “adaptado de” Messer, 1999 [128]
Legenda: Máscara binária resultante do algoritmo de segmentação de imagem. Pixels brancos nas imagens das máscaras representam valores iguais a um e pixels pretos valores iguais a zero. Locutores 000 e 002

Em alguns casos, como pode ser observado na figura 45 para o Locutor 001, sessão 1, frase 1, é possível notar que detecções de outras regiões como os olhos podem ocorrer devido a movimentos de fechamento das pálpebras do locutor. Uma possível solução para este problema é atribuir uma dimensão contínua mínima da região desejada, sendo uma proporcionalidade da região da face detectada. Outra possível solução é aumentar a janela temporal de análise, intensificando a média na região labial.

Figura 45 - Detecção da região labial. Locutor 001

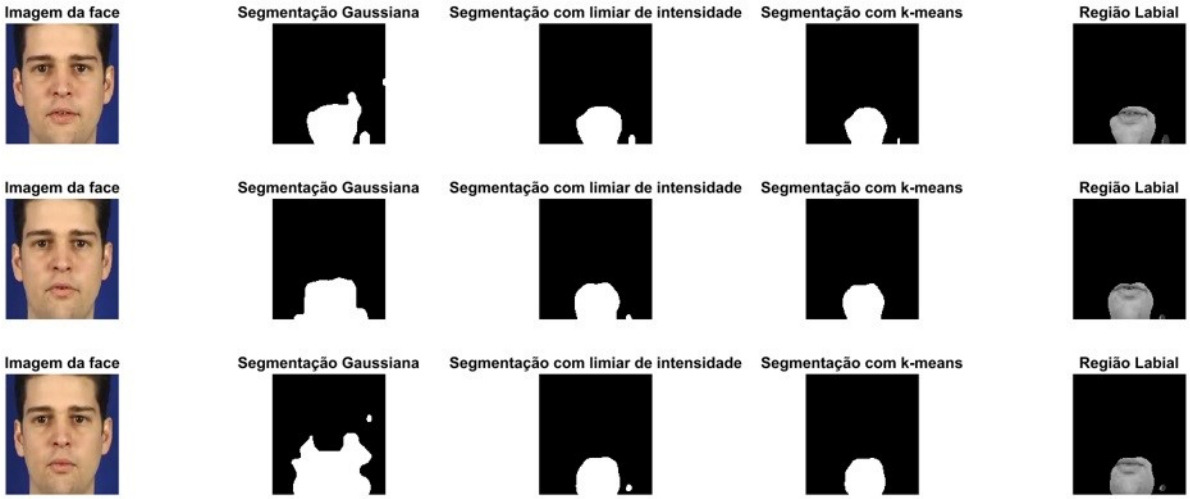


Fonte: Autor “adaptado de” Messer, 1999 [128]
Legenda: Máscara binária resultante do algoritmo de segmentação de imagem. Pixels brancos nas imagens das máscaras representam valores iguais a um e pixels pretos valores iguais a zero. Aplicado ao locutor 001

O método de segmentação por limiar conforme descrito na seção anterior foi utilizado dado a simplicidade na implementação e qualidade dos resultados comparado a outras técnicas. Para realizar esta avaliação, outras técnicas de segmentação foram aplicadas e avaliadas, sendo elas: *k*-means, conforme descrito na seção 3.2 e segmentação Gaussiana com maximização da esperança, conforme descrito na seção 2.3.1.

A figura 46 exibe uma comparação dos resultados obtidos para o locutor 000, sessão 1, frase 1. Neste caso é possível observar uma semelhança entre os resultados obtidos com *k*-means ($k = 2$) e separação através do limiar. A separação com uso da segmentação Gaussiana apresentou um pior resultado uma vez que a máscara ficou mais dispersa. A máscara binária resultante da técnica de segmentação por limiar foi multiplicada ponto-a-ponto pela imagem do canal de luminância do último quadro da janela temporal observada, de forma a resultar em uma figura com pixels igual a zero (pretos) onde o lábio não foi detectado, conforme pode ser observado na figura 46.

Figura 46 - Comparação dos algoritmos de segmentação



Fonte: Autor “adaptado de” Messer, 1999 [128]
Legenda: Máscara binária resultante do algoritmo de segmentação de imagem por três métodos diferentes: Segmentação Gaussiana, Segmentação por limiar de intensidade e segmentação *k*-means. Pixels brancos nas imagens das máscaras representam valores iguais a um e pixels pretos valores iguais a zero. Resultado da aplicação para o locutor 001

10 CONCLUSÃO

As contribuições desta tese são voltadas para o desenvolvimento de novos algoritmos aplicados ao *front-end* de sistemas de verificação automática de locutor, sendo baseados na combinação de parâmetros extraídos do áudio e de movimentos extraídos do vídeo, resultando em contribuições para um sistema biométrico multimodal de reconhecimento de locutor, sendo eles: detecção audiovisual de atividade vocal; nova fusão de parâmetros Mel-cepstrais e vetores de movimento para verificação de locutor; novo método para detecção da região labial. Os algoritmos propostos são implementados de forma a extrair os vetores de movimento diretamente do vídeo codificado em MPEG, aproveitando parâmetros que já são extraídos para compressão. Os resultados foram avaliados com a base de dados XM2VTS em diferentes condições de relação sinal-ruído no áudio, sendo a base modificada pelo autor com adição de diferentes tipos de ruído. Desta forma, os melhores parâmetros e ajustes dos algoritmos puderam ser feitos em função de condições mais próximas do uso real destes sistemas com a consequente validação da proposta, seguindo protocolo de avaliação Lausanne.

Os resultados do método de detecção da atividade vocal foram comparados com uma observação manual dos instantes de ocorrência de movimento e com uso de um detector baseado somente no sinal de áudio, mostrando que o algoritmo proposto encontra de forma satisfatória os instantes de fala, sendo mais vantajoso que o algoritmo baseado somente no sinal de áudio, quando este é capturado sob a presença de ruído ambiente.

