PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

MAYARA FERREIRA DA SILVA

APLICAÇÃO DO MÉTODO DE FUSÃO PARA VERIFICAÇÃO DE LOCUTOR INDEPENDENTE DE TEXTO

Porto Alegre
2015

MAYARA FERREIRA DA SILVA

APLICAÇÃO DO MÉTODO DE FUSÃO PARA VERIFICAÇÃO DE LOCUTOR INDEPENDENTE DE TEXTO

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica da Pontifícia Universidade Católica do Rio Grande do Sul, como requisito fundamental para conclusão do curso e obtenção do grau de Mestre em Engenharia Elétrica.

Orientadora: Profª. Maria Cristina Felippeto de Castro

Co-orientador: Prof. Dênis Fernandes

Porto Alegre
2015

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

APLICAÇÃO DO MÉTODO DE FUSÃO PARA VERIFICAÇÃO DE LOCUTOR INDEPENDENTE DE TEXTO

CANDIDATA: MAYARA FERREIRA DA SILVA

Esta Dissertação de Mestrado foi julgada para obtenção do título de MESTRE EM ENGENHARIA ELÉTRICA e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia Elétrica da Pontifícia Universidade Católica do Rio Grande do Sul.

DRA. MARIA CRISTINA F. DE CASTRO - ORIENTADORA

DR. DÊNIS FERNANDES - CO-ORIENTADOR

BANCA EXAMINADORA

DRA. CLÁUDIA REGINA BRESCANCINI - DA FACULDADE DE LETRAS - PUCRS

DR. FERNANDO CESAR COMPARSI DE CASTRO - PPGEE - FENG - PUCRS

PUCRS
Campus Central
Av. Ipiranga, 6681 - Prédio 30 - Sala 103 - CEP: 90619-900
Telefone: (51) 3320.3540 - Fax: (51) 3320.3625
E-mail: engenharia.pg.eletrica@pucrs.br
www.pucrs.br/feng

Dedico este trabalho à minha família,
que sempre acredita em mim e me apoia em
todos os projetos que assumo.

AGRADECIMENTOS

A Deus, por nos dar saúde e forças para vencer os desafios.

Aos meus pais, Jorge Ferreira da Silva Filho e Nelci Rodrigues da Silva, que são meus exemplos, a quem admiro muito, por sempre me mostrarem o melhor caminho e por seu apoio incondicional em todos os momentos de minha vida.

Ao meu irmão, Rogers Ferreira da Silva, de quem me orgulho muito, e que, mesmo longe, deu suas contribuições para o aperfeiçoamento do trabalho e sempre me motivou.

Ao meu noivo, Roberto Moraes Brondani, que sempre me incentivou e acreditou que eu conseguiria.

Ao colega, Endrigo Rosa de Carvalho, que me ajudou em vários momentos, e a todos os colegas do mestrado, pelo bom convívio durante o mestrado e por todas as trocas de experiências.

Ao professor Dênis Fernandes, meu co-orientador, por ter confiado em mim e dedicado muito do seu tempo, sempre me orientando de forma positiva e compartilhando comigo os seus conhecimentos.

À professora Maria Cristina Felippeto de Castro, minha orientadora, por ter me dado esta oportunidade e por todo o aprendizado.

Ao LAFA (Laboratório de Áudio e Fonética Acústica), pelos conhecimentos compartilhados e pelo companheirismo dos membros do grupo.

À Rádio Guaíba, em especial à Sinara Félix, pelo material cedido, que foi de extrema importância para execução deste trabalho, e pela receptividade.

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), pela concessão da bolsa durante o período de mestrado.

À PUCRS (Pontifícia Universidade Católica do Rio Grande do Sul), por sua excelente trajetória como universidade.

A todos que de alguma forma contribuíram para o meu crescimento pessoal e acadêmico, e aos amigos e parentes que sempre acreditaram e torceram por mim.

RESUMO

Este trabalho apresenta uma visão geral acerca de verificação de locutor independente de texto, demonstrando o funcionamento básico do sistema e as principais referências de métodos já utilizados ao longo de anos para extração de características da fala e modelamento do locutor. Detectado um ponto a ser trabalhado dentro da etapa de extração de características, objetiva-se determinar coeficientes ou um conjunto destes relevantes para discriminação do locutor, com o intuito de minimizar a EER (*Equal Error Rate*). A proposta consiste em substituir os coeficientes delta(Δ) e *double*-delta(Δ^2) por coeficientes de um preditor LPC (*Linear Predictor Coding)* o qual realiza a predição dos coeficientes MFCC (*Mel Frequency Cepstral Coeficients)*. Além disso, aplica-se uma fusão a nível de *score* em função de sistemas baseados em MFCC e LPC. Outra análise discutida no trabalho é a fusão de um sistema MFCC com Δ e Δ^2. Um tópico também avaliado é com relação a variações de SNRs (*Signal to Noise Ratios*) nos áudios testados. Além disso, é elaborado um banco de falas em português brasileiro. Por fim, são expostos os resultados obtidos e é feita a análise dos mesmos, a fim de refletir sobre o que era esperado e levantar alguns comentários. Enfim, são feitas as considerações a respeito do trabalho, e elencadas as perspectivas futuras em torno das pesquisas de verificação de locutor independente de texto. Com este trabalho atingiu-se uma redução de 4% na taxa de erro igual (EER) em comparação ao sistema de referência, sendo que os melhores resultados foram apresentados pelo sistema que realiza um fusão do sistema MFCC com o Δ e Δ^2.

Palavras-chave: Reconhecimento de Voz; Verificação de Locutor; Fusão de Escores; Modelo de Mistura Gaussiana.

ABSTRACT

This work presents an overview of text independent speaker verification, describing the basic operation of the system and the reviewing some important developments in speaker modeling and feature extraction from speech. Following, a point of improvement identified within the feature extraction stage leads to the main objective of this work: to determine one or more sets of coefficients relevant to speaker discrimination while minimizing the equal error rate (EER). The proposal is to replace the delta(Δ) and double-delta($Δ^2$) coefficients by a linear predictor code (LPC) for the mel frequency cepstral coefficients (MFCC). In addition, score level fusion is employed to combine the ouputs of MFCC-only and MFCC-LPC systems, as well as MFCC-only and MFCC-Δ-$Δ^2$ systems. In all cases, performance is evaluated with respect to variations of the signal to noise-ratio (SNR) in the tested audio. In addition, the work introduces a new Brazilian Portuguese speech repository containing free-speech from 155 males. Results and discussions are presented with a reflection on the expected outcomes, as well as general comments and observations. Finally, concludings remarks are made about the work, featuring future prospects regarding text independent speaker verification research. This work attained a 4% reduction in the EER compared to the reference system (MFCC-only), with best results occuring in the case fusion of MFCC-only and MFCC-Δ-$Δ^2$ scores.

Keywords: Voice Recognition; Speaker Verification; Scores Fusion; Gaussian Mixture Models

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

ANNs – *Artificial Neural Networks*

CMN – *Cepstral Mean Normalization*

DCF – *Decision Cost Function*

DET – *Decision Error Tradeoff*

DTW – *Dynamic Time Warping*

EER – *Equal Error Rate*

EM – *Expectation Maximization*

FA – *Factor Analysis*

FAR – *False Acceptance Rate*

FRR – *False Rejection Rate*

GMM – *Gaussian Mixture Models*

HMM – *Hidden Markov Model*

HTK – *Hidden Markov Model Toolkit*

LPC – *Linear Predictive Coding*

LPCCs – *Linear Predictive Cepstral Coefficients*

LR – *Likelihood Ratio*

LSFs – *Line Spectral Frequencies*

MAP – *Maximum a Posteriori*

MFCC – *Mel Frequency Cepstral Coeficients*

ML – *Maximum Likelihood*

NIST – *National Institute of Standards and Technology*

PCA – *Principal Components Analysis*

PDA – *Pitch Determination Algorithm*

PLP – *Perceptual Linear Prediction*

SMS – *Short Message Service*

SVM – *Suport Vector Machines*

UBM – *Universal Background Model*

VQ – *Vector Quantization*

LISTA DE SÍMBOLOS

MFCC-Δ-Δ² – Sistema baseado num vetor de coeficientes MFCC concatenado com um vetor de coeficientes Δ e Δ²

MFCC-LPC – Sistema baseado num vetor de coeficientes MFCC concatenado com um vetor de coeficientes LPC

MFCC+LPC – Fusão de Escores dos sistemas MFCC e LPC

MFCC+Δ+Δ² – Fusão de Escores dos sistemas MFCC e Δ Δ²

cm – centímetro

ms – milissegundo

min – minuto

SUMÁRIO

1 Introdução

Atualmente, uma das principais áreas de pesquisa é o reconhecimento biométrico. O reconhecimento biométrico possibilita identificar as pessoas através de características únicas de cada indivíduo, conforme o recurso que se tenha disponível, como imagem, voz ou impressão digital. Os sinais biométricos caracterizam o indivíduo, logo, não podem ser facilmente falsificados.

Esta ferramenta já é utilizada por diversos setores, como em bancos, nos caixas eletrônicos; empresas, no ponto dos funcionários e para acessos a áreas restritas; no governo, para autenticação do voto nas urnas eletrônicas; na área forense, por peritos e investigadores, entre outras aplicações.

Os procedimentos biométricos mais conhecidos são baseados em face, impressões digitais e voz. O reconhecimento através da face procura identificar pontos do rosto e calcular algumas medidas entre eles, como distância entre orelhas, nariz e boca, arcada dentária, crânio, com o cuidado de desprezar traços que variam constantemente, como maquiagem e penteado. Por outro lado, o reconhecimento através de impressões digitais identifica as cristas e vales de fricção de cada dedo, visto que os sulcos das impressões digitais não são retos e contínuos, e sim partidos, bifurcados e curvos (AMORIM, 2005). Logo, o que diferencia uma impressão da outra é o conjunto desses detalhes, suas posições, tamanhos e quantidade. O reconhecimento de voz, por sua vez, analisa características da fala de um indivíduo, tanto físicas como emocionais e comportamentais. Em muitos casos, a voz é o único recurso disponível (como em uma ligação telefônica).

Muitos acreditam que elaborar técnicas de reconhecimento biométrico seja algo recente, porém já data de anos atrás. Conforme (CANEDO, 2010):

> O uso da impressão digital para assinar documentos foi prática entre os antigos Assírios, Babilônios, Japoneses e Chineses. No Leste da Ásia, artesãos da cerâmica usavam a impressão digital como marca pessoal para seus produtos. Negociantes do vale do Nilo, no Egito antigo, eram identificados pela altura, cor dos olhos e compleição. Essa informação ajudava a identificar negociantes com os quais os mercadores já tinham feito negócios com sucesso no passado. O

> explorador João de Barros relatou que os mercadores chineses estampavam mãos e pés de crianças com papel e tinta para distinguir uma criança da outra.

Na década de 70 começaram a surgir os primeiros sistemas para biometria automatizados. No Brasil, a biometria por impressão digital se deu somente no início do século XX. Aos poucos, o próprio governo passou a ter interesses nessa área e diversas aplicações foram surgindo de forma crescente. Hoje em dia até os aparelhos celulares já têm embarcados alguns sistemas de biometria.

A fala é a forma mais natural de comunicação entre as pessoas, incorporando tanto o que está sendo dito como características específicas de quem está falando. Cada indivíduo possui diferentes formas e tamanhos dos órgãos produtores da voz, além de maneiras distintas de falar, como ritmo, estilo, pronúncia, entre outros. Dentro da área de reconhecimento de voz, podemos destacar três diferentes ênfases: o reconhecimento da locução (o que está sendo dito), o reconhecimento do locutor (quem está falando) e o reconhecimento do idioma falado (inglês, português, alemão, chinês). Neste trabalho serão abordadas técnicas dirigidas ao reconhecimento do locutor.

Ao pensar em reconhecimento de locutor também obtemos algumas ramificações, como identificação, verificação, detecção e segmentação, cada uma com suas particularidades. A identificação e a verificação muitas vezes se confundem: a identificação consiste de identificar um locutor dentro de um grupo, calculando quem é ele, se ele pertence ao grupo ou não; já na verificação, pretende-se confirmar a identidade do locutor desconhecido, confrontando o seu áudio com o do locutor alvo (locutor que acredita-se ter dado origem ao áudio do locutor desconhecido). Já a detecção, detecta a presença de locutores no áudio e a segmentação separa os trechos de áudio de determinado locutor. Como escopo deste trabalho, será abordada a verificação de locutor, ou seja, os dados de uma declaração desconhecida serão comparados com os dados de um locutor específico, objetivando descobrir se a fala desconhecida pertence a tal locutor.

As principais aplicações de verificação de locutor são autenticações, incluindo assim transações bancárias, telecomunicações, aplicações *on-site* (controle de acesso a locais e facilidades), aplicações remotas (controle de acesso a serviços) e aplicações forenses.

Em serviços telefônicos como os disponibilizados por bancos e agências de telefonia, o único recurso disponível para biometria é a voz, possibilitando a autenticação do usuário para o uso de determinados serviços restritos a ele. Um exemplo interessante e pouco conhecido de autenticação telefônica é no monitoramento de prisão domiciliar e de chamadas realizadas dentro do presídio (REYNOLDS, 2002; LI; JAIN, 2009). Em controles de acesso, o recurso de voz também já vem se expandindo, na maioria das vezes como um complemento a outro mecanismo já utilizado (chaves ou senhas) a fim de abrir portas, liberar acesso a computadores ou celulares.

Nas aplicações forenses, o uso da verificação de locutor auxilia os investigadores a direcionarem sua investigação, servindo como um recurso auxiliar, não podendo ser utilizado como prova final de culpa ou inocência. As aplicações forenses são de elevada importância, já que, há muito tempo, delegados, juízes e advogados utilizam procedimentos de verificação (autenticação) da voz para investigação de um suspeito ou mesmo confirmação em um julgamento (CAMPBELL *et al.*, 2009).

Focando as aplicações forenses, o trabalho desenvolve o uso de verificação de locutor independente de texto, visto que o conteúdo das gravações (o que foi dito) não será de conhecimento prévio, não usará uma senha ou frase determinada, e provavelmente será uma fala espontânea.

A verificação de locutor independente de texto já vem sendo estudada ao longo de mais de 30 anos (KINNUNEN; LI, 2010) e já tem uma longa base de dados e resultados, conforme literatura. Os procedimentos para implementação de um sistema de verificação de locutor são constituídos basicamente de duas fases: a fase de treino e a fase de teste. A fase de treino consiste no modelamento dos locutores e de um modelo universal para futura comparação (*background* e/ou impostores) e a fase de teste, em modelar a declaração desconhecida e compará-la com a do locutor alvo, decidindo se a fala desconhecida pertence a tal locutor (aceita) ou não pertence (rejeita). Juntas, as fases de treino e teste constituem um processo de classificação. Na fase de treino, o locutor disponibiliza amostras de sua fala que serão usadas para treinar o modelo deste locutor.

Para o modelamento do locutor é necessário extrair suas características, o que consiste do processamento dos sinais de voz do locutor, com vistas a identificar e

selecionar uma representação mais sucinta dos mesmos, e que ainda contenha informação suficiente para conduzir o processo de verificação. Buscando um modelamento flexível do locutor, a extração de características pode ser executada de diferentes maneiras. Isto permite a obtenção de diversas formas de características, entre elas, características da fonte da voz, características espectrais de tempo curto, características espectro-temporais, características prosódicas e características de alto nível (KINNUNEN; LI, 2010), que serão todas elas detalhadas na Seção 3.3. O procedimento de extração de características mais utilizado atualmente é o MFCC (*Mel Frequency Cepstral Coeficients*) (HOSSAN; MEMON; GREGORY, 2010).

Ao nível do modelamento estatístico encontram-se diversos estudos, tais como, *Vector Quantization* (VQ) (SOONG; ROSENBERG, 1988), *Dynamic Time Warping* (DTW) (FURUI, 1981), *Gaussian Mixture Models* (GMM) (REYNOLDS; QUATIERI; DUNN, 2000), *Hidden Markov Model* (HMM) (BENZEGHIBA; BOURLAND, 2006), *Suport Vector Machines* (SVM) (CAMPBELL; STURIM; REYNOLDS, 2006) e *Artificial Neural Networks* (ANNs) (FARRELL; MAMMONE; ASSALEH, 1994). O mais utilizado, conforme a literatura, é o GMM, porém, o uso do SVM, por apresentar melhores resultados, tem sido a tendência das pesquisas mais recentes. Além disso, existe uma forte busca por modelos híbridos.

No capítulo 3 será detalhado todo o processo de treino e de teste, bem como os tipos de extração de características e os principais métodos de modelamento do locutor.

Dentro deste cenário, encontra-se a extração de características como um ponto a ser melhor desenvolvido, já que a utilização de apenas um tipo de característica pode não ser capaz de representar o locutor com o melhor desempenho. Pelas pesquisas mais recentes (LIU; HUANG, 2009; LI *et al.*, 2012; NAKAGAWA; WANG; OHTSUKA, 2012; LI; GUO; DAI, 2012), evidencia-se a tendência ao uso de mais de uma informação, de naturezas diferentes, no vetor de características (por exemplo, incluindo coeficientes que representem ritmo da fala e/ou entonação).

A proposta deste trabalho vem a ser a de selecionar e combinar características, através dos coeficientes do vetor característico, analisando a taxa de erros resultante, a

fim de buscar um melhor desempenho do sistema, dado que, em aplicações forenses, é necessário focar na minimização drástica do erro.

No próximo capítulo são elencados os objetivos gerais e específicos do sistema proposto. Em seguida, é apresentada a fundamentação teórica do trabalho, detalhando como a voz é produzida pelos humanos e como é realizado o processamento da voz a fim de extrair informações da mesma. É também contextualizado o assunto abordado no trabalho, evidenciando-se em que área de estudo o mesmo se encontra e também os principais métodos e práticas já implementadas, salientando o que vem sendo apontado como estado-da-arte em Verificação de Locutor.

Logo, explica-se com mais profundidade alguns conceitos importantes como *background*, método da fusão e medidas de desempenho do sistema. Como referência para posterior comparação são apresentados alguns resultados de outros pesquisadores, com o intuito de avaliar o desempenho do sistema proposto frente aos já existentes.

Além disso, é determinada a metodologia aplicada, descrevendo todos os passos necessários para a obtenção dos resultados. Por fim, destacam-se os recursos utilizados na implementação do sistema proposto, os desafios enfrentados e os procedimentos realizados, assim como os resultados obtidos com cada configuração do sistema e a análise dos valores. Então, seguem-se as conclusões referentes ao trabalho proposto.

2 Objetivos

Este capítulo estabelece os objetivos a serem alcançados com este trabalho. O objetivo principal será desenvolver métodos para verificação de locutor independente de texto, buscado selecionando um método para determinação dos coeficientes do vetor de características do locutor, através de características do sinal de voz que são relevantes para discriminar o locutor. Este vetor característico será usado na geração do modelo do locutor para futura comparação.

Além disso, pretende-se implementar um sistema de verificação de locutor independente de texto, com os métodos mais desenvolvidos pela literatura, e realizar testes com um banco de falas em português (a ser elaborado), possibilitando avaliar o desempenho do sistema.

Projeta-se também desenvolver uma fusão entre um sistema básico de verificação de locutor e um sistema com o método proposto a fim de avaliar os resultados desta configuração.

2.1 Objetivo Geral

O objetivo principal do trabalho é selecionar características do sinal de voz relevantes para discriminar o locutor, através de um conjunto de coeficientes a ser determinado, a fim de obter melhores resultados no desempenho de um sistema de verificação de locutor independente de texto.

2.2 Objetivos Específicos

- Implementar um sistema de verificação de locutor independente de texto;
- Implementar um banco de falas em português;
- Selecionar um conjunto de coeficientes relevantes para discriminar o locutor;
- Validar as técnicas com diferentes níveis de relação sinal-ruído nos áudios;
- Desenvolver diferentes métodos de combinação das características;

- Avaliar o desempenho das diferentes configurações do sistema;

3 Fundamentação Teórica

Neste capítulo são descritos os principais tópicos do embasamento teórico do trabalho, a saber: primeiramente, um apanhado geral de reconhecimento de voz, até o ponto a ser desenvolvido no trabalho, a verificação de locutor independente de texto; em seguida, os principais métodos já utilizados no processo, dentre eles, diferentes métodos de extração de características e os principais modelamentos do locutor.

3.1 Produção da Voz

As principais fontes de características da voz humana são as físicas, socioculturais e psicológicas. Com relação às características físicas, a produção da fala humana dá-se através do funcionamento do chamado aparelho fonador, dado pelos sistemas articulatório, fonatório e respiratório. das cordas vocais é o trato vocal, que será detalhado a seguir.

Conforme LI e JAIN (2009), a produção de fala é o resultado da execução de comandos neuromusculares que expelem ar dos pulmões, causando a vibração das cordas vocais (no caso dos sons vozeados), ou as deixando inativas (no caso dos sons desvozeados).

Na Figura 1 temos o aparato vocal, incluindo as três cavidades básicas de ressonância (cavidade nasal, bucal e faríngea). Conforme a posição dos articuladores (lábios, língua, mandíbula, palato, etc), é possível a obtenção de diferentes conformações do trato vocal. O tamanho estimado de um trato vocal masculino adulto é de aproximadamente 17cm (FLANAGAN, 1972).

Figura 1 - Trato Vocal Humano

Fonte: Adaptado de LI e JAIN (2009, p.1265)

Para a produção de sons de voz, as cordas vocais, localizadas na laringe, quando tensionadas, abrem e fecham a laringe, controlando o fluxo de ar. O ar é cortado e pulsado no aparato vocal a uma dada frequência chamada *pitch*. Quando as cordas vocais não estão vibrando, o ar pode passar livremente através da laringe, e então pode-se produzir dois tipos de sons: sor io vozeados e sons plosivos transientes. Os não vozeados são gerados quando o ar começa turbulento no ponto de constrição e os plosivos transientes, quando a pressão é acumulada e abruptamente liberada no ponto de total fechamento (oclusão) no trato vocal.

Como a onda acústica passa pelo trato vocal, suas componentes de frequência (espectro) têm amplitude modificada pelas ressonâncias que encontra, sendo possível estimar a forma do trato vocal através da forma do espectro do sinal de voz (localização das formantes). As formantes são as ressonâncias do trato vocal, sendo as frequências onde há maior transmissão de energia (PETRY, 2002). Conforme a fonte de excitação do fluxo de ar, a mesma pode ser caracterizada como fonação, sussurro, fricção, compressão, vibração ou uma combinação dessas (CAMPBELL JR, 1997).

Conforme CAMPBELL JR (1997), a excitação de fonação ocorre quando o fluxo de ar é modulado pelas pregas vocais. Quando elas se fecham a pressão acumula-se abaixo das mesmas até que elas explodam. Em seguida, as pregas são puxadas novamente para trás por sua tensão, elasticidade e o efeito de Bernoulli (lei que explica como as pregas vocais entram em vibração e se mantém). Este fluxo de ar pulsado, decorrente

das pregas vocais oscilantes, excita o trato vocal. A frequência de oscilação é chamada de frequência fundamental, e isso depende do comprimento, tensão e magnitude das pregas vocais. Assim, a frequência fundamental é uma característica física distintiva de um locutor.

A excitação de sussurro é produzida pelo fluxo de ar correndo através de uma pequena abertura triangular entre as cartilagens aritenóides na parte de trás das pregas vocais quase fechadas. Isto resulta num fluxo de ar turbulento, que tem uma característica de ruído de banda larga.

Excitação de fricção é produzida por constrições no trato vocal. O lugar, forma e grau de constrição determinam a forma do ruído de excitação de banda larga. À medida que a constrição se move para frente, a concentração espectral em geral aumenta em frequência. Os sons gerados por fricção são chamados fricativas sendo algumas sibilantes. Fricção pode ocorrer com ou sem vozeamento.

Excitação de compressão resulta da liberação de um trato vocal completamente fechado e pressurizado. Isto indica um silêncio (durante a acumulação de pressão), seguido de uma soltura abrupta de ruído. Se o lançamento é súbito, uma plosiva é gerada.

