UNIVERSIDADE FEDERAL DE PERNAMBUCO

GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO

CENTRO DE INFORMÁTICA

**Reconhecimento de Locutor**

TRABALHO DE GRADUAÇÃO

ALUNO: Daniel Brito de Oliveira Carvalho {dboc@cin.ufpe.br}

ORIENTADOR: Tsang Ing Ren {tir@cin.ufpe.br}

UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE INFORMÁTICA

**Reconhecimento de Locutor**

Daniel Brito de Oliveira Carvalho

*Monografia apresentada ao Centro de*

*Informática da Universidade Federal de*

*Pernambuco como requisito parcial*

*para a obtenção do título de Engenheiro*

*da Computação.*

*Orientador: Tsang Ing Ren*

**Recife, Julho de 2011**

*“O sofrimento é passageiro, desistir é para sempre.”*

**Bernardo Fonseca**

Agradecimentos

Gostaria de agradecer primeiramente a Deus, acima de tudo, Àquele que me concedeu saúde necessária e oportunidade de chegar nessa última fase do curso. Acreditar Nele fez com que eu tivesse estímulo para driblar os desafios, por vezes, tão recorrentes.

Agradeço imensamente a meus pais, José Regivaldo de Carvalho e Carlota Brito de O. Carvalho, por terem me apoiado e me dado, dia após dia, as condições de chegar até aqui. Sem o esforço deles, de forma alguma eu teria conseguido. Além deles, meus sinceros agradecimentos a meus irmãos com quem sempre pude contar em todos os momentos, tanto nos de felicidade quanto naqueles mais difíceis.

Aos meus amigos, tanto aqueles da faculdade que construíram junto a mim os conhecimentos que aprendi; quanto àqueles que, ainda que não tenhamos um contato diário, sinto com plena convicção, estão sempre dispostos a dividir comigo minhas conquistas e também os meus problemas.

Por fim, mas com a mesma importância, agradeço ao meu orientador Tsang Ing Ren pela paciência e confiança que em mim depositou e, consequentemente, possibilitou a conclusão deste trabalho.

Obrigado

Resumo

A maioria de nós está consciente do fato de que vozes de indivíduos diferentes não soam iguais. É nesse contexto que a área de reconhecimento de locutor se insere: reconhecer pessoas através da voz. Assim, reconhecimento de locutor é o processo de reconhecer pessoas usando informações específicas contidas no sinal de voz.

Ao longo das últimas décadas presenciou-se um grande avanço na área de sistemas de reconhecimento automático de locutor, onde são usadas técnicas computacionais para reconhecer sua respectiva voz. Dentro dessas se destacam o método de extração de coeficientes cepstrais de frequências mel (*Mel Frequency Cepstral Coefficient* - MFCC) e o método geração de modelos e de classificação baseados em modelos de misturas gaussianas ( *Guassian Mixture Models -* GMM ).

Este trabalho tem como objetivo implementar e analisar um sistema de reconhecimento automático de locutor utilizando GMM, enumerando seus pontos fracos e fortes. Será exposto todo o processo de criação desde a aquisição da base de dados, passando pela extração de características até chegar ao processo de treinamento e classificação.

Abstract

Most of us are aware of the fact that different individual’s voices do not sound the same. In this context, the speaker recognition area works in: to recognize people by voice. Thus, speaker recognition is the process of recognizing people using specific information contained in the voice signal.

Over the past decades witnessed a great advance in automatic speaker recognition systems, where computational techniques are used to recognize speakers. Within these stand out the method of extracting Mel Frequency Cepstral Coefficient (MFCC) and the method of generation models and classification based on Guassian Mixture Models (GMM).

This study aims to analyze and implement a system of automatic speaker recognition using GMM, listing their strengths and weaknesses. Will be exposed the whole process of creation from the acquisition of the database, through the feature extraction process to reach the training and classification.

Sumário

[Lista de Figuras 9](#_Toc297807523)

[Lista de Tabelas 10](#_Toc297807524)

[1. Introdução 11](#_Toc297807525)

[1.1. Definindo o Problema 11](#_Toc297807526)

[1.2. Aplicações 12](#_Toc297807527)

[1.3. Visão geral do trabalho 13](#_Toc297807528)

[2. Reconhecimento de Locutor 14](#_Toc297807529)

[2.1. Breve Histórico 14](#_Toc297807530)

[2.2. Reconhecimento automático de locutor 18](#_Toc297807531)

[3. Extração de Características 21](#_Toc297807532)

[3.1. Digitalização 21](#_Toc297807533)

[3.2. Coeficientes Cepstrais de Frequência Mel 23](#_Toc297807534)

[3.2.1. Pré-ênfase 23](#_Toc297807535)

[3.2.2. Segmentação 25](#_Toc297807536)

[3.2.3. Janelamento 25](#_Toc297807537)

[3.2.4. DFT 26](#_Toc297807538)

[3.2.5. Banco de Filtros Mel 26](#_Toc297807539)

[3.2.6. Log e DCT 28](#_Toc297807540)

[4. Modelos de Misturas Gaussianas 29](#_Toc297807541)

[4.1. GMM para Reconhecimento de Locutor 31](#_Toc297807542)

[4.1.1. Treinamento 31](#_Toc297807543)

[4.1.1.1. Expectation-Maximization(EM) 32](#_Toc297807544)

[4.1.2. Identificação com GMM 33](#_Toc297807545)

[4.1.3. Verificação com GMM 33](#_Toc297807546)

[5. O Sistema de Identificação de Locutor 36](#_Toc297807547)

[5.1. VAD 36](#_Toc297807548)

[5.2. MFCC + LogEnergy 36](#_Toc297807549)

[5.3. Algoritmo EM 37](#_Toc297807550)

[5.3.1. Inicialização 37](#_Toc297807551)

[5.3.2. Critério de parada 38](#_Toc297807552)

[6. Experimentos e Resultados 39](#_Toc297807553)

[6.1. Base de dados 39](#_Toc297807554)

[6.1.1. ELSDSR 39](#_Toc297807555)

[6.1.2. MIT 41](#_Toc297807556)

[6.2. Experimentos 43](#_Toc297807557)

[6.2.1. ELSDSR 43](#_Toc297807558)

[6.2.2. MIT 46](#_Toc297807559)

[7. Conclusão e Trabalhos Futuros 48](#_Toc297807560)

[Referências 49](#_Toc297807561)

[APÊNDICE A – Textos de Treinamento e Testes ELSDSR 53](#_Toc297807562)

[Texto de Treinamento da Base de dados ELSDSR 53](#_Toc297807563)

[Parte A: 53](#_Toc297807564)

[Parte B: 53](#_Toc297807565)

[Parte C: 53](#_Toc297807566)

[Parte D: 53](#_Toc297807567)

[Parte E: 53](#_Toc297807568)

[Parte F: 53](#_Toc297807569)

[Parte G: 53](#_Toc297807570)

[Texto de Teste da Base de Dados ELSDSR 54](#_Toc297807571)

# Lista de Figuras

[Figura 1 - Engenheiro Lawrence G. Kersta um dos inventores do espectrógrafo de som (sound spectrograph).[Forensic] 15](#_Toc297807630)

[Figura 2 - Estrutura Básica de Sistemas de Reconhecimento Automático de Locutor 19](#_Toc297807631)

[Figura 3 - Estrutura Básica de Sistemas de Identificação de Locutor 19](#_Toc297807632)

[Figura 4 - Estrutura Básica de Sistemas de Verificação de Locutor 20](#_Toc297807633)

[Figura 5 - Fluxograma de Digitalização 22](#_Toc297807634)

[Figura 6 - Fluxograma para obtenção dos MFCCs 23](#_Toc297807635)

[Figura 7 - Resposta em frequência do filtro de pré-ênfase 24](#_Toc297807636)

[Figura 8 - Espectrograma do Sinal Original (a) e do Sinal pré-enfatizado (b). 24](#_Toc297807637)

[Figura 9 - Segmentação do sinal de voz 25](#_Toc297807638)

[Figura 10 - Janela de Hamming 26](#_Toc297807639)

[Figura 11 – Gráfico de Frequência Mel x Frequência Hetz 27](#_Toc297807640)

[Figura 12 - Banco de filtros triangulares escalados linearmente na escala Mel 28](#_Toc297807641)

[Figura 13 – Distribuição de dados em torno de um ponto médio e modelagem através de uma gaussiana 29](#_Toc297807642)

[Figura 14 – Distribuição de dados num nível mais complexo e modelagem através de uma gaussiana 29](#_Toc297807643)

[Figura 15 – Distribuição de dados num nível mais complexo e modelagem por várias misturas gaussianas 30](#_Toc297807644)

[Figura 16 – Estrutura básica para um sistema de verificação do locutor baseado em modelos de misturas gaussianas. 35](#_Toc297807645)

[Figura 17 – Diagrama do Sistema de identificação de locutor 36](#_Toc297807646)

[Figura 18 – Gráfico: Número de Gaussianas x Taxa de Acerto – Primeiro experimento ELSDSR 44](#_Toc297807647)

[Figura 19 - Gráfico: Número de Gaussianas x Taxa de Acerto – Segundo experimento ELSDSR 45](#_Toc297807648)

[Figura 20 - Gráfico: Número de Gaussianas x Taxa de Acerto – Terceiro experimento ELSDSR 46](#_Toc297807649)

[Figura 21 - Gráfico: Número de Gaussianas x Taxa de Acerto – experimento MIT 47](#_Toc297807650)

# Lista de Tabelas

[Tabela 1 – Cronologia reconhecimento de locutor [29] 17](#_Toc297807651)

[Tabela 2 – Descrição de Idade e Nacionalidade na Base de Dados ELSDSR (Feminino) 39](#_Toc297807652)

[Tabela 3 - Descrição de Idade e Nacionalidade na Base de Dados ELSDSR (Masculino) 40](#_Toc297807653)

[Tabela 4 – Duração da leitura do texto de treinamento e teste 41](#_Toc297807654)

[Tabela 5 – Exemplo da lista de frases faladas em cada sessão 42](#_Toc297807655)

[Tabela 6 – Primeiro Experimento ELSDSR 43](#_Toc297807656)

[Tabela 7 - Segundo Experimento ELSDSR 44](#_Toc297807657)

[Tabela 8 - Terceiro Experimento ELSDSR 45](#_Toc297807658)

[Tabela 9 – Experimento MIT 47](#_Toc297807659)

# Introdução

Sabe-se que cada pessoa possui diversas características únicas que os diferencia das outras pessoas, como as digitais, a íris, a voz, entre outros. A partir disso, surgem diversos sistemas que tentam diferenciar cada pessoa por uma dessas características intrínsecas a cada um. No caso do reconhecimento de locutor é usado o sinal produzido pelo aparelho fonador humano, a voz, como discriminante entre pessoas.

A voz humana é um sinal complexo, um resultado da influência de vários fatores fisiológicos, psicológicos e ambientais. Assim, o sinal produzido pelo humano é determinado, dentre outros fatores, pela fisiologia do aparelho fonador, pelas características regionais como o sotaque e manifestações sentimentais como a raiva. Com isso é possível extrair vários tipos de informações do sinal de voz, como a língua falada, a emoção e a identidade do locutor.

## Definindo o Problema

O reconhecimento automático de locutor é o processo de reconhecer automaticamente uma pessoa, utilizando meios computacionais, com base nas características do sinal de voz. A motivação de usar o sinal de voz como biometria deriva do fato de que a voz, pela atual tecnologia, é um sinal de fácil obtenção. De tal modo, os sistemas biométricos baseados em voz podem ser considerados os sistemas mais naturais e econômicos [2]. Esse tipo de sistema inclui dois tipos de tarefas:

* Verificação

Tem o objetivo de verificar a identidade de uma pessoa. Nesse caso o sistema sabe a priori a identidade da pessoa, através de um *login* ou senha, e sua função é somente verificar se aquela voz é da pessoa que se diz ser.