A combinação de parâmetros Mel-cepstrais e coeficientes da transformada dos cossenos aplicada aos vetores de movimento extraídos do vídeo codificado em MPEG na região labial, foram utilizados para avaliar o desempenho de um sistema de reconhecimento de locutor com modelagem *i-vector* em diferentes condições de relação sinal-ruído do áudio, resultando em um método para a escolha dos parâmetros biométricos mais discriminativos. Os resultados obtidos mostram um desempenho superior do sistema de reconhecimento do que com o uso de parâmetros extraídos apenas do áudio. Os valores de EER_T e HTER avaliados para a base de dados original, sem presença de ruído no áudio, foram utilizados como comparativo dos resultados obtidos, mostrando que o desempenho do sistema biométrico com a fusão de parâmetros foi superior a outras propostas multimodais apresentadas anteriormente, também baseadas em parâmetros extraídos dos lábios e do áudio.

O método proposto para segmentação da região labial da imagem foi baseado na média de movimentos entre quadros consecutivos do vídeo, agrupando as regiões onde maior quantidade de movimento é detectada. Este método pode ser combinado com os demais

algoritmos propostos de forma a resultar em um *front-end* completo baseado em parâmetros extraídos do vídeo comprimido.

Todos os algoritmos propostos foram avaliados separadamente, implementados em MATLAB® e seu código se faz disponível para download em <https://github.com/maparada/biometria-multimodal> de forma a tornar possível a implementação e eventual melhoria do código por outros pesquisadores interessados, agilizando o desenvolvimento de novas pesquisas.

11 TRABALHOS FUTUROS

As contribuições obtidas nesta tese serão exploradas em pesquisas futuras, sendo elas:

a) Verificação do desempenho da técnica de detecção da região labial desenvolvida, aplicada em conjunto com os algoritmos de detecção de atividade vocal, bem como na extração dos parâmetros de movimento a ser utilizado para verificação automática de locutor;

b) Combinação de parâmetros extraídos do áudio codificado em MPEG com os parâmetros extraídos do vídeo codificado neste mesmo formato, para finalidades biométricas. Propondo desta forma, uma integração completa da extração de parâmetros biométricos multimodais com os parâmetros que são extraídos para compressão;

c) Aplicação de algoritmos de rede neural DNN para modelagem dos parâmetros combinados extraídos através do MFCC e vetores de movimento;

d) Extração dos parâmetros da média de movimento ao longo do tempo com criação de uma imagem integral e aplicação da transformada Haar, resultando em uma aplicação do algoritmo Viola-Jones, porém, aplicado ao movimento da imagem para reconhecimento de regiões da face.

e) Proposta de detecção *liveness* baseada na correlação de movimento e áudio.

f) Extração de parâmetros geométricos do formato da região labial, resultante da aplicação do algoritmo de segmentação proposto para reconhecimento de indivíduo.

g) Análise do desempenho do sistema ASV multimodal com a deterioração da imagem do vídeo.

REFERÊNCIAS

1. PATO, J. N.; MILLETT, L. I. **Biometric Recognition:** challenges and opportunities. Washington, DC: The National Academy Press, 2010.

2. JAIN, A.K.; ROSS, A.; PRABHAKAR, S. An Introduction to Biometric Recognition. **IEEE Transactions on Circuits and Systems for Video Technology,** v. 14, n. 1, p.4-20, Jan. 2004. Disponível em: <http://ieeexplore.ieee.org/document/1262027/>. Acesso em: 10 dez. de 2017.

3. ORTEGA-GARCIA, J. et al. Authentication gets personal with biometrics. **IEEE Signal Processing Magazine,** v. 21, n. 2, p.50-62, Mar. 2004. Disponível em: <http://ieeexplore.ieee.org/document/1276113/>. Acesso em: 10 dez. de 2017.

4. HE, D.; WANG, D. Robust Biometrics-Based Authentication Scheme for Multiserver Environment. **IEEE Systems Journal,** v. 9, n. 3, p.816-823, Sep. 2015. Disponível em: <http://ieeexplore.ieee.org/document/6733264/>. Acesso em: 15 dez. de 2017.

5. KUMAR, A.; ZHOU, Y. Human Identification Using Finger Images. **IEEE Transactions on Image Processing,** v. 21, n. 4, p.2228-2244, Apr. 2012. Disponível em: <http://ieeexplore.ieee.org /document/6044711/>. Acesso em: 15 dez. de 2017.

6. SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. FaceNet: A unified embedding for face recognition and clustering. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2015, Boston, p.815-823, Jun. 2015. IEEE. Disponível em: <http://ieeexplore.ieee.org/document/7298682/>. Acesso em: 15 dez. de 2017.

7. MENOTTI, D. et al. Deep Representations for Iris, Face, and Fingerprint Spoofing Detection. **IEEE Transactions on Information Forensics and Security,** v. 10, n. 4, p.864-879, Apr. 2015. Disponível em: <http://ieeexplore.ieee.org/document/7029061/>. Acesso em: 15 dez. de 2017.

8. CAMPBEL, J.P. Speaker recognition: a tutorial. **Proceedings of the IEEE,** v. 85, n. 9, p.1437-1462, 1997. Disponível em: <http://ieeexplore.ieee.org/document/628714/>. Acesso em: 20 dez. de 2017.

9. HANSEN, J.H.L. et al. Speaker Recognition by Machines and Humans: A tutorial review. **IEEE Signal Processing Magazine**, v. 32, n. 6, p.74-99, Nov. 2015. Disponível em: <http://ieeexplore.ieee.org/document/7298570/>. Acesso em: 20 jun. de 2016.