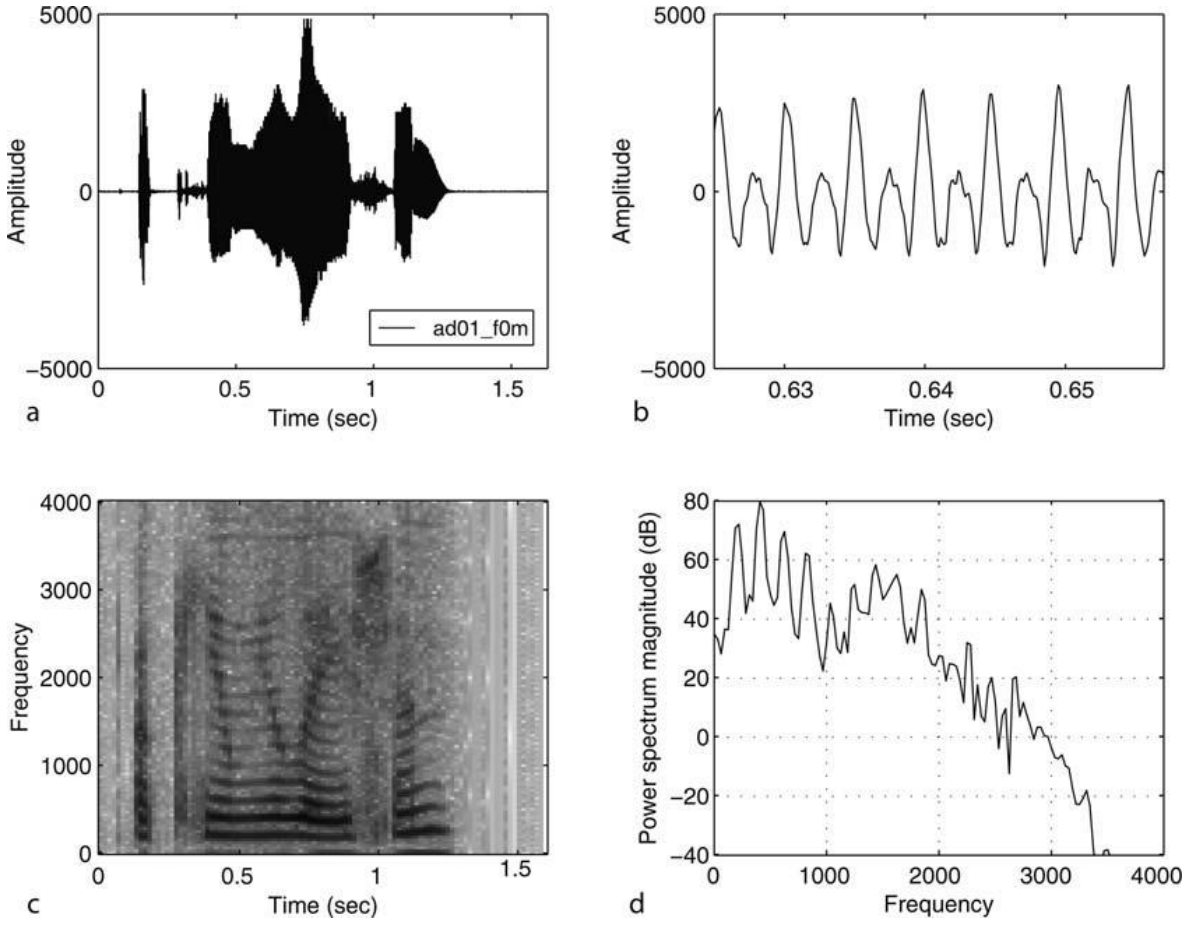
A fala produzida por uma excitação de fonação é chamada vozeada, a produzida por fonação mais fricção é chamada mista e a fala produzida pelos outros tipos de excitação é chamada de não-vozeada. Devido às diferenças na produção dos sons de fala, é razoável esperar que alguns modelos sejam mais precisos para certas classes de excitação do que outros. Logo, é interessante o uso de diferentes modelos dependendo da região do discurso a ser analisado.

As amostras do sinal de voz podem ser analisadas em curtos períodos de tempo ou em longos períodos de tempo. Quando as amostras são extraídas num período curto (5-100ms), o sinal de voz é considerado um sinal estacionário; e, para um período longo (>200ms), não-estacionário. É possível perceber a característica estacionária do sinal através da Figura 2.

Na Figura 2, o primeiro gráfico (a) mostra a curva de um sinal de voz de 0 a 1,5s (período de 1,5s), onde percebe-se a característica não-estacionária do sinal de voz, ou

seja, o sinal se modifica ao longo do tempo. Já no segundo gráfico (b), verifica-se que a curva do sinal de voz de 0,63s a 0,65s (período de 20ms) se mantém estacionária neste período curto, ou seja, mantém determinada forma ao longo deste tempo. No terceiro gráfico (c), apresenta-se um espectograma que representa uma série de Transformadas de Fourier do sinal de voz ao longo do tempo. Estas transformadas são identificadas analisando-se pequenos trechos horizontais (0,25s) na direção vertical, e demonstram a característica não estacionário do sinal de voz. E o último gráfico (d) é a Transformada de Fourier de um desses pequenos trechos em escala expandida.

Figura 2 - Característica estacionária do sinal de voz



Fonte: LI e JAIN (2009, p.1266)

3.2 Processamento de Voz

Esta seção apresenta um apanhado geral na área de processamento de voz, seus principais conceitos e áreas de utilização já existentes e futuras. As áreas de estudo dentro do tópico processamento de voz são: análise/síntese de voz, reconhecimento de voz e codificação de voz. Neste trabalho será destacada a área de reconhecimento de voz, que será detalhada a seguir.

3.2.1 Reconhecimento de Voz

A fala é a forma mais natural de comunicação entre os seres humanos e é produzida como resultado de muitas diferenças ao nível semântico linguístico, articulatório e acústico. Estas diferenças, tanto características anatômicas do trato vocal como características culturais de cada indivíduo, aparecem nas propriedades do sinal de fala. O reconhecimento de locutor procura discriminar estas diferenças entre os locutores.

Reconhecimento de locutor está inserido dentro de uma área mais ampla, de reconhecimento de voz. O reconhecimento de voz está presente em diferentes aplicações, entre elas reconhecimento da fala, identificação de idioma e reconhecimento do locutor.

O reconhecimento da fala é o trabalho de reconhecer o discurso do locutor, ou seja, o texto pronunciado. Algumas utilizações comuns são: transformação do texto falado em palavras (exemplo, no celular, pronunciar a mensagem que quer enviar e o aparelho automaticamente convertê-la em texto para ser enviado via SMS), acionamentos por comando por voz (exemplo, em uso residencial, acionar as luzes, ligar ou desligar, através de um comando falado), identificação de senhas de acesso (exemplo, controle de acesso a uma sala restrita, através de uma senha falada), entre muitas outras.

Já a identificação de idioma define-se por identificar a língua falada pelo locutor, ou seja, se o indivíduo está pronunciando um texto em inglês, português, italiano, russo, ou qualquer outra língua.

Enfim, o reconhecimento do locutor, que é o escopo do trabalho, trata de dizer quem é o locutor que pronuncia o discurso, ou seja, reconhecer a pessoa pela sua voz. Os seres humanos são capazes de reconhecer humanos apenas escutando sua voz (TEJA;

CHAITRA, 2011). Cada indivíduo possui características singulares, devido, principalmente, às diferenças físicas dos órgãos produtores de voz de cada um, como o trato vocal, tamanho da laringe, entre outros. Além disso, cada ser possui traços de fala característicos de seu dialeto[1] (ritmo, entonação, pronúncia), muitas vezes influenciado pela região onde vive, e também um dialeto próprio (vocabulário, palavras que costuma pronunciar).

Esta tarefa de reconhecimento de locutor pode ser de diferentes formas, entre elas, a identificação do locutor e a verificação do locutor que, apesar de muito similares, têm diferenças importantes. A principal diferença entre identificação e verificação de locutor é que, a primeira decide se o locutor é uma pessoa específica ou se está em determinado grupo de pessoas (vide Figura 3 e Figura 4). A segunda, por sua vez, define se o locutor é alguém em especial com base em quem o mesmo afirma ser, ou em quem se suspeita que ele seja (vide Figura 5 e Figura 6), ou seja, verifica se é verdade. Como exemplo, na Figura 3 “identificamos” que a voz do locutor desconhecido pertence ao grupo especificado e na Figura 4, não pertence. Na Figura 5 “verificamos” que o locutor desconhecido que afirma ser o Carlos, realmente é o Carlos, e, na Figura 6, que ele não é o Carlos, é um impostor.

Recentemente, novas pesquisas focam também em segmentação de fala contendo vários locutores, separando os trechos de fala de cada locutor. A segmentação também pode separar trechos de música, trechos de fala, trechos de silêncio e outros.

Dentro dessas ideias, o conceito que será abordado aqui é o da Verificação de Locutor.

[1] Descreve os hábitos da fala (pronúncia, léxica, gramática, pragmática), características de uma área geográfica ou região, ou de um grupo social específico. (SWANN, 2004)

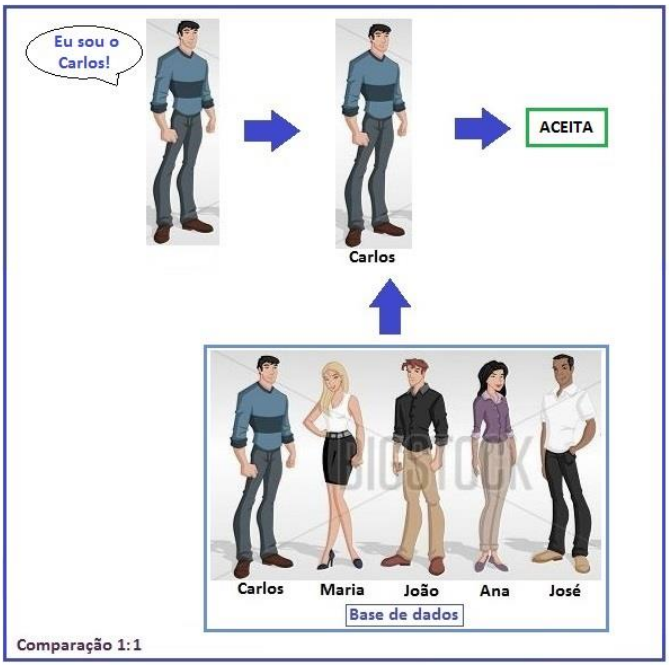
Figura 3 - Identificação de Locutor - Resultado Positivo

Fonte: Ferreira (2013).

Figura 4 - Identificação de Locutor - Resultado Negativo

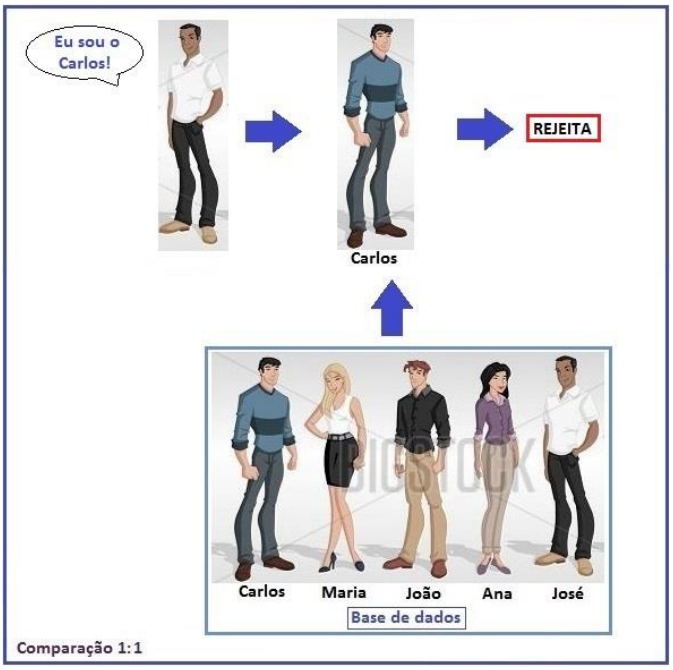
Fonte: Ferreira (2013).

Figura 5 - Verificação de Locutor - Resultado Positivo



Fonte: Ferreira (2013).

Figura 6 - Verificação de Locutor - Resultado Negativo



Fonte: Ferreira (2013).

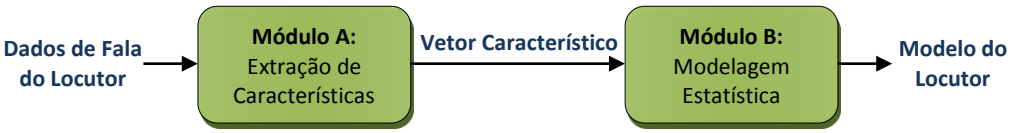
3.2.2 Verificação de Locutor

A verificação de locutor é uma tarefa binária (LIMA, 2001), onde existem apenas duas opções, aceita ou rejeita, sendo a primeira se o sistema detecta que a declaração desconhecida pertence ao locutor, e a segunda, caso contrário.

Para a tarefa de verificação de locutor encontramos duas fases distintas (BIMBOT *et al.*, 2004): a fase de treino (Figura 7) e a fase de teste (Figura 8).

Durante a fase de treino, é realizada a modelagem dos locutores, a fim de identificar os parâmetros específicos de cada locutor e armazená-los no banco de dados do sistema. Conforme Figura 7, a partir dos dados de fala de cada locutor é feita a extração de características, ou seja, a busca dos coeficientes que melhor distinguem o indivíduo, gerando assim um vetor característico $(\vec{x}_i)$. Esses coeficientes, por sua vez, servirão para o modelamento estatístico do locutor. O modelo do *background* também é determinado através do mesmo método. O *background* vem a ser a junção de falas de vários locutores, através de amostras de cada locutor, criando um modelo único e universal (TOGNERI; PULLELLA, 2011).

Figura 7 - Fase de treino do sistema de verificação de locutor



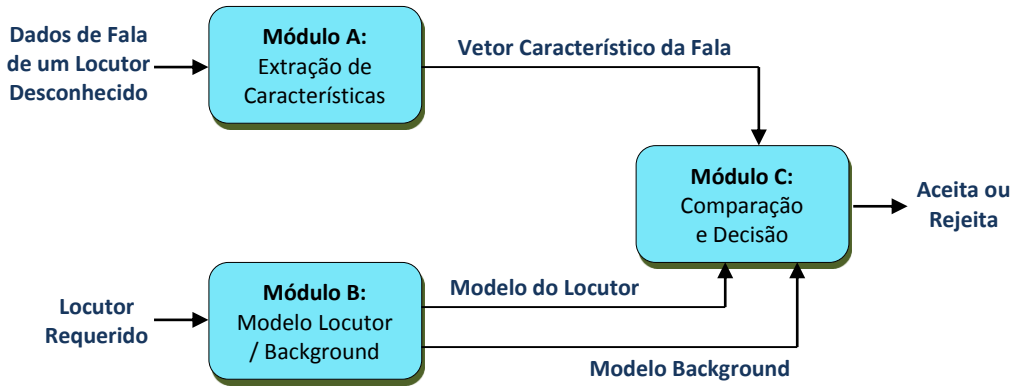
Fonte: Adaptado de BIMBOT *et al.* (2004, p.431)

Já na fase de teste, também conhecida como fase de verificação, é realizada a extração de características da fala de um locutor desconhecido. Como mostra a Figura 8, os coeficientes gerados para este indivíduo serão comparados com o modelo do locutor alvo (locutor a ser verificado) e do *background* (modelo universal), indicando se está mais próximo de serem do locutor (aceita) ou do *background* (rejeita), realizando um casamento de padrões.

Em termos práticos, cada um gera um valor de *likelihood*, que será avaliado através de um *score*, com um valor limiar determinado previamente, que irá apoiar a decisão (aceita ou rejeita) do sistema, baseado num problema de teste de hipóteses. A diferença

entre as duas medidas (*likelihood* do locutor alvo e *likelihood* do *background*) é comparada com o valor limiar. Estando acima deste limiar, o locutor alvo é aceito como verdadeiro e, estando abaixo, o locutor é rejeitado, como falso.

Figura 8 - Fase de teste do Sistema de Verificação de Locutor



Fonte: Adaptado de BIMBOT *et al.* (2004, p.431)

As técnicas de extração dos coeficientes serão detalhadas na Seção 3.3 e os modelamentos estatísticos na Seção 3.4.

A verificação de locutor é realizada de duas maneiras: uma dependente do texto que está sendo pronunciado, e a outra, independente do texto. A primeira melhora as condições do sistema através de comparações mais precisas e mais confiáveis. Porém, nem em todos os casos isto é possível, o que exige um sistema livre de dependência do texto falado, isto é, um sistema mais flexível.

Em um sistema de verificação de locutor independente de texto, não é necessária a cooperação do usuário, ou seja, o mesmo não é obrigado a falar um “determinado” texto, o que facilita em certas situações em que o locutor alvo é um suspeito que está sendo investigado e não quer ser reconhecido e, portanto, não estará de acordo com as exigências.

3.2.3 Verificação de Locutor Independente de texto

Primeiramente, é preciso conhecer os tipos de declarações usadas para autenticação e são elas: dependente de texto, independente de texto e texto solicitado. Os sistemas dependentes de texto usam o mesmo trecho de fala (mesmo pedaço de

texto) nas sessões de treino como nas de teste. Em geral, a performance do sistema é boa e as características extraídas são mais estáveis, ficando como principal inconveniente o fato de uma simples reprodução de uma amostra de voz pré-gravada do usuário ser o suficiente para burlar o sistema (quando um sistema de autenticação de usuário).

A fim de evitar tal inconveniente, existe o sistema baseado em texto solicitado (*text prompted*), onde o usuário desconhece anteriormente o texto a ser declarado e o sistema escolhe aleatoriamente a sequência de palavras a ser pronunciada pelo usuário. Então o sistema verifica primeiramente se a sequência pronunciada está correta e somente após irá realizar a verificação do locutor, o que torna o sistema robusto a ataques e mais seguro.

O sistema independente de texto, por sua vez, não exige um texto específico a ser pronunciado pelo locutor e, assim como sistema de texto solicitado, não exige senhas específicas, deixando o sistema livre para solicitar mais dados até que atinja determinado nível de confiabilidade. Porém pode ocorrer o uso de alguma gravação do usuário a fim de quebrar o sistema.

Enfim, para aplicações na área forense, os sistemas de verificação de locutor dependentes de texto acabam exigindo uma colaboração do usuário, a fim de pronunciar determinado texto. Em algumas situações isso pode ser negado pelo mesmo, tornando-se mais comum o uso de sistemas de verificação de locutor independentes de texto, facilitando assim a abordagem do usuário e tornando o processo menos rigoroso. Sistemas dependentes de texto, além de exigirem sistemas mais complexos (devido ao processo de comparação do texto), têm sua aplicação prática limitada.

Tendo em vista o exposto, será abordado no trabalho a verificação de locutor independente de texto.

3.3 Extração de Características

O módulo de extração de características é responsável por transformar o sinal de fala em um vetor característico. O sinal de fala pode ser representado por uma sequência

de vetores característicos. Nem todas as características são importantes para diferenciar os locutores e também o número de componentes do vetor deve ser relativamente baixo, para diminuir a capacidade de processamento necessário (KINNUNEN; LI, 2010). As características são utilizadas para gerar os modelos de locutores.

A extração de característica consiste em estimar variáveis de um conjunto de outras variáveis, como no caso do sinal de voz ao longo do tempo. O objetivo da seleção de características é encontrar uma transformação para um espaço de características de dimensão relativamente baixa que preserve as informações pertinentes, permitindo comparações significativas por meio de medidas simples de similaridade (CAMPBELL JR, 1997). Assim, produz-se um vetor de características a partir das variáveis do sinal de voz.

Um cuidado a ser tomado é com relação à “maldição da dimensionalidade” (JAIN; DUIN; MAO, 2000), evitando selecionar muitas características, pois este aumento implicará em maior recurso computacional, tanto em processamento, como armazenamento. A fim de evitar tal prejuízo, utilizam-se os métodos de Análise de Componentes Principais ou *Principal Components Analysis* (PCA) (SMITH, 2002) e Análise Fatorial ou *Factor Analysis* (FA) (GORSUCH, 1983) para encontrar uma representação com menor dimensão.

Existem diversos métodos para selecionar ou estimar estes parâmetros, conhecidos como seleção ou extração de características (*features*). A identidade do locutor pode ser constatada de diferentes maneiras, conforme a característica a ser extraída: através da configuração do trato vocal, prosódia, fonética ou contexto social e educação.

O *software* aberto mais utilizado para extração de características é o HTK (HTK Toolkit), assim como o UNIANAL (UNIANAL Universal Speech Analysis and Synthesis) para determinação de *pitch*, energia e detecção de atividade vocal.

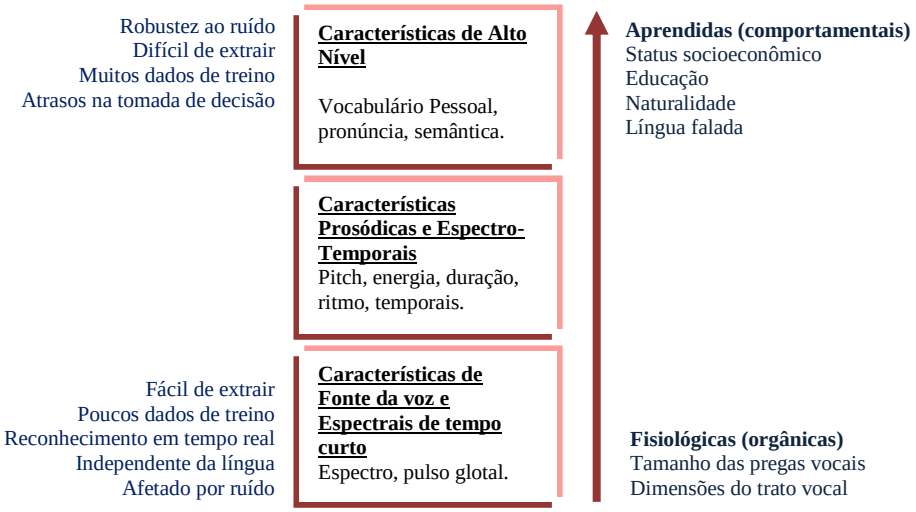
3.3.1 Tipos de Extração de Características

Existem diversos modos de categorizar características e, do ponto de vista de sua interpretação física, elas são divididas em: características espectrais de tempo curto (*short-term spectral features*), características de fonte da voz (*voice source features*), características espectro-temporais (*spectro-temporal features*), características

prosódicas (*prosodic features*) e características de alto-nível (*high-level features*), conforme

São características de alto nível (linguísticas) o uso de fonemas, de palavras e de aspectos dependentes das condições sociolinguísticas do indivíduo, e características prosódicas e espectro-temporais, parâmetros como energia instantânea, entonação, taxa de fala. Além dessas, são características de baixo nível (espectrais) a informação extraída do espectro de frequências e em janelas de tempo curto que procuram a dinâmica do trato vocal específica para cada locutor.

Figura 9 - Resumo das categorias de características e suas particularidades



Fonte: Adaptado de KINNUNEN e LI (2010, p.14)

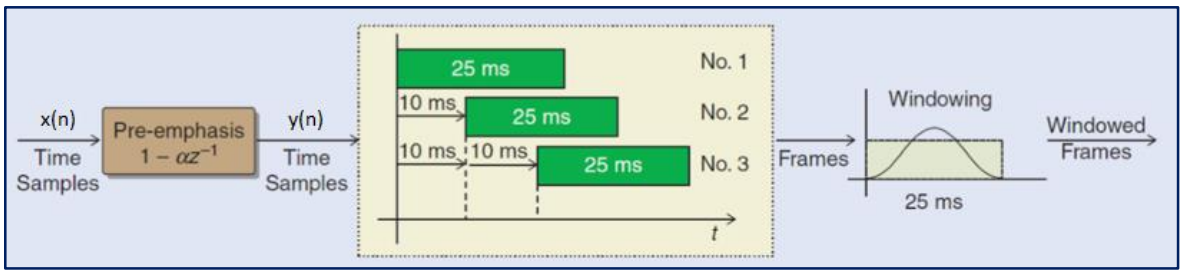
3.3.1.1 Características Espectrais de Tempo Curto

O sinal de voz se modifica continuamente conforme os movimentos articulatórios do trato vocal, caracterizando um sinal do tipo não-estacionário, isto é, um sinal cujo espectro varia ou muda com o passar do tempo. Portanto, a utilização da totalidade de uma declaração para o estudo das características espectrais de tais sinais não é recomendada, visto que tal abordagem não é capaz de capturar a dinâmica das variações espectrais. As características espectrais de tempo curto, por outro lado, são capazes de capturar tal dinâmica, levando a uma melhor descrição das propriedades de ressonância do trato vocal. Em aplicações práticas, uma maneira simples de obter boas estimativas

das características espectrais de curto tempo de uma porção do sinal amostrado é pela aplicação de janelas sobre o mesmo (vide Figura 10). Isto minimiza os efeitos oscilatórios da resposta em frequência de sinais truncados no tempo, reduzindo, dessa forma, a distorção espectral devido à *ripples* (ANDRADE; SOARES, 2000). A média de duração de cada janela ou frame do sinal varia de 20-30ms, pois, neste intervalo, assume-se que o sinal seja estacionário, e então extraem-se os coeficientes, como demonstrado na Figura 11.