* Identificação

Tem o objetivo de identificar uma pessoa dentro de um conjunto conhecido de várias pessoas. Nesse caso o sistema não tem nenhuma informação a priori e sua função é identificar, num grupo cadastrado de N locutores, de quem pertence à voz.

Em qualquer um dos casos, verificação ou identificação, os sistemas de reconhecimento de locutor podem ser divididos quanto ao texto em:

* Dependente de Texto

Onde o sistema depende das palavras que são ditas. Nesse caso o sistema requer que o locutor forneça locuções, ou palavras-chave, com o mesmo texto para ambos o treinamento e o reconhecimento.

* Independente de Texto

O sistema não tem nenhuma dependência para com o texto falado. Nesse caso a locução fornecida para treinamento não precisa ser a mesma para o teste.

Diante da variabilidade na construção de um sistema de reconhecimento de locutor é preciso avaliar o desempenho desse baseado nos seguintes aspectos:

* Qualidade do Sinal de Voz

Características do microfone e do canal de transmissão, tipo e nível de ruído ambiente e variação entrem os sinais de treinamento e teste.

* Modalidade de texto

Dependente ou independe de texto.

* Duração das locuções

Duração e número de sessões de treinamento e teste.

* População de locutores

Número de locutores e similaridade entre os locutores.

## Aplicações

Sistemas de reconhecimento de locutor têm aplicações em diversas áreas. Na realidade, qualquer aplicação que envolva biometria pode potencialmente utilizar esses sistemas. Dentre as aplicações mais comuns encontra-se:

* Autenticação de transações comerciais

Esses sistemas visam autenticar o locutor em chamadas telefônicas, validar transações comerciais com pagamento através de cartão de credito e validar aplicações financeiras por telefone.

* Controle de acesso

Esses sistemas visam controlar o acesso de pessoas a ambientes físicos e sistemas, via identificação ou verificação utilizando a voz.

* Monitoramento

Sistemas de identificação podem ser usados no monitoramento de chamadas. Assim pode-se monitorar *call-centers*, registrando o atendente e o tempo gasto nas chamadas; monitorar as chamadas feitas dentro da prisão, identificando os contados dos prisioneiros; e entre outros tipos de monitoramento.

* Fonte de informações

Sistemas de identificação que buscam identificar os trechos de cada locutor em arquivos que contém conversas de vários locutores.

## Visão geral do trabalho

Dentro desse contexto, buscou-se neste trabalho, investigar sistemas de reconhecimento automático de locutor, obtendo como resultado da pesquisa um protótipo de identificação de locutores independente de texto. Como base para construção desse protótipo foram escolhidos algoritmos que transmitem a atual tendência dos sistemas de reconhecimento automático de locutor. Assim foram escolhidos o algoritmo de extração de características, MFCC, e o algoritmo de reconhecimento de padrões, GMM, que serão explicados ao longo deste trabalho.

O trabalho está dividido em sete capítulos descritos a seguir:

* O capítulo 2 tem por finalidade discorrer sobre o histórico e os conceitos e estrutura básica dos sistemas de reconhecimento de locutor.
* O capítulo 3 apresenta as principais etapas do processamento inicial do sinal de fala, que objetiva a extração de parâmetros do sinal de voz capazes de diferenciar de forma eficiente as características de cada pessoa.
* O capítulo 4 apresenta a teoria dos modelos misturas gaussianas, juntamente com aspectos relacionados com a implementação de um sistema de reconhecimento de locutor baseado nestes modelos.
* O capítulo 5 mostra particularidades na implementação de um sistema de identificação de locutor.
* O capítulo 6 apresenta os experimentos e os resultados obtidos com a construção do sistema de identificação de locutor
* O capítulo 7 apresenta uma conclusão a respeito do trabalho realizado e possibilidades futuras a partir do conhecimento gerado por este.

# Reconhecimento de Locutor

A maioria de nós está consciente do fato de que vozes de indivíduos diferentes não soam iguais. Foi a partir disso que os estudos sobre reconhecimento de locutor começaram. Assim o reconhecimento de locutor busca diferenciar pessoas através da voz e é uma área interdisciplinar que envolve diferentes tipos de conhecimento como processamento de sinais, reconhecimento de padrões, física acústica, linguística, matemática, inteligência artificial, entre outros. Neste capítulo será abordado um breve histórico nessa área e mostrados os conceitos e estrutura usados para construção de reconhecimento automático de locutor.

## Breve Histórico

Durante muito tempo vêm sendo realizadas pesquisas em reconhecimento de pessoas através da voz. Um dos primeiro estudos realizados nessa área retomam a década de 1930, onde o estudo em questão realizado por Dr. Francis McGehee, professor de psicologia da Universidade Johns Hopkins, buscou explorar o quão bem uma pessoa pode identificar pessoas, ainda quenão sejam familiares, apenas escutando suas vozes[3].

Durante a Segunda Guerra Mundial houve um interesse maior em identificar pessoas pela voz. Com esse intuito, o Bells Laboratories inventou, em 1941, uma máquina capaz de fazer o espectrograma da voz, onde era esperado que esse espectrograma ajudasse a identificar vozes de alemães interceptadas por rádio [4]. Os resultados adquiridos naquela época para identificação de locutores com espectrogramas não foram satisfatórios e foram deixados por um bom tempo de lado.

Em 1962, Lawrence Kersta, um dos inventores do espectrógrafo de som, publicou na revista *Nature* o trabalho “Voiceprint Identification” [5]. Nesse trabalho ele defendia a infalibilidade de seu método e relatava taxas de identificação corretas de 99%. Contudo, seus métodos e resultados foram controvertidos e sua aceitação na comunidade científica em geral foi restrita [6]. Apesar das controvérsias, seus estudos foram o pontapé inicial para área de reconhecimento de locutor.



Figura 1 - Engenheiro Lawrence G. Kersta um dos inventores do espectrógrafo de som (sound spectrograph).[Forensic]

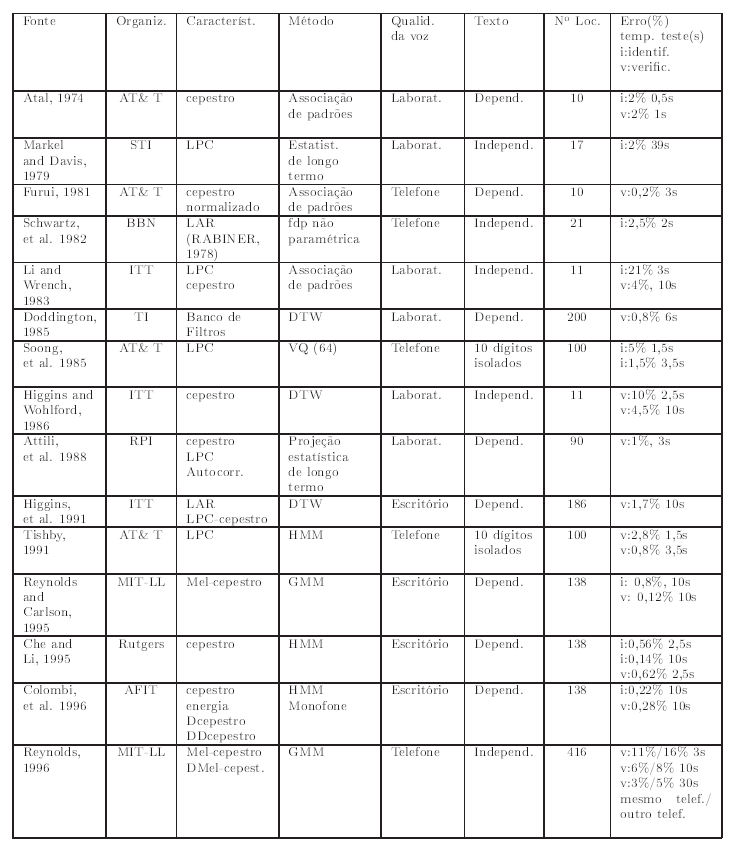
Até agora pôde se observar que o processo de reconhecimento de locutor era uma tarefa executada manualmente, através da comparação visual de espectrogramas por um especialista treinado. A primeira técnica de reconhecimento automática de locutor surgiu em 1963, com Pruzansky no Bell Laboratories [7]. Nesse sistema foi usado banco de filtros e dois espectrogramas digitais para medir a similaridade. Ainda durante a década de 1960 e 1970 surgiram vários outros sistemas baseados em analisar a evolução temporal de certos parâmetros da voz, especialmente da frequência fundamental, formantes, intensidade, e coeficientes do preditor linear ([8],[9],[10],[12],[13]). Vale salientar que esses sistemas eram dependentes do texto, onde o reconhecimento era feito com base na pronúncia de um texto pré-definido. Ainda nessa época surgiu a abordagem de utilizar a análise das médias dos parâmetros de trechos longos de voz, como em S. Furui et al em 1972 [14], S. Furui em 1974[15] ; Markel e Davis em 1979[16], sendo os primeiros métodos independentes de texto. Dentro dessa época podemos dar ênfase aos trabalhos de Atal em 1974 [10] e Furui em 1981 [17] que demonstraram a superioridade da representação cepstral (cepstrum) frente a diversos outros tipos de parâmetros, tabela 1.

Com o aumento do poder computacional, na década de 1980, as técnicas de reconhecimento de locutor ficaram progressivamente mais complexas, proporcionando melhorias de desempenho dos sistemas. Em 1985 Song et al [18], após verificar o sucesso da utilização de técnicas de Quantização Vetorial (Vector Quantization – VQ) no reconhecimento de fala, propuseram um sistema baseado nesse tipo de técnica. Ainda que os experimentos com VQ demonstrassem bons resultados, em geral, havia uma restrição quanto ao tamanho do vocabulário, devido à própria característica da modelagem [19]. Visando suprir isto surgiram as modelagens probabilísticas.

A partir da década de 1990, houve uma grande popularização dos sistemas baseados em modelagens probabilísticas. Entre eles se destacam os modelos ocultos de markov (Hidden Markov Models – HMM) e os modelos de misturas gaussianas (Gaussian Mixture Models – GMM). O HMM é um modelo estatístico baseado em cadeias de markov que incorporam informações sobre a evolução temporal dos parâmetros, sendo bastante utilizado tanto para o reconhecimento de voz como para o reconhecimento de locutor em modo texto dependente. Sua utilização em reconhecimento de locutor pode ser vistos em Rosemberg et al 1990[22], Webb et al 1993[23] , Che and Lin 1995[24] e Colombi et al 1995[25]. Apesar de o HMM obter bons resultados para o reconhecimento em modo texto dependente, a informação temporal incorporada nesses modelos não mostrou vantagem nos sistemas de reconhecimento automático de locutor independente do texto.

O GMM também é um modelo estatístico, mas diferente do HMM, ele não leva em consideração a relação temporal, basicamente os GMM são HMM desprovidos da informação temporal. Assim sendo, o GMM está sendo amplamente utilizado em sistemas de reconhecimento automático de locutor independente do texto. A utilização do GMM em sistemas de reconhecimento automático de locutor foi introduzida por Reynolds, em 1992[20], na sua tese de doutorado e posteriormente publicado em 1995[21].

Tabela 1 – Cronologia reconhecimento de locutor [29]



## Reconhecimento automático de locutor

O reconhecimento automático de locutor (RAL) é o processo de reconhecer automaticamente uma pessoa, utilizando meios computacionais, com base nas características do sinal de voz. Mas antes de discutir sobre RAL, é preciso conhecer os conceitos de verificação de locutor e identificação do locutor.

Nos sistemas RAL, há dois tipos de tarefas: Identificação e Verificação. A identificação é o processo de identificar uma pessoa dentro de um conjunto de N locutores, onde, por definição, o conjunto é considerado fechado quando envolve N decisões ou aberto quando houver N + 1 decisões (decidindo se a voz pertence a um dos N locutores ou a nenhum deles)[20]. O desempenho de um sistema de identificação é degradado com o acréscimo do número de locutores.