10. ANIKIN, I.V.; ANISIMOVA, E.S. Handwritten signature recognition method based on fuzzy logic. **2016 Dynamics of Systems, Mechanisms and Machines (dynamics),** Nov. 2016. Disponível em: <http://ieeexplore.ieee.org/document/7818968/>. Acesso em: 20 dez. de 2017.

11. JASWAL, G.; NATH, R.; KAUL, A. Texture based palm Print recognition using 2-D Gabor filter and sub space approaches. In: INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING, COMPUTING AND CONTROL (ISPCC), 2015, Waknaghat, p.344-349. Disponível em: <http://ieeexplore.ieee.org/document/7375053/>. Acesso em: 20 dez. de 2017.

12. HUANG, J.; HOU, D.; SCHUCKERS, S. A practical evaluation of free-text keystroke dynamics. In: IEEE INTERNATIONAL CONFERENCE ON IDENTITY, SECURITY AND BEHAVIOR ANALYSIS (ISBA), 2017, New Delhi. Disponível em: <http://ieeexplore.ieee.org/document/7947695/>. Acesso em: 20 dez. de 2017.

13. YAO, L. et al. Robust Gait Recognition under Unconstrained Environments Using Hybrid Descriptions. In: INTERNATIONAL CONFERENCE ON DIGITAL IMAGE COMPUTING: TECHNIQUES AND APPLICATIONS (DICTA), 2017, Sydney. Disponível em: <http://ieeexplore.ieee.org/document/8227486/>. Acesso em: 03 jan. de 2018.

14. RONG, L. et al. Identification of Individual Walking Patterns Using Gait Acceleration. In: 1ST INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICAL ENGINEERING, 2007, Wuhan, p.543-546,. Disponível em: <http://ieeexplore.ieee.org/document/4272626/>. Acesso em: 03 jan. de 2018.

15. MINAMIDANI, T.; SAI, H.; WATABE, D. Improving ear recognition robustness from single-view-based images rotated in depth for forensic observations. In: INTERNATIONAL CONFERENCE ON BIOMETRICS AND KANSEI ENGINEERING (ICBAKE), 2017, Kyoto, p.90-93,. Disponível em: <http://ieeexplore.ieee.org/document/8090643/>. Acesso em: 03 jan. de 2018.

16. WAN, Haipeng et al. Dorsal hand vein recognition based on convolutional neural networks. In: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE (BIBM), 2017, Kansas City, p.1215-1221. Disponível em: <http://ieeexplore.ieee.org/document/8217830/>. Acesso em: 20 dez. de 2017.

17. WAHEED, Z. et al. Robust extraction of blood vessels for retinal recognition. In: SECOND INTERNATIONAL CONFERENCE ON INFORMATION SECURITY AND CYBER FORENSICS (INFOSEC), 2015, Cape Town, p.1-4. Disponível em: <http://ieeexplore.ieee.org/document/7435497/>. Acesso em: 20 dez. de 2017.

18. OLOYEDE, M.O. et al. Unimodal and Multimodal Biometric Sensing Systems: A Review. **IEEE Access,** v. 4, p.7532-7555, 2016. Disponível em: <http://ieeexplore.ieee.org/document/7580649/>. Acesso em: 03 jan. de 2018.

19. HEZIL, N.; BOUKROUCHE, A. Multimodal biometric recognition using human ear and palmprint. **IET Biometrics,** v. 6, n. 5, p.351-359, Sep. 2017. Institution of Engineering and Technology (IET).

20. PRIMORAC, R. et al. Audio-visual biometric recognition via joint sparse representations. In: 23RD INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION (ICPR), p.3031-3035, 2016, Cancun. Disponível em: <http://ieeexplore.ieee.org/document/7900099/>. Acesso em: 03 jan. de 2018.

21. LOIZOU, P.C. **Speech Enhancement:** Theory and Practice. 2. ed. Boca Raton: CRC Press, 2013.

22. TAVIN. Imagem do trato vocal com numeração, CC-BY-3.0. Disponível em: <https://commons.wikimedia.org/wiki/File:VocalTract_withNumbers.svg>. Acesso 01 jan. 2018.

23. ATAL, B.S. Automatic recognition of speakers from their voices. **Proceedings of the IEEE,** v. 64, n. 4, p.460-475, 1976. Disponível em: <http://ieeexplore.ieee.org/document/1454424/>. Acesso em: 20 abr. de 2016.

24. MIAO, Y.; ZHANG, H.; METZE, F. Speaker Adaptive Training of Deep Neural Network Acoustic Models Using I-Vectors. **IEEE/ACM Transactions on Audio, Speech, And Language Processing,** v. 23, n. 11, p.1938-1949, Nov. 2015. Disponível em: <http://ieeexplore.ieee.org/document/7160703/>. Acesso em: 03 jan. de 2018.

25. SENIOR, A.; LOPEZ-MORENO, I. Improving DNN speaker independence with I-vector inputs. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2014, Florence, p.225-229,. Disponível em: <http://ieeexplore.ieee.org/document/6853591/>. Acesso em: 16 jan. de 2018.

26. LIU, G.; HANSEN, J.H.L. An Investigation into Back-end Advancements for Speaker Recognition in Multi-Session and Noisy Enrollment Scenarios. **IEEE/ACM Transactions on Audio, Speech, And Language Processing,** v. 22, n. 12, p.1978-1992, Dec. 2014. Disponível em: <http://ieeexplore.ieee.org/document/6883142/>. Acesso em: 03 jan. de 2018.

27. AL-ALI, A.K.H. et al. Enhanced Forensic Speaker Verification Using a Combination of DWT and MFCC Feature Warping in the Presence of Noise and Reverberation Conditions. **IEEE Access,** v. 5, p.15400-15413, 2017. Disponível em: <http://ieeexplore.ieee.org/document/7984791/>. Acesso em: 03 jan. de 2018.