Na Figura 10, as amostras do sinal de voz x(n) passam por uma pré-enfase, que será descrita em seguida, sendo sua saída o sinal y(n). O sinal y(n) é dividido em *frames* de 25ms de duração, deslocando 10ms a cada *frame*. A janela utilizada no exemplo é uma Janela de Hamming, que é multiplicada por cada *frame*, a fim de suavizar as bordas do sinal truncado.

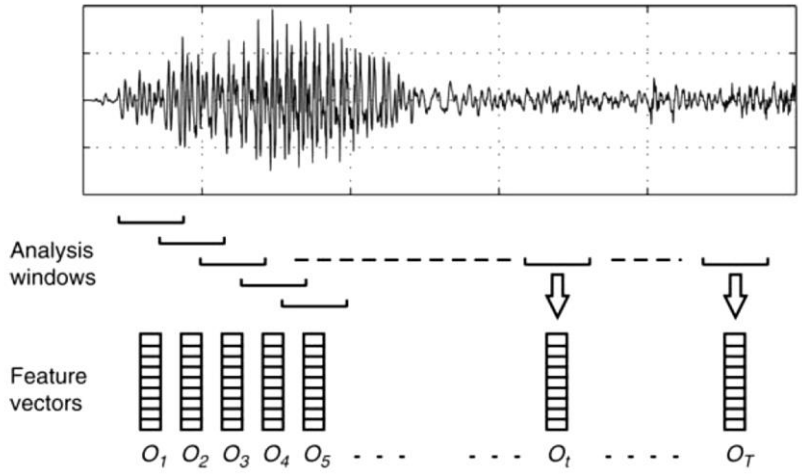
Figura 10 - Janelamento do Sinal (*frames*)



Fonte: TOGNERI e PULLELLA (2011, p.27)

Na Figura 11, observa-se no primeiro nível o sinal de voz ao longo do tempo e no segundo nível a divisão do sinal em *frames*. A cada *frame* é extraído um vetor de coeficientes, conforme o terceiro nível na figura, que mostra os vetores de O_1 até O_T.

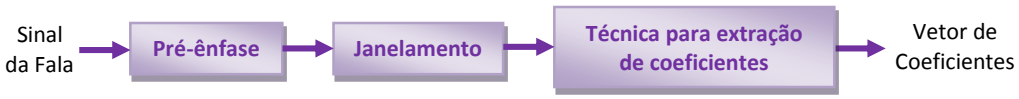
Figura 11 - Extração de *Features*



Fonte: LI e JAIN (2009, p.1255)

Basicamente, segue-se a sequência descrita na Figura 12. A pré-enfase nada mais é do que a aplicação de um filtro a fim de acentuar as altas frequências, no sentido de tê-las com pesos iguais às de baixa. Esta etapa nem sempre é realizada. A ênfase é feita nas altas frequências, a fim de compensar o processo de produção da fala humana, que tende a atenuar as altas frequências. Depois, uma janela é aplicada ao sinal, com o intuito de suavizar o efeito do uso de segmentos finitos, sendo as mais comuns Hamming e Hanning. Em seguida, define-se a técnica para extração dos coeficientes de interesse, baseando-se em modelos de produção e percepção da fala, sendo os coeficientes MFCC (*Mel Frequency Cepstral Coeficients*) os mais utilizados nas pesquisas (LI; JAIN, 2009; BHATTACHARJEE; SARMAH, 2012).

Figura 12 - Diagrama de Blocos de Aquisição dos Coeficientes



Fonte: Ferreira (2013)

A extração dos vetores de *features* continua sendo objeto de pesquisa, apesar de muitas técnicas já terem sido desenvolvidas. Além de MFCC, outros coeficientes

normalmente utilizados também são: *Linear Predictive Cepstral Coefficients* (LPCCs) (HUANG; ACERO; HON, 2001), *Line Spectral Frequencies* (LSFs) (HUANG; ACERO; HON, 2001) e *Perceptual Linear Prediction* (PLP) (HERMANSKY, 1990). Na seção 3.3.2 serão detalhados alguns dos principais métodos.

3.3.1.2 Características de Fonte da Voz

Conforme NETO *et al.* (2012), no modelo fonte-filtro, o aparelho fonador humano é separado em dois componentes distintos: um filtro linear, cuja função de transferência está relacionada às frequências de ressonância das cavidades supra-glotais do trato vocal humano (boca, faringe, fossas nasais), e uma fonte geradora que excitará esse filtro.

As características de fonte da voz descrevem as propriedades da fonte (do modelo fonte-filtro) da voz ou o fluxo glotal. Caracterizam o sinal de excitação da glote (parte da laringe que se fecha para a passagem de alimentos e se abre para a passagem do ar), algumas características vocais, tais como a forma do pulso glotal e a frequência fundamental. Logo, é possível assumir que essas características carregam informações específicas do locutor.

O cálculo dos coeficientes não é direto, devido ao efeito de filtragem do trato vocal. Assumindo que os dois são independentes um do outro, podem-se estimar os parâmetros do trato vocal através de uma das técnicas mencionadas na Seção 3.3.1.1 e aplicar a filtragem inversa para obter uma estimativa da fonte do sinal (KINNUNEN; ALKU, 2009).

Outros métodos alternativos podem ser considerados, como *closed-phase covariance anaysis* (GUDNASON; BROOKES, 2008). Conforme pesquisas, a fonte da voz não é tão discriminativa quanto os parâmetros do trato vocal, porém, a fusão das mesmas pode melhorar a precisão (ZHENG, 2007).

3.3.1.3 Características Espectro-Temporais

As características espectro-temporais descrevem as propriedades da fala dependentes de tempo, como entonação, ritmo e duração. Analisando a transição de

formantes e modulações energéticas, é possível extrair informações úteis para especificação do locutor.

Os métodos mais conhecidos são os coeficientes delta (Δ) e *double*-delta (Δ^2) (HOSSAN; MEMON; GREGORY, 2010) e eles representam as diferenças temporais entre os vetores característicos adjacentes. Normalmente são anexados aos coeficientes espectrais de tempo curto. Outros métodos também já vêm sendo estudados ao longo dos anos (MAGRIN-CHAGNOLLEAU; DUROU; BIMBOT, 2002; MALAYATH *et al.*, 2000).

3.3.1.4 Características Prosódicas

A prosódia é o estudo dos elementos da cadeia da fala que se acrescentam aos segmentos fonéticos ou fones, como o acento, a duração, o tom e a entonação (WEISS, 1988). Assim como as características espectro-temporais, as características prosódicas também descrevem as propriedades da fala como entonação, taxa de fala e ritmo, aspectos não segmentais da fala. Elas se estendem por longos segmentos, como sílabas, palavras e declarações, refletindo diferenças no estilo de fala e emoções.

As características prosódicas são baseadas no *pitch* (F0), intensidade (energia) e duração, modelando diferentes níveis de informação prosódica para captura de dados distintos para cada locutor. O parâmetro mais importante é a frequência fundamental (F0) que, em conjunto com as características espectrais de tempo curto, demonstra ser bastante efetivo quanto ao ruído. Um dos *softwares* utilizados para cálculo da F0 é o PRAAT (LIESHOUT, 2003). Um método para modelar uma sequência de símbolos prosódicos é o Bi-gram ou N-gram (DRGAS; CETNAROWICZ; DABROWSKI, 2008).

3.3.1.5 Características de Alto-Nível

As características de alto nível capturam informações ao nível de conversação dos locutores, como palavras usadas repetidamente e a tendência de frases e palavras a serem declaradas pelos locutores durante uma conversa. A ideia é analisar, ao longo da declaração, a reincidência de um determinado padrão, podendo determinar diferenças entre locutores.

3.3.2 Métodos de Extração de Características

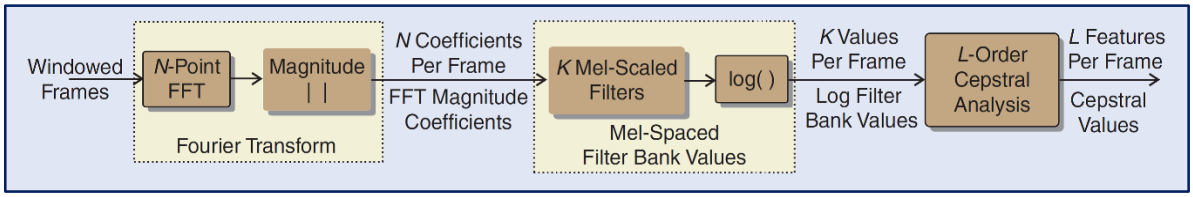
A seguir são detalhados alguns dos principais métodos usados para extração de características, entre eles MFCC (*Mel Frequency Cepstral Coeficients*), coeficientes delta(Δ) e *double*-delta(Δ^2) e LPC (*Linear Predictive Coding*).

3.3.2.1 MFCC

Os coeficientes MFCC são os mais utilizados em verificação de locutor. O diagrama de blocos completo para determinação dos coeficientes MFCC pode ser analisado na Figura 13, extraída de TOGNERI e PULLELLA (2011).

Como visto na Figura 13, uma transformada de Fourier é aplicada a cada *frame* do sinal de voz, sendo ignorada a informação de fase e permanecendo somente o espectro de magnitude. Em seguida, os coeficientes são convertidos para escala MEL (STEVENS; VOLKMANN; NEWMAN, 1937; HARTMANN, 1997), através de um banco de filtros e é aplicado o logaritmo. Por fim, são gerados os coeficientes cepstrais por meio de uma transformação cepstral em cada saída do filtro, através da Transformada Discreta do Cosseno, e então se obtêm os coeficientes MFCC.

Figura 13 - Diagrama de Blocos para determinação dos Coeficientes MFCC



Fonte: TOGNERI e PULLELLA (2011, p.28)

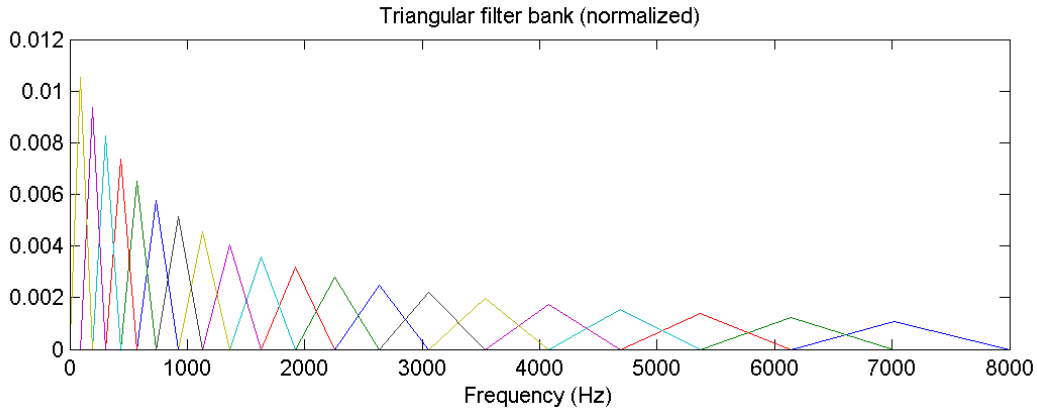
É comum converter a amplitude do sinal sonoro (variações de pressão de ar) para a escala logarítmica de decibéis por duas razões: uma porque a escala logarítmica converte valores muito grandes (da diferença na variação da pressão do som entre o som mais silencioso e o som mais alto) para outros mais manejáveis, e também porque o

julgamento humano de variação de intensidade sonora corresponde mais a uma escala logarítmica do que a uma escala linear (HARRINGTON; CASSIDY, 1999).

Converte-se os coeficientes espectrais do vetor característico para a escala MEL (similar à escala de frequência do ouvido humano), cujas frequências são definidas através de análises psico-acústicas, onde a faixa de frequência mais baixa é normalmente representada com maior resolução. Para obter estes novos coeficientes, aplica-se um banco de filtros com diferentes larguras de banda e frequências centrais determinadas pela conversão das frequências para a escala MEL (vide Equação (3.1)), alocando mais filtros com larguras de banda estreita nas frequências mais baixas (vide Figura 14).

$$f_{MEL} = 1000 \cdot \frac{\log(1 + f_{LIN}/1000)}{\log 2} \qquad (3.1)$$

Figura 14 - Banco de Filtros na Escala MEL



Fonte: http://mirlab.org

Na equação (3.1), f_{MEL} representa a frequência na escala MEL e f_{LIN}, a frequência linear. A aplicação do banco de filtros se faz necessária a fim de obter um vetor cujas componentes correspondem as saídas dos filtros.

As *features* espectrais são altamente correlacionadas, enquanto as *features* cepstrais produzem uma representação compacta mais descorrelacionada (TOGNERI; PULLELLA, 2011). É comum a extração de apenas 12 coeficientes MFCC por *frame* (TOGNERI; PULLELLA, 2011). Aumentar o número de coeficientes extraídos não traz mudanças significativas para o sistema.

Os vetores característicos são gerados a cada 10ms porque se assume que neste período de tempo o sinal de voz é estacionário. Para cada *frame* é importante descartar todo o silêncio e manter as amostras de fala. Uma das característicos extraídas é energia, que corresponde a intensidade da voz do locutor. A fim de evitar maus resultados devido ao uso desta informação, a componente energia é removida (MODI; SAUL, 2006).

3.3.2.2 DELTA E DELTA/DELTA

A análise independente dos coeficientes cepstrais pode perder informações importantes, como a coarticulação. Alguns processamentos são capazes de capturar e modelar a informação temporal (dinâmica) entre os *frames*, concatenando *features* de aproximação das derivadas de primeira, segunda e terceira ordem dos coeficientes MFCC. Essas informações trazem evidências da natureza da fala e características de estilo e duração da fala (TOGNERI; PULLELLA, 2011).

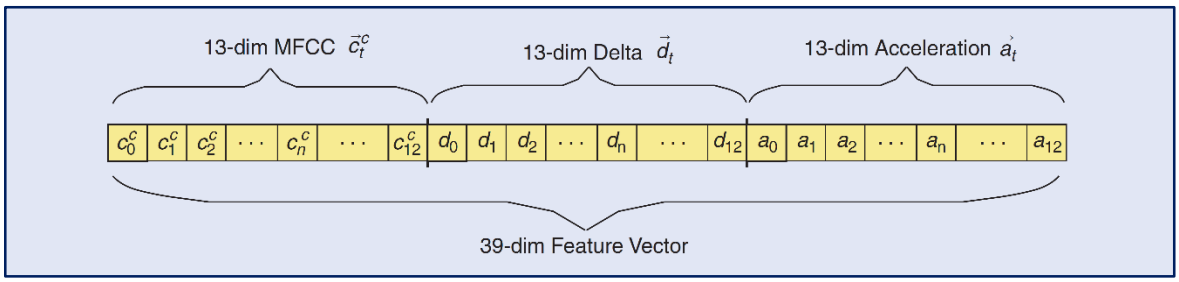
A primeira derivada, ou delta (Δ), pode ser aproximada pela Equação (3.2), onde tipicamente P=2 e $\vec{\mathrm{x}}_{\mathrm{t}}$ é o vetor em cada *frame* t (TOGNERI; PULLELLA, 2011, p. 29):

$$\overrightarrow{d_t} = \frac{\sum_{p=1}^{P} p(\vec{\mathrm{x}}_{\mathrm{t+p}} - \vec{\mathrm{x}}_{\mathrm{t-p}})}{2\sum_{p=1}^{P} p^2} \qquad (3.2)$$

Substituindo-se o valor de $\vec{\mathrm{x}}_{\mathrm{t}}$ por $\overrightarrow{d_t}$, obtém-se a segunda derivada, (*double*-delta) (Δ²) ou parâmetros de aceleração. Esses resultados são concatenados com o vetor original MFCC, resultando num aumento do vetor característico. O exemplo da Figura 15 mostra um vetor característico de 13 dimensões (12 coeficientes + C_0) concatenado com

13 coeficientes delta e mais 13 coeficientes *double*-delta. O coeficiente C_0 é a componente energia mencionada anteriormente.

Figura 15 - Concatenação MFCC, Delta e Delta-Delta



Fonte: TOGNERI e PULLELLA (2011, p. 29)

Concluindo, utilizam-se os coeficientes delta e *double*-delta, acrescentando-se assim a velocidade e a aceleração da variação temporal dos vetores cepstrais.

3.3.2.3 LPC

O método LPC se baseia no modelo de um preditor linear. O preditor linear modela um sinal $s[n]$ através da combinação linear dos valores passados do sinal e uma entrada atual conforme a Equação (3.3):

$$s[n] = -\sum_{k=1}^{p} a[k]s[n-k] + G.u[n] \qquad (3.3)$$

onde $s[n]$ é o valor de saída atual, p é a ordem do preditor, $a[k]$ são os coeficientes do preditor, $s[n-k]$ são as saídas passadas, G é um fator de ganho e $u[n]$ é a entrada atual. Em aplicações de fala, a entrada $u[n]$ é normalmente desconhecida, sendo ignorada, e o preditor linear acaba dependendo apenas das amostras passadas, conforme a Equação (3.4), sendo $s^*[n]$ o sinal predito:

$$s^*[n] = -\sum_{k=0}^{p} a[k]s[n-k] \quad (3.4)$$

A análise LPC assume que as amostras de fala podem ser aproximadas por uma soma ponderada linearmente de um determinado número de amostras passadas.

A predição do sinal de voz é $s^*[n]$ e $a[k]$ são os coeficientes LPC que minimizam o erro, erro este representado pelo somatório ao quadrado da diferença entre o valor real e o predito.

Definindo-se o erro do preditor como $e[n]$ (também conhecido como resíduo), sendo a diferença entre o valor atual e o valor predito, conforme equação, ele seria idêntico à entrada de sinal $G.u[n]$. Sendo E o erro mínimo quadrático, temos:

$$e[n] = s[n] - s^*[n] \quad (3.5)$$

$$E = \sum_{n} e[n]^2 \quad (3.6)$$

Assim, separamos a fonte (entrada da glote) do filtro (trato vocal) , sendo que a fonte não é modelada pelos coeficientes do preditor linear. É claro que existem informações dependentes do locutor na excitação (como a frequência fundamental), e, ao ser ignorada, informações discriminantes seriam perdidas.

Segundo LI e JAIN (2009), os coeficientes LPC são altamente correlacionados entre si, o que não representa uma característica desejável, sendo, portanto, necessária uma transformação (*cepstrum transform),* gerando os coeficientes LPCC *(Linear Prediction Cepstral Coefficients)*.

3.3.2.4 Contorno do Pitch

Uma técnica para extração de características prosódicas da voz é através do contorno do *pitch*. Essa técnica busca modelar as diferenças de entonação da voz através de um polinômio cúbico que estima o contorno do *pitch* (LI *et al.*, 2012).

3.3.2.5 Pitch

Conforme PADIARAJ *et al*. (2011), a frequência de vibração das cordas vocais é uma *feature* importante para distinguir a voz. Uma característica importante do *pitch* é sua robustez quanto a ruído e distorções do canal. O uso exclusivo do *pitch* para reconhecimento de locutor tem bons resultados somente quando o número de locutores é pequeno. Quando aumenta o número de locutores a performance reduz drasticamente. A informação de *pitch* é importante para distinção de gênero.

O *pitch* é estimado através de uma pequena porção do sinal de voz, determinando-se a frequência dominante. Para tal, é necessário descobrir o menor intervalo periódico do sinal de voz. Uma técnica para determinar o *pitch* é o PDA (*Pitch Determination Algorithm*) (SUN, 2002) baseado na taxa de sub-harmônica-para-harmônica. O valor do *pitch* é zero nas partes do sinal de voz não-vozeadas (MARKOV; NAKAGAWA, 1999).

3.3.3 Silêncio

Os dados de silêncio não contêm informações específicas do locutor e, além disso, prejudicam o treinamento do GMM (Modelo de Mistura Gaussiana, que será detalhado na seção 3.4.2), na proporção da quantidade de silêncio versus a quantidade de fala (TOGNERI; PULLELLA, 2011). Um passo fundamental para o desenvolvimento de um sistema de reconhecimento de locutor é a separação das porções da fala que sejam de silêncio e não-vozeadas. Isto porque a maior parte dos atributos específicos do locutor estão presentes na parte vozeada do sinal de voz (PADIARAJ *et al.*, 2011).

Uma técnica para retirada de silêncio dos áudios dos locutores, ficando apenas com porções de fala do sinal, é aplicar um modelamento bi-Gaussiano da componente energia, detectando atividades de voz. A gaussiana com menor média corresponde ao silêncio e a gaussiana com maior média corresponde a porções de fala. Então, os vetores cepstrais são normalizados, a fim de terem média zero e variância unitária. Finalmente, os coeficientes energia são descartados do vetor e os *frames* correspondentes ao silêncio são deletados (DIKICI; SARAÇLAR, 2009).

Outro método bastante utilizado é remover o silêncio após a extração de características. Isto é possível devido ao fato de que os segmentos de silêncio resultam

num vetor de características cujos coeficientes MFCC são todos iguais a zero. Então todos os vetores que são compostos apenas por coeficientes iguais a zero são removidos do conjunto (KOMLEN *et al.*, 2011).

3.4 Sistemas de Classificação

Em suma, para efeitos de classificação, se fazem necessários alguns passos: a determinação de um “modelo do locutor”, a comparação do áudio desconhecido com este “modelo do locutor”, gerando assim um *score*, mais conhecido como *likelihood ratio* (razão de verossimilhança), e um processo de decisão, através do uso deste resultado, podendo este ser comparado com um “modelo de potenciais impostores”.

O problema da classificação pode ser destacado de duas maneiras: (i) os modelos gerativos (como GMMs, *Gaussian Mixture Models*) que exigem apenas amostras de dados de treino dos locutores alvo e constroem um modelo estatístico que descreve a distribuição dos locutores alvo; (ii) os classificadores discriminativos, que requerem dados de treino tanto para locutores alvo como para impostores e obtêm uma ótima separação entre os diferentes locutores, sendo o mais popular as SVMs (*Suport Vector Machines*) (KINNUNEN e LI, 2010).

O modelo GMM adaptado tem sido o modelo de aproximação dominante em verificação de locutor independente de texto (REYNOLDS; QUATIERI; DUNN, 2000; BHATTACHARJEE; SARMAH, 2012; BIMBOT *et al.*, 2004). Segundo a literatura, as SVMs atingem performances compatíveis ou até superiores que os GMMs com uma quantidade de dados de treino muito menor.

O *software* aberto mais utilizado para modelamento GMM/UBM (Modelo de Mistura Gaussiana com o uso de um Modelo Universal) é o software BECARS (BLOUET *et al.*, 2004) e para implementar a SVM, a biblioteca LIBSVM (CHANG; LIN, 2014).

3.4.1 Quantização Vetorial

Quantização Vetorial ou *Vector Quantization* (VQ) é usada para compressão de informações de modo a obter uma redução no número de vetores de *features* (armazenados dentro de um *codebook*) sem que as características importantes da distribuição (função densidade de probabilidade) dos mesmos se perca. VQ é uma técnica de quantização clássica em processamento de sinais. (CHAUHAN; SONI; ZAFAR, 2013).

Conhecido como modelo centróide, VQ é um dos mais simples modelos de locutor independentes de texto, com técnicas computacionais de alta velocidade (KINNUNEN; LI, 2010). A terminologia 'centroide' é devida ao fato de que, após o treinamento, os vetores presentes no *codebook* (as *codewords*) representam as áreas do espaço de *features* com maior concentração de amostras.

No treino, um *codebook* é estabelecido para cada um dos N locutores, resultando em N *codebooks*. Cada *codebook* é gerado com dados de treino (*features*) de apenas um locutor. Portanto, os *codebooks* não são sobrepostos uns aos outros no espaço de *features*. O processo de treino gira em torno da redução da distância mínima média entre um dos vetores de *features* e todos os vetores do *codebook*.

Na fase de reconhecimento, um grupo de vetores da fala a ser reconhecido é utilizado para cálculo da distância mínima média em relação a cada um dos N *codebooks* presentes no sistema. A fala é então associada com o locutor de menor distância mínima média no espaço de *features*. (YUJIN; PEIHUA; QUN, 2010).