A verificação consiste em verificar se uma dada voz pertence ou não a uma pessoa. Essa decisão é feita no denominado conjunto aberto de locutores[20], pois o reconhecimento é efetuado em um grupo de locutores desconhecidos, impostores. Segundo Jayant [26], o desempenho de um sistema de verificação não depende do número de locutores.

Quanto ao texto, o RAL pode ser dependente ou independente de texto. Onde, como já explicado, os sistemas dependente do texto requer que o locutor forneça locuções, ou palavras-chave, com o mesmo texto para ambos o treinamento e o reconhecimento. Enquanto que para os sistemas independentes de texto não é requerido que a locução fornecida para treinamento e teste seja o mesmo texto, ou palavra-chave.

Na construção de um sistema de reconhecimento de locutor é importante destacar o acontecimento de duas fases distintas. A primeira é chamada fase de treinamento ou de cadastramento, enquanto que a segunda é referida como fase de operação ou teste. Na fase de treinamento são coletadas amostras de voz de cada locutor para que o sistema possa gerar modelos referentes a cada locutor. Durante a fase de operação, as amostras de voz de testes são “comparadas” com os modelos gerados na fase de treinamento e a partir disso o sistema toma uma decisão.

Sendo assim, no mais alto nível, todos os sistemas de reconhecimento de locutor contêm dois módulos principais: Extração de características e Classificação [27], Figura 2. O módulo de extração de características é o responsável por extrair um conjunto de características do sinal de voz. Dentre as técnicas mais usadas se destaca a extração de coeficientes mel-cepstrais (*Mel-Frequency cepstral coefficient* – MFCC) [28], nesse trabalho foi usado o MFCC como método de extração de características e será descrito com mais detalhes na seção 3.

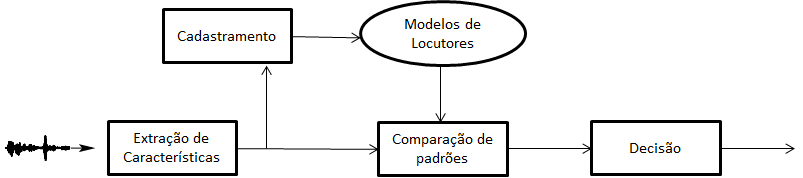


Figura 2 - Estrutura Básica de Sistemas de Reconhecimento Automático de Locutor

O módulo de classificação é responsável tanto por gerar os modelos dos locutores, fase de treinamento, quanto por classificar estes, fase de operação. Na fase operação, o módulo de classificação pode ser divido em dois blocos: Comparação de Padrões e Decisão. O modulo de comparação de padrões é responsável por comparar as características estimadas com os modelos dos locutores. Já o módulo de decisão é encarregado de tomar a decisão que varia dependendo do tipo de tarefa, verificação ou identificação. Na Figura 3 e 4, podemos ver uma estrutura básica para sistemas de identificação e verificação de locutor.

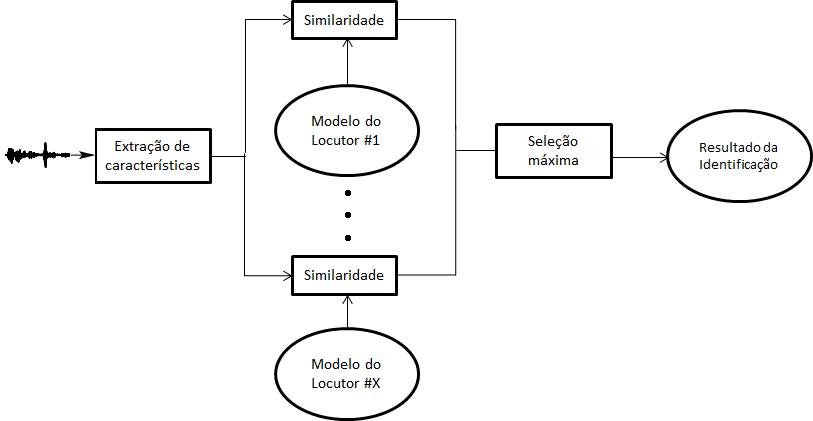


Figura 3 - Estrutura Básica de Sistemas de Identificação de Locutor

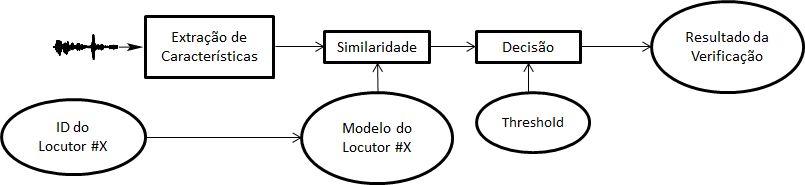


Figura 4 - Estrutura Básica de Sistemas de Verificação de Locutor

Nesse trabalho, foi utilizado como método de extração o MFCC e como método de classificação e geração de modelos, o GMM. Esses dois algoritmos foram escolhidos pela sua alta aceitação da comunidade acadêmica e por demonstrarem bons resultados. Ambos serão explicados ao longo desse trabalho.

# Extração de Características

A fim de se obter as características da voz referentes a cada locutor, a voz passa por dois processos: Digitalização e Processamento. Na Digitalização é obtido o sinal de voz no formato digital, a fim de possibilitar o processamento digital do mesmo. Já no Processamento, o objetivo é extrair as características do sinal a fim de obter um conjunto de parâmetros que possam seguir as seguintes características [QinJin]:

•Ocorrer naturalmente e com frequência no discurso normal

•Ser facilmente mensurável

•Ter alta variabilidade entre locutores

•Ser consistente para cada locutor

•Não mudar com o tempo ou ser afetado pela saúde do locutor

•Não ser afetado pelo ruído de fundo, nem ser dependente de um tipo de transmissão específica.

•Mostrar resistência para distinguir imitações.

Na prática, nem todas essas características são satisfeitas pelas atuais extrações de características. Nas próximas seções serão abordados o processo de digitalização e processamento de voz utilizando os Coeficientes Cepstrais de Frequência Mel ( Mel Frequency Cepstral Coefficient – MFCC).

## Digitalização

O processo de digitalização é responsável pela captura e digitalização (discretização do tempo e da amplitude) do sinal de voz, figura 5. Uma vez amostrado e discretizado, o sinal de voz pode ser processado digitalmente.

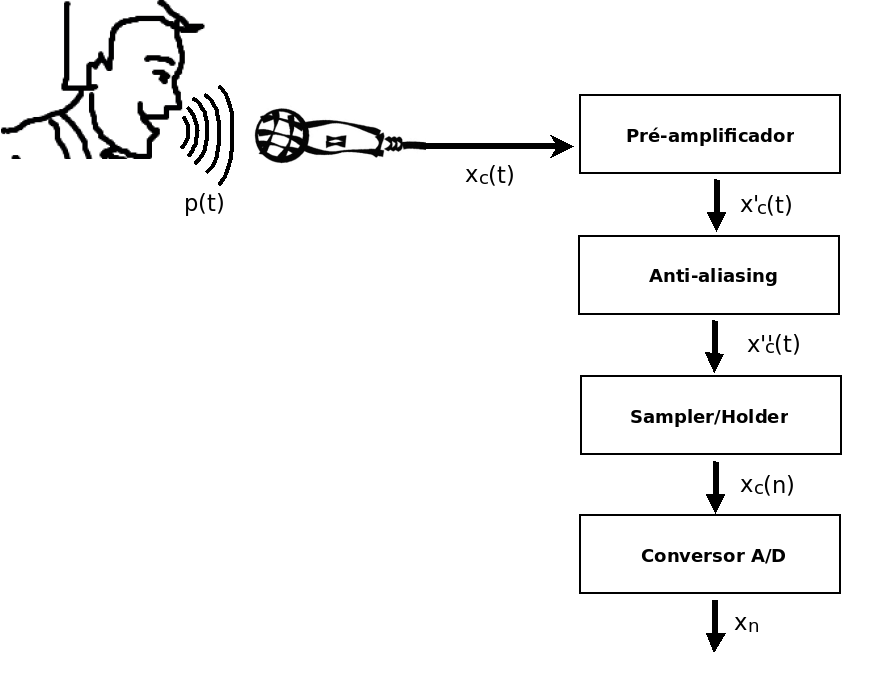


Figura 5 - Fluxograma de Digitalização

**Microfone:** Equipamento usado para captação de ondas sonoras transformando-as num sinal elétrico analógico. Assim, os sinais emitidos pelo locutor são convertidos em sinais analógicos , onde t representa o tempo.

**Pré-amplificador**: Filtro analógico de ganho positivo na entrada .

**Anti-aliasing**: Filtro analógico que limita o sinal em frequência evitando assim a ocorrência de *aliasing*, limitando as frequências superiores à metade da frequência da amostragem, critério de Nyquist [32].

**Sample/Holder**: Amostra (*Sampler*) o sinal em intervalos Tc, com frequência de amostragem fc = 1 / Tc. O instante de amostragem t é discretizado, sendo dado por t = nTc, onde n é o índice de tempo discreto da amostra. O sinal é mantido estável (*holder*) dutante o intervalo necessário para conversão A/D. O sinal, a partir de então, é representado por .

**Conversor A/D**: Recebe os sinais amostrados e os quantiza com uma determinada resolução, geralmente 8,12 ou 16 bits, gerando o sinal x(n).

## Coeficientes Cepstrais de Frequência Mel

Os Coeficientes Cepstrais de Frequência Mel (MFCC) é um método de extração de características que provê uma maneira eficiente de representar as características de frequência dos sinais de fala. Segundo O'Shaughnessy [33], uma das principais técnicas de extração de características é a extração MFCC. Os parâmetros MFCC são baseados na percepção do sistema auditivo humano, em que a percepção para as frequências sonoras não segue um escala linear. O processo de extração dos MFCC é resumido de acordo com a figura 6.

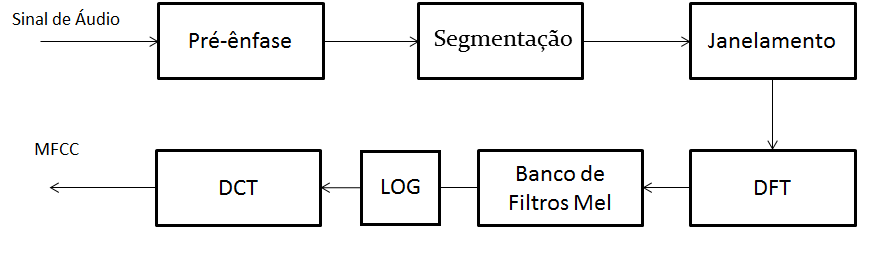


Figura 6 - Fluxograma para obtenção dos MFCCs

### Pré-ênfase

Quando falamos, os sinais de altas frequências são atenuados pelos lábios e pela glote. Em função dessa perda é preciso reverter esse processo; para isso utiliza-se um filtro de pré-ênfase que incrementa as altas frequências. Isso pode ser feito através de um filtro digital cuja função transferência é:

,

No domínio do tempo, o filtro é implementado da seguinte maneira:

,

Usualmente, o valor de  é igual a 0.95 [1]. Na figura 7 é ilustrada a resposta em frequência desse filtro com . Nele podemos confirmar a ação do filtroao enfatizar as altas frequências.