28. BRATOSZEWSKI, P.; SZWOCH, G.; CZYZEWSKI, A. Comparison of acoustic and visual voice activity detection for noisy speech recognition. **Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)**, p.287-291, Sep. 2016. Disponível em: <http://ieeexplore.ieee.org/document/7763629/>. Acesso em: 03 jan. de 2018.

29. KUMARI, M. et al. An efficient un-supervised Voice Activity Detector for clean speech. **2015 Communication, Control and Intelligent Systems (CCIS),** p.227-232, Nov. 2015. Disponível em: <http://ieeexplore.ieee.org/document/7437913/>. Acesso em: 20 jan. de 2018.

30. REYNOLDS, D.A. **A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification**. 1992. 308 f. Tese (Doutorado) - Curso de Engenharia, Georgia Institute of Technology, Georgia, 1992.

31. REYNOLDS, D.A.; ROSE R.C. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. **Speech and Audio Processing**, IEEE Trans., v.3, n.1, p. 72-83, 1995. Disponível em: <http://ieeexplore.ieee.org/document/365379/>. Acesso em: 15 jan. de 2018.

32. REYNOLDS, D.A.; QUATIERI, T.F.; DUNN, R.B. Speaker Verification Using Adapted Gaussian Mixture Models. **Digital Signal Processing**, v. 10, n.1, p.19-41, 2000. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1051200499903615>. Acesso em 05 jan. 2018

33. CAMPBELL, W.M.; STURIM, D.E.; REYNOLDS, D.A. Support vector machines using GMM supervectors for speaker verification. **IEEE Signal Processing Letters,** v. 13, n. 5, p.308-311, May 2006. Disponível em: <http://ieeexplore.ieee.org/document/1618704/>. Acesso em: 03 jan. de 2018.

34. KENNY, P. et al. A Study of Interspeaker Variability in Speaker Verification. **IEEE Transactions on Audio, Speech, And Language Processing,** v. 16, n. 5, p.980-988, Jul. 2008. Disponível em: <http://ieeexplore.ieee.org/document/4531370/>. Acesso em: 03 jan. de 2018.

35. DEHAK, N. **Discriminative and Generative Approaches for Long- And Short-Term Speaker Characteristics Modelling: Application to Speaker Verification.** 2009. 183 f. Tese (Doutorado) - Curso de Engenharia, École de Technologie Supérieure, Montreal, 2009.

36. DEHAK, N. **et** al. Front-End Factor Analysis for Speaker Verification. **IEEE Transactions on Audio, Speech, And Language Processing,** v. 19, n. 4, p.788-798, May 2011. Disponível em: <http://ieeexplore.ieee.org/document/5545402/>. Acesso em: 03 jan. de 2018.

37. BRIDLE, J.S.; BROWN, M.D. **An Experimental Automatic Word-Recognition System**. JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England, 1974.

38. MERMELSTEIN, P. Distance measures for speech recognition, psychological and instrumental. **Pattern Recognition and Artificial Intelligence**, Ed. C. H. Chen, p. 374–388. Academic, New York, 1976.

39. DAVIS, S.B.; MERMELSTEIN, P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, v. 28, n.4, p. 357-366, 1980. Disponível em: <http://ieeexplore.ieee.org/document/1163420/>. Acesso em: 05 jan. de 2018.

40. JO, J.; YOO, H.; PARK, I. Energy-Efficient Floating-Point MFCC Extraction Architecture for Speech Recognition Systems. **IEEE Transactions on Very Large Scale Integration (VLSI) Systems,** v. 24, n. 2, p.754-758, Feb. 2016. Disponível em: <http://ieeexplore.ieee.org/document/7072490/>. Acesso em: 05 jan. de 2018.

41. KUCUKBAY, S.E.; SERT, M. Audio-based event detection in office live environments using optimized MFCC-SVM approach. **Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015),** p.475-480, fev. 2015. Disponível em: <http://ieeexplore.ieee.org/document/7050855/>. Acesso em: 07 jan. de 2018.

42. AL-KALTAKCHI, M.T.S. et al. Study of fusion strategies and exploiting the combination of MFCC and PNCC features for robust biometric speaker identification. In: 4TH INTERNATIONAL CONFERENCE ON BIOMETRICS AND FORENSICS (IWBF), Mar. 2016, Limassol. Disponível em: <http://ieeexplore.ieee.org/document/7449685/>. Acesso em: 03 jan. de 2018.

43. BOGERT, B.P.; HEALY, M.J.R.; TUKEY, **The Quefrency Alanysis [sic] of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking**. Proceedings of the Symposium on Time Series Analysis (M. Rosenblatt, Ed), Chapter 15, p. 209-243. New York: Wiley, 1963.

44. Noll, A.M. Cepstrum Pitch Determination. **Journal of the Acoustical Society of America**. [S.l.], vol. 41, no. 2, p. 293-309, 1967.

45. BEIGI, H. **Fundamentals of Speaker Recognition.** Ed. Springer US, 2011.

46. TOGNERI, R.; PULLELLA, D. An Overview of Speaker Identification: Accuracy and Robustness Issues. **IEEE Circuits and Systems Magazine,** v. 11, n. 2, p.23-61, 2011. Disponível em: <http://ieeexplore.ieee.org/document/5871484/>. Acesso em: 03 jan. de 2018.

47. STEVENS, S.S.; VOLKMANN, J.; NEWMAN, E.B. A scale for the measurement of the psychological magnitude pitch. **Journal of the Acoustical Society of America**. [S.l.], v. 8, n.3, p. 185-190, 1937.

48. STEVENS, S. S.; VOLKMANN, J. The relation of pitch to frequency: A revised scale. **The American Journal of Psychology**. [S.l.], v. 53, n.3, p. 329-353, 1940.