3.4.2 GMM-UBM

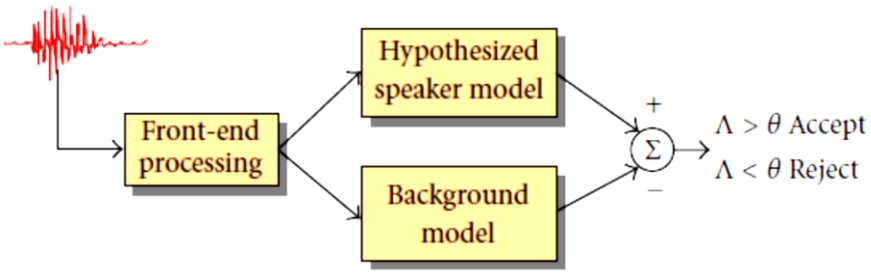
Basicamente um método para modelamento do locutor segue um Teste de Hipóteses Bayesiano (*Bayesian Hypothesis Test*) com as seguintes premissas: dado um segmento de fala Y de um locutor desconhecido e um locutor alvo S, H0 representa a hipótese de a fala Y ser do locutor S e H1 representa a hipótese de não ser. O teste “ótimo” para decidir entre estas duas hipóteses é o *likelihood ratio* (LR) dado por

$$\mathrm{LR} = \frac{\mathrm{p(Y|H0)}}{\mathrm{p(Y|H1)}} \begin{cases} > \theta, \text{aceita H0} \\ < \theta, \text{aceita H1} \end{cases} \qquad (3.7)$$

onde $\mathrm{p(Y|H0)}$ é a função densidade de probabilidade dada a hipótese H0 e $\mathrm{p(Y|H1)}$, dada a hipótese H1. O limiar de decisão para aceitar ou rejeitar H0 é θ.

Como visto na Figura 16, a fala do locutor desconhecido passa por um processamento e sua saída é uma sequência de vetores $\mathrm{X} = \{\vec{\mathrm{x}}_1, \ldots, \vec{\mathrm{x}}_\mathrm{T}\}$, onde cada vetor característico é uma amostra no tempo discreto do sinal. Esses vetores são utilizados para calcular as taxas de verossimilhança de H0 e H1, através do Modelo de Misturas Gaussianas, gerando o modelo do locutor alvo e do *background.* Os valores logarítmicos de taxa de verossimilhança encontrados são subtraídos um do outro e o resultado encontrado (Λ) é comparado com o limiar (θ). Se este for maior que o limiar, o sistema aceita que a fala é do locutor alvo e caso contrário, rejeita esta hipótese.

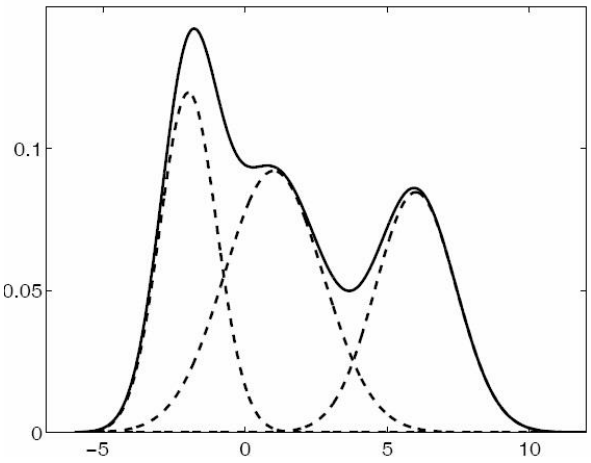
Figura 16 - Sistema de verificação de locutor baseado na taxa de verossimilhança



Fonte: BIMBOT *et al.* (2004, p. 434)

O GMM, representado por λ, nada mais é que o uso de uma mistura finita de distribuições gaussianas para aproximação (modelamento) da função densidade de probabilidade de interesse. O objetivo é modelar o locutor por meio de um modelo de distribuição estatística das features do locutor, através de uma mistura de gaussianas, conforme a Figura 17, onde a função é modelada por 3 gaussianas.

Figura 17 - GMM com três componentes



Fonte: SINITH *et al.* (2010, p. 294)

Conforme BIMBOT *et al.* (2004), para um vetor de características D-dimensional, a densidade da mistura para posterior obtenção de uma função de verossimilhança é definida como segue:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} w_i \cdot p_i(\vec{x}) \qquad (3.8)$$

A densidade é, portanto, uma combinação linear ponderada de M densidades gaussianas unimodais $p_i(\vec{x})$ cada uma parametrizada por um vetor média $\vec{\mu}_i$ $(Dx1)$ e uma matriz covariância Σ_i (DxD):

$$p_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} e^{-(1/2)(\vec{x}-\vec{\mu}_i)'(\Sigma_i)^{-1}(\vec{x}-\vec{\mu}_i)} \qquad (3.9)$$

A soma dos pesos da mistura, w_i, deve satisfazer $\sum_{i=1}^{M} w_i = 1$. Coletivamente, os parâmetros do modelo de densidade são simbolizados por $\lambda = (w_i, \vec{\mu}_i, \Sigma_i), i = (1, \dots, M)$. Normalmente, utilizam-se apenas matrizes de covariância diagonais, principalmente por serem mais eficientes computacionalmente.

Para estimar os parâmetros do GMM, usa-se a técnica de *maximum likelihood* (maximização da verossimilhança, ML) através do algoritmo iterativo *Expectation*

Maximization (EM) (BISHOP, 2006). Este algoritmo aperfeiçoa iterativamente os parâmetros do GMM a fim de aumentar a verossimilhança (*likelihood*) do modelo estimado a partir dos vetores de características observados. Geralmente são necessárias de 5 a 10 iterações para o algoritmo convergir (REYNOLDS; QUATIERI; DUNN, 2000; REYNOLDS, 1995).

Assumindo a independência dos vetores característicos, para a sequência de vetores $\mathrm{X} = \{\vec{\mathrm{x}}_1, \ldots, \vec{\mathrm{x}}_\mathrm{T}\}$, obtém-se a distribuição conjunta a partir do *likelihood* do modelo λ, dado por:

$$\mathrm{l}(\lambda) = \prod_{1}^{\mathrm{T}} \mathrm{p}(\vec{\mathrm{x}}_\mathrm{t}|\lambda) \qquad (3.10)$$

Normalizando pelo número total de vetores T e aplicando o logaritmo, chega-se ao *log-likelihood* do modelo λ a seguir:

$$L(\lambda) = \log p(X \mid \lambda) = \frac{1}{T}\sum_{1}^{t} \log p(\vec{x}_t|\lambda) \qquad (3.11)$$

Enquanto o modelo de H0 é bem definido e pode ser estimado usando trechos de fala do locutor alvo S, o modelo para H1 pode ser aproximado através de um conjunto de outros modelos de locutores a fim de contemplar o espaço da hipótese alternativa (REYNOLDS; QUATIERI; DUNN, 2000). Dado um conjunto de N modelos de locutores para formar este modelo único (*background)*, representando H1, o seu modelo pode ser representado por:

$$\mathrm{p}(\,\mathrm{X} \mid \lambda_{ubm}\,) = \mathcal{F}(\mathrm{p}(\mathrm{X} \mid \lambda_1), \mathrm{p}(\mathrm{X} \mid \lambda_2), \ldots \mathrm{p}(\mathrm{X} \mid \lambda_N)) \qquad (3.12)$$

onde $\mathcal{F}(\,)$ é alguma função como média ou máximo, dos valores de *likelihood* do conjunto de locutores do *background*.

A *score function* do sistema determina o valor a ser comparado com o limiar de decisão (θ) para determinar se a declaração pertence ao locutor alvo. Abaixo, a equação da *score function* (Λ). Sendo $\Lambda > \theta$, o sistema determina que a fala seja do locutor alvo (aceita), e, sendo $\Lambda < \theta$, o sistema determinada que não seja do locutor alvo (rejeita).

$$\Lambda = \log p(X|\lambda_{alvo}) - \log p(X|\lambda_{ubm}) \qquad (3.13)$$

Na Equação (3.13) acima, λ_{alvo} representa o modelo do locutor alvo e λ_{ubm} o modelo universal, denominado *background* (UBM – *Universal Background Model*), um modelo único e universal de supostos impostores.

Uma técnica comumente utilizada é gerar o modelo do locutor a partir do modelo universal, adaptando os parâmetros do UBM através de adaptação Bayesiana. Esta técnica é chamada de MAP (*Maximum a Posteriori*) e normalmente são adaptadas apenas as médias, permanecendo iguais os outros parâmetros. O procedimento é dado pela seguinte fórmula (REYNOLDS; QUATIERI; DUNN, 2000):

$$\mu_{map}^{i} = \frac{n_i}{n_i + r} \cdot \mu_{emp}^{i} + \left(1 - \frac{n_i}{n_i + r}\right) \cdot \mu_{ubm}^{i} \qquad (3.14)$$

onde μ_{map}^{i} é a média adaptada para a componente gaussiana i, μ_{emp}^{i}é a média empírica correspondente (obtida usando os dados de registro do locutor e o algoritmo EM), μ_{ubm}^{i} é a média do UBM, n_i é o taxa de ocupação da componente (obtida também com a ajuda do algoritmo EM, usando o UBM e os dados do locutor) e r é o fator de regulação.

TOGNERI e PULLELLA (2011) cita alguns motivos pelos quais o GMM faz tanto sucesso em reconhecimento de locutor. O GMM usa todos os dados de fala disponíveis de um único locutor e busca modelar todas as possíveis variações acústicas de fala do mesmo, independente do que esteja sendo dito. Apesar de ser uma tarefa difícil, com um número de misturas suficiente (da ordem de 64 ou mais), as densidades componentes podem conseguir representar a ampla distribuição fonética específica do locutor, desde que o número de fonemas da língua em questão seja menor que o número de misturas. Outra vantagem é o seu poderoso e versátil algoritmo para estimação dos parâmetros: o *Expectation Maximization* (EM) ou Maximização do Valor Esperado). O algoritmo EM garante uma convergência monotônica para o conjunto de parâmetros ótimos (com máxima verossimilhança) em apenas 5 ou mais iterações (TOGNERI; PULLELLA, 2011).

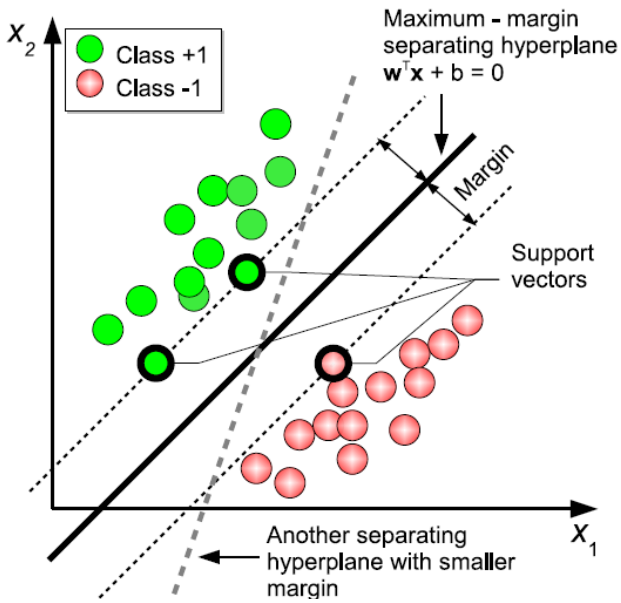
Assim também cita dois inconvenientes. O primeiro é que necessita de uma quantidade de dados de treino grande para estimar os parâmetros do modelo. Uma sugestão é reduzir o número de parâmetros a serem estimados, por exemplo, usar a matriz de covariância na forma diagonal ao invés da completa. Isto é aceitável devido aos *features* MFCC serem naturalmente descorrelacionados (valores de correlação baixos para os elementos da matriz não-diagonais). Além de reduzir os recursos computacionais, está comprovado que a performance permanece a mesma com o uso dessa técnica (TOGNERI; PULLELLA, 2011). O segundo é que tipos de dados não vistos na fase de treino podem aparecer durante a fase de teste, gerando baixas medidas de verossimilhança e degradando a performance do sistema. A solução óbvia seria aumentar e variar os dados de treino. Porém, na prática, isto pode ser inviável.

3.4.3 SVM

Segundo KINNUNEN e LI (2010), *Suport Vector Machine* (SVM) é um classificador discriminante muito potente, que tem sido adotado recentemente em reconhecimento de locutor. Atualmente, SVM é um dos mais robustos classificadores para verificação de locutor e tem muito sucesso combinado ao GMM com o intuito de aumentar a precisão (CAMPBELL *et al.*, 2006).

Como explicitado na Figura 18, o SVM é um classificador binário, que modela o limite de decisão entre duas classes como um hiperplano de separação. Para verificação de locutor, uma classe são os vetores treinados do locutor alvo (classificados como +1) e a outra classe são os vetores treinados do *background* (classificados como –1). Com o objetivo de otimizar o sistema, o SVM encontra, durante a fase de treino, um hiperplano de separação que maximiza a margem de separação entre essas duas classes.

Figura 18 - Princípio de Funcionamento do SVM



Fonte: KINNUNEN e LI (2010, p. 22)

O hiperplano ótimo é escolhido através do critério de margem máxima, ou seja, de tal maneira que maximize a distância Euclidiana entre os pontos de dados mais próximos em cada lado do plano (TOGNERI; PULLELLA, 2011). Os dados mais próximos são conhecidos como *support vectors*.

A performance do SVM depende da função kernel escolhida (LIU *et al.*, 2006). Algumas considerações para uso do SVM podem ser feitas. Os locutores não são linearmente separáveis e o SVM básico deve ser aumentado através do uso de *slack variables* e uma função kernel que projeta os dados separados de forma não-linear em linearmente separáveis com dimensão maior (TOGNERI; PULLELLA, 2011). Como transformar uma sequência de vetores característicos em um único vetor de dados adequado para ser classificado por um SVM? Algumas soluções incluem o uso de classificadores polinomiais, funções kernel e supervetores GMM.

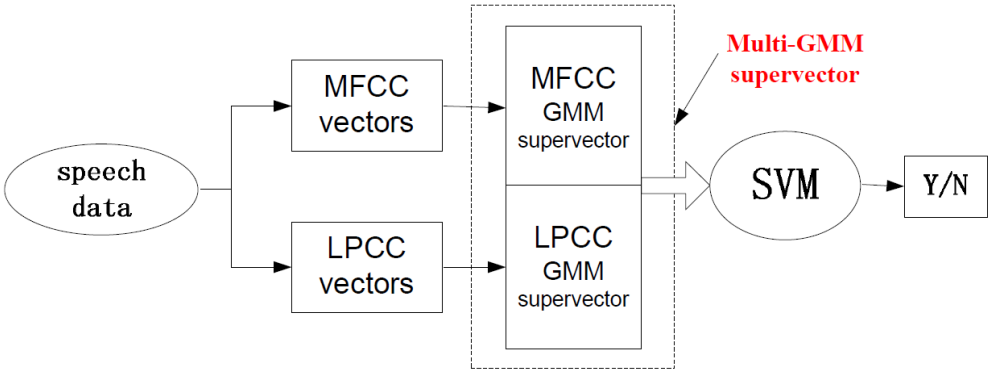
3.4.4 GMM-SVM

A fim de obter uma declaração inteira em um único vetor característico utiliza-se a técnica mais popular, conhecida como GMM *supervector*. Este “supervetor” é construído

através do empilhamento das componentes médias da mistura do GMM, concatenando uma após a outra em um único vetor de alta dimensão. Para um modelo GMM de M misturas e vetor D dimensional, o supervetor GMM correspondente terá dimensão MDx1 (DIKICI; SARAÇLAR, 2009).

A Figura 19 exemplifica o uso do modelo GMM-SVM. Os dados de fala passam pelo módulo de extração de características e geram dois vetores característicos distintos, um vetor com coeficientes MFCC e outro com coeficientes LPCC. Da mesma forma, no GMM, são gerados dois modelos, o GMM-MFCC, usando o vetor característico MFCC e o GMM-LPCC, usando o vetor característico LPCC. De ambos os modelos (GMM-MFCC e GMM-LPCC) permanece apenas o vetor de médias, e então são todas elas concatenadas gerando apenas um vetor, que passa a ser chamado de supervetor. Esse supervetor é que será utilizado para classificação no SVM.

Figura 19 - Sistema GMM Supervector / SVM



Fonte: LIU e HUANG (2009, p. 3)

3.5 *Background*

Num sistema de verificação de locutor, o resultado é obtido por meio de um *score* do áudio do locutor desconhecido contra o modelo do locutor alvo e um *score* do áudio do locutor desconhecido contra um modelo de impostor. Este modelo de impostor é mais conhecido como *Universal Background Model* (UBM). Conceitualmente, o UBM representa a distribuição das *features* independente de locutor através de todos os dados

dos locutores. Este modelo é usado para representar os espaços acústico, fonético e linguístico.

Segundo TOGNERI e PULLELLA (2011), este modelo de impostor é formado por todos os locutores exceto o locutor alvo. Na prática, este modelo é treinado com todos os dados de locutores (incluindo os do locutor alvo), assumindo-se que os dados específicos do locutor alvo serão atenuados pela presença de outros locutores. Para tanto, é necessário um mínimo de locutores para gerar o UBM. A vantagem é que se pode usar o mesmo UBM para qualquer tarefa de verificação de locutor.

O ideal é aumentar gradativamente a quantidade de dados de treino do UBM, a fim de estimá-lo com maior segurança. Segundo pesquisas, em torno de uma hora de fala do total dos locutores do *background* é o suficiente (REYNOLDS; QUATIERI; DUNN, 2000). A seleção, tamanho e combinação do conjunto de locutores tem sido objeto de muitas pesquisas. Basicamente, se a tarefa é verificar um locutor numa conversa telefônica, usa-se um *background* de locutores em ligações telefônicas, se o gênero (homem/mulher) do locutor é previamente conhecido, usa-se um *background* de locutores homens ou mulheres, e assim para outros casos.

O número de gaussianas indicado é da ordem de 256 para cima. Segundo VARCHOL, LEVICKY e JUHAR (2008), o melhor modelamento do UBM é feito com 1.024 gaussianas (UBM com total de 60min de fala dos locutores).

Para o modelamento GMM é necessária uma quantidade mínima de dados de treino a fim de gerar o modelo do locutor. Quanto maior a quantidade de dados utilizados no treino e teste, menor a taxa de erro. Uma alternativa é treinar um UBM e então gerar o modelo do locutor adaptando este UBM, através do algoritmo MAP, com os dados do locutor. Assim, para eficiência do modelo não é necessária uma grande quantidade de dados do locutor. O algoritmo MAP foi detalhado na seção 3.4.2.

No UBM, os dados são estimados com segurança e com uma quantidade de dados suficiente. Sendo o GMM do locutor individual treinado a partir do UBM, consegue-se igual segurança mesmo com pouca quantidade de dados, muito maior do que se o GMM fosse treinado diretamente com os poucos dados do locutor. Devido à grande quantidade dos dados de treino do UBM, o número de misturas para treiná-lo é maior que o número

necessário para treinar o locutor individualmente. Fazendo-se a adaptação do locutor através do UBM, tem-se o mesmo número de misturas tanto no modelo do locutor como no *background* (TOGNERI; PULLELLA, 2011).

O modelo do *background* é primordial para um bom desempenho do sistema. Ele atua como uma normalização para ajudar a minimizar a variabilidade devida a informações que não são dependentes do locutor na decisão, como ambiente, microfone, ruído (REYNOLDS, 2002).

3.6 Normalização e Fusão

Nesta seção será detalhada uma técnica de normalização conhecida como CMN (*Cepstral Mean Normalization*) e também será explicado como funciona um método conhecido como Fusão.

3.6.1 CMN

Assumindo que os efeitos de canal e ambiente são invariantes na duração do trecho de fala e também constantes em todas as declarações, tanto de treino como de teste, é possível utilizar um processo chamado *Cepstral Mean Normalization* (CMN), a fim de realizar compensações com relação às *features* MFCC (TOGNERI; PULLELLA, 2011).

O processo dá-se através da subtração do vetor média MFCC de cada vetor MFCC individual. Primeiro calcula-se a média em função de todos os vetores MFCC:

$$\overrightarrow{\mu_T} = \frac{1}{T} \cdot \sum_{t=1}^{T} X \qquad (3.15)$$

onde $\mathrm{X} = \{\vec{\mathrm{x}}_1, \ldots, \vec{\mathrm{x}}_\mathrm{T}\}$, sendo $\vec{\mathrm{x}}_\mathrm{t}$ cada *frame* da declaração. Em seguida, subtrai-se este vetor de média em cada frame:

$$\vec{\mathrm{x}}_\mathrm{t} = \vec{\mathrm{x}}_\mathrm{t} - \overrightarrow{\mu_T} \qquad (3.16)$$

A ideia é remover qualquer efeito de canal invariante no tempo e permanecerem apenas as variações dinâmicas importantes que caracterizam o locutor.

3.6.2 Fusão

A combinação de diferentes *features* pode trazer melhores performances para o sistema de verificação de locutor independente de texto (LIU; HUANG, 2009)

Diferentes *features* extraídos da mesma declaração traduzem características diferentes. Como a taxa de *frames* de diferentes *features* pode não ser a mesma e algumas *features* também podem ser perdidas algumas vezes (como o *pitch,* que não existe em sons não-vozeados), surge uma dificuldade em combiná-los diretamente, a nível de *features*. Além disso, a maldição da dimensionalidade (já mencionada na seção 3.3) é outro fator. Para tanto, muitos trabalhos realizam esta fusão a nível de *score* (LIU; HUANG, 2009).

3.7 Medidas de desempenho do sistema

O principal órgão avaliador de desempenho de sistemas de reconhecimento de locutor é o NIST (*National Institute of Standards and Technology*) (U.S. DEPARTMENT OF COMMERCE, 2010), agência do Departamento de Comércio dos Estados Unidos, patrocinada por agências de defesa do governo.

O descasamento entre os trechos de fala da fase de treino e os trechos de fala da fase de teste são os principais problemas encontrados em reconhecimento de locutor. Alguns fatores como ambiente de gravação (estúdio, pessoas, carro, TV, ...), o microfone utilizado, o canal de transmissão, condições emocionais (stress ou coação), o contexto fonético e linguístico e aspectos patológicos da idade vocal do locutor (ao longo da vida a voz vai se modificando) contribuem para um pior desempenho do sistema.

Esses fatores ficam fora do escopo dos algoritmos e devem ser corrigidos através de outros meios. Esses fatores humanos podem afetar o desempenho de bons algoritmos de verificação de locutor.

Além da voz, os áudios carregam ruído ambiental e o microfone pode absorver ecos da voz do locutor com atrasos (devido a reflexões da voz em alguma superfície). O ideal seria captar a voz do locutor em ambientes com condições supervisionadas, para um

melhor modelamento. Outro fator que deve ser melhor estudado é quanto a imitação da voz.

A seguir alguns indicadores utilizados para comparação de resultados, entre eles a curva DET (*Decision Error Tradeoff*), a DCF (*Decision Cost Function*) e a EER (*Equal Error Rate*).

3.7.1 Taxa de Erro

Existem dois tipos de erros que podem ocorrer nos sistemas de verificação de locutor. Um deles é a falsa rejeição (ou não detectar), ou seja, dizer que a fala desconhecida não é do locutor alvo, quando na verdade ela era. Em termos de decisão, estando o locutor alvo presente, declará-lo como sendo “FALSO”. O outro é a falsa aceitação (ou falso alarme), quando aceita um impostor, declarando que a fala desconhecida pertence ao locutor alvo. Novamente, em termos de decisão, não estando o locutor alvo presente, e sim um impostor, declará-lo como “VERDADEIRO”. O primeiro é conhecido como perda (*miss*), e o segundo, como falso alarme (*false alarm*). Então, para testes de locutores alvos temos a taxa de perdas e, para testes de impostores, temos a taxa de falsos alarmes.

A FRR (taxa de falsa rejeição) é o número de vezes que um locutor verdadeiro é rejeitado incorretamente e a FAR (taxa de falsa aceitação) é o número de vezes que um locutor impostor é aceito incorretamente. Mudanças nos valores de *threshold* (θ) alteram estas taxas. É interessante determinar o valor de *threshold*, a partir dos valores das taxas de erro FAR e FRR, definindo o *threshold* como o valor em que a taxa de falsa aceitação é igual a taxa de falsa rejeição, ou seja, quando FAR=FRR. Este cálculo é conhecido como taxa de erro igual, ou EER (*Equal Error Rate*). O objetivo é minimizar esta taxa de erro igual.