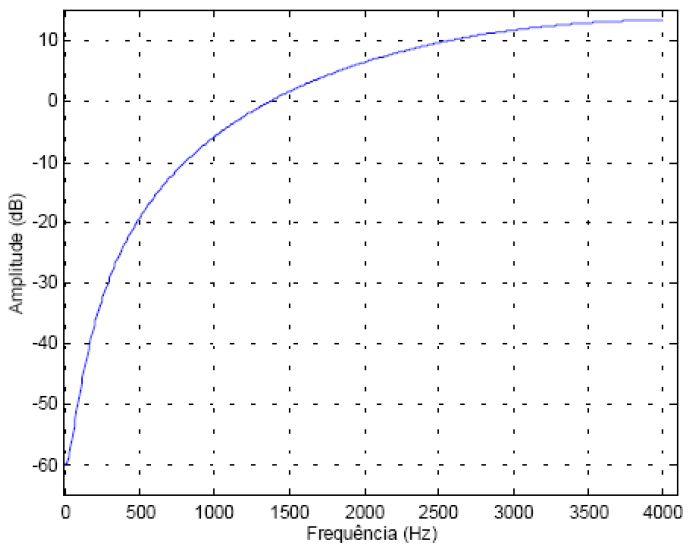


Figura 7 - Resposta em frequência do filtro de pré-ênfase

A figura 8 mostra o espectrograma de sinal de fala “mint chocalate chip” antes e depois da pré-ênfase.

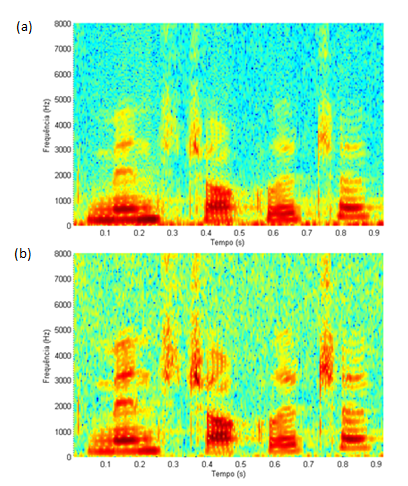


Figura 8 - Espectrograma do Sinal Original (a) e do Sinal pré-enfatizado (b).

### Segmentação

Os parâmetros do sinal de voz podem ser considerados invariantes no tempo para curtos intervalos de tempo, da ordem de 10 a 30ms[23]. Para isso, o sinal é divido em segmentos (quadros) de tamanho fixo N com superposição entre os segmentos M, figura 9, onde essa superposição geralmente é da ordem de 25% a 50% de N. Essa superposição é feita, pois podem surgir descontinuidades bruscas em suas extremidades, o que viria a prejudicar a avaliação dos elementos próximos a essa extremidade.

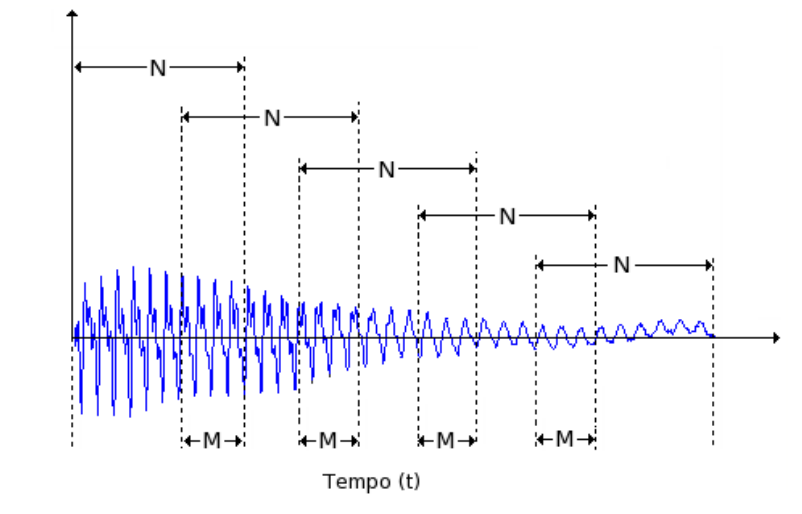


Figura 9 - Segmentação do sinal de voz

### Janelamento

Logo após o processo de segmentação e antes de utilizar a análise espectral é preciso utilizar a técnica de janelamento que busca minimizar as descontinuidades do sinal no começo e no final de cada segmento e admitir que ele seja aproximadamente estacionário nesse intervalo. Essa técnica consiste em multiplicar cada amostra de um quadro por uma função janela, assim se o quadro apresentar amostras x(n) e a função janela for dada por w(n), temos:

Há diversas funções de janelamento como Hamming, Hanning, Triangular, Kaisser-Bessel e etc. No entanto, a mais frequentemente utilizada é janela de Hamming na qual pode ser definida pela a seguinte equação:

sendo N o tamanho do quadro. Na figura 10 é mostrada a função de janelamento de Hamming no domínio do tempo e da frequência.

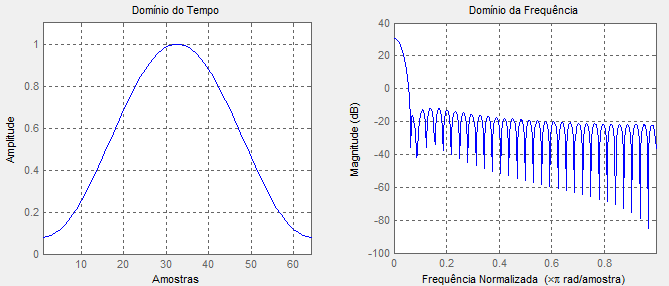


Figura 10 - Janela de Hamming

### DFT

Após o janelamento, converte-se cada segmento do domínio do tempo para o domínio da frequência através do algoritmo FFT(*Fast Fourrier Transform*) que é um algoritmo rápido para calcular a Transformada Discreta de Fourrier(DFT) definida por:

Ao aplicar a DFT para cada segmento são produzidos segmentos contendo N/2 componentes de espectro de potência. A explicação para ser apenas metade é devido ao sinal de entrada ser composto por números reais e o espectro tornar-se simétrico à frequência de *Nyquist*, que é metade da frequência de amostragem (fc/2).

### Banco de Filtros Mel

A ideia principal do processo de extração de características MFCC é inserida nessa etapa, onde o espectro é escalado na escala *mel*. A escala *mel* foi definida por [30] como uma escala psicoacústica da sensibilidade do ouvido para diversas frequências do espectro audível. Assim um *mel* é uma unidade de frequência percebida  para uma determinada frequência recebida , onde essa relação entre a e pode ser observada na figura 11 e pode ser aproximada por:

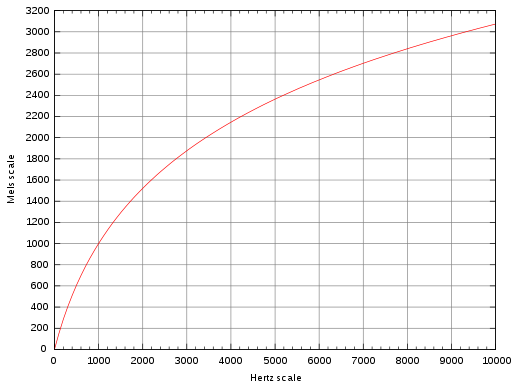


Figura 11 – Gráfico de Frequência Mel x Frequência Hetz

A conversão de cada segmento do domínio da frequência para a escala mel geralmente é implementada por um banco de filtros triangulares. Esse banco de filtros triangulares é projetado de tal forma que as frequências centrais dos filtros triangulares estão espaçadas linearmente segundo a escala mel, Figura 12.

Um sinal, quando submetido a um filtro triangular, tem as componentes que estão próximas ao centro deste filtro enfatizadas e as demais, atenuadas. Dessa forma, ao se empregar um banco de filtros o que está se fazendo é enfatizar as frequências mel, assim escalando o espectro na escala mel.

Há diversas propostas de implementação de bancos de filtros triangulares. Essas implementações diferem principalmente no número de filtros, na forma dos filtros, na forma como os filtros são espaçados, na largura de banda dos filtros, e na maneira em que o espectro é escalado. Diante disso foi feito um estudo[31] para verificar várias propostas de implementações de banco de filtros. No banco de filtro que sobressaiu foram utilizados 40 filtros triangulares, sendo os 13 primeiros espaçados linearmente até 1000hz e os 27 seguintes, espaçados logaritmicamente entre si. As frequências centrais de cada filtro são dadas:

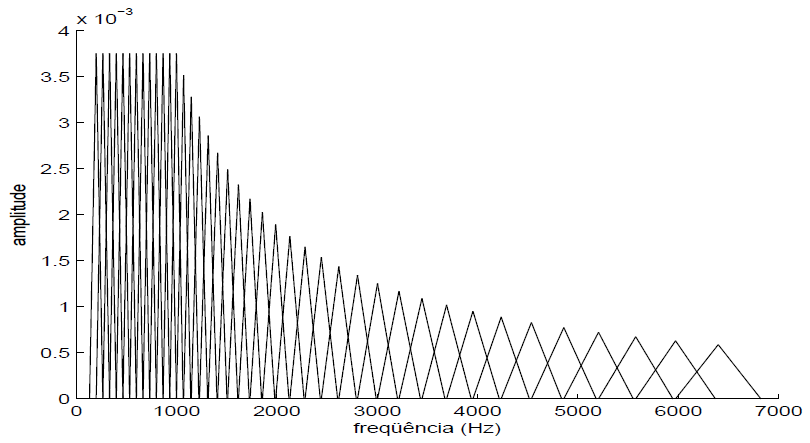


Figura 12 - Banco de filtros triangulares escalados linearmente na escala Mel

Cada filtro triangular é dimensionado de forma a apresentar uma mesma área. Assim pode-se observar que a altura é a mesma para os filtros linearmente espaçados e diferentes para os filtros logarítmicos.

### Log e DCT

Nessa etapa final, converte-se o logaritmo do espectro mel de volta ao domínio do tempo. Como o logaritmo do espectro mel são números reais, eles podem ser convertidos usando a transformada discreta do cosseno (*Discrete Cosine Transform* -DCT) que pode ser calculada através de:

X_k =
 \sum_{n=0}^{N-1} x_n \cos \left[\frac{\pi}{N} \left(n+\frac{1}{2}\right) k \right] \quad \quad k = 0, \dots, N-1.

Ao final de todo esse processo temos, enfim, os coeficientes cepstrais de frequência mel (MFCC).

# Modelos de Misturas Gaussianas

Os Modelos de Misturas Gaussianas (*Gaussian Mixture Models* – GMM) é um modelo estocástico que se baseia num conjunto ponderado de funções de densidade de probabilidade, no caso a gaussiana. A distribuição gaussiana é apropriada para modelagem de eventos que possuem um comportamento definido em torno de um ponto médio, como na Figura 13.

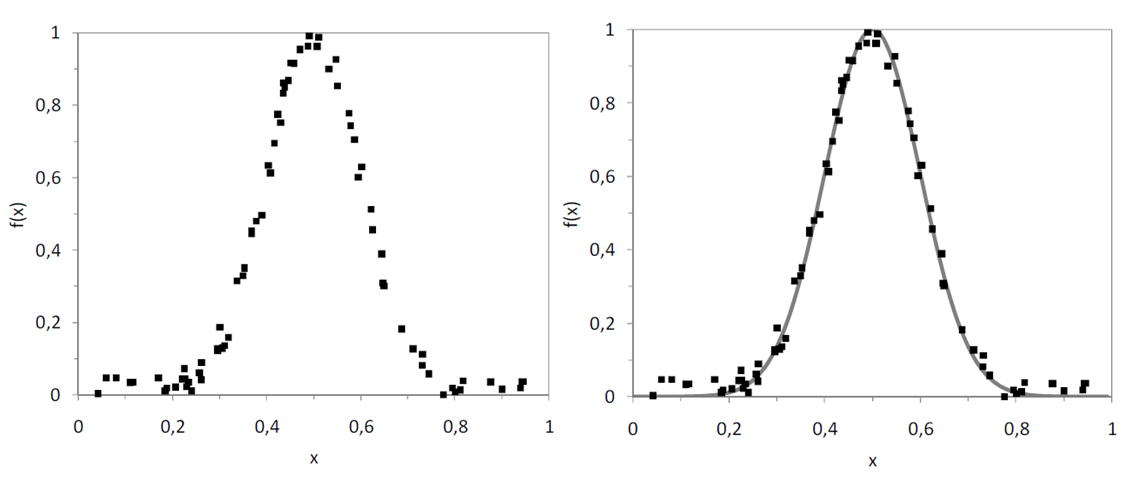


Figura 13 – Distribuição de dados em torno de um ponto médio e modelagem através de uma gaussiana

Contudo para problemas mais complexos a utilização de apenas uma única gaussiana não é suficiente para modelagem, como na figura 14.