49. BERANEK, L.L. **Acoustic Measurements**. Ed. New York: Wiley, 1949.

50. O'SHAUGHNESSY, D. **Speech communication:** human and machine**.** Ed. Addison-Wesley, 1987.

51. ZWICKER, E. Subdivision of the audible frequency range into critical bands. **The Journal of the Acoustical Society of America**. [S.l.], v. 33, n. 2, p. 248-248, 1961.

52. SOHN, J.; KIM, N.S.; SUNG, W. A statistical model-based voice activity detection. **IEEE Signal Processing Lett**., vol. 6, n. 1 p. 1-3, 1999. Disponível em: <http://ieeexplore.ieee.org/document/4907484/>. Acesso em: 03 jan. de 2018.

53. BERITELLI, F.; SPADACCINI, A. The role of Voice Activity Detection in forensic speaker verification. In: 17TH INTERNATIONAL CONFERENCE ON DIGITAL SIGNAL PROCESSING (DSP), Jul. 2011, Corfu. Disponível em: <http://ieeexplore.ieee.org/document/6004980/>. Acesso em: 03 jan. de 2018.

54. MOATTAR, M.H., HOMAYOUNPOUR M.M. A Simple but Efficient Real-Time Voice Activity Detection Algorithm, **Proceedings of European Signal Processing Conference**, p. 2549-2553, 2009. Disponível em: <https://www.eurasip.org/Proceedings/Eusipco/Eusipco2009/contents/Titles-g.html>. Acesso em 01 fev. de 2018.

55. SADJADI, S.O.; HANSEN, J.H.L. Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux. **IEEE Signal Processing Letters,** v. 20, n. 3, p.197-200, Mar. 2013. Disponível em: <http://ieeexplore.ieee.org/document/6403507/>. Acesso em: 03 jan. de 2018.

56. MCLOUGHLIN, I.V. The use of low-frequency ultrasound for voice activity detection. **Proc. Interspeech**, p. 1553-1557, 2014.

57. ANEEJA, G.; YEGNANARAYANA, B. Single Frequency Filtering Approach for Discriminating Speech and Nonspeech. **IEEE/ACM Transactions on Audio, Speech, And Language Processing,** v. 23, n. 4, p.705-717, Apr. 2015. Disponível em: <http://ieeexplore.ieee.org/document/7042263/>. Acesso em: 03 jan. de 2018.

58. YOO, I.; LIM, H.; YOOK, D. Formant-Based Robust Voice Activity Detection. **IEEE/ACM Transactions on Audio, Speech, And Language Processing,** v. 23, n. 12, p.2238-2245, Dec. 2015. Disponível em: <http://ieeexplore.ieee.org/document/7239555/>. Acesso em: 03 jan. de 2018.

59. ZHANG, X.; WANG, D. Boosting Contextual Information for Deep Neural Network Based Voice Activity Detection. **IEEE/ACM Transactions on Audio, Speech, And Language Processing,** v. 24, n. 2, p.252-264, Feb. 2016. Disponível em: <http://ieeexplore.ieee.org/document/7347379/>. Acesso em: 03 jan. de 2018.

60. BRATOSZEWSKI, P. et al. Comparison of acoustic and visual voice activity detection for noisy speech recognition. **2016 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA),** p.287-291, Sep. 2016. Disponível em: <http://ieeexplore.ieee.org/document/7763629/>. Acesso em: 20 jan. de 2018.

61. DOV, D. et al. Audio-Visual Voice Activity Detection Using Diffusion Maps. **IEEE/ACM Transactions on Audio, Speech, and Language Processing,** v. 23, n. 4, p.732-745, Apr. 2015. Disponível em: < http://ieeexplore.ieee.org/document/7045572/ >. Acesso em: 20 nov. de 2017.

62. AYAZ, A. et al. Lipreading using optical flow and support vector machines. In: 3RD INTERNATIONAL CONGRESS ON IMAGE AND SIGNAL PROCESSING, 2010, Yantai, p. 327-330,. Disponível em: <http://ieeexplore.ieee.org/document/5646264/>. Acesso em: 20 nov. de 2017.

63. MONTAZZOLLI, S. et al. Audiovisual voice activity detection using off-the-shelf cameras. In: IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP), p.3886-3890, 2015, Quebec City. Disponível em: <http://ieeexplore.ieee.org/document/7351533/>. Acesso em: 07 jan. de 2018.

64. SONG, T. et al. Robust visual voice activity detection using chaos theory under illumination varying environment. In: IEEE INTERNATIONAL CONFERENCE ON CONSUMER ELECTRONICS (ICCE), Jan. 2014, Las Vegas, p.562-563,. Disponível em: <http://ieeexplore.ieee.org/document/6776134/>. Acesso em: 07 jan. de 2018.

65. MCLACHLAN, G.; PEEL, D. **Finite Mixture Models**. Nova Jersey: Wiley, 2000.

66. TUSKE, Z. et al. Speaker adaptive joint training of Gaussian mixture models and bottleneck features. In: IEEE WORKSHOP ON AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING (ASRU), 2015, Scottsdale, p.596-603. Disponível em: <http://ieeexplore.ieee.org/document/7404850/>. Acesso em: 07 jan. de 2018.

67. DAT, T.T. et al. Robust Speaker Verification Using Low-Rank Recovery under Total Variability Space. In: 5TH INTERNATIONAL CONFERENCE ON IT CONVERGENCE AND SECURITY (ICITCS), Aug. 2015, Kuala Lumpur. Disponível em: <http://ieeexplore.ieee.org/document/7293016/>. Acesso em: 07 jan. de 2018.

68. BAUM, L. et al. A Maximization Technique Occurring in The Statistical Analysis of Probabilistic Functions of Markov Chains. **The Annals of Mathematical Statistics**. [S.l.], v. 41, n.1, p.164-171, 1970.

69. SOONG, F. et al. A vector quantization approach to speaker recognition. In: ICASSP 85. IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, Speech, And Signal Processing, 1985, Tampa, p.387-390,. Disponível em: <http://ieeexplore.ieee.org/document/1168412/>. Acesso em: 20 jan. de 2018.