3.7.2 DCF

A DCF (*Decision Cost Function*), ou Função Custo de Decisão leva em consideração uma combinação linear dos dois tipos de erros (perda e falso alarme). A função é definida como:

$$\mathrm{DCF} = \mathrm{C}_{\mathrm{Miss}} \times \mathrm{P}_{\mathrm{Miss|Target}} \times \mathrm{P}_{\mathrm{Target}} + \mathrm{C}_{\mathrm{FalseAlarm}} \times \mathrm{P}_{\mathrm{FalseAlarm|NonTarget}} \times (1 - \mathrm{P}_{\mathrm{Target}}) \tag{3.17}$$

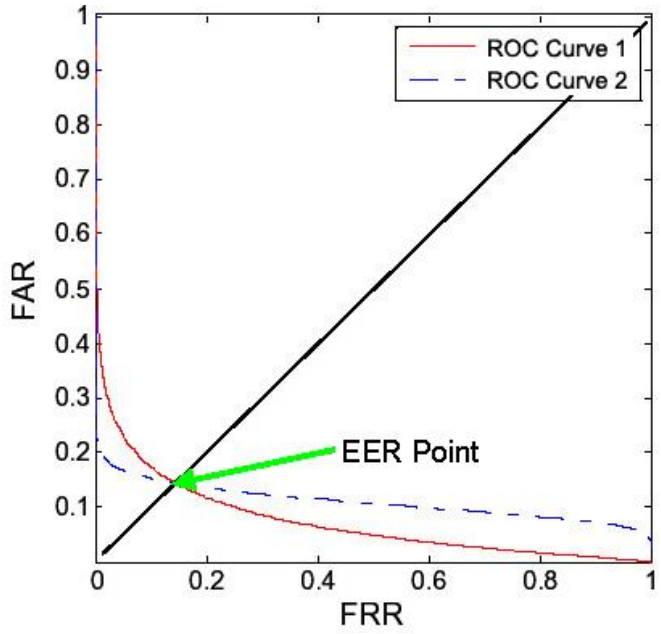
Nesta função $\mathrm{C}_{\mathrm{Miss}}$ representa o custo de uma perda (*miss*), $\mathrm{C}_{\mathrm{FalseAlarm}}$ o custo de um falso alarme (*false alarm*) e $\mathrm{P}_{\mathrm{Target}}$ a probabilidade a priori de um teste de alvo. O NIST usa os seguintes valores como parâmetro: $\mathrm{P}_{\mathrm{Target}} = 0.01$, $\mathrm{C}_{\mathrm{FalseAlarm}} = 1$ e $\mathrm{C}_{\mathrm{Miss}} = 10$.

O DCF é a medida de performance oficial das Campanhas de Avaliação do NIST para reconhecimento de locutor.

3.7.3 DET

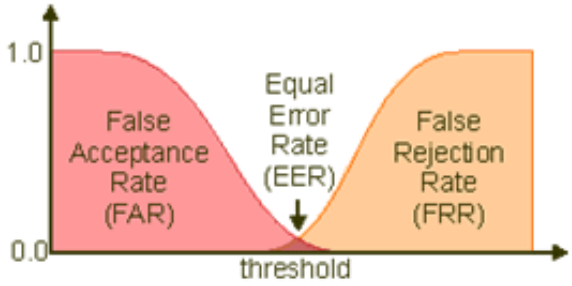
O DET (*Decision Error Tradeoff*) é uma curva obtida ao plotar ambos os erros (perda e falso alarme), conforme Figura 20. Estes erros dependem do limiar de decisão (θ) utilizado. O *threshold* deve ser o mesmo para todos os locutores alvos, a fim de avaliar a performance do sistema. Ambos os erros são relevantes e a diminuição de um pode levar ao aumento do outro, como pode ser visto na Figura 21. Portanto, usa-se a EER (*Equal Error Rate*), que corresponde ao ponto onde as duas taxas de erro (falsa aceitação e falsa rejeição) são iguais, como apresentado na Figura 20, para duas curvas de erro diferentes, é possível obter a mesma EER. Essa é uma medida da eficiência do sistema em separar impostores de locutores verdadeiros.

Figura 20 - Curva ROC



Fonte: (DU e CHANG, 2007)

Figura 21 - Equal Error Rate



Fonte: (MCCARTHY, 2008)

Conforme a literatura, os valores de performance variam de 0,1% a 30%, com a relação a taxa de erro EER. Para sistemas de texto-dependente usando áudios de alta qualidade, obtém-se muito baixas EER, em torno de 0,1% a 2%. Para aplicações dependentes de texto através de canal telefônico tem-se performances de 2% a 5%. Para sistemas de texto-independente, baseados em áudios de ótima qualidade obtêm-se taxas de 7% a 15%. E, finalmente, para aplicações independentes de texto, em canais com alto ruído, tem-se performances entre 20% e 35% (LI; JAIN, 2009).

3.8 Banco de Dados

Na década de 1980 iniciou-se a formação dos primeiros bancos de falas para aplicações em processamento de voz. É importante que um banco de dados de falas contenha um número razoável de diferentes locutores, além de diferentes sessões para cada um. É interessante que haja também variabilidade nas sessões. Alguns bancos comumente utilizados (Tabela 1) são: o TIMIT, da *Texas Instruments* (TI) e *Massachussetts Institute of Technology* (MIT), o KING e o YOHO (LI e JAIN, 2009).

O TIMIT é composto por um corpus de fala lida, através de 10 sentenças foneticamente diferentes, pronunciadas por 630 locutores, na maioria coletadas de um microfone de alta qualidade. O corpus KING é composto de 10 sessões, de 30 segundos cada, de 51 locutores masculinos, e as falas são coletadas através de canal telefônico. Já o corpus YOHO é composto de 138 locutores, cada um com 4 sessões de treino e 10 de verificação, e é utilizado para reconhecimento dependente de texto (LARCHER *et al.*, 2014).

Tabela 1 - Corpora usado em reconhecimento de locutor

Year	Corpus	Size	Types of speech
Early 1980s	TIMIT	630 speakers of eight major US English dialects, 10 sentences each; alternative versions run original wideband data through other specified channels	Read speech of phonetically rich sentences
1987	KING	51 male speakers (25 New Jersey, 26 San Diego), 10 sessions each recorded on both a wide-band and a narrow-band channel	Sessions contain 30 s of speech on an assigned topic
1989	YOHO	138 speakers with 4 enrollment sessions (24 phrases) and 10 test sessions (4 phrases)	"Combination lock" phrases

Fonte: LI e JAIN (2009, p. 1247)

Ao longo dos anos também foi desenvolvido o corpora Switchboard (Tabela 2), com variações de canais entre as sessões de treino e teste, o que possibilitou uma melhor avaliação de performance dos sistemas, através do NIST (*National Institute of Standards and Technology*). Além desse, o corpora Mixer (Tabela 3), mais recente, incluiu conversas em múltiplas línguas, ou seja, mudanças na língua da fala de treino para a de

teste, assim como conversas em que os participantes foram gravados simultaneamente através do telefone e de 8 microfones diferentes.

Tabela 2 - Switchboard Corpora

Year	Corpus	Size	Types of speech
1990/1991	SWBD I	543 speakers, 2400 two-sided conversations	USA conversational telephone speech on assigned topics
1996	SWBD II phase 1	657 speakers, 3638 conversations	Primarily US Mid-Atlantic, conversational telephone
1997	SWBD II phase 2	679 speakers, 4472 conversations	Primarily US Mid-West, conversational telephone
1997/1998	SWBD II phase 3	640 speakers, 2728 conversations	Primarily US South, conversational telephone
1999/2000	SWBD cellular p1	254 speakers, 1309 conversations	Primarily cellular GSM, USA conversational
2000	SWBD cellular p2	419 speakers, 2020 conversations	Cellular, largely CDMA, USA conversational

Fonte: LI e JAIN (2009, p. 1248)

Tabela 3 - Mixer Corpora

Year	Corpus	Size	Types of speech
2003	MIXER p1 and p2	600 speakers with 10 or more calls 200 with 4 cross-channel calls	Conversational, some calls in four non-English languages
2005	MIXER p3	1,867 speakers with 15 or more calls	Conversational, includes calls in 19 languages
2007	MIXER p4	200 speakers making 10 calls including 4 cross-channel	Conversational, primarily English
2007	MIXER p5	300 speakers doing 6 interviews and generally 10 phone calls	Conversational in interview setting, some read speech

Fonte: LI e JAIN (2009, p. 1249)

No Brasil também existem diversos bancos de fala de importância relevante. O banco de dados Iboruna foi desenvolvido pela equipe técnica do Projeto ALIP (Amostra Linguística do Interior Paulista). O banco é composto por Amostra Comunidade e Amostra de Interação Dialógica, conforme Tabela 4.

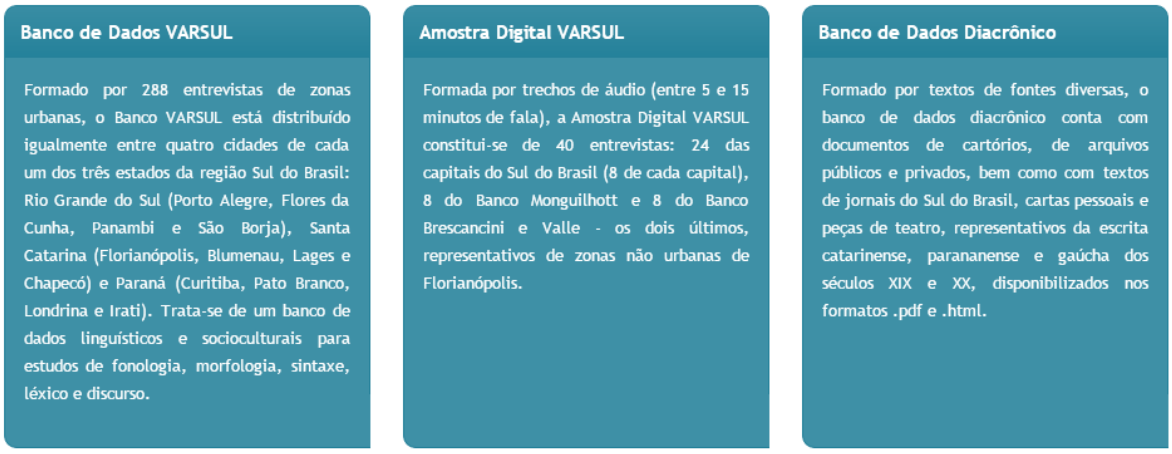
Tabela 4 - Banco de Falas IBORUNA

Banco	Tipo	A cada amostra de fala coletada:
Amostra Comunidade (ou Amostra Censo)	amostras de fala controladas sociolingüisticamente	• cinco arquivos sonoros • um com dados da ficha social do informante • um arquivo de transcrição • um com registros do diário de campo
Amostra de Interação Dialógica	amostras de fala coletadas secretamente em situações livres de interação social	• um arquivo sonoro • um com dados da ficha social dos informantes • um arquivo de transcrição • um com registros do diário de campo

Fonte: (PROJETO ALIP)

Outro banco de falas de elevada importância é o do projeto VARSUL (Variação Linguística na Região Sul do Brasil), que foi desenvolvido por quatro universidades brasileiras e conta com três bancos, cujas característica serão apresentadas na Figura 22.

Figura 22 – Banco de Falas VARSUL



Fonte: (PROJETO VARSUL)

Além destes, existe o banco de falas do Projeto Variação Linguística no Estado da Paraíba – VALPB (HORA, 1993) – que consta de uma amostra de 60 entrevistas, estratificadas em função das características sociais constantes na Tabela 5. Os informantes que constituem o corpus foram selecionados de forma aleatória. As falas foram captadas de forma que fosse minimizado o efeito negativo da presença do entrevistador-pesquisador. São narrativas de experiência pessoal, ocasionando, dessa forma, uma maior desenvoltura e espontaneidade no ato de fala.

Tabela 5 - Banco de Falas VALPB

Projeto Variação Linguística no Estado da Paraíba	
SEXO	Masculino – 30 informantes Feminino – 30 informantes
ANOS DE ESCOLARIZAÇÃO	Nenhum – 12 informantes 1 a 4 anos – 12 informantes 5 a 8 anos – 12 informantes 9 a 11 anos – 12 informantes Mais de 11 anos – 12 informantes
FAIXA ETÁRIA	15 – 25 anos – 20 informantes 26 – 49 anos – 20 informantes 50 anos ou mais – 20 informantes

Fonte: (PROJETO VALPB)

Acima foram apresentados os principais bancos de falas utilizados em pesquisas de processamento da fala e os mais relevantes bancos de fala do português brasileiro. Dentre estes bancos de fala, não foi encontrado um que contemplasse as características esperadas para o desenvolvimento deste trabalho, sendo então desenvolvido um novo banco de falas, como será detalhado no capítulo 5.

4 Comparativos

Este capítulo contempla alguns resultados obtidos por outros pesquisadores, identificando as técnicas utilizadas por cada um e as taxas de erro encontradas. A partir dos resultados apresentados poderá ser feita uma comparação entre as taxas de erros dos métodos já utilizados em sistemas de verificação de locutor independente de texto e dos métodos propostos neste trabalho.

Conforme o artigo "*Investigating the Effect of data Partitioning for GMM Supervector Based Speaker Verification*", de DIKICI e SARAÇLAR (2009), os autores usaram um banco de dados de conversações telefônicas, com 90 locutores (44 homens, 46 mulheres) e realizaram 12 sessões (em diferentes dias) num período de 2 anos. Os locutores deveriam executar as seguintes tarefas: respostas para perguntas curtas, repetição de palavras, números e frases e fala espontânea de curta duração. Em cada sessão foram cerca de 100 declarações de aproximadamente 4 minutos cada em média. Foram constituídos 3 grupos (UBM, *background* e conjunto de usuários). Para o UBM usaram 12 sessões de 20 locutores. Para o background usaram 6 sessões para treino e as outras 6 para teste de outros 20 locutores. E o conjunto de usuários usou 6 sessões para treino e as outras 6 para teste dos outros 50 locutores restantes. O método de extração de características foi o MFCC, com *frames* de 20ms, deslocando 10ms, janela de Hamming, gerando 16 coeficientes MFCC + 16 delta + 16 *double*-delta. Para modelamento do locutor foi implementado GMM com 256 gaussianas (fator 14) e suas médias foram concatenadas em supervetores com dimensão da ordem de 33x256 = 8448x1. Os resultados podem ser acompanhados na Tabela 6.

A pesquisa em questão pretendia investigar o efeito da repartição dos áudios de treino e teste. A duração dos áudios de teste variou entre 4min, 1min e 10s e os melhores resultados encontrados foram com o uso de 4min de gravação, que continham mais informações já que quanto maior a duração, mais informações são esperadas. Também quanto aos áudios de treino, novamente observa-se o aumento da EER conforme diminui a duração de cada trecho de teste, visto que pouca informação (dados do locutor) acaba sendo injetada no sistema.

Tabela 6 - Valores de EER

Test Duration (Partitioning: ×1)	Training Duration	Training Data Partitioning	EER
4min	24min	4min×6	**1.67**
		1min×24	2.53
		10sec×144	3.02
	4min	4min×1	**6.87**
		1min×4	6.92
		10sec×24	7.51
1min	24min	4min×6	10.68
		1min×24	**4.66**
		10sec×144	5.26
	4min	4min×1	16.32
		1min×4	**12.75**
		10sec×24	12.91
	1min	1min×1	18.12
		10sec×6	**17.60**
10sec	24min	4min×6	32.38
		1min×24	21.25
		10sec×144	**8.13**
	4min	4min×1	34.14
		1min×4	24.87
		10sec×24	**18.40**
	1min	1min×1	32.80
		10sec×6	**23.96**
	10sec	10sec×1	**33.63**

Fonte: DIKICI e SARAÇLAR (2009, p. 468)

Já no artigo "*Novel variable length teager energy based features for person recognition from their hum*", de PATIL e PARHI (2010), os autores criaram um banco de dados de 51 locutores sussurrando alguns sons (35 homens e 16 mulheres), a uma taxa de amostragem de 22050Hz. As gravações foram realizadas em estúdio com microfone. Os trechos de fala para treinamento foram de 30s e 60s. Os trechos para teste variavam de 1s a 15s. Foram realizados um total de 2.907 testes verdadeiros e 145.350 testes de impostores. O objetivo dos autores foi criar um novo método de extração de características e compará-lo com os MFCC, além de realizar uma fusão entre os dois. Os resultados encontram-se na Tabela 7, onde é visto que a aplicação do método da fusão entre os dois sistemas apresentados diminui a EER do sistema.

Tabela 7 - Tabela de EER (%)

	MFCC	**VTMFCC(DI=9)**	**Fusão**
EER (%)	14,25	13,89	12,52

Fonte: PATIL e PARHI (2010, p. 4529)

Analisando o artigo "*Multi-feature fusion using multi-GMM supervector for SVM speaker verification*", de (LIU e HUANG, 2009), os autores usaram um subconjunto do corpus Switchboard mencionado na seção 3.8, incluindo 370 locutores alvo com 10s de fala de treino. A duração dos segmentos de teste é de 10s. São gerados os resultados para um sistema usando 16 coeficientes MFCC, incluindo 16 Δ (vetor total de 32 coeficientes) e um sistema usando 12 coeficientes LPCC, incluindo 12 Δ e 12 Δ^2 (total de 36 coeficientes). Para o treinamento do GMM são utilizadas 256 gaussianas. Os autores compararam 3 tipos de classificações diferentes: a primeira usando exclusivamente o GMM; outra utilizando as médias do GMM, gerando um supervetor e aplicando numa SVM; e a última agregando o supervetor MFCC e o supervetor LPCC num único supervetor a ser aplicado na SVM, realizando uma fusão a nível de supervetor. Os resultados mostrados na Tabela 8 demonstram a eficiência do método da fusão.

Tabela 8 - Comparação EER(%)

Systems	EER(%)		
	MFCC	*LPCC*	*Fusion*
GMM-UBM	32.35	30.45	**28.60**
GMM supervector/SVM	31.66	29.62	**28.03**
Multi-GMM supervector/SVM			**27.12**

Fonte: LIU e HUANG (2009, p. 3)

Outro artigo interessante é o "*Multi-layered features with SVM for text-independent speaker verification*", de LI *et al.* (2012). Os autores não mencionam no artigo a estruturação do banco de falas usado. Este artigo compara um sistema utilizando apenas os coeficientes MFCC, outro sistema utilizando apenas o contorno do *pitch* e um terceiro utilizando o método da fusão entre os anteriores. A Tabela 9 apresenta os resultados do artigo, sendo que o uso do contorno do *pitch* sozinho não traz grandes vantagens, porém em fusão produz redução na EER.

Tabela 9 - EER dos Sistemas

Feature	EER
MFCC	**2.7%**
Pitch contour	**46.9%**
MFCC+Pitch contour	**1.4%**

Fonte: LI *et al.* (2012, p. 380)

Conforme verifica-se nas pesquisas realizadas, a comparação dos resultados de cada autor é feita através da EER, ou taxa de erro igual, mencionada na seção 3.7, e será igualmente utilizada neste trabalho.

5 Sistema de Verificação de Locutor Independente de Texto Proposto

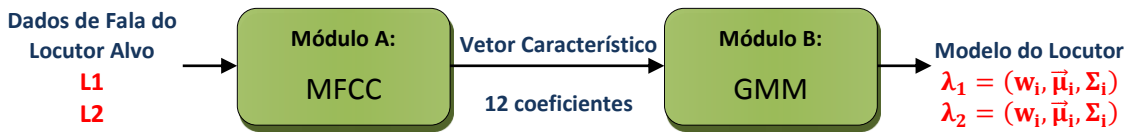
Diante das inúmeras pesquisas já realizadas acerca de verificação de locutor independente de texto e focando em uma área que ainda pudesse ser desenvolvida e melhorada, decidiu-se trabalhar na melhoria do sistema através dos coeficientes utilizados para discriminar o locutor. Nesta seção será apresentada a proposta do trabalho, assim como o detalhamento da metodologia escolhida, a fim de demonstrar todos os passos necessários para a realização deste trabalho.

Dentro desta seção, é importante salientar que será desenvolvido um sistema padrão de verificação de locutor independente de texto, referenciado pelo estado da arte até então pesquisado e implementado, e um novo sistema, com algumas modificações a nível de coeficientes discriminantes do locutor, que será a proposta da dissertação. Com os sistemas implementados será possível realizar comparações entre os mesmos e avaliar os resultados, ressaltando as melhorias encontradas e descrevendo as diferenças entre ambos.

5.1 Proposta

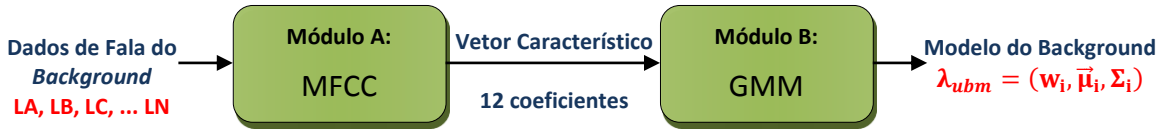
Para atingir o objetivo proposto será implementado, em *software* específico, um sistema de verificação de locutor independente de texto, composto de duas fases distintas: uma de treino e a outra de teste. Num primeiro momento, serão utilizados os coeficientes MFCC para extração das características de cada locutor e o GMM como modelo estatístico. Para cada locutor (L1, L2,...) é gerado um modelo (λ_1, λ_2,...) a partir dos seus dados de fala, conforme Figura 23, e o modelo do *background* (λ_{ubm}) é gerado a partir dos dados de N locutores (L1, L2, ..., LN), criando um único modelo em função dos dados de vários locutores, como mostra a Figura 24. Todos estes modelos ficam gravados no banco de dados, vide Figura 25.

Figura 23 - Fase de Treino do Modelo dos Locutores



Fonte: Ferreira (2015)

Figura 24 - Fase de Treino do Modelo do Background



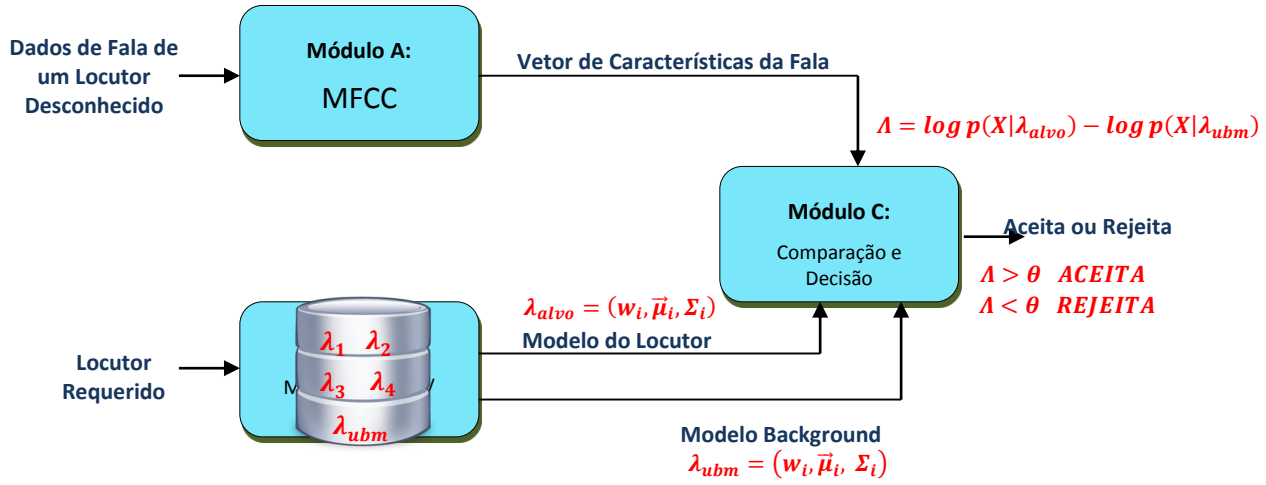
Fonte: Ferreira (2015)

Figura 25 - Banco de Dados com Modelos

Fonte: Ferreira (2015)

Em seguida, é realizada a fase de teste (Figura 26), onde são extraídos os coeficientes MFCC dos dados do locutor desconhecido, e então, esses coeficientes são aplicados no modelo do locutor alvo e do *background* (modelos gerados na fase de treino e que se encontram no banco de dados do sistema), gerando um valor de *likelihood* para cada um [$p(X|\lambda_{alvo})$ e $p(X|\lambda_{ubm})$]. Então, através da razão entre estes valores de *likelihood*, obtém-se um valor de *score* (Λ), cujo valor é comparado com um limiar de decisão, determinando a resposta do sistema. A partir de então, o sistema poderá ser testado a fim de obter resultados que servirão de parâmetro para futura comparação, de acordo com os métodos mais utilizados nas pesquisas mais recentes.

Figura 26 - Fase de Treino (Verificação)



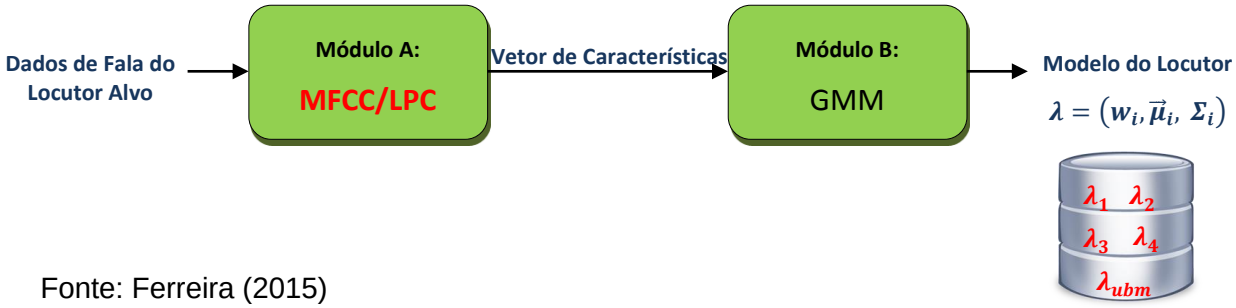
Fonte: Ferreira (2015)

Assim, será possível desenvolver um método que, em conjunto com os coeficientes MFCC, obtenha melhores resultados frente ao mencionado anteriormente. Para tanto, faz-se indispensável o conhecimento das *features* que podem ser extraídas do sinal de voz, conseguindo, então, avaliar um modo de gerar coeficientes que, em conjunto com os MFCCs, evidenciem uma melhor performance ao sistema.