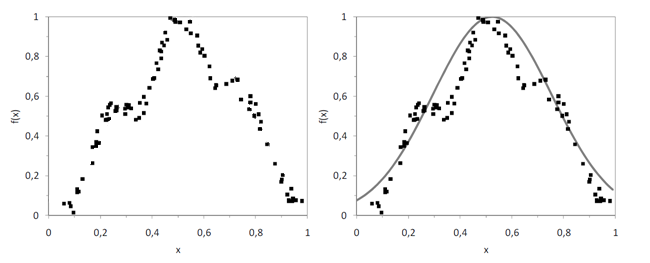


Figura 14 – Distribuição de dados num nível mais complexo e modelagem através de uma gaussiana

Assim em problemas mais complexos a utilização de conjunto de gaussianas gera melhores resultados, como na figura 15.

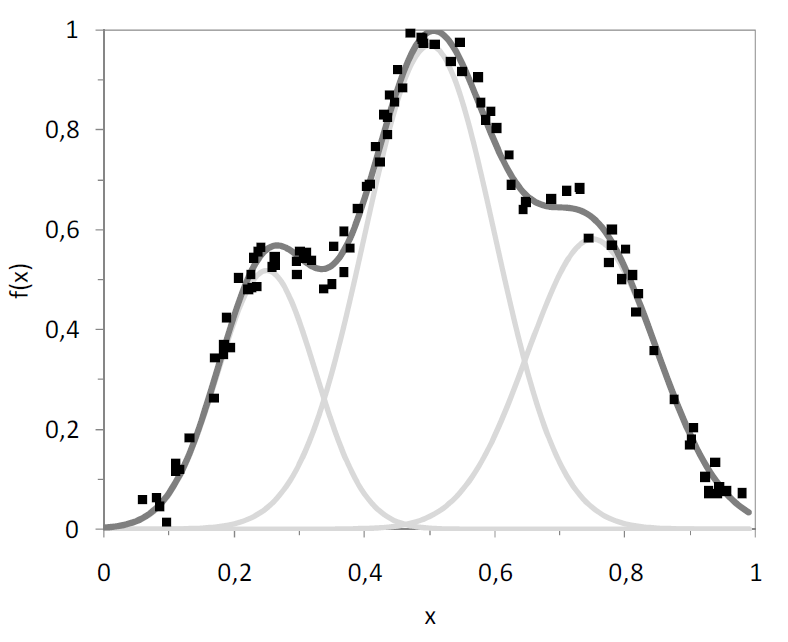
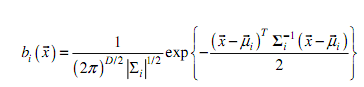


Figura 15 – Distribuição de dados num nível mais complexo e modelagem por várias misturas gaussianas

Assim o GMM é definido como uma densidade formada pelo somatório ponderado de M densidades gaussianas e pode ser representado por:

em que i = 1,... , M, é o vetor de características de D dimensional, é o peso da componente i e é a função de densidade de probabilidade da componente i. Cada densidade componente do GMM, , é uma densidade gaussiana D-dimensional definida por:



com vetor de média e matriz de covariância .

Por definição as funções de densidade de probabilidade quando integradas no intervalo apresentam um resultado igual a 1. De tal modo, é possível afirmar que a soma dos pesos devem ser unitária, conforme o desenvolvimento da equação:

Assim o GMM pode ser representado pela média, matriz de covariância e pesos de cada componente.

## GMM para Reconhecimento de Locutor

A utilização do GMM em sistemas de reconhecimento automático de locutor foi introduzida por Reynolds em 1992[11]. Existem dois motivos principais para sua utilização em reconhecimento de locutor. O primeiro deles é dado pela noção intuitiva de que as componentes individuais de cada densidade gaussiana podem modelar conjuntos não distinguíveis de classes acústicas. Estas classes acústicas, representadas por eventos fonéticos, tais como, sons vogais, nasais e fricativos, que refletem a dependência do locutor com o seu trato vocal. O segundo motivo é uma das características mais poderosas do GMM que é sua capacidade de aproximação para formar densidades de probabilidades desconhecidas, o que pode ser observado nos exemplos anteriores, Figura 14 e 15. Assim o GMM representa de forma geral a dependência das características espectrais da voz associadas ao locutor, em conjunto com a capacidade de modelar densidades de probabilidades desconhecidas[11].

Portanto nos sistemas RAL, a voz de cada locutor é modelada por um GMM distinto, dando origem a um modelo , no qual o universo de locutores pode ser representado por:

{

### Treinamento

Para cada locutor a ser modelado, é preciso que cada um possua um conjunto de vetores de características de treinamento. Esses são obtidos através de arquivos de áudio contendo gravações de voz, onde para cada arquivo são calculados os diversos vetores de características , para diferentes instantes de tempo t. Assim podemos representar esse conjunto de vetores por:

Diante desse conjunto de vetores extraídos do sinal de voz, o objetivo do treinamento é estimar os parâmetros do GMM de modo que melhor corresponda a distribuição desses vetores. Essa estimação é feita de forma a maximizar a verossimilhança :

Considerando para simplificação matemática a independência entre os vetores de características temos:

Assim através de , seria dado os subsídios necessários para obtenção do modelo de máxima verossimilhança. Porém a expressão é uma função não linear dos parâmetros de , oque impede uma maximização direta. Então para tal é utilizado um algoritmo iterativo de estimação de maximização da verossimilhança, o Expectation-Maximization (EM).

### Expectation-Maximization(EM)

O algoritmo Expectation-Maximization foi primeiramente descrito por Dempster, Laird e Rubin (1977). Esse algoritmo garante que o modelo atual esteja mais correlacionado com o conjunto de observações X em comparação ao modelo da iteração anterior , assim observa-se a propriedade:

O algoritmo EM é comumente descrito como contendo duas etapas. Na primeira etapa E (expectation) são calculados os valores da probabilidade, chamada de probabilidade a posteriori:

Na segunda etapa M (maximization) é feita a maximização dos modelos atuais a partir dos parâmetros do modelo anterior, que é feita pelas seguintes equações:

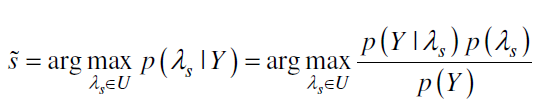
onde o índice [n+1] denota os parâmetros atualizados, modelo atual, e o índice [n] denota os parâmetros antigos, modelo anterior.

Visando melhorar a performance, são utilizado apenas os elementos da diagonal principal da matriz de covariância . Essa simplificação é feita pois é muito custoso calcular a matriz inversa da matriz de covariância na equação . Segundo [1], essa simplificação não afeta o rendimento do sistema. Assim o cálculo para atualização da matriz de covariância fica:

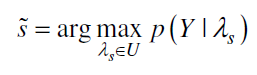
onde representa o vetor de variância que compõe a diagonal da matriz de covariância.

### Identificação com GMM

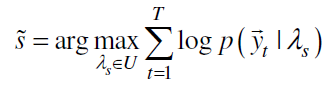
Na identificação, deseja-se descobrir o modelo de locutor mais verossímil dada uma amostra de teste Y = . Para isso é necessário determinar qual dos modelos,, do universo, U, tenha a máxima probabilidade a posteriori:



onde a segunda parcela da equação corresponde a regra de Bayes. Supondo que todos os locutores são igualmente prováveis (p() = 1/S) e que p(Y) é a mesma para todos os locutores, temos:



Admitindo a independência entre os elementos da amostra de testes e usando o logaritmo para evitar problemas numéricos, temos :



Por fim, o locutor que maximiza esta expressão, será o locutor identificado e será considerado correto se de fato o locutor reconhecido é o locutor identificado.

### Verificação com GMM

Na verificação, deseja-se descobrir se uma dada amostra de teste Y = é ou não pertencente a um determinado locutor, . Então o sistema tem que fazer uma escolha entre , onde:

Para que se decida entre ou , é preciso calcular um score que exprima o quão bem a amostra de teste pertence ao locutor. Esse score é calculado através da razão entre a verossimilhança da amostra “pertencer ao locutor” sobre a verossimilhança da amostra “não pertencer ao locutor”.

Usualmente para obter tal razão de verossimilhança, é empregado um modelo do universo de possibilidades falsas, denominado *background*, que é composto por um conjunto de falsos locutores (impostores).

Assim dada uma amostra de teste Y = , com um modelo do pretenso locutor e um modelo não pertencente ao pretenso locutor, a razão de verossimilhança é dada por:

Aplicando a regra de Bayes e supondo a independência entre os elementos de Y, temos que o logaritmo da razão de verossimilhança é:

onde é a verossimilhança da amostra de teste do pretenso locutor e é a verossimilhança dado um modelo de impostores(background).

No caso da verificação, o cálculo do log da verossimilhança para qualquer modelo é dado por:

onde é utilizada a normalização, , de acordo com a duração da locução, para a duração da locução não afetar o score.

Para obter a verossimilhança dado um modelo de impostores(background), é preciso primeiramente gerar o modelo de background, que segundo [39] pode ser montado de duas maneiras:

* Utilizando vários modelos individuais de locutores (GMM), escolhidos de acordo com certas especificações, para formar o conjunto de impostores. Supondo um conjunto com K locutores, o log da verossimilhança para os impostores ficará:

onde é utilizada a normalização para normalizar a verossimilhança de acordo com o número de modelos utilizado no background.

* Utilizando vários locutores em um único modelo. Assim são utilizadas as características de vários locutores para gerar um GMM, . Esse é denominado modelo Universal de Background.

Após o cálculo da razão de verossimilhança , esse é comparado a um limiar ( *threshold* ) onde é aceito se >  e rejeitado se < . Esse limiar pode ser dependente de locutor, quando há um limiar para cada locutor, ou pode ser um limiar global, quando há um limiar global no sistema. A escolha do limiar é puramente experimental e é feita na fase de treinamento e validação. Um diagrama para um sistema de verificação baseado em modelos de misturas gaussianas é mostrado abaixo:

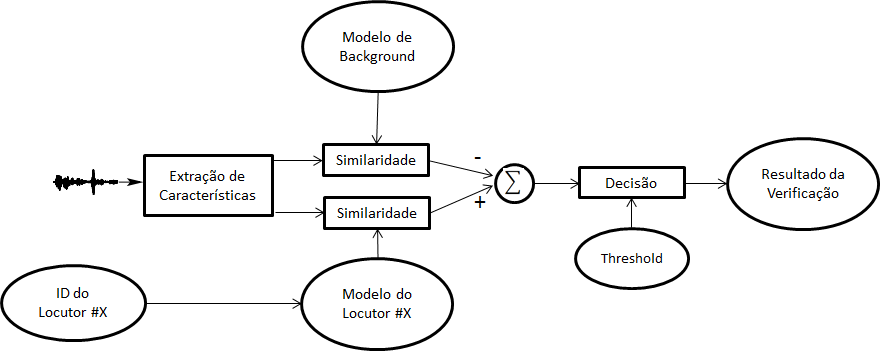


Figura 16 – Estrutura básica para um sistema de verificação do locutor baseado em modelos de misturas gaussianas.