70. GAUVAIN, J.-L.; LEE, C.-H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. **IEEE Transactions on Speech and Audio Processing,** v. 2, n. 2, p.291-298, Apr. 1994. Disponível em: <http://ieeexplore.ieee.org/document/279278/>. Acesso em: 22 dez. de 2017.

71. REYNOLDS, D. **Gaussian Mixture Models**. MIT Lincoln Laboratory, USA.

72. KENNY, P. Joint factor analysis of speaker and session variability: Theory and algorithms. **Tech. Rep. CRIM-06/08-13**, CRIM, 2005.

73. KUHN, R. et al. Rapid speaker adaptation in eigenvoice space. **IEEE Transactions on Speech and Audio Processing,** v. 8, n. 6, p.695-707, 2000. Disponível em: <http://ieeexplore.ieee.org/document/876308/>. Acesso em: 04 jan. de 2018.

74. KENNY, P.; BOULIANNE, G.; DUMOUCHEL, P. Eigenvoice modeling with sparse training data. **IEEE Transactions on Speech and Audio Processing,** v. 13, n. 3, p.345-354, May 2005. Disponível em: <http://ieeexplore.ieee.org/document/1420369/>. Acesso em: 13 jan. de 2018.

75. LEI, Yun; BURGET, Lukas; SCHEFFER, Nicolas. A noise robust i-vector extractor using vector taylor series for speaker recognition. In**:** IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, May 2013, Vancouver, p.6788-6791,. Disponível em: <http://ieeexplore.ieee.org/document/6638976/>. Acesso em: 03 jan. de 2018.

76. ZHANG, S.; ZHENG, R.; XU, B. An iVector extractor using pre-trained neural networks for speaker verification. **The 9th International Symposium on Chinese Spoken Language Processing,** p.73-77, set. 2014. Disponível em: <http://ieeexplore.ieee.org/document/6936722/>. Acesso em: 08 jan. de 2018.

77. TSUJIKAWA, M.; NISHIKAWA, T.; MATSUI, T. I-vector-based speaker identification with extremely short utterances for both training and testing. In: 2IEEE 6TH GLOBAL CONFERENCE ON CONSUMER ELECTRONICS (GCCE), Oct. 2017, Nagoya. Disponível em: <http://ieeexplore.ieee.org/document/8229389/>. Acesso em: 03 jan. de 2018.

78. JIANG, Y.; LEE, K.A.; WANG, L. PLDA in the i-supervector space for text-independent speaker verification. **Eurasip Journal on Audio, Speech, And Music Processing,** v. 29, n. 1, p.1-13, 15 jul. 2014. Disponível em: <http://asmp.eurasipjournals.com/content/2014/1/29>. Acesso em 10 dez. de 2017.

79. SREE, S., RADHA, N. A Survey on Fusion Techniques for Multimodal Biometric Identification. **International Journal of Innovative Research in Computer and Communication Engineering**, v. 2, n. 12, p. 7493-7497, Dec. 2014.

80. GEETHA, K., RADHAKRISHNAN, V. Multimodal biometric system: A feature level fusion approach. **International journal of computer applications**, v.71, n.4, 2013.

81. ALMAHAFZAH, H.; ZAID, M. Feature Level Fusion the Performance of Multimodal Biometric Systems. **International Journal of Computer Applications,** v. 123, n. 11, p.37-43, Aug. 2015. Foundation of Computer Science.

82. HAGHIGHAT, M.; ABDEL-MOTTALEB, M.; ALHALABI, W. Discriminant Correlation Analysis: Real-Time Feature Level Fusion for Multimodal Biometric Recognition. **IEEE Transactions on Information Forensics and Security,** v. 11, n. 9, p.1984-1996, Sep. 2016. Disponível em: <http://ieeexplore.ieee.org/document/7470527/>. Acesso em: 03 jan. de 2018.

83. PRABHAKAR, S.; PANKANTI, S.; JAIN, A.K. Biometric recognition: security and privacy concerns. **IEEE Security & Privacy Magazine,** v. 1, n. 2, p.33-42, Mar. 2003. Disponível em: <http://ieeexplore.ieee.org/document/1193209/>. Acesso em: 03 jan. de 2018.

84. RUIZ-ALBACETE, V. et al. Direct Attacks Using Fake Images in Iris Verification. **Lecture Notes in Computer Science,** p.181-190, 2008. Springer Berlin Heidelberg.

85. PAN, G. et al. Eyeblink-based Anti-Spoofing in Face Recognition from a Generic Webcamera. In: IEEE 11TH INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2007, Rio de Janeiro, p.1-8. Disponível em: <http://ieeexplore.ieee.org/document/4409068/>. Acesso em: 11 jan. de 2018.

86. GALBALLY, J.; MARCEL, S.; FIERREZ, J. Image Quality Assessment for Fake Biometric Detection: Application to Iris, Fingerprint, and Face Recognition. **IEEE Transactions on Image Processing,** v. 23, n. 2, p.710-724, Feb. 2014. Disponível em: <http://ieeexplore.ieee.org/document/6671991/>. Acesso em: 03 jan. de 2018.

87. CZAJKA, A. Pupil Dynamics for Iris Liveness Detection. **IEEE Transactions on Information Forensics and Security,** v. 10, n. 4, p.726-735, Apr. 2015. Disponível em: <http://ieeexplore.ieee.org/document/7029052/>. Acesso em: 03 jan. de 2018.

88. AKHTAR, Z. et al. Evaluation of serial and parallel multibiometric systems under spoofing attacks. In: IEEE FIFTH INTERNATIONAL CONFERENCE ON BIOMETRICS: THEORY, APPLICATIONS AND SYSTEMS (BTAS), set. 2012, Arlington, p.283-288. Disponível em: <http://ieeexplore.ieee.org/document/6374590/>. Acesso em: 22 dez. de 2017.