O método proposto para gerar coeficientes é através de um filtro auto-regressivo, que utiliza o método da covariância modificada, gerando coeficientes que modelam a variação dos coeficientes MFCC através do tempo. O método da covariância modificada (GONÇALVES, 2007) estima os parâmetros, por via da minimização dos erros preditivos posterior e anterior.

Com a obtenção destes coeficientes é possível implementar o sistema de outras duas diferentes formas: apenas com os novos coeficientes sendo utilizados para gerar o modelo do locutor, e, também, concatenando os coeficientes MFCC com os novos coeficientes e através deste único vetor gerando o modelo do locutor (Figura 27). Assim, os resultados encontrados serão comparados com os do sistema básico de verificação de locutor independente de texto.

Figura 27 - Fase de Treino com os coeficientes LPC



Fonte: Ferreira (2015)

Além disso, pretende-se realizar uma fusão, a nível de *score*, do sistema, utilizando apenas coeficientes MFCC e do sistema utilizando os novos coeficientes. Os resultados desta fusão serão também comparados com os outros já realizados.

Outra análise prevista no trabalho é com relação a variações de relação sinal-ruído nas gravações dos locutores, tanto dos arquivos que geram o *background*, como dos arquivos de teste dos locutores. A proposta é variar estes níveis de relação sinal-ruído, realizando novos testes, verificando a performance de cada um dos métodos descritos acima e o comportamento dos sistemas em diferentes situações.

Por fim, e não menos importante, é a elaboração de um banco de falas em português brasileiro, pois não foram encontrados bancos de falas em português com as características necessárias para o desenvolvimento deste trabalho. Isto foi feito em parceria com a Rádio Guaíba, que cedeu as gravações de um de seus programas de entrevistas. Os áudios foram editados e então selecionadas as vozes de cada locutor em separado (num total de 155 locutores), gerando diversos trechos de fala para cada locutor.

5.2 Banco de Falas

Os bancos de falas em português encontrados, na sua maioria oferecem frases prontas repetidas por vários locutores, o que não é compatível com a necessidade da pesquisa. Para verificação de locutor independente de texto, é fundamental um banco de falas onde os locutores falem espontaneamente (como em uma conversa telefônica). A

ideia inicial foi de uma entrevista pessoal, onde os locutores fossem interrogados sobre sua vida, suas atividades, seus *hobbies*, deixando o locutor descontraído para falar de forma natural. Por fim, foi aprovada uma parceria com a Rádio Guaíba, emissora de rádio do Rio Grande do Sul, com sede em Porto Alegre, que cedeu material de suas transmissões para este trabalho. O programa escolhido foi o “Esfera Pública”, apresentado por Juremir Machado da Silva e Taline Oppitz, programa voltado para debates sobre temas atuais, como política, cultura, economia e temas sociais, com descontração e informalidade, e trazendo sempre convidados diferentes (Estréia Programa Esfera Pública, 2015). Este programa foi interessante para o trabalho devido a ter sempre 2 a 3 convidados diferentes, o que enriquecia a quantidade de locutores do banco de dados, e também por ser um programa de debate onde cada convidado apresentava sua opinião sobre determinado tema em debate, de forma muito natural também, sendo possível extrair dos locutores uma fala espontânea. O fato de os temas em debate serem diversificados também colaborou bastante.

Os áudios recebidos foram da gravação diária do programa. A frequência de amostragem variava de um áudio para outro e alguns estavam gravados em mono e outros em estéreo. Para padronização dos áudios optou-se pela frequência de amostragem de 22050 Hz e, realizando a média entre os dois canais estéreo, converte-los para mono. Outro ajuste necessário foi com relação a amplitude do sinal, que foi normalizado entre -1 e 1, para que todos os áudios tivessem o mesmo peso (mesma amplitude máxima e mínima). Após estas considerações, foi necessário recortar a fala de cada locutor da gravação, identificando cada trecho de maneira correta para as futuras comparações. Foi possível extrair vários trechos de voz do mesmo locutor para a maioria dos locutores considerados, o que permitiu que os trechos de fala da fase de treino do sistema fossem diferentes dos trechos de fala da fase de teste.

No programa foram entrevistadas algumas mulheres, porém selecionou-se somente os locutores homens para composição do banco de falas. O número de mulheres entrevistadas era bastante reduzido com relação ao número de homens. Para compor o *background* necessitar-se-ia uma composição equilibrada entre o número de homens e mulheres. Por este fato que se escolheu trabalhar somente com os locutores homens.

5.3 Metodologia

Esta seção descreve a metodologia empregada durante o projeto:

- **Pesquisa acerca do assunto:** Busca de referências acerca de verificação de locutor independente de texto, com uma visão geral do assunto, identificando as técnicas já utilizadas e aquelas que têm melhores resultados e destacando os mais recentemente publicados.
- **Leitura e Interpretação:** Estudo aprofundado dos tipos de características presentes na fala do locutor e as diferentes técnicas de extração dessas características; estudo dos algoritmos de modelamento estatístico e classificação, entre eles, o GMM e o SVM; estudo do modo como costumam ser feitas as avaliações de desempenho para verificação de locutor.
- **Implementação da fase de treinamento:** Desenvolvimento, através de recurso de *software* específico, do bloco de treino para modelar os locutores e o *background*. Neste caso, utilizaram-se os MFCC para gerar o vetor característico e o GMM para modelamento dos locutores.
- **Implementação da fase de teste/verificação:** Desenvolvimento, através de recurso de *software* específico, do bloco de teste para modelar a declaração desconhecida e compará-la com o locutor alvo. Neste caso, são utilizados os MFCC para gerar o vetor característico da declaração desconhecida e o algoritmo GMM para modelagem, tendo sido adicionado também o bloco de *score* e decisão.
- **Criação do Banco de Falas em Português:** Obtenção de um banco de falas em português, através de gravações de falas de 155 locutores (gravações extraídas de um programa de entrevistas da Rádio Guaíba, através de fala espontânea dos locutores (Estréia Programa Esfera Pública, 2015)).
- **Realização de testes:** Verificação da eficácia de um sistema padrão de verificação de locutor, o qual servirá como parâmetro para os testes futuros, com as devidas modificações.

- **Extração de um novo conjunto de coeficientes:** Implementação de um filtro auto-regressivo, que utiliza o método da covariância modificada, para obtenção de novos coeficientes para modelo do locutor, através do *software* MATLAB.
- **Realização de testes utilizando apenas os novos coeficientes:** Utilização destes coeficientes para gerar o modelo do locutor e verificar os resultados obtidos.
- **Realização de testes concatenando o vetor de coeficientes MFCC e o vetor com os novos coeficientes:** Utilização de um único vetor, concatenando os coeficientes MFCC e os novos coeficientes e, com este único vetor, gera-se o modelo dos locutores, verificando os resultados obtidos.
- **Realização de testes através da fusão:** A fusão é realizada a nível de *score*. Utilizam-se os *scores* obtidos no sistema que usa apenas os coeficientes MFCC e os *scores* obtidos no sistema que utiliza somente os novos coeficientes. Através de uma ponderação entre estes *scores* gera-se um novo *score,* fruto da fusão entre os dois sistemas.
- **Realização de todos os testes com diferentes níveis de relação sinal-ruído:** Aplicação de uma série de testes variando-se a relação sinal-ruído dos áudios usados no *background* e dos usados nos áudios de testes.
- **Comparação do desempenho de cada método proposto:** Comparação do desempenho de cada método com os outros e com o modelo de parâmetro que utiliza apenas MFCC, através das taxas de erro EER (*Equal Error Rate*).
- **Análise de comportamento em cada situação:** Realização de análise dos resultados de cada método em relação a cada situação proposta, descrevendo as peculiaridades percebidas através dos resultados.
- **Definição do método que apresenta melhor desempenho:** Determinação do modelo que apresenta menor taxa de erro, assim como seleção das características que trouxeram melhores resultados para descrever o locutor.

Todos os passos acima descritos foram executados a fim de implementar um sistema de verificação de locutor independente de texto e desenvolver métodos para melhoria da taxa de erro do sistema.

6 Simulação e Resultados

Esta pesquisa propôs determinar um conjunto de coeficientes para representar os indivíduos, conseguindo assim diminuir as taxas de erro do sistema, levando em conta que um número maior de informações deve ser útil para distinguir (tornar único) o locutor.

6.1 Simulação

Primeiramente, realizou-se uma vasta pesquisa sobre o tópico verificação de locutor independente de texto, a fim de encontrar o estado-da-arte dentro do tema proposto. É importante desenvolver este estudo, entendendo o que já foi implementado por outros pesquisadores e obteve os melhores resultados, além de verificar quais os principais métodos vem se revelando como perspectiva futura dentro do assunto abordado. Outro detalhe relevante refere-se aos resultados obtidos por outros pesquisadores que foram utilizados para comprovação da eficiência do sistema implementado.

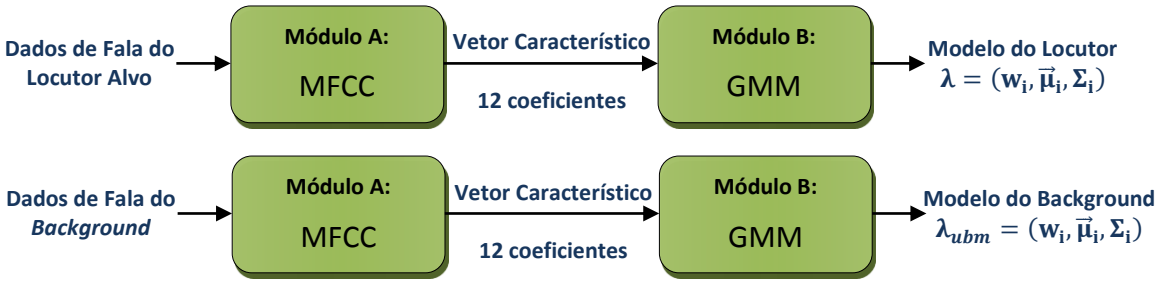
Tomou-se como base o sistema definido como estado-da-arte em verificação de locutor independente de texto, que neste trabalho será chamado de sistema MFCC. Nesse sistema, somente os coeficientes MFCC compõem o vetor característico e o método usado para gerar o modelo do locutor é o GMM. Este é o sistema de referência e sua taxa de erro serve de base para comparação com os outros sistemas.

Em seguida, fez-se necessária a construção de um banco de falas para que fosse possível a realização dos testes. Todo o processo para construção do banco de falas e a descrição de suas características será descrito na próxima seção.

O *software* utilizado para construção dos sistemas foi importante porque proporciona um ótimo ambiente gráfico para análise dos resultados, além de ter uma linguagem simples e de fácil entendimento para o usuário. A fim de construir o sistema de referência, o sistema MFCC, projetaram-se as duas fases: o treino e o teste. Na fase de treino foram modelados o b*ackground* e os locutores alvo, conforme a Figura 28. Para cada locutor foram gerados os coeficientes MFCC (12 coeficientes por *frame*), cada *frame* de 24 ms deslocando a cada 12 ms. Para o *background* utilizaram-se 30s de cada locutor, de um total de 120 locutores que participam do *background*. Estes coeficientes foram

processados, através do Modelo de Misturas Gaussianas (GMM), gerando um modelo específico para cada locutor e um modelo específico para o *background*. Os dados armazenados como modelo são um vetor de médias, uma matriz de covariância e um vetor de pesos. Para criar o modelo foram utilizadas 256 gaussianas e matriz na forma diagonal.

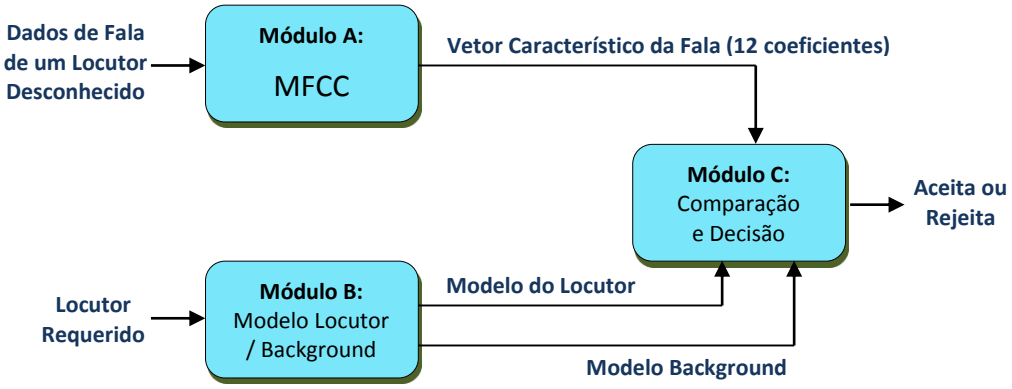
Figura 28 - Fase de Treino do Sistema MFCC



Fonte: Ferreira (2015)

Em havendo armazenado os modelos para cada locutor e o modelo do *background*, é possível passar para a fase de testes. Projeta-se novamente o módulo de geração dos coeficientes (agora do locutor desconhecido) e os módulos de *score* e decisão, conforme Figura 29. Do áudio referente ao locutor desconhecido são extraídos os coeficientes MFCC. No módulo de *score* estes coeficientes são aplicados no modelo do locutor a ser verificado (locutor alvo) e no modelo do *background*, estimando um valor de taxa de verossimilhança para cada modelo. Em se obtendo estes níveis de *score*, um para o locutor alvo e um para o *background*, calcula-se a diferença entre eles, no módulo de decisão. O resultado define se o áudio do locutor desconhecido pertence ao locutor alvo ou não. Esta decisão foi feita através da comparação com um valor de limiar: em estando acima do nível do limiar a afirmação é verdadeira (o locutor desconhecido é o locutor alvo), e em estando abaixo do nível do limiar a afirmação é falsa (o locutor desconhecido não é o locutor alvo).

Figura 29 - Fase de Testes Sistema MFCC



Fonte: Ferreira (2015)

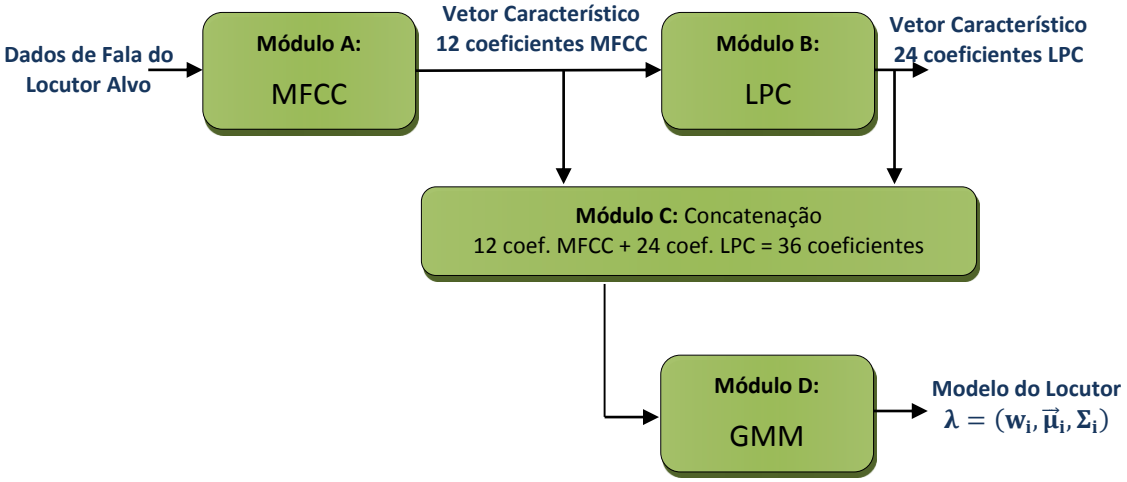
Com o sistema completo, passou-se à realização de vários testes, comparando os locutores com eles mesmos e com outros locutores. Para cada teste realizado, determinou-se o resultado que deveria ser obtido, assim observando-se o número de erros e acertos do sistema. Quantos aos erros, eles se categorizavam em dois tipos: erros de falsa aceitação e erros de falsa rejeição. O erro de falsa rejeição ocorre quando o locutor desconhecido era o locutor alvo, porém o sistema não o reconhecia como tal. Já o erro de falsa aceitação ocorria quando o locutor desconhecido não era o locutor alvo, mas o sistema o reconhecia como sendo. Visto que esses dois tipos de erros são ruins para o sistema, objetivou-se encontrar a taxa de erro igual (EER), ou seja, o ponto em que o sistema tem quantidade de erros de falsa aceitação igual à taxa de erros de falsa rejeição, não pendendo o sistema para nenhum dos lados. Com isso conseguiu-se determinar um limiar de decisão que apresentava a menor EER do sistema.

Todos estes testes foram realizados também variando-se os níveis de relação sinal-ruído (SNR) dos áudios do background e dos áudios dos locutores. Os áudios com diferentes SNRs foram obtidos pela função awgn (*additive white gaussian noise*), que adiciona ruído branco gaussiano ao sinal. Os níveis de variação foram sem ruído e com SNRs de 60dB, 40dB e 20dB. Desta maneira, observou-se o comportamento do sistema quando a relação sinal-ruído do *background* era igual ou próxima à do locutor desconhecido e quando a SNR era diferente, podendo chegar a algumas conclusões quanto ao melhor uso do sistema.

Passada esta etapa e estando com o sistema de referência de acordo com o estado-da-arte em verificação de locutor independente de texto e ainda com os resultados dos testes protocolados, partiu-se para a obtenção de um novo conjunto de coeficientes para modelar o locutor. O conjunto de coeficientes escolhido foi chamado de coeficientes LPC, com o objetivo de estimar a variação dos coeficientes MFCC ao longo do tempo. A técnica utilizada foi o método da covariância modificada, que realiza a predição e visa minimizar o erro preditivo, através da minimização dos erros preditivos posterior e anterior. Esses novos coeficientes são gerados a partir da análise dos coeficientes MFCC. Os testes realizados foram com preditores de segunda e terceira ordem e a janela de coeficientes MFCC escolhidos para realizar a predição variou de 7 a 12 coeficientes.

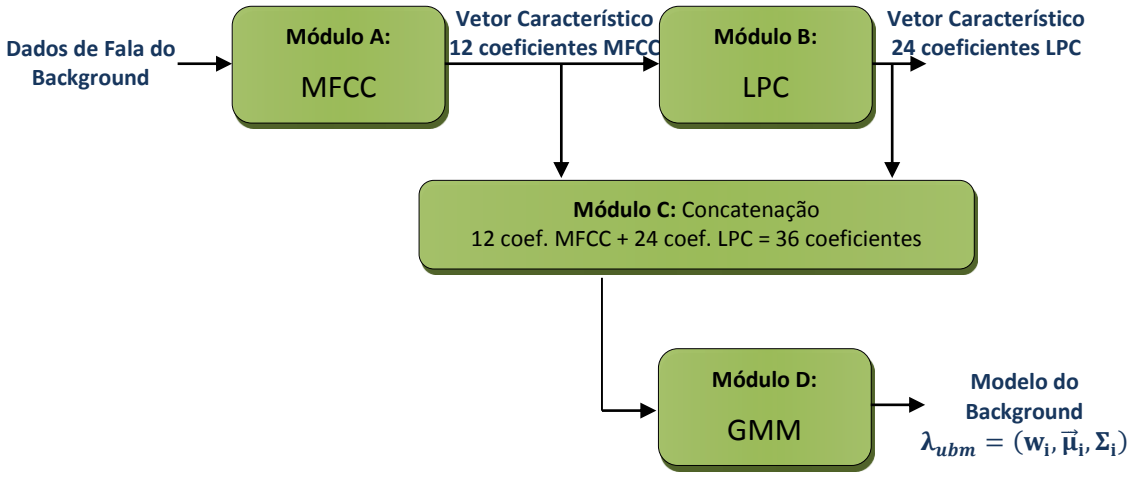
Assim, foi produzido o sistema MFCC-LPC, com as mesmas fases de treino e teste, porém agregando os coeficientes LPC aos coeficientes MFCC, conforme Figura 30 e Figura 31. Como vetor de coeficientes para modelamento dos locutores foi então feita uma concatenação do vetor de coeficientes MFCC (12 coeficientes) com o vetor de coeficientes LPC de segunda ordem (24 coeficientes), ficando então um vetor único de 36 coeficientes. Os resultados obtidos foram analisados e comparados com o sistema de referência, o sistema MFCC.

Figura 30 - Fase de Treino dos Locutores Alvo do Sistema MFCC-LPC



Fonte: Ferreira (2015)

Figura 31 - Fase de Treino do *background* do Sistema MFCC-LPC



Fonte: Ferreira (2015)

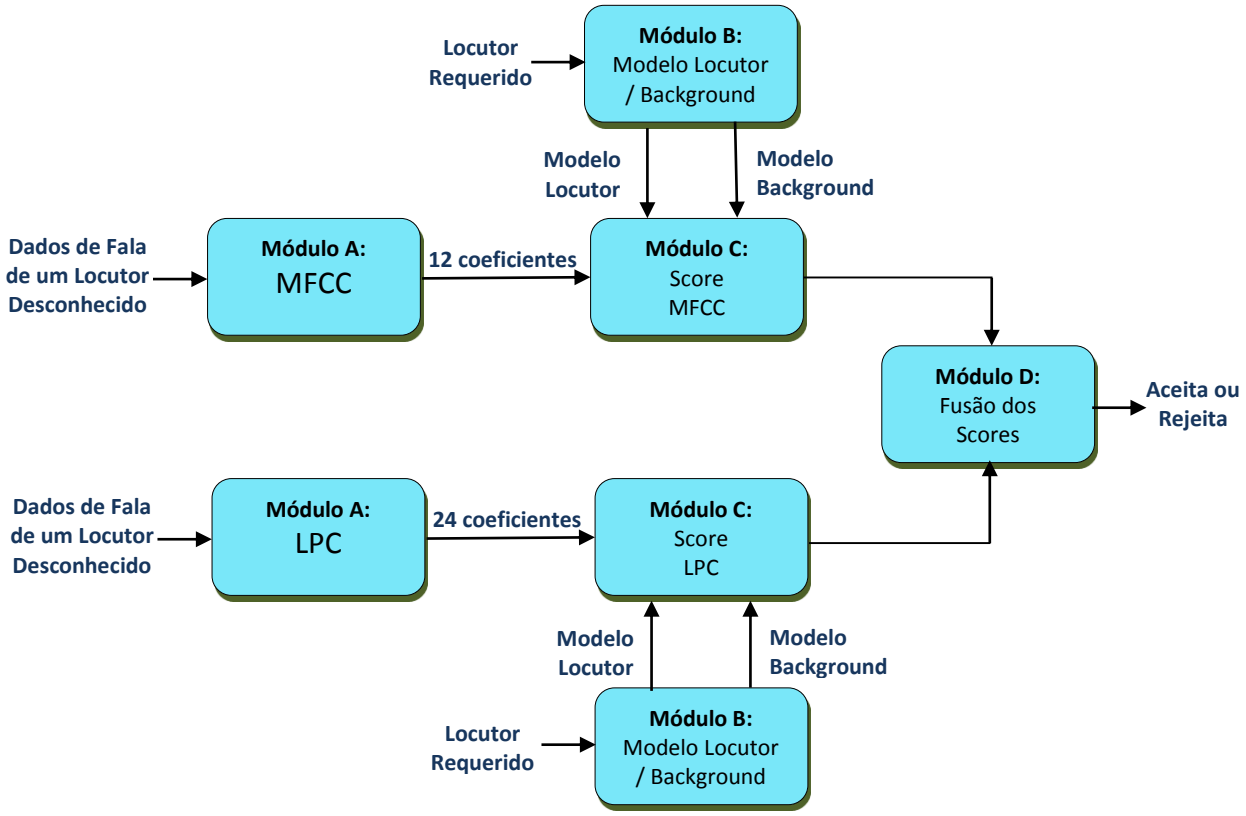
O próximo sistema construído foi o sistema MFCC-Δ-Δ^2, novamente com as fases de treino e teste, e, igualmente, como no sistema MFCC-LPC, realizando a concatenação dos coeficientes MFCC com os coeficientes Δ e Δ^2, num total de 36 coeficientes. Novamente foram apontados os resultados e comparados com o sistema de referência.