# O Sistema de Identificação de Locutor

Diante da teoria apresentada, anteriormente, este trabalho se propôs a fazer um protótipo de um sistema de identificação de locutor baseados nas técnicas MFCC e GMM. Assim podemos representar o seguinte protótipo através do diagrama:

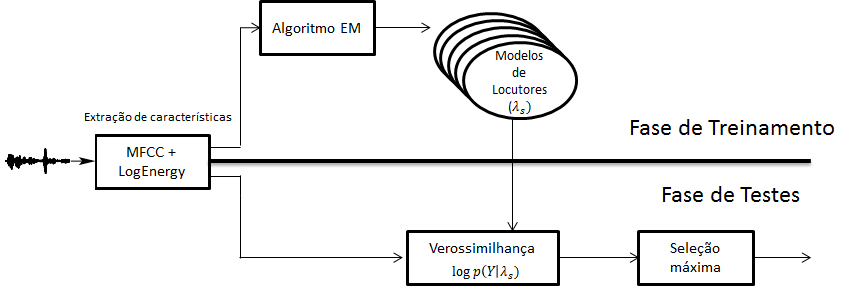


Figura 17 – Diagrama do Sistema de identificação de locutor

A construção do protótipo foi baseada na teoria apresentada nos capítulos anteriores e sua implementação foi feita no MATLAB.

Neste capítulo serão apresentadas algumas particularidades de implementação que não foram abordadas na teoria. A primeira está relacionada com a aquisição do sinal de fala. A segunda se refere aos parâmetros MFCC e a característica adicional extraída. A terceira está relacionada ao algoritmo EM: inicialização e critério de parada.

## VAD

Geralmente quando é adquiro o sinal de voz, há uma quantidade incerta de silêncio no inicio e no final de cada aquisição. Segundo [35], o silêncio pode degradar o processo de treinamento, ou seja, o processo de ajustes do GMM. Assim faz-se necessária a retirada do silêncio, que pode ser feito através de um algoritmo de ativação por voz ( *Voice Activion Detection* - VAD). Apesar do VAD conseguir tirar uma boa parte do silêncio, uma pequena porção é inevitável. Porém, segundo [35], um GMM é capaz de lidar com pequenas quantidades de silêncio sacrificando algumas misturas gaussianas para modelar a classe silêncio. Neste trabalho foi utilizado um VAD baseado num modelo estatístico para detecção da voz[34], disponibilizado pelo *toolbox* VOICEBOX[36].

## MFCC + LogEnergy

Para gerar o conjunto de vetores de características do sinal de voz, foi utilizado o algoritmo de extração dos Coeficientes Cepstrais de Frequência Mel. Esse foi implementado de acordo com o Capítulo 3, gerando 12 MFFCs a cada 20ms. Visando o acréscimo de mais uma característica é comum utilizar uma informação de energia. Assim neste trabalho foi utilizado o logaritmo da energia que é definida por:

## Algoritmo EM

O Expectation-Maximization (EM) é um algoritmo iterativo, em que é preciso um modelo da iteração anterior para calcular o modelo atual. Assim o EM precisa de uma inicialização (iteração 0), ou seja, precisa-se inicializar um modelo , com seus pesos, médias e covariâncias. Após inicializado, o algoritmo é executado até atingir um número máximo de iterações ou quando é estabelecido algum critério de parada. Resumidamente o EM é implementado pelos seguintes passos:

Passo 1 : Inicialize o modelo e vá para o passo 2

Passo 2 : Calcule todas probabilidades a posteriori e vá para o passo 3

Passo 3: Utilize o cálculo das probabilidades a posteriori e atualize o modelo , através da seguintes fórmulas:

Passo 4 : Se o número máximo de iterações foi estabelecido ou se algum critério de parada foi estabelecido, pare o algoritmo, senão faça e vá para o passo 2.

### Inicialização

A inicialização pode ser feita através de duas formas: aleatoriamente ou utilizando algum algoritmo de clusterização. Na primeira forma, as médias de cada componente são inicializadas aleatoriamente, escolhendo-se aleatoriamente G vetores de Coeficientes Cepstrais de Frequência Mel, derivados da base de treinamento. Para inicializar a matriz de covariância é utilizada a matriz identidade e os pesos são inicializados uniformemente. Na segunda forma é utilizado um algoritmo de clusterização, usualmente o algoritmo LBG-VQ. Nesse tipo de inicialização as médias do GMM são inicializadas a partir do centro de massa de cada grupo. A matriz de covariâncias é inicializada calculando-se as variâncias através do centro de massa e amostras de cada grupo e os pesos são inicializados uniformemente.

Neste trabalho foram feitos os dois tipos de inicialização, mas optou-se por utilizar a forma randômica visto que os resultados encontrados inicialmente, para um número pequeno de gaussianas, não demonstraram uma expressiva diferença. Assim foi utilizada a forma com o menor custo computacional. Segundo [21], a explicação para a inicialização não ter gerado resultados expressivamente diferentes é que os diferentes modelos iniciais podem ter convergido para diferentes máximos locais. Essa diferença, porém, entre os modelos finais é insignificante em termos de desempenho de identificação do locutor.

### Critério de parada

O EM é um algoritmo de maximização iterativa que garante encontrar um máximo local, o que nem sempre é um máximo global. Assim existem dois critérios de paradas: quando o algoritmo atinge um número máximo de iterações ou quando atinge o critério de parada que, usualmente, é calculado através da diferença relativa do log da verossimilhança entre o modelo estimado e o modelo da iteração, definida pela seguinte equação:

Para uma escolha do valor de um *threshold* mais significativo, optou-se por calcular o valor absoluto da equação acima, definida pela equação:

onde nesse trabalho foi utilizado para os treinamentos o valor de *threshold* igual a 0.0001.

# Experimentos e Resultados

Neste capítulo apresenta-se os testes realizados e os resultados obtidos em cada uma das experimentações usando o GMM implementado para identificação de locutor. Com esses experimentos e resultados podemos chegar a um resultado prático mais conclusivo do que a teoria aqui apresentada. Para tal experimentos, foram utilizadas duas bases de dados de voz, a base de dados ELSDSR e a base do MIT.

## Base de dados

### ELSDSR

A base de dados ELSDSR (English Language Speech Database for Speaker Recognition)[38] foi criada para fornecer dados de voz para o desenvolvimento e avaliação de sistema automático de reconhecimento de locutor. Ela foi criada pelos os docentes do departamento de informática e modelagem matemática (Informatics and Mathematical Modeling – IMM) da Universidade Técnica da Dinamarca (Technical University of Denmark –DTU).

ELSDSR contém 23 locutores: 10 do sexo feminino e 13 do sexo masculino, com idades no intervalo de 24 a 63 anos. A língua falada na base de dados é o inglês, falado por 21 dinamarqueses, um islandês e um canadense, assim a maioria deles são falantes não-nativos de Inglês. Não foi obrigado que a pronúncia das palavras fosse perfeita nem tão pouco corrigidas, visto que objetivo é utilizar fatores realistas e que podem ajudar a identificar a pessoa.

Tabela 2 – Descrição de Idade e Nacionalidade na Base de Dados ELSDSR (Feminino)

|  |  |  |
| --- | --- | --- |
| Locutor Feminino | Idade | Nacionalidade |
| FAML | 48 | Dinamarquês |
| FDHH | 28 | Dinamarquês |
| FEAB | 58 | Dinamarquês |
| FHRO | 26 | Irlandês |
| FJAZ | 25 | Canadense |
| FMEL | 38 | Dinamarquês |
| FMEV | 46 | Dinamarquês |
| FSLJ | 24 | Dinamarquês |
| FTEJ | 50 | Dinamarquês |
| FUAN | 63 | Dinamarquês |
| Média de Idade | 40.6 |  |

Tabela 3 - Descrição de Idade e Nacionalidade na Base de Dados ELSDSR (Masculino)

|  |  |  |
| --- | --- | --- |
| Locutor Masculino | Idade | Nacionalidade |
| MASM | 27 | Dinamarquês |
| MCBR | 26 | Dinamarquês |
| MFKC | 47 | Dinamarquês |
| MKBP | 30 | Dinamarquês |
| MLKH | 47 | Dinamarquês |
| MMLP | 27 | Dinamarquês |
| MMNA | 26 | Dinamarquês |
| MNHP | 28 | Dinamarquês |
| MOEW | 37 | Dinamarquês |
| MPRA | 29 | Dinamarquês |
| MREM | 29 | Dinamarquês |
| MRKO | 26 | Dinamarquês |
| MTLS | 28 | Dinamarquês |
| Média de Idade | 31.3 |  |

As locuções foram pronunciadas através da leitura de textos de treinamento e de testes. No texto de treinamento buscou-se capturar todas as possíveis pronúncias do idioma Inglês, que inclui as vogais, consoantes e ditongos. Assim foram criados sete parágrafos e consequentemente sete arquivos de treinamento (A,B,C,D,E,F e G). Para a base de testes foram pronunciados de 4 a 6 sentenças, retiradas do texto “Nova Home”, vide apêndice A, onde cada locutor pronuncia diferentes partes do texto. Um dos textos de treinamento é apresentado abaixo, para os outros, vide o apêndice A.

A

“Chicken Little was in the woods one day when an acorn fell on her head. It scared her so much she trembled all over. The poor girl shook so hard, half her feathers fell out.”

Em média, a duração para a leitura do texto de treinamento é: 78,7 segundos para o sexo masculino; 88,3 segundos para as mulheres; 82.9 segundos para todos. E a duração para a leitura do texto de testes, em média, é: 16,1 segundos (masculino); 19,6 segundos (feminino); 17,6 segundos (para todos). A Tabela 3 mostra o tempo gasto em leitura, para tanto o texto de treinamento e quanto para o texto de teste.

Tabela 4 – Duração da leitura do texto de treinamento e teste

|  |  |  |
| --- | --- | --- |
| Masculino | Treino (segundos) | Teste (segundos) |
| MASM | 81.2 | 20.9 |
| MCBR | 68.4 | 13.1 |
| MFKC | 91.6 | 15.8 |
| MKBP | 69.9 | 15.8 |
| MLKH | 76.8 | 14.7 |
| MMLP | 79.6 | 13.3 |
| MMNA | 73.1 | 10.9 |
| MNHP | 82.9 | 20.3 |
| MOEW | 88.0 | 23.4 |
| MPRA | 86.8 | 9.3 |
| MREM | 79.1 | 21.8 |
| MRKO | 79.7 | 15.8 |
| MTLS | 66.2 | 14.05 |
| FAML | 99.1 | 18.7 |
| FDHH | 77.3 | 12.7 |
| FEAB | 92.8 | 24.0 |
| FHRO | 86.6 | 21.2 |
| FJAZ | 79.2 | 18.0 |
| FMEL | 76.3 | 18.2 |
| FMEV | 99.1 | 24.1 |
| FSLJ | 80.2 | 18.4 |
| FTEJ | 102.9 | 15.8 |
| FUAN | 89.5 | 25.1 |
| Tempo Médio | 78.7 | 88.3 |

### MIT

The MIT Mobile Device Speaker Verification, aqui referida como MIT, é uma base de dados de voz que foi coletada por meio de pequenos dispositivos móveis usando vários microfones em diversos ambientes. Essa base de dados de voz, que está disponível ao público pelo MIT [37], tem o objetivo de ajudar pesquisas em verificação de locutor para dispositivos móveis em diversos ambientes.

Na base MIT, os dados foram coletados por protótipos de dispositivos móveis fornecidos pela Intel. No intuito de simular os diversos cenários encontrados nos sistemas de verificação de voz pelo mundo real, os dados de fala coletadas consistem em dois conjuntos: um conjunto de usuários inscritos e um conjunto diferente de dedicados impostores. Para o conjunto de usuários inscritos, os dados de voz foram coletados ao longo do curso de duas diferentes sessões.

A fim de captar a variabilidade esperada das condições ambientais e acústicas inerentes ao uso de um dispositivo móvel, foram variadas as condições ambientais e as condições do microfone durante a coleta de dados. Para cada sessão, os dados foram coletados em três locais diferentes (um escritório silencioso, um hall de entrada barulhenta, e um cruzamento de rua movimentado), bem como com dois microfones diferentes (o microfone interno embutido do dispositivo e um fone de ouvido auriculare externo) resultando em seis condições experimentais distintas. Ao gravar em ambientes ruidosos, essa base de dados contém o efeito Lombard (ou seja, alto-falantes alteram seu estilo de fala em condições mais ruidosas em uma tentativa de melhorar a inteligibilidade). O efeito Lombard não se verifica em base de dados que simplesmente adicionam ruído eletronicamente para dados coletados em ambientes silenciosos.