89. SADJADI, S. O.; SLANEY, M.; HECK, L. MSR identity toolbox v1. 0: A MATLAB toolbox for speaker-recognition research, **Speech and Language Processing Technical Committee Newsletter**, vol. 1, no. 4, 2013.

90. SLANEY, M. Auditory Toolbox for Matlab, **Apple Computer Technical Report#45**, 1994.

91. BROOKES, M. **Voicebox**. Disponível em: <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>. Acesso em 14 jan. 2018.

92. TAIGMAN, Y. et al. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, Jun. 2014, Columbus, p.1701-1708. Disponível em: <http://ieeexplore.ieee.org/document/6909616/>. Acesso em: 03 jan. de 2018.

93. JUEFEI-XU, F. et al. A preliminary investigation on the sensitivity of COTS face recognition systems to forensic analyst-style face processing for occlusions. In: 2015 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS (CVPRW), Jun. 2015, Boston, p.25-33. Disponível em: <http://ieeexplore.ieee.org/document/7301316/>. Acesso em: 03 jan. de 2018.

94. OUYANG, S. et al. ForgetMeNot: Memory-Aware Forensic Facial Sketch Matching. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), Jun. 2016, Las Vegas, p.5571-5579. Disponível em: <http://ieeexplore.ieee.org/document/7780970/>. Acesso em: 03 jan. de 2018.

95. VIOLA, P.; JONES, M. Rapid object detection using a boosted cascade of simple features. **Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001,** p.511-518, 2001. Disponível em: <http://ieeexplore.ieee.org/document/990517/>. Acesso em: 03 jan. de 2018.

96. HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques.** 3. ed. USA: Elsevier, 2011.

97. BELHUMEUR, P.N.; HESPANHA, J.P.; KRIEGMAN, D.J. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. **IEEE Transactions on Pattern Analysis and Machine Intelligence,** v. 19, n. 7, p.711-720, Jul. 1997. Disponível em: <http://ieeexplore.ieee.org/document/598228/>. Acesso em: 03 jan. de 2018.

98. FENGYU, Z. et al. Image segmentation algorithm of Gaussian mixture model based on map/reduce. In: CHINESE AUTOMATION CONGRESS (CAC), 2017, Jinan, p.1520-1525. Disponível em: <http://ieeexplore.ieee.org/document/8243008/>. Acesso em: 05 jan. de 2018.

99. HAN, K. et al. Speech recognition and lip shape feature extraction for English vowel pronunciation of the hearing-impaired based on SVM technique. In: INTERNATIONAL CONFERENCE ON BIG DATA AND SMART COMPUTING (BIGCOMP), 2016, Hong Kong, p.293-296. Disponível em: <http://ieeexplore.ieee.org/document/7425931/>. Acesso em: 21 nov. de 2017.

100. MATHULAPRANGSAN, Seksan et al. A survey of visual lip reading and lip-password verification. In: INTERNATIONAL CONFERENCE ON ORANGE TECHNOLOGIES (ICOT), Dec. 2015, Hong Kong, p.22-25. Disponível em: <http://ieeexplore.ieee.org/document/7498485/>. Acesso em: 21 nov. de 2017.

101. FRISKY, A.Z.K. et al. Lip-based visual speech recognition system. In: INTERNATIONAL CARNAHAN CONFERENCE ON SECURITY TECHNOLOGY (ICCST), Sep. 2015, Taipei, p.315-319,. Disponível em: <http://ieeexplore.ieee.org/document/7389703/>. Acesso em: 21 nov. de 2017.

102. JAIN, S. et al. Lip contour detection for estimation of mouth opening area. In: FIFTH NATIONAL CONFERENCE ON COMPUTER VISION, PATTERN RECOGNITION, IMAGE PROCESSING AND GRAPHICS (NCVPRIPG), Dec. 2015, Patna, p.1-4. Disponível em: <http://ieeexplore.ieee.org/document/7490009/>. Acesso em: 21 nov. de 2017.

103. ARSIC, A.; JORDANSKI, M.; TUBA, M. Improved lip detection algorithm based on region segmentation and edge detection. In: 23RD TELECOMMUNICATIONS FORUM (TELFOR), Nov. 2015, Belgrade, p.472-475. Disponível em: <http://ieeexplore.ieee.org/document/7377509/>. Acesso em: 22 nov. de 2017.

104. CETINGUL, H.E. et al. Robust Lip-Motion Features for Speaker Identification. **Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech and Signal Processing,** v. 1, p.509-512, May 2005. Disponível em: <http://ieeexplore.ieee.org/document/1415162/>. Acesso em: 03 jan. de 2018.

105. CETINGUL, H.E. et al. Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading. **IEEE Transactions on Image Processing,** v. 15, n. 10, p.2879-2891, Oct. 2006. Disponível em: <http://ieeexplore.ieee.org/document/1703580/>. Acesso em: 03 jan. de 2018.

106. FARAJ, M.-I. et al. Synergy of Lip-Motion and Acoustic Features in Biometric Speech and Speaker Recognition. **IEEE Transactions on Computers,** v. 56, n. 9, p.1169-1175, Sep. 2007. Disponível em: <http://ieeexplore.ieee.org/document/4288084/>. Acesso em: 03 jan. de 2018.

107. YUAN, Y. et al. SALM: Smartphone-Based Identity Authentication Using Lip Motion Characteristics. In: IEEE INTERNATIONAL CONFERENCE ON SMART COMPUTING (SMARTCOMP), 2017, Hong Kong, p.1-8. Disponível em: <http://ieeexplore.ieee.org/document/7947043/>. Acesso em: 03 jan. de 2018.