Uma tendência bastante recente em verificação de locutor independente de texto é a realização de um método chamado Fusão, que permite o uso de diferentes tipos de coeficientes para discriminar o locutor. Este método é realizado da seguinte forma, conforme a Figura 32: constroem-se os sistemas independentes, cada um utilizando um modo de extração de coeficientes. Com os valores de *score* de cada teste, ou seja, de taxa de verossimilhança de cada teste, procede-se à fusão, que nada mais é que uma ponderação entre os resultados de *score* obtidos para cada sistema. Neste caso, com os resultados de *score* do sistema MFCC e os resultados de *score* do Sistema LPC, realizou-se tal fusão, através da seguinte equação:

$$Score_{fusão} = w \cdot Score_{MFCC} + (1 - w) \cdot Score_{LPC} \qquad (6.1)$$

onde w é o fator de ponderação cujo valor foi determinado variando-se w entre 0 e 1 e analisando-se os resultados obtidos.

Figura 32 - Fase de Teste do Sistema MFCC+LPC



Fonte: Ferreira (2015)

No final destes testes surgiu a dúvida de qual efeito teria a fusão dos coeficientes MFCC com os coeficientes delta (Δ), e projeta-se um novo sistema, chamado de Sistema MFCC+Δ+Δ^2.

Por fim, fez-se uma análise de todos os resultados obtidos, considerando-se os diferentes sistemas implementados, o Sistema MFCC, o sistema MFCC-LPC, o sistema MFCC-Δ-Δ^2, o sistema MFCC+LPC e o sistema MFCC+Δ+Δ^2, e também os diferentes níveis de relação sinal-ruído incorporados ao *background* e aos áudios dos locutores. Após a avaliação, elegeu-se o sistema que apresentou a melhor performance.

6.2 Resultados

Através dos dados coletados foram realizadas análises dos resultados obtidos, buscando a menor taxa de erro do sistema, o que traduz um sistema com melhor desempenho.

Os dados coletados em cada sistema, ou seja, os valores de EER apresentados pelo programa MATLAB, foram recolhidos e aplicados em tabelas de EXCEL (MICROSOFT CORPORATION, 2010) para armazenamento dos dados, a fim de que fosse possível realizar a comparação entre os sistemas ao final de todos os testes.

Os testes foram feitos com os 5 diferentes sistemas citados (MFCC, MFCC-LPC, MFCC+LPC, MFCC-Δ-Δ^2 e MFCC+Δ+Δ^2) e para cada sistema foram realizados 16 diferentes testes (com situações diferentes em relação à presença de ruído nos áudios do *background* e nos áudios de teste). Essas diferentes configurações de SNR foram executadas com o intuito de analisar o comportamento dos sistemas na presença de ruído.

A seguir, serão expostos os resultados de cada sistema. Os valores das tabelas são a EER (*Equal Error Rate*), a taxa de erro igual, que foi detalhada na seção 3.7. A comparação entre os sistemas foi feita através dessa taxa de erro. Nas duas primeiras colunas das tabelas encontram-se as diferentes configurações testadas, variando-se o ruído presente nos áudios do *background* e variando os valores de ruído presentes nos áudios de teste. Como visto nas tabelas, mantém-se um valor de ruído do *background* constante e varia-se o ruído dos áudios de teste, sendo que um dos testes é com o valor de ruído do *background* igual ao valor de ruído dos áudios de teste.

O total são 16 testes, divididos em 4 grupos, dos quais, o primeiro grupo mantém o *background* sem ruído e varia a relação sinal-ruído (SNR) nos áudios de teste para 20dB, 40dB, 60dB e sem ruído; o segundo grupo mantém o *background* com relação sinal-ruído de 60dB e varia a SNR nos áudios de teste para 20dB, 40dB, 60dB e sem ruído; o terceiro grupo mantém o *background* com relação sinal-ruído de 40dB e varia a SNR dos áudios de teste em 20dB, 40dB, 60dB e sem ruído; e o quarto grupo mantém o *background* com 20dB de relação sinal-ruído e varia a SNR dos áudios de teste para 20dB, 40dB, 60dB e sem ruído.

O sistema de referência construído com base no estado-da-arte em verificação de locutor independente de texto, chamado sistema MFCC, apresentou os resultados da terceira coluna da Tabela 10. Neste sistema o vetor característico é composto pelos coeficientes MFCC, tendo o vetor 12 coeficientes MFCC. Os modelos do *background* e dos locutores são gerados a partir de 256 gaussianas no Modelo de Mistura Gaussiana (GMM), sendo que o modelo dos locutores é adaptado do modelo do *background*, adaptando suas médias.

Observando a Tabela 10, os valores estão de acordo com a literatura, onde a taxa de erro é baixa para áudios de ótima qualidade (conforme a 3ª coluna da 3ª linha da tabela a EER ficou em 1,68% na situação em que os áudios do *background* e os áudios de teste tem alta qualidade) e aumenta conforme as diferenças na relação sinal ruído presentes nos áudios.

As pesquisas realizadas também afirmavam que quanto mais próximo fossem os áudios do *background* dos áudios de teste, melhores seriam os resultados do sistema. Isso se comprova na tabela, ao avaliarmos cada grupo: no Grupo 1, onde os áudios do *background* não possuem inserção de ruído, a menor taxa de erro é quando os áudios de teste não possuem ruído também (EER = 1,68%); no Grupo 2, onde os áudios do *background* possuem relação sinal ruído de 60dB, a menor taxa de erro é quando os áudios de teste também possuem relação sinal-ruído de 60dB (EER = 1,54%); no Grupo 3, onde os áudios do *background* possuem relação sinal ruído de 40dB, a menor taxa de erro é quando os áudios de teste também possuem relação sinal-ruído de 40dB (EER = 2,91%); e no Grupo 4, onde os áudios do *background* possuem relação sinal ruído de 20dB, a menor taxa de erro é quando os áudios de teste também possuem relação sinal-ruído de 20dB (EER = 7,45%).

Tabela 10 - Comparação Sistema MFCC e Sistema MFCC-Δ-Δ²

	Situação		EER(%) MFCC	EER(%) MFCC-Δ-Δ²
	Background	Teste		
1	Sem Ruído	Sem Ruído	1,68	1,45
	Sem Ruído	60dB	2,10	1,78
	Sem Ruído	40dB	6,32	5,60
	Sem Ruído	20dB	50,25	48,94
2	60dB	Sem Ruído	1,83	1,45
	60dB	60dB	1,54	1,45
	60dB	40dB	3,80	3,52
	60dB	20dB	48,13	45,76
3	40dB	Sem Ruído	4,88	4,09
	40dB	60dB	4,53	3,88
	40dB	40dB	2,91	2,26
	40dB	20dB	39,38	41,09
4	20dB	Sem Ruído	28,36	26,09
	20dB	60dB	28,03	25,28
	20dB	40dB	21,94	21,39
	20dB	20dB	7,45	8,87

Fonte: Ferreira (2014).

Analisando especialmente o Grupo 4, pode-se perceber uma enorme diferença, já que 20dB de SNR, são áudios de baixíssima qualidade e a taxa de erro diminui de 21,94% para 7,45%, simplesmente porque os áudios do *background* têm a mesma relação sinal-ruído dos áudios de teste. A partir da perspectiva real, onde se cria o *background* a partir dos áudios de teste, tendo os áudios de teste 20dB de SNR, verifica-se que ao usar um *background* sem ruído, a taxa de erro é de 50,25%, ao usar um *background* com 60dB

de SNR, a taxa de erro cai para 48,13%, ao usar um *background* com 40dB de SNR, a taxa de erro já cai para 39,38%, e ao usar um *background* com a mesma relação dos áudios de teste, com 20dB de SNR, a taxa de erro reduz drasticamente para 7,45%. É como querer verificar um áudio desconhecido de uma gravação telefônica com ruído de canal e ter a opção de escolher entre um *background* de gravações com alta qualidade, um *background* de gravações realizadas em microfones e um *background* de ligações telefônicas. O correto seria escolher o *background* de ligações telefônicas que estaria o mais próximo do áudio desconhecido a ser verificado, reduzindo assim a sua taxa de erro.

Uma variação do sistema de referência bastante utilizada em verificação de locutor independente de texto e que tem bons resultados é o uso dos coeficientes delta(Δ) e delta-delta(Δ^2), ou seja, das derivadas primeira e segunda dos coeficientes MFCC. Esses coeficientes são anexados ao vetor característico que já possui os coeficientes MFCC. Implementou-se este sistema concatenando-se 12 coeficientes delta e 12 coeficientes delta-delta ao vetor característico original com 12 coeficientes MFCC. Os resultados encontram-se na quarta coluna da Tabela 10.

Para comparar os resultados do Sistema MFCC-Δ-Δ^2 com os resultados do sistema de referência, analisa-se a Tabela 10. Comparando os resultados, 14 dos 16 testes realizados apresentam melhoria na taxa de erro do sistema. Este sistema é amplamente utilizado na literatura e ao analisarmos a Tabela 10, entende-se o porquê, visto que ele realmente gera melhores resultados.

O método de extração de características desenvolvido foi o sistema MFCC-LPC. Ele usa um preditor linear de 2ª ordem, que estima o comportamento temporal dos coeficientes MFCC. Sendo 12 os coeficientes MFCC, o LPC gera 24 coeficientes que são concatenados ao vetor característicos com os coeficientes MFCC. O novo vetor característico tem um total de 36 coeficientes. Na Tabela 11 encontram-se os resultados deste sistema. O preditor utiliza 12 amostras passadas para calcular o coeficiente atual.

Tabela 11 - EER Sistema MFCC-LPC

	Situação		EER (%) MFCC-LPC
	Background	Teste	
1	Sem Ruído	Sem Ruído	6,82
	Sem Ruído	60dB	7,45
	Sem Ruído	40dB	15,55
	Sem Ruído	20dB	44,39
2	60dB	Sem Ruído	7,57
	60dB	60dB	6,57
	60dB	40dB	14,29
	60dB	20dB	43,11
3	40dB	Sem Ruído	11,50
	40dB	60dB	11,18
	40dB	40dB	8,75
	40dB	20dB	39,40
4	20dB	Sem Ruído	43,76
	20dB	60dB	41,16
	20dB	40dB	38,18
	20dB	20dB	20,29

Fonte: Ferreira (2014).

Analisando a Tabela 11, verifica-se que o Sistema MFCC-LPC apresenta piores resultados, muito acima do sistema de referência, o sistema MFCC. Isto se justifica pelo fato de os coeficientes MFCC e os coeficientes LPC possuírem média e desvio padrão diferentes, ou seja, seus valores não estão normalizados.

Uma alternativa é agregar os coeficientes a nível de *score*, após a geração dos modelos através do GMM que fará esta normalização. Este método é conhecido como fusão. Trabalha-se com os dois sistemas separados, o MFCC e o LPC, gerando os

modelos do *background* e os modelos de teste, realizando-se a comparação e, ao obter-se os valores de *score* (*likelihood ratio*), aplica-se uma ponderação, criando um novo valor de *score* único para os dois sistemas. Este valor é usado para determinar se o locutor desconhecido é o locutor alvo ou não.

O sistema que realiza esta fusão é o sistema MFCC+LPC, o fator de ponderação é w e segue a seguinte equação:

$$Score_{fusão} = w \cdot Score_{MFCC} + (1 - w) \cdot Score_{LPC} \quad (6.2)$$

Os resultados são apresentados na Tabela 12, conforme a variação de w, de 0 a 1, incrementando 0,05. Quando w =0, o resultado refere-se a um sistema MFCC-LPC, pois utiliza somente o $Score_{LPC}$ que é criado a partir do vetor com 12 coeficientes MFCC + 24 coeficientes LPC. Quando w =1, o resultado refere-se a um sistema MFCC, pois utiliza somente o $Score_{MFCC}$ que é criado a partir do vetor característico unicamente com os coeficientes MFCC.

Analisando os resultados conforme a variação de w, determina-se w=0,6 como o de melhor performance. Na Tabela 12 está destacado em negrito os resultados com w=0,6 e em vermelho os campos com menor valor de taxa de erro.

Tabela 12 - Sistema MFCC+LPC

Situação		EER(%) - Sistema MFCC+LPC (LPC com janela de 12 coeficientes)										
Backg.	Teste	*w*=0	*w*=0.05	*w*=0.1	*w*=0.15	*w*=0.2	*w*=0.25	*w*=0.3	*w*=0.35	*w*=0.4	*w*=0.45	*w*=0.5
Sem ruído	Sem ruído	6,82	5,35	4,45	3,24	2,44	2,10	1,94	1,78	1,62	1,45	**1,45**
Sem ruído	60dB	7,4	5,96	4,81	3,34	2,91	2,59	2,59	2,43	2,15	1,94	**1,94**
Sem ruído	40dB	15,55	14,01	12,48	11,18	9,98	8,62	8,11	7,18	6,80	6,43	**6,10**
Sem ruído	20dB	44,57	44,57	44,57	44,57	44,57	44,57	44,57	44,57	44,57	44,57	**44,57**
60dB	Sem ruído	7,14	5,71	4,53	3,56	3,08	2,59	2,40	2,19	1,86	1,78	**1,62**
60dB	60dB	8,10	6,10	4,7	3,72	3,05	2,43	2,26	1,94	1,94	1,90	**1,78**
60dB	40dB	14,26	12,61	11,06	9,62	8,11	6,86	6,03	5,24	4,63	4,09	**3,89**
60dB	20dB	43,11	42,92	43,11	42,63	42,79	42,63	42,82	42,96	43,27	43,11	**43,27**
40dB	Sem ruído	11,51	8,91	6,80	5,35	5,35	5,02	4,56	4,37	4,21	4,21	**4,09**
40dB	60dB	11,35	8,91	6,48	5,53	5,18	4,59	4,34	4,21	4,05	4,05	**3,89**
40dB	40dB	8,42	6,48	5,34	4,53	4,131	3,40	2,75	2,59	2,26	2,10	**2,10**
40dB	20dB	39,48	39,22	38,9	38,9	38,58	38,41	37,97	37,44	37,93	38,25	**38,4**
20dB	Sem ruído	41,38	39,55	37,25	34,85	32,83	31,65	30,35	29,34	28,53	28,2	**27,95**
20dB	60dB	41,27	38,83	36,21	34,36	32,41	30,71	29,82	29,17	28,77	27,71	**26,9**
20dB	40dB	40,73	38,41	35,82	33,71	31,44	30,31	28,2	26,9	25,77	24,8	**24,14**
20dB	20dB	21,62	20,08	19,12	17,64	16,69	15,88	15,07	14,26	13,61	12,97	**11,96**

Situação		EER(%) - Sistema MFCC+LPC (LPC com janela de 12 coeficientes)									
Background	Teste	*w*=0.55	*w*=0.6	*w*=0.65	*w*=0.7	*w*=0.75	*w*=0.8	*w*=0.85	*w*=0.9	*w*=0.95	*w*=1
Sem ruído	Sem ruído	1,45	**1,45**	1,45	1,45	1,45	1,58	1,65	1,72	1,76	1,68
Sem ruído	60dB	1,94	**1,94**	1,94	1,94	1,94	1,94	1,94	1,94	1,94	2,01
Sem ruído	40dB	5,67	**5,38**	5,24	5,34	5,51	5,46	5,67	5,96	6,28	6,32
Sem ruído	20dB	44,57	**44,57**	44,57	44,57	44,57	44,57	44,57	44,57	44,57	44,57
60dB	Sem ruído	1,58	**1,61**	1,62	1,50	1,62	1,62	1,62	1,65	1,72	1,78
60dB	60dB	1,65	**1,62**	1,62	1,62	1,50	1,45	1,45	1,45	1,54	1,61
60dB	40dB	3,48	**3,40**	3,40	3,24	3,26	3,12	3,41	3,41	3,66	4,05
60dB	20dB	43,18	**43,6**	44,25	45,15	46,03	47,23	47,99	48,14	47,97	48,62
40dB	Sem ruído	4,05	**4,05**	4,052	4,052	4,37	4,53	4,7	4,86	4,86	5,02
40dB	60dB	3,89	**3,72**	3,72	3,89	3,98	4,09	4,31	4,31	4,37	4,53
40dB	40dB	2,10	**2,10**	2,10	2,26	2,26	2,29	2,51	2,75	2,87	2,91
40dB	20dB	39,04	**39,38**	40,16	40,8	40,23	40,19	39,62	39,38	39,71	39,06
20dB	Sem ruído	27,73	**27,87**	27,88	27,98	27,88	28,38	28,52	28,81	28,85	29,06
20dB	60dB	26,54	**26,26**	26,69	26,58	27,07	27,55	28,04	28,2	28,53	28,69
20dB	40dB	23,66	**23,02**	22,67	22,59	22,53	22,2	22,53	22,49	22,69	22,85
20dB	20dB	10,88	**10,21**	9,56	9,23	8,59	8,42	8,26	7,75	7,57	7,45

Fonte: Ferreira (2014).

Na Tabela 13 compara-se os resultados obtidos pelo Sistema MFCC+LPC com os do sistema de referência. A partir desses dados confirma-se que o sistema desenvolvido, através do método da fusão, traz melhor performance ao sistema, diminuindo a taxa de erro numa média de 1% a menos. A performance do sistema MFCC+LPC se equivale à do sistema MFCC-Δ-Δ^2.

Tabela 13 - Comparação Sistema MFCC e Sistema MFCC+LPC

	Situação		EER(%) MFCC	EER(%) MFCC+LPC
	Background	Teste		
1	Sem Ruído	Sem Ruído	1,68	1,45
	Sem Ruído	60dB	2,10	1,94
	Sem Ruído	40dB	6,32	5,38
	Sem Ruído	20dB	50,25	44,57
2	60dB	Sem Ruído	1,83	1,61
	60dB	60dB	1,54	1,62
	60dB	40dB	3,80	3,40
	60dB	20dB	48,13	43,59
3	40dB	Sem Ruído	4,88	4,05
	40dB	60dB	4,53	3,72
	40dB	40dB	2,91	2,10
	40dB	20dB	39,38	39,38
4	20dB	Sem Ruído	28,36	27,87
	20dB	60dB	28,03	26,25
	20dB	40dB	21,94	23,02
	20dB	20dB	7,45	10,21

Fonte: Ferreira (2014).

Por fim, ainda foi realizado um último teste. Como já mencionado, segundo a literatura pesquisada, sempre que se faz uso dos coeficientes delta(Δ) e delta-delta (Δ^2), concatenam-se esses coeficientes (ver Figura 15) a um vetor característico que já possui os coeficientes MFCC. Em nenhum material foi encontrado esta junção ao nível de *score*, assim como foi feito com os coeficientes LPC, no Sistema MFCC+LPC. Para tanto, foi realizado este teste, a fim de verificar como seria o comportamento do sistema.

Na Tabela 14 encontram-se os resultados para as variações de w. Pode-se observar que quando w=0, o sistema já apresenta bons resultados, sendo que w=0 representa o sistema usando apenas os coeficientes delta e delta-delta. É intrigante que um método tão conhecido como é o delta e delta-delta não tenha sido usado em fusão ainda.

É selecionado como w que gera os melhores resultados, o valor de w=0,35, cujos valores da taxa de erro estão destacados em negrito. Os menores valores de cada linha são destacados em vermelho.

Tabela 14 - Sistema MFCC+Δ+Δ^2

Situação		EER(%) - MFCC+Δ+Δ^2																
Backg.	Teste	w=0	w=0.05	w=0.1	w=0	w=0.05	w=0.1	w=0	w=0.05	w=0.1	w=0	w=0.05	w=0.1	w=0	w=0.05	w=0.1	w=0	w=0.05	w=0.1
Sem ruído	Sem ruído	2,431	2,19	1,94	2,443	2,26	1,94	4,993	4,59	4,41	22,85	21,07	20,47	4,538	4,38	4,38			
Sem ruído	60dB	2,658	2,33	2,04	2,431	2,19	1,86	4,862	4,53	4,23	22,04	20,26	18,64						
Sem ruído	40dB	3,736	3,56	3,52	3,079	2,80	2,69	1,904	1,86	1,78	17,34	16,21	14,91						
Sem ruído	20dB	34,16	**33,94**	34,63	35,98	**35,66**	35,98	23,6	23,01	22,99									
60dB	Sem ruído	2,443	2,26	1,94	4,993	4,59	4,41	22,85	21,07	20,47	4,538	4,38	4,38						
60dB	60dB	2,431	2,19	1,86	4,862	4,53	4,23	22,04	20,26	18,64									
60dB	40dB	3,079	2,80	2,69	1,904	1,86	1,78	17,34	16,21	14,91									
60dB	20dB	35,98	**35,66**	35,98	23,6	23,01	22,99												
40dB	Sem ruído	4,993	4,59	4,41	22,85	21,07	20,47	4,538	4,38	4,38									
40dB	60dB	4,862	4,53	4,23	22,04	20,26	18,64												
40dB	40dB	1,904	1,86	1,78	17,34	16,21	14,91												
40dB	20dB	23,6	23,01	22,99															
20dB	Sem ruído	22,85	21,07	20,47	4,538	4,38	4,38												
20dB	60dB	22,04	20,26	18,64															
20dB	40dB	17,34	16,21	14,91															
20dB	20dB	4,538	4,38	4,38															

Situação		EER(%) - Sistema MFCC+Δ+Δ^2									
Background	Teste	*w*=0.15	*w*=0.2	*w*=0.25	*w*=0.3	*w*=0.35	*w*=0.4	*w*=0.45	*w*=0.5	*w*=0.55	*w*=0.6
Sem ruído	Sem ruído	1,62	1,47	1,45	1,36	**1,29**	1,29	1,40	1,45	1,45	**1,45**
Sem ruído	60dB	1,78	1,62	1,47	1,45	**1,45**	1,45	1,45	1,45	1,54	**1,47**
Sem ruído	40dB	3,40	3,44	3,40	3,40	**3,24**	3,40	3,62	3,7	3,59	**3,89**
Sem ruído	20dB	36,82	38,57	39,66	40,59	**41,98**	42,64	43,5	44,83	45,94	**47,49**
60dB	Sem ruído	1,78	1,65	1,61	1,54	**1,45**	1,45	1,36	1,29	1,22	**1,18**
60dB	60dB	1,68	1,54	1,45	1,45	**1,43**	1,29	1,29	1,18	1,13	**1,13**
60dB	40dB	2,43	2,26	2,19	2,22	**2,10**	2,10	2,08	2,01	1,94	**1,976**
60dB	20dB	36,53	35,85	36,3	37,12	**37,93**	39,87	41,88	43,57	44,57	**45,19**
40dB	Sem ruído	4,05	3,91	3,84	3,98	**4,05**	4,02	3,95	3,91	3,98	**4,05**
40dB	60dB	3,89	3,62	3,62	3,56	**3,52**	3,56	3,66	3,72	3,62	**3,77**
40dB	40dB	1,79	1,72	1,62	1,65	**1,62**	1,62	1,62	1,62	1,76	**1,78**
40dB	20dB	23,67	25	25,61	26,26	**27,07**	28,04	29,34	31	32,4	**33,48**
20dB	Sem ruído	19,76	20,1	20,11	20,26	**20,75**	21,12	21,88	22,84	23,34	**24,31**
20dB	60dB	18,8	19,25	19,11	20,08	**20,69**	20,94	21,56	22,41	23,38	**23,99**
20dB	40dB	15,24	15,24	15,88	15,95	**16,27**	16,95	17,03	17,49	17,85	**18,79**
20dB	20dB	4,37	4,45	4,49	4,49	**4,34**	4,37	4,37	4,45	4,56	**4,7**

Situação		EER(%) - Sistema MFCC+Δ+Δ²								
Background	Teste	*w*=0.65	*w*=0.7	*w*=0.75	*w*=0.8	*w*=0.85	*w*=0.9	*w*=0.95	*w*=1
Sem ruído	Sem ruído	1,45	1,45	1,61	1,62	1,62	1,62	1,76	1,68
Sem ruído	60dB	1,58	1,65	1,65	1,79	1,90	1,94	1,94	2,10
Sem ruído	40dB	4,05	4,21	4,52	4,7	4,86	5,35	5,96	6,32
Sem ruído	20dB	48,46	48,62	49,27	49,57	49,76	49,8	50,08	50,57
60dB	Sem ruído	1,32	1,40	1,47	1,54	1,72	1,78	1,79	1,78
60dB	60dB	1,22	1,25	1,29	1,43	1,54	1,58	1,54	1,62
60dB	40dB	2,26	2,43	2,58	2,80	2,91	3,19	3,48	4,05
60dB	20dB	45,65	46,41	46,77	47,56	48,14	48,31	48,49	48,62
40dB	Sem ruído	3,98	4,13	4,09	4,27	4,53	4,70	4,86	5,02
40dB	60dB	3,80	3,89	3,98	4,05	4,09	4,20	4,23	4,53
40dB	40dB	1,78	1,94	2,10	2,10	2,40	2,59	2,83	2,917
40dB	20dB	34,52	35,17	35,7	36,53	37	37,6	38,47	39,06
20dB	Sem ruído	24,86	25,45	26,09	26,47	27,01	28,04	28,66	29,06
20dB	60dB	24,5	25,25	25,61	26,26	26,76	27,71	28,27	28,69
20dB	40dB	19,29	19,79	20,42	20,58	21,07	21,72	22,27	22,85
20dB	20dB	4,86	5,18	5,46	5,96	6,46	6,48	7,00	7,45

Fonte: Ferreira (2014).