Dentro de cada sessão de coleta de dados, o usuário recitou uma lista de nomes (sabores de sorvetes) que foram exibidos no dispositivo. Exemplos de frases são mostrados na tabela 5. No total, 12 listas diferentes foram criadas para usuários inscritos, enquanto 7 listas foram criadas para impostores. Usuários inscritos recitaram duas listas de frases, que eram idênticas, diferindo apenas na localização das frases com os sabores dos sorvetes nas listas. A primeira lista de frases foi lida na sessão de coleta de dados inicial dos usuários inscritos, enquanto que a segunda lista de frases foi lida num dia posterior à sessão inicial.

Tabela 5 – Exemplo da lista de frases faladas em cada sessão

|  |  |  |
| --- | --- | --- |
| Escritório/Mic.Externo | Hall/Mic.Externo | Cruzamento/Mic.Externo |
| alex park | alex park | alex park |
| rocky road | chocolate fudge | mint chocolate chip |
| kensteele | ken steele | ken steele |
| rocky road | chocolate fudge | mint chocolate chip |
| thomas cronin | thomas cronin | thomas cronin |
| rocky road | chocolate fudge | mint chocolate chip |
| saiprasad | Saiprasad | saiprasad |
| rocky road | chocolate fudge | mint chocolate chip |
| trenton young | trenton young | trenton young |
| Escritório/Mic.Interno | Hall/Mic.Interno | Cruzamento/Mic.Interno |
| alex park | alex park | alex park |
| peppermint stick | pralines and cream | chunky monkey |
| ken steele | ken steele | ken steele |
| peppermint stick | pralines and cream | chunky monkey |
| thomas cronin | thomas cronin | thomas cronin |
| peppermint stick | pralines and cream | chunky monkey |
| saiprasad | Saiprasad | saiprasad |
| peppermint stick | pralines and cream | chunky monkey |
| trenton young | trenton young | trenton young |

No total, em cada sessão foram coletadas 54 amostras de voz por usuário. Isso rendeu 5.184 exemplos de usuários inscritos (2.592 por sessão) e 2700 exemplos de impostores de usuários fora do grupo de inscrição. Com o conjunto de inscritos de 48 locutores, 22 eram do sexo feminino, enquanto que 26 eram do sexo masculino. Para o conjunto de impostor com 40 locutores, 17 eram do sexo feminino, enquanto 23 eram do sexo masculino.

## Experimentos

### ELSDSR

O intuito da utilização da base de dados ELSDSR foi testar o GMM num ambiente controlado com pouco ruído e com um número mediano de locutores. Para isso, os arquivos de testes foram divididos em segmentes de 3 segundos, gerando 111 arquivos de testes. Ademais, buscou-se nos experimentos avaliar o desempenho do GMM do ponto de vista de números de gaussianas e duração do treinamento.

No primeiro experimento, foi utilizada a parte A do treinamento para gerar os modelos dos locutores. Nesse experimento buscou-se variar o número de gaussianas indo de 2 a 1024 com um passo de 2 elevado a 2. Os resultados obtidos foram:

Tabela 6 – Primeiro Experimento ELSDSR

|  |  |
| --- | --- |
| # Gaussianas | Taxa de Acerto |
| 2 | 82,88% |
| 4 | 88,29% |
| 8 | 90,99% |
| 16 | 91,89% |
| 32 | 90,99% |
| 64 | 90,99% |
| 128 | 87,39% |
| 256 | 65,77% |
| 512 | 53,11% |
| 1024 | 26,13% |

Figura 18 – Gráfico: Número de Gaussianas x Taxa de Acerto – Primeiro experimento ELSDSR

Através do gráfico da figura 18, podemos ver que houve um aumento crescente na taxa de acerto com aumento do número de gaussianas no sistema. Porém, a partir de um certo número de gaussianas, no caso 64, é possível visualizar um queda brusca no desempenho dos sistemas. Isso pode ser explicado devido a um problema relacionado ao modelamento através de misturas gaussianas, pois o GMM necessita de um número expressivo de amostras para ajustar suas gaussianas. Assim quanto maior o número de gaussianas, maior será preciso o número de amostras para treinamento.

No segundo experimento, foram utilizadas as partes A, B e C do treinamento para gerar os modelos dos locutores. Para termos uma comparação entre o primeiro experimento e este, buscou-se variar o número de gaussianas da mesma forma em que foi feito o primeiro experimento, indo de 2 a 1024. Os resultados obtidos foram:

Tabela 7 - Segundo Experimento ELSDSR

|  |  |
| --- | --- |
| # Gaussianas | Taxa de Acerto |
| 2 | 95,50% |
| 4 | 96,40% |
| 8 | 98,20% |
| 16 | 98,20% |
| 32 | 99,10% |
| 64 | 99,10% |
| 128 | 98,20% |
| 256 | 98,20% |
| 512 | 94,59% |
| 1024 | 89,19% |

Figura 19 - Gráfico: Número de Gaussianas x Taxa de Acerto – Segundo experimento ELSDSR

Podemos observar que com o acréscimo de mais dados de treinamento houve uma melhora na taxa de reconhecimento. Isso pode ser explicado através do fato de que o conjunto de treinamento utilizado nesse experimento, partes A,B e C, contém um número maior de pronúncias de ditongos, vogais e consoantes do idioma Inglês, assim contendo também mais características relacionadas ao modo de falar de cada pessoa. Com isso, obteve-se uma taxa de acerto de até 99,10%, muito maior do que a obtida no experimento anterior que é de 90,99%. Pode-se observar também que houve uma degradação no reconhecimento quando utilizado muitas gaussianas, mas bem menos expressiva do que a apresentada no experimento anterior.

No terceiro e último experimento, foi utilizado todo o conjunto de treinamento para gerar os modelos dos locutores. Também nesse experimento variou o número de gaussianas da mesma forma do primeiro e segundo experimentos. Os resultados obtidos foram:

Tabela 8 - Terceiro Experimento ELSDSR

|  |  |
| --- | --- |
| # Gaussianas | Taxa de Acerto |
| 2 | 98,20% |
| 4 | 98,20% |
| 8 | 99,10% |
| 16 | 100% |
| 32 | 100% |
| 64 | 100% |
| 128 | 100% |
| 256 | 100% |
| 512 | 98,20% |
| 1024 | 99,10% |

Figura 20 - Gráfico: Número de Gaussianas x Taxa de Acerto – Terceiro experimento ELSDSR

Através desse último experimento foram conseguidas taxas de acerto de 100%. Houve diversos fatores que influenciaram para conseguir esta taxa; um deles é fato do conjunto de treinamento conter todas as possíveis pronúncias do idioma inglês, assim constituindo um bom conjunto de características que pudessem discriminar pessoas. Outro fator é referente à qualidade do sinal de voz, pois a base de dados ELSDSR é praticamente livre de ruído. O último fator é referente à população de locutores, pois há um número pequeno de locutores e o grau de similaridade entre eles é pequeno, pois, como descrito anteriormente, essa base é constituída de diferentes locutores de diferentes países, com idades distintas e a maioria deles são falantes não-nativos do Inglês.

## MIT

A partir dos resultados obtidos com a base de dados ELSDSR, verificou-se a necessidade de aplicar o algoritmo numa base de dados mais complexa. O MIT, como descrito na seção anterior, é um base com o dobro do número de locutores da ELSDSR que simula um ambiente real de operação.

Para o experimento realizado foram utilizadas ambas as seções de treinamento, totalizando um total de 80 arquivos de treinamento para cada locutor com aproximadamente 120 segundos de fala e para os testes foram utilizados 28 arquivos para cada locutor onde cada arquivo tem aproximadamente 1 segundo de fala.

Nesse experimento buscou-se variar o número de gaussianas para tentar chegar a uma relação entre número de gaussianas e a taxa de acerto. Assim foram utilizadas 2 a 256 gaussianas, com um passo de 2 elevado a 2. A utilização de mais gaussianas não foi possível dado o tempo de treinamento associado a esse experimento. Assim os resultados obtidos foram:

Tabela 9 – Experimento MIT

|  |  |
| --- | --- |
| # Gaussianas | Taxa de Acerto |
| 2 | 47,62% |
| 4 | 53,13% |
| 8 | 66,82% |
| 16 | 66,82% |
| 32 | 82,14% |
| 64 | 86,53% |
| 128 | 89,29% |
| 256 | 90,18% |

Figura 21 - Gráfico: Número de Gaussianas x Taxa de Acerto – experimento MIT

Analisando os resultados obtidos, podemos observar, novamente, o aumento crescente na taxa de acerto com aumento do número de densidades gaussianas, chegando a taxas de até 90,18% com 256 gaussianas. Observa-se também que mesmo a base de dados sendo bastante ruidosa, foi possível conseguir um taxa bastante aceitável. Diferente dos experimentos com a base de dados ELSDSR, entretanto, esse taxa só foi atingida com um número maior de densidades gaussianas, o que ocasiona um custo computacional maior devido ao cálculo de um grande número de pontos flutuantes.

# Conclusão e Trabalhos Futuros

O trabalho realizado visa à pesquisa e ao desenvolvimento da tecnologia de reconhecimento de locutor, tendo como enfoque principal a implementação de um sistema de identificação de locutor. O enfoque adotado foi explorar e entender as principais técnicas usadas no reconhecimento de locutor, a fim de exaltar suas limitações.

Primeiramente, foi apresentada uma introdução à área de reconhecimento de locutor, discutindo as suas aplicações, relatando o seu histórico e estrutura básica. Depois, descreveu-se o processo de extração de características do sinal de voz, desde sua aquisição até a geração dos MFCCs.

Discutiu-se ainda neste trabalho os modelos de misturas gaussianas ou GMM, descrevendo sua definição e uso desses no reconhecimento automático do locutor, tanto para verificação como para identificação.

Apresentou-se então o desenvolvimento em MATLAB de um sistema de identificação de locutor usando modelos de misturas gaussianas. Foram abordadas algumas informações relevantes ao desenvolvimento de um sistema desse tipo.

Por último foram apresentados os experimentos e os resultados alcançados por esse sistema, utilizando duas bases de dados: uma base de dados próxima do ideal, ELSDSR, e uma base de dados próxima do real, MIT.

A partir dos resultados obtidos, podemos concluir que o GMM é uma técnica bastante eficiente para modelagem de características de locutores, tanto para um ambiente ideal quanto para um ambiente real. Num ambiente real é necessário um número maior de gaussianas para atingir uma taxa de acerto aceitável. Porém, o aumento de gaussianas é limitado à base de dados de treinamento, visto que o GMM necessita de um número expressivo de amostras para serem treinados. Ademais, quanto maior o número de gaussianas maior o custo computacional, visto que é necessário calcular muitos parâmetros de ponto flutuante para estimar um GMM.

Os resultados adquiridos foram satisfatórios, mas o sistema aqui proposto pode ser ampliado e melhorado de várias maneiras:

* Utilização de outras formas de extração de características ou acréscimos de outras características.
* A utilização de outros métodos de Estimação de Máxima Verossimilhança.
* A utilização de *Support Vector Machines* (SVM) para hibridização do GMM, como feito em [32], no intuito de estimar o GMM através de uma base de dados limitada.
* A utilização de lógica *fuzzy* para Hibridização do GMM.

# Referências

[1] L. RABINER and B. JUANG, Fundamentals of speech recognition, Prentice Hall, 1993.

[2] D.A. REYNOLDS, *An overview of automatic speaker recognition technology*, Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on, Vol. 4 (2002), pp. IV-4072-IV-4075 vol.4.