108. ICHINO, M. et al. Lip-Movement Based Speaker Recognition Focused on the Distributed Structure of Lip-Movement Data. In: 5TH IIAI INTERNATIONAL CONGRESS ON ADVANCED APPLIED INFORMATICS (IIAI-AAI), 2016, Kumamoto, p.1132-1135. Disponível em: <http://ieeexplore.ieee.org/document/7557784/>. Acesso em: 03 jan. de 2018.

109. SHI, X.-X. et al. Visual speaker authentication by ensemble learning over static and dynamic lip details. In: IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP), Sep. 2016, Phoenix, p.3942-3946. Disponível em: <http://ieeexplore.ieee.org/document/7533099/>. Acesso em: 03 jan. de 2018.

110. PARADA, M.; SANCHES, I. Visual Voice Activity Detection Based on Motion Vectors of MPEG Encoded Video. In: EUROPEAN MODELLING SYMPOSIUM (EMS2017), Manchester, Nov. 2017. IEEE. A ser publicado.

111. ZAFEIRIOU, S.; ZHANG, C.; ZHANG, Z. A survey on face detection in the wild: Past, present and future. **Computer Vision and Image Understanding,** v. 138, p.1-24, Sep. 2015. Elsevier BV.

112. AGRAWAL, Samiksha; KHATRI, Pallavi. Facial Expression Detection Techniques: Based on Viola and Jones Algorithm and Principal Component Analysis. In: FIFTH INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING & COMMUNICATION TECHNOLOGIES, Feb. 2015, Haryana, p.108-112. Disponível em: <http://ieeexplore.ieee.org/document/7079062/>. Acesso em: 03 jan. de 2018.

113. CROW, Franklin C. Summed-area tables for texture mapping. **ACM Siggraph Computer Graphics,** v. 18, n. 3, p.207-212, Jul. 1984. Association for Computing Machinery (ACM).

114. PAPAGEORGIOU, C.P.; OREN, M.; POGGIO, T. A general framework for object detection. **Sixth International Conference on Computer Vision,** p.555-562, 1998. Narosa Publishing House.

115. FREUND, Y.; SCHAPIRE, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. **Journal of Computer and System Sciences,** v. 55, n. 1, p.119-139, Aug. 1997. Elsevier BV.

116. LIENHART, R.; KURANOV, A.; PISAREVSKY, V. Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection. **Lecture Notes in Computer Science,** p.297-304, 2003. Springer Berlin Heidelberg.

117. TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Boston: Addison-wesley Longman, 2005.

118. MARPE, D. et al. The H.264/MPEG4 advanced video coding standard and its applications. **IEEE Communications Magazine,** v. 44, n. 8, p.134-143, Aug. 2006. Disponível em: <http://ieeexplore.ieee.org/document/1678121/>. Acesso em: 03 jan. de 2018.

119. SULLIVAN, G.J. et al. Overview of the High Efficiency Video Coding (HEVC) Standard. **IEEE Transactions on Circuits and Systems for Video Technology,** v. 22, n. 12, p.1649-1668, Dec. 2012. Disponível em: <http://ieeexplore.ieee.org/document/6316136/>. Acesso em: 03 jan. de 2018.

120. ALENCAR, Marcelo S. **Televisão digital**. São Paulo: Érica, 2011.

121. ROBIN, Michael; POULIN, Michel. **Digital television fundamentals**. Chicago: Mcgraw-Hill, 2000.

122. FOGG, Chad et al. **MPEG video compression standard**. New York: Springer US, 2000.

123. OPPENHEIM, Alan; SCHAFER, Ronald. **Processamento em tempo discreto de sinais**. São Paulo: Pearson, 2013.

124. FFmpeg Developers. (2016). **Biblioteca FFmpeg**. Disponível em: <http://ffmpeg.org/>. Acesso em 14 jan. 2018.

125. KANTOROV, V.; LAPTEV, I. Efficient Feature Extraction, Encoding, and Classification for Action Recognition. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2014, Columbus, p.2593-2600. Disponível em: <http://ieeexplore.ieee.org/document/6909728/>. Acesso em: 03 jan. de 2018.

126. **International Standard ISO/IEC 2382-37**: Information technology, Vocabulary Part 37: Biometrics, 1ªed., Dec. 2012.

127. MARTIN, A., et al. The DET Curve in Assessment of Detection Task Performance, **Proc. Eurospeech 97**, vol. 4, p. 1895-1898, Sep. 1997.

128. MESSER, K., et al. XM2VTSdb: The Extended M2VTS Database, **Proceedings of the 2nd Conference on Audio and Video-base Biometric Personal Verification (AVBPA99)**, Springer Verlag, New York, 1999.

129. LUETTIN, J.; MAITRE, G. Evaluation Protocol for the XM2FDB (Lausanne Protocol). **IDIAP Com**., Oct. 1998.

130. CHAN, C.H. et al. Local Ordinal Contrast Pattern Histograms for Spatiotemporal, Lip-Based Speaker Authentication. **IEEE Transactions on Information Forensics and Security**, [s.l.], v. 7, n. 2, p.602-612, Apr. 2012. Disponível em: < http://ieeexplore.ieee.org/document/6081928/>. Acesso em 07 abr. de 2018.

131. BROUN, C.C. et al. Automatic speechreading with application to speaker verification. IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING, 2002, Orlando, p.685-688. Disponível em: <http://ieeexplore.ieee.org/document/5743810/>. Acesso em 07 abr. de 2018.

132. FARAJ, M.I.; BIGUN, J. Motion Features from Lip Movement for Person Authentication. 18TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION (ICPR'06), 2006, Hong Kong, p.1059-10626. Disponível em: <http://ieeexplore.ieee.org/document/1699708/>. Acesso em 07 abr. de 2018.

133. SANCHEZ, R. **Aspects of Facial Biometrics for Verification of Personal Identity**. 2000. 143 f. Tese (Doutorado) - Curso de Engenharia, Universidade de Surrey, Surrey, U.K., 2000.