Na Tabela 15, comparam-se os resultados obtidos com o sistema MFCC+Δ+Δ² com os do sistema de referência. A partir desses dados conclui-se que os coeficientes delta (Δ) e *double*-delta (Δ²) são muito mais eficientes em conjunto com os coeficientes

MFCC se forem combinados a nível de *score*, ao invés de serem concatenados no vetor característico, como normalmente é feito. A taxa de erro diminui numa média de 4% a menos. A performance do sistema MFCC+Δ+Δ^2 mostra-se a melhor dentre os sistemas discutidos até aqui.

Tabela 15 - Comparação Sistema MFCC e Sistema MFCC+Δ+Δ^2

	Situação		EER(%) MFCC	EER(%) MFCC+Δ+Δ^2
	Background	Teste		
1	Sem Ruído	Sem Ruído	1,68	1,29
	Sem Ruído	60dB	2,10	1,45
	Sem Ruído	40dB	6,32	3,24
	Sem Ruído	20dB	50,25	41,97
2	60dB	Sem Ruído	1,83	1,45
	60dB	60dB	1,54	1,43
	60dB	40dB	3,80	2,10
	60dB	20dB	48,13	37,93
3	40dB	Sem Ruído	4,88	4,05
	40dB	60dB	4,53	3,52
	40dB	40dB	2,91	1,62
	40dB	20dB	39,38	27,06
4	20dB	Sem Ruído	28,36	20,74
	20dB	60dB	28,03	20,68
	20dB	40dB	21,94	16,27
	20dB	20dB	7,45	4,34

Fonte: Ferreira (2014).

Ainda realizou-se mais um teste, modificando um parâmetro do GMM, o número de misturas, passando de 256 gaussianas para 1.024 gaussianas. Este teste foi feito no sistema de referência e no Sistema MFCC+LPC. Para a comparação, utiliza-se o mesmo w escolhido para o teste com 256 gaussianas, pois assim, varia-se apenas o número de gaussianas de um resultado para o outro. Os resultados encontram-se na Tabela 16.

Tabela 16 - Comparação Sistema MFCC e Sistema MFCC+LPC (1.024 gaussianas)

Situação		**EER(%)**	**EER(%)**	**EER(%)**	**EER(%)**
Background	**Teste**	**MFCC (256 gaussianas)**	**MFCC (1024 gaussianas)**	**MFCC+LPC (256 gaus.)**	**MFCC+LPC (1024 gaus.)**
Sem Ruído	**Sem Ruído**	1,68	1,29	1,45	1,13
Sem Ruído	**60dB**	2,10	1,29	1,94	1,29
Sem Ruído	**40dB**	6,32	5,24	5,38	4,59
Sem Ruído	**20dB**	50,25	49,59	44,57	45,21
60dB	**Sem Ruído**	1,83	1,29	1,61	0,97
60dB	**60dB**	1,54	1,13	1,62	0,97
60dB	**40dB**	3,80	3,41	3,40	2,75
60dB	**20dB**	48,13	47,48	43,59	45,86
40dB	**Sem Ruído**	4,88	4,23	4,05	3,91
40dB	**60dB**	4,53	3,88	3,72	3,40
40dB	**40dB**	2,91	2,26	2,10	2,10
40dB	**20dB**	39,38	36,30	39,38	38,32
20dB	**Sem Ruído**	28,36	25,89	27,87	23,98
20dB	**60dB**	28,03	27,55	26,25	26,50
20dB	**40dB**	21,94	23,77	23,02	22,70
20dB	**20dB**	7,45	7,61	10,21	11,34

Fonte: Ferreira (2014).

Conforme a Tabela 16, verifica-se uma redução na EER quando alteramos o número de gaussianas de 256 para 1.024, tanto para o sistema MFCC, quanto para o sistema MFCC+LPC, porém o tempo de processamento aumenta consideravelmente.

Para finalizar, apresenta-se a Tabela 17, com o resultado final para cada sistema, que oferece uma visão geral para a avaliação sobre qual deles apresentou melhor desempenho. A pior performance está no sistema MFCC-LPC, que trouxe aumento da taxa de erro com relação ao sistema de referência (sistema MFCC), aumento este de 6,6% em média. Os sistemas MFCC-Δ-Δ^2 e MFCC+LPC apresentaram resultados bem próximos, o primeiro com uma redução média de 0,65% na taxa de erro, e o segundo, com uma redução média de 0,81% na taxa de erro. Ambos apresentam bons resultados. Enfim, o sistema MFCC+Δ+Δ^2 superou os outros sistemas, com uma redução média de 4% na taxa de erro.

Tabela 17 - Comparativo Final

Situação		**EER(%) MFCC**	**EER(%) MFCC-LPC**	**EER(%) MFCC+LPC**	**EER(%) MFCC-Δ-Δ^2**	**EER(%) MFCC+Δ+Δ^2**
Background	**Teste**					
Sem Ruído	**Sem Ruído**	1,68	6,82	1,45	1,45	1,29
Sem Ruído	**60dB**	2,10	7,45	1,94	1,78	1,45
Sem Ruído	**40dB**	6,32	15,55	5,38	5,60	3,24
Sem Ruído	**20dB**	50,25	44,39	44,57	48,94	41,97
60dB	**Sem Ruído**	1,83	7,57	1,61	1,45	1,45
60dB	**60dB**	1,54	6,57	1,62	1,45	1,43
60dB	**40dB**	3,80	14,29	3,40	3,52	2,10
60dB	**20dB**	48,13	43,11	43,59	45,76	37,93

Situação		EER(%) MFCC	EER(%) MFCC-LPC	EER(%) MFCC+LPC	EER(%) MFCC-Δ-Δ²	EER(%) MFCC+Δ+Δ²
Background	Teste					
40dB	Sem Ruído	4,88	11,50	4,05	4,09	4,05
40dB	60dB	4,53	11,18	3,72	3,88	3,52
40dB	40dB	2,91	8,75	2,10	2,26	1,62
40dB	20dB	39,38	39,40	39,38	41,09	27,06
20dB	Sem Ruído	28,36	43,76	27,87	26,09	20,74
20dB	60dB	28,03	41,16	26,25	25,28	20,68
20dB	40dB	21,94	38,18	23,02	21,39	16,27
20dB	20dB	7,45	20,29	10,21	8,87	4,34

Fonte: Ferreira (2014).

Todos os dados compilados são utilizados para realizar a conclusão final do trabalho, que será apresentada no próximo capítulo.

7 Conclusão e Perspectivas Futuras

Este capítulo encerra o trabalho, elencando as conclusões extraídas ao final do projeto e expondo as perspectivas futuras na área de Verificação de Locutor Independente de Texto, inclusive estimulando a continuação das pesquisas.

7.1 Conclusão

A tarefa de verificação de locutor independente de texto vem sendo desenvolvida há diversos anos. Muitos avanços já foram alcançados, muitas mudanças já aconteceram. Em virtude do material disponível até o momento, objetivou-se contribuir com as pesquisas dentro de uma área que realmente precisa de uma maior atenção, pois ainda tem muito a desenvolver.

O objetivo principal do trabalho foi alcançado através da seleção dos coeficientes LPC, proporcionando características do sinal de voz, da variação temporal dos coeficientes MFCC, e o conjunto desses coeficientes, a partir do método da fusão, possibilitou obter melhores resultados que o estado-da-arte em verificação de locutor independente de texto.

Implementou-se um sistema de verificação de locutor independente de texto onde o vetor característico foi composto apenas por coeficientes MFCC. Este foi o sistema de referência e foi chamado de sistema MFCC. Todos os outros sistemas desenvolvidos foram comparados com este.

Criou-se um banco de falas em português, por meio de um programa de rádio que recebia diversos convidados. O banco de dados de fala foi de extrema importância para a realização do trabalho. Foi necessária uma grande quantidade de dados para a implementação do sistema. Apenas para o *background* foram necessários áudios de 120 diferentes locutores, a fim de gerar um modelo padrão. Os bancos de falas na língua portuguesa no Brasil encontrados não são adequados as necessidades de trabalhos em verificação de locutor independente de texto, logo estes áudios servirão para outros trabalhos futuros. Os arquivos ficarão disponíveis no LAFA, Laboratório de Áudio e Fonética Acústica, da Faculdade de Engenharia da PUCRS (PUCRS, 2015).

Como resultado direto do trabalho, identificou-se nos testes realizados, assim como foi comentado em publicações, que o ideal é que o *background* escolhido seja o mais próximo possível do áudio desconhecido. No trabalho os testes foram feitos com diferenças na relação sinal-ruído presente nos áudios e constatou-se, através das simulações, que a melhor situação é quando temos os áudios no *background* com a mesma relação sinal-ruído do áudio testado, comprovando-se o que foi sugerido nas publicações.

Durante o processo de simulação dos sistemas, observou-se que o método inicialmente proposto, o sistema MFCC-LPC, não forneceu os resultados esperados. Contudo, ao implementá-lo através do método da fusão, como feito no sistema MFCC+LPC, o sistema se equivaleu aos métodos mais avançados já desenvolvidos, como o caso do sistema MFCC-Δ-Δ^2.

Enfim, como um ganho adicional, aplicou-se a fusão a um método já consagrado na literatura, o delta-delta, realizando a fusão do sistema MFCC com o delta, o sistema MFCC+Δ+Δ^2. Esta diversificação do método permitiu implementar um sistema com uma performance muito melhor, reduzindo consideravelmente a taxa de erro e aumentando a confiabilidade do sistema.

7.2 Perspectivas Futuras

Considerando-se os resultados obtidos no presente trabalho, propõe-se como trabalhos futuros a implementação de um sistema de detecção de níveis de sinal ruído no sinal gravado para que seja utilizado o mesmo nível na composição do *background*.

Além disso, a inserção de informações linguísticas da fala junto aos métodos de modelamento dos locutores já existentes seria de grande relevância, pois características específicas do locutor encontram-se presentes ali. Essas características também poderiam ser conjugadas através do método da fusão apresentado neste trabalho.

Todas estas possibilidades futuras indicam que esta área de verificação de locutor independente de texto tem um caminho vasto a ser percorrido e muitas perspectivas de ganhos com trabalhos futuros.

Referências Bibliográficas

AMORIM, P. R. F. Biometria, Recife, 19 Dezembro 2005. Disponivel em: <www.cin.ufpe.br/~prfa/Monografia.doc>. Acesso em: Dez. 2014.

ANDRADE, A. O.; SOARES, A. B. **Técnicas de Janelamento de Sinais**. III Seminário dos Estudantes de Engenharia Elétrica da UFU. Uberlândia: [s.n.]. 2000.

BENZEGHIBA, M.; BOURLAND, H. User-customized password speaker verification using multiple reference and. **Speech Communication**, 2006.

BHATTACHARJEE, U.; SARMAH, K. **A Multilingual Speech Database for Speaker Recognition**. IEEE International Conference on Signal Processing, Computing and Control (ISPCC). Waknaghat Solan: IEEE. 2012. p. 1-5.

BIMBOT, F. et al. A tutorial on Text-Independent Speaker Verification. **Eurasip Journal on Applied Signal Processing**, 2004. 4, 430-451.

BISHOP, C. M. **Patern Recognition and Machine Learning**. 4. ed. New York: Springer, v. 4, 2006.

BLOUET, R. et al. **Becars:** a free software for speaker verification. Speaker and Language Recognition Workshop. [S.l.]: [s.n.]. 2004.

CAMPBELL JR, J. P. Speaker Recognition: a tutorial. **Proceedings of the IEEE**, v. 85, n. 9, Setembro 1997. p. 1437-1462.

CAMPBELL, J. P. et al. Forensic speaker recognition. **Signal Processing Magazine, IEEE**, v. 26, 27 Março 2009. p. 95-103.

CAMPBELL, W. M. et al. **Support vector machines for speaker and language recognition**. The speaker and language recognition workshop. [S.l.]: Elsevier. 2006. p. 210-229.

CAMPBELL, W. M.; STURIM, D. E.; REYNOLDS, D. A. Support vector machines using GMM supervectors for speaker verification. **Signal Processing Letters**, v. 13, n. 5, 2006. p. 308-311.

CANEDO, J. A. História da Biometria. **Fórum da Biometria**, 2010. Disponivel em: <http://www.forumbiometria.com/fundamentos-de-biometria/118-historia-da-biometria.html>. Acesso em: 17 Julho 2014.

CHANG, C.-C.; LIN, C.-J. **LIBSVM - A Library for Support Vector Machines**, 2014. Disponivel em: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>. Acesso em: 15 Dez. 2014.

CHAUHAN, T.; SONI, H.; ZAFAR, S. A Review of Automatic Speaker Recognition. **International Journal of Soft Computing and Engineering**, v. 3, n. 4, 2013. p. 132-135.

DIKICI, E.; SARAÇLAR, M. Investigating the Effect of data Partitioning for GMM Supervector Based Speaker Verification. **Computer and Information Sciences**, Setembro 2009. 465 - 470.

DRGAS, S.; CETNAROWICZ, D.; DABROWSKI, A. Speaker verification based on prosodic features. **Signal Processing Algorithms, Architectures, Arrangements, and Applications (SPA)**, Poznan, Poland, 2008. p. 79-82.

DU, Y.; CHANG, C.-I. Rethinking the effective assessment of biometric systems. **SPIE Newsroom**, 2007.

ESTRÉIA Programa Esfera Pública. **Correio do Povo**, 26 Março 2015. Disponivel em: <http://www.correiodopovo.com.br/ArteAgenda/118333/Programa-%22Esfera-Publica%22-estreia-nesta-segunda-feira-na-Radio-Guaiba>.

FARRELL, K. R.; MAMMONE, R. J.; ASSALEH, K. T. Speaker recognition using neural networks and conventional classifiers. **IEEE Transactions on Speech and Audio Processing**, v. 2, n. 1, 1994. p. 104-205.

FLANAGAN, J. **Speech Analysis Synthesis and Perception**. New York and Berlin: Springer-Verlag, 1972.

FURUI, S. Cepstral Analysis Technique for Automatic Speaker Verification. **IEEE Transactions**, v. ASSP-29, n. 2, 1981. p. 254-272.

GONÇALVES, A. C. **Processamento Digital de Sinais – Estimação Paramétrica**. UFPR - Universidade Federal do Paraná. [S.l.]. 2007.

GORSUCH, R. L. **Factor Analysis**. 2ª. ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1983.

GUDNASON, J.; BROOKES, M. **Voice source cepstrum coefficients for speaker identification**. IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2008. Las Vegas, NV: IEEE. 2008. p. 4821-4824.

HARRINGTON, J.; CASSIDY, S. **Techniques in speech acoustics**. [S.l.]: Springer Science & Business Media, v. v. 8, 1999.

HARTMANN, W. M. **Signals, sound and sensation**. [S.l.]: Springer Science & Business Media, 1997. Disponivel em: <http://kom.aau.dk/group/04gr742/pdf/MFCC_worksheet

.pdf>. Acesso em: 15 Maio 2015.

HERMANSKY, H. Perceptual linear predictive (PLP) analysis of speech. **Journal of the Acoustical Society of America**, v. 87, n. 4, 1990. p. 1738-1752.

HOSSAN, M. A.; MEMON, S.; GREGORY, M. A. **A novel approach for MFCC feature extraction**. 4th International Conference on Signal Processing and Communication Systems (ICSPCS). Gold Coast: IEEE. 2010. p. 1-5.

HTK Toolkit. **Hidden Markov Model Toolkit (HTK)**. Disponivel em: <http://htk.eng.cam.ac.uk/>. Acesso em: 20 Janeiro 2014.

HUANG, X.; ACERO, A.; HON, H. W. **Spoken Language Processing:** a Guide to Theory, Algorithm, and System Development. New Jersey: Editora Prentice-Hall, 2001.

JAIN, A. K.; DUIN, R. P. W.; MAO, J. Statistical Pattern Recognition: A Review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 1, 2000. p. 4-37.

KINNUNEN, T.; ALKU, P. **On separating glottal source and vocal tract information in telephony speaker verification**. Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2009). Taipei, Taiwan: [s.n.]. 2009. p. 4545-4548.

KINNUNEN, T.; LI, H. An Overview of Text-Independent Speaker Recognition: from Features to Supervectors. **Speech Communication**, v. 52, 31 Jan. 2010. p.12-40.

KOMLEN, D. et al. **Text Independent Speaker Recognition Using LBG Vector Quantization**. Proceedings of the 34th International Convention MIPRO. Opatija, Croacia: IEEE. 2011. p. 1652-1657.

LARCHER, A. et al. Text-dependent speaker verification: Classifiers, databases and RSR2015. **Speech Communication**, v. 60, p. 56-77, Maio 2014.

LI, J.; GUO, W.; DAI, L.-R. **Total Variability Factors Combination for Speaker Verification**. International Conference on Audio, Language and Image Processing. Shanghai: IEEE. 2012. p. 1001-1004.

LI, S. Z.; JAIN, A. K. **Encyclopedia of Biometrics**. [S.l.]: Springer, 2009.

LI, Y.-G. et al. **Multy-layered Features with SVM for text-independent speaker verification**. International Conference on Conputer Science and Service System. [S.l.]: IEEE Explorer. 2012. p. 378-380.

LIESHOUT, P. V. **PRAAT Short Tutorial. A basic introdution.** University of Toronto, Graduate Department of SpeechLanguage. Toronto, Canadá. 2003.

LIMA, C. B. D. **Sistema de Verificação de Locutor Independente do Texto baseados em GMM e AR-Vetorial utilizando PCA**. Instituto Militar de Engenharia. Rio de Janeiro. 2001.

LIU, M. et al. A New Hybrid GMM/SVM for Speaker Verification. **The 18h International Conference on Pattern Recognition**, 2006. 314-317.

LIU, M.; HUANG, Z. Multi-feature fusion using Multi-GMM Supervector for SVM Speaker Verification, 2009.

LIU, M.; HUANG, Z. **Multi-feature fusion using Multi-GMM Supervector for SVM Speaker Verification**. International Congress on Image and Signal Processing. Tianjin: IEEE. 2009. p. 1-4.

MAGRIN-CHAGNOLLEAU, I.; DUROU, G.; BIMBOT, F. Application of time-frequency principal component analysis to text-independent speaker identification. **IEEE Transactions on Speech and Audio Processing**, v. 10, n. 6, 2002. 371-378.

MALAYATH, N. et al. Data-Driven Temporal Filters and Alternatives to GMM in Speaker Verification, v. 10, p. 55-74, 2000.

MARKOV, K. P.; NAKAGAWA, S. Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition. **Journal of the Acoustical Society of Japan**, v. 20, n. 4, p. 281-291, 1999.

MATHWORKS. MATLAB - The language of technical computing. **MathWorks**, 2015. Disponivel em: <http://www.mathworks.com/products/matlab/>. Acesso em: 2015 Abril 07.

MCCARTHY, J. Re-rethinking Recommendation Engines: Psychology and the Influence of False Negatives. **Gumption**, 2008. Disponivel em: <http://gumption.typepad.com/blog/2008/02/re-rethinking-r.html>. Acesso em: 2015 Julho 20.

MICROSOFT CORPORATION. Office Excel. **Microsoft**, 2010. Disponivel em: <https://products.office.com/pt-br/excel>. Acesso em: 2015 Maio 12.

MODI, K.; SAUL, L. **Text Independent Speaker Verification System**. [S.l.]. 2006.

NAKAGAWA, S.; WANG, L.; OHTSUKA, S. Speaker Identification and Verification by Combining. **IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING,** v. 20, n. 4, p. 1085-1095, 2012.

NETO, M. U. et al. **Análise Paramétrica de Sinais de Voz Baseada em Estimação Conjunta do Modelo Fonte-Filtro**. XXX Simpósio Brasileiro de Telecomunicações. Brasilia, DF: [s.n.]. 2012.

PADIARAJ, S. et al. A Confidence Measure based - Score Fusion Technique to integrate MFCC and Pitch for Speaker Verification. **IEEE Explorer**, 2011.

PATIL, H. A.; PARHI, K. K. **Novel variable length teager energy based features for person recognition from their hum**. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). Dallas, Texas, USA: IEEE. 2010. p. 4526-4529.

PETRY, A. **Reconhecimento Automático de Locutor Utilizando Medidas de Invariantes Dinâmicas Não-Lineares**. UFRGS. Porto Alegre, p. 43. 2002.

PROJETO ALIP. Banco de Dados Iboruna. **ALIP. Amostra Linguística do Interior Paulista**. Disponivel em: <http://www.iboruna.ibilce.unesp.br>. Acesso em: 30 Agosto 2015.

PROJETO VALPB. Projeto VALPB. **Corpus Linguístico VALPB**. Disponivel em: <http://projetovalpb.com.br>. Acesso em: 30 Agosto 2015.

PROJETO VARSUL. Banco de Dados Varsul. **Projeto Varsul**. Disponivel em: <http://www.varsul.org.br>. Acesso em: 30 Agosto 2015.

PUCRS. Pontifícia Universidade Católica do Rio Grande do Sul. **PUCRS**, 2015. Disponivel em: <http://www.pucrs.br/portal/>. Acesso em: 2015 Maio 12.

REYNOLDS, D. A. Speaker identification and verification using Gaussian mixture speaker models. **Speech Communication**, 1995. 91-108.

REYNOLDS, D. A. **An Overview of Automatic Speaker Recognition Technology**. IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, Flórida. USA: IEEE. 2002. p. 4072-4075.

REYNOLDS, D. A.; QUATIERI, T. F.; DUNN, R. B. Speaker Verification Using Adapted Gaussian Mixture Models. **Digital Signal Processing**, v. 10, 2000. p. 19-41.

SINITH, M. S. et al. **A novel method for text-independent spaker identification using MFCC and GMM**. International Conference on Audio Language and Image Processing (ICALP). Shangai: IEEE. 2010. p. 292-296.

SMITH, L. I. **A tutorial on Principal Component Analysis**. [S.l.]. 2002.

SOONG, F. K.; ROSENBERG, A. E. On the use of instantaneous and transitional spectral information in speaker recognition. **IEEE Transactions on Acoustics Speech and Signal Processing**, v. 36, n. 6, 1988. p. 871-879.

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

SUN, X. **Pitch Determination and vocie quality analysis using subharmonic-to-harmonic ratio**. Department of Communication Sciences and Disorders, Northwestern University. Evanston, IL, USA. 2002.

SWANN, J. E. A. A Dictionary of Sociolinguistics, Tuscaloosa: The University of Alabama Press, p. 76, 2004.

TEJA, M. H.; CHAITRA, N. **Computationally Efficient Speaker Identification System using AMDF and LPC**. 3rd International Conference on Electronics Computer Technology (ICECT). Kanyakumari: [s.n.]. 2011. p. 288-291.

TOGNERI, R.; PULLELLA, D. An overview of Speaker Identification: Accuracy and Robustness Issues. **IEEE Circuits and Systems Magazine**, p. 23-58, 2011.

U.S. DEPARTMENT OF COMMERCE. NIST. **National Institute of Standards and Technology**, 2010. Disponivel em: <http://www.nist.gov/>. Acesso em: 2015 Maio 07.

UNIANAL Universal Speech Analysis and Synthesis. **Speech Processing Group**. Disponivel em: <http://speech.fit.vutbr.cz/files/software/unianal/unianal.tar.gz>. Acesso em: 2014 Maio 10.

VARCHOL, P.; LEVICKY, D.; JUHAR, J. **Optimalization of GMM for Text Independent Speaker Verification System**. Radioelektronika, 2008 18th International Conference. [S.l.]: [s.n.]. 2008. p. 1-4.

WEISS, U. Fonética Articulatória, Brasília: Summer Institute of Linguistics, p. 69, 1988.

YUJIN, Y.; PEIHUA, Z.; QUN, Z. **Research of Speaker Recognition Based on Combination of LPCC and MFCC**. IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS). Xiamen: IEEE. 2010. p. 765-767.

ZHENG, N. L. T. E. C. P. C. Integration of Complementary Acoustic Features for Speaker Recognition. **IEEE Signal Processing Letters**, v. 14, n. 3, 2007. p. 181-184.