[3] H.F. HOLLIEN*, Forensic Voice Identification*, Academic Press, 2002, 240 p.

[4] L. YOUNT, *Forensic science: from fibers to fingerprints,* Infobase Publishing, 2006, 206 p.

[5] L. G. KERSTA, Voiceprint Identification , Nature, 1996, pp. 1253-1257

[6] R. VANDERSLICE AND P. LANDEFOGED, The Voiceprint Mystique, UCLA Working Papers in Phonetics, 7 , pp. 126-142

[7] S. FURUI, 50 Years of Progress in Speech and Speaker Recognition Research, SPECOM conference, Patras, Greece, October 2005.

[8] G.R. DODDINGTON, J.L. FLANAGAN AND R.C. LUMMIS, Automatic Speaker Verification by Nonlinear Alignment of Acoustic Parameters”, U.S Patent 3, 700, 815.

[9] B.S. ATAL, Automatic Speaker Recognition Based on Pitch Contours, Journal Acoustic Society of America, 1972, vol.52, pp. 1687-1697.

[10] B.S ATAL, Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification, Journal Acoustic Society of America, 1974, vol. 55, no. 6, pp. 1304-1312.

[11] B.S ATAL, Automatic Recognition of Speaker From Their Voices, Proceedings IEEE,1976, vol.64, no.4, pp. 460-475.

[12] R.C. LUMMIS, Speaker Verification by Computer Using Speech Intesity for Temporal Registration, IEEE Trans. Audio Electroacoust., 1973, vol. AU-21, pp 80-89

[13] A.E. ROSEMBERG, Automatic speaker verification: a review, Proceedings of the IEEE, vol.64, no.4: 475-486.

[14] S. FURUI; F. ITAKURA AND S. SAITO, Talker recognition by Long-Time Averaged Speech Spectrum”, Electronics Communications of Japan, 1972,vol. 55A, pp. 54-61.

[15] S. FURUI, An Analysis of Long-Term Variation of Feature Parameters of Speech and its Application to Talker Recognition, Electronics Communications, 1974, vol.57-A, pp. 34-42.

[16] J.D. MARKEL and S.B. DAVIS, Text-Independent speaker recognition from a large linguistically unconstrained time-spaced data base, IEEE Trans. Acoust., Speech, Signal Processing, 1979, vol. ASSP-27, no.1, pp.74-82.

[17] S. FURUI, Cepstral analisys technique for automatic speaker verification, IEEE Trans. Acoust., Speech, Signal Processing, 1981, vol.29(2), pp. 254-272.

[18] F.K. SOONG, A.E. ROSEMBERG, L.R RABINER and B.H. JUANG, A vector quantization approach to speaker recognition, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.387-390.

[19] F.Q. D’ALMEIDA, Técnicas Eficientes de Identificação Automática de Locutores, Tese de Doutorado em Engenharia Elétrica, Publicação PPGENE.TD-037, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 146p.

[20] D.A. REYNOLDS, A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification, 1992, Ph. D. Thesis, Georgia Institute of Technology, Department of Eletrical Engineering

[21] D.A. REYNOLDS, Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1. January 1995.

[22] A.E ROSEMBER, C.H. LEE and F.K. SOONG, Sub-word unit talker verification using hidden markov models, Proc. Intl. Conf. on Acoustics, Speech, and Signal Processing, 1990, vol. 1, pp. 269-272.

[23] J.J. WEBB and E.L. RISSANEN, Speaker identification experiments using HMMs, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1993, vol.2, pp. 387-390.

[24] C. CHE and Q. LIN, Speaker recognition using HMM with experiments on the YOHO database, Proc. EUROSPEECH, Madrid, Italy, pp. 625–628, 1995.

[25] J. COLOMBI, D. RUCK, S. ROGERS, M. OXLEY, and T. ANDERSON, Cohort selection and word grammer effects for speaker recognition, Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Atlanta, GA, 1996, pp. 85–88.

[26] M.N. JAYANT, Speaker Verification: A Tutorial, IEEE Communications Magazine,1990, p. 42-47.

[27] Q. JIN, Robust Speaker Recognition, Thesis of Doctor of Philosophy in Language and Information Technologies, Language Technologies Institute School of Computer Science Carnegie Mellon University, Pittsburgh, 2007.

[28] T. GANCHEV; N. FAKOTAKIS and G. KOKKINAKIS, Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task, Proc. of the SPECOM-2005 , Patras, Greece, 2005, Vol. 1, pp. 191-194.

[29] J. P. CAMPBELL, *Speaker Recognition: A Tutorial*, Proceedings of the IEEE, Vol. 85, No. 9. September 1997.

[30] S.S. STEVENS, J. VOLKMAN and E.B. NEWMAN, [A scale for the measurement of the psychological magnitude pitch](http://scitation.aip.org/getabs/servlet/GetabsServlet?prog=normal&id=JASMAN000008000003000185000001) , 1937, Journal of the Acoustical Society of America, vol. 8 , pp. 185–190.

[31] T. GANCHEV, N. FAKOTAKIS and G. KOKKINAKIS, Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task, Proc. of the SPECOM-2005, October 17-19, 2005. Patras, Greece. Vol. 1, pp. 191-194.

[32] A.V. OPPENHEIN and R. W. SCHAFER, Discrete Time Signal Processing, 2.ed. New York: Prentice Hall, 2002.

[33] D. O'SHAUGHNESSY, Automatic speech recognition: History, methods and challenges, 2008, Pattern Recognition 41 (2008) 2965-2979

[34] J. S, N.S. KIM, AND W. SUNG, A Statistical Model-Based Voice Activity Detection IEEE SIGNAL PROCESSING LETTERS, VOL. 6, NO. 1, JANUARY 1999

[35] R. TOGNERI AND D. PULLELLA, [Circuits and Systems Magazine, IEEE](http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=7384) An Overview of Speaker Identification: Accuracy and Robustness Issues, 2011, vol 11 , pp. 23-61

[36] VOICEBOX: Speech Processing Toolbox for MATLAB, Disponível em < <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>> Último acesso em 06/07/2011

[37] The MIT Mobile Device Speaker Verification Corpus, Disponível em <<http://groups.csail.mit.edu/sls/mdsvc/index.cgi> > Último acesso em 07/07/2011

[38] English Language Speech Database for Speaker Recognition, Disponível em < <http://www2.imm.dtu.dk/~lf/elsdsr/> > Último acesso em 07/07/2011

[39] C.B. LIMA, Sistemas de verificação de locutor independente do texto baseados em GMM e AR-Vetorial utilizando PCA, 2001, Tese de Mestrado, Engenharia Elétrica do Instituo Militar.

# APÊNDICE A – Textos de Treinamento e Testes ELSDSR

## Texto de Treinamento da Base de dados ELSDSR

### Parte A:

Chicken Little was in the woods one day when an acorn fell on her head. It scared her so much she trembled all over. The poor girl shook so hard, half her feathers fell out.

### Parte B:

Billions of black, shrimp-size bugs with transparent wings and beady red eyes are beginning to carpet trees, buildings, poles, and just about anything else vertical in the U.S. from the eastern seaboard west through Indiana and south to Tennessee.

### Parte C:

Oymyakon, in Siberia, is the coldest permanently inhabited place on Earth. Now geographer and adventurer Nick Middleton reveals the locals' secrets for coping with the cold.

### Parte D:

Few shores are immune from the tide of plastic soda bottles, bags, cartons, and other trash floating on the ocean today. Now a new study suggests the problem runs deeper: Microscopic bits of plastic permeate the world's beaches and marine environment.

### Parte E:

One hundred years later, the life of the Negro is still sadly crippled by the manacles of segregation and the chains of discrimination.

### Parte F:

People are finding medieval toys in Britain's Thames River—and these toys have been changing how historians view the lives of medieval kids.

### Parte G:

My friend Tricia suggests me to drive to the woods to watch the poor bear being hunted for pleasure. And I say yes.

## Texto de Teste da Base de Dados ELSDSR

**Ancient Egypt**

There are days when the sand blows ceaselessly, blanketing the remains of a powerful dynasty that ruled Egypt 5,000 years ago.(1) When the wind dies down and the sands are still, a long shadow casts a wedge of darkness across the Sahara, creeping ever longer as the north African sun sinks beyond the horizon.(2)

Five thousand years ago, the fourth dynasty of Egypt's Old Kingdom was a highly advanced civilization where the kings, known as pharaohs, were believed to be gods.(3) They lived amidst palaces and temples built to honor them and their deified ancestors.(4) "Pharaoh" originally meant "great house," but later came to mean king.(5) This web site will show you science in action -- bringing you face to face with the evidence archaeologists use to understand the meaning of Giza's pyramids, and to the process of evaluating the finds they will uncover beneath the sands of the plateau.(6)

Before looking closely at pharaonic society and the beginning of the Pyramid Age, one first has to step into Egypt's landscape and take a look around.(7) Ancient Egyptians called their land "Kemet," which meant "black," after the black fertile silt-layered soil that was left behind each year during the annual innundation, when the Nile flooded the fields.(8) The most prevalent color of the desert, however, is a decidedly reddish-yellow ochre.(9) The Egyptians called the desert "deshret," meaning "red," and this endless carpet of sand covers an estimated 95 % of Egypt, interrupted only by the narrow band of green carved by the waters of the Nile.(10)

It was at this time that hieroglyphic writing made its first appearance, in the tombs and treasures of the pharaohs.(11) To seal the unification of Upper and Lower Egypt, Menes founded the capital city of the kingdom at the place where the two met: at the apex of the Nile, where it fans out onto the fertile silt plain.(12) The fortress city was named "White Walls" by Menes, but it is known today by its Greek name, Memphis.(13) For much of the 3,000 years of ancient Egypt, it remained the capital seat of the pharaohs.(14) Only 20 miles to the north of Memphis is the modern capitol, Cairo, still situated near the juncture of the Nile valley and the delta.(15)

How does the pyramid fit into early Egyptian life?(16) Pyramids today stand as a reminder of the ancient Egyptian glorification of life after death, and in fact, the pyramids were built as monuments to house the tombs of the pharaohs.(17) Death was seen as merely the beginning of a journey to the other world.(18) In this society, each individual's eternal life was dependent on the continued existence of their king, a belief that made the pharaoh's tomb the concern of the entire kingdom.(19)

Pictures on the walls of tombs tell us about the lives of the Kings and their families.(20) We know pyramids were built during a king's lifetime because heiroglyphs on tomb walls have been found depicting the names of the gangs who built the pyramids for their kings.(21) Furniture and riches were buried with the king so he would have the familiar comforts of his lifetime buried near him.(22) Whole subdivisions of tombs of those in high positions in the court of a king can be found surrounding the pyramids of Giza.(23) These are primarily mastabas, or covered rectangular tombs that consist of a deep burial shaft, made of mud brick and half-buried by the drifts of sand on the plateau.(24)

The first pyramid was the Step Pyramid at Saqqara, built for King Zoser in 2750 BC.(25) This first application of large scale technology, however, is often attributed to Imhotep, the architect of the Step Pyramid.(26) He was not a pharaoh, but was the Director of Works of Upper and Lower Egypt.(27) The superstructure of the pyramid was made of small limestone blocks and desert clay.(28) Inside, the burial chamber and storage spaces for Zoser's grave goods were carved out of the earth and rock beneath the structure.(29) Imhotep's intent was to mimic the basic structure of King Zoser's palatial home in the burial chamber.(30) The tomb, like those that followed, was meant to be a replica of the royal palace.(31) In early tombs, the central area was always the burial place.(32)

It is thought that in 816 AD Caliph al-Mamun first ordered workers to blast through the blocked stone entrance in order to explore within Khufu's pyramid.(33) But looters, probably from dynastic Egyptian times, had already absconded with King Khufu's burial treasures and his body.(34) This is true of all of the pyramids at (1)-(58): Notations of the sentence number for test text. Giza, so very little is known about Khufu or any of his successors who were buried at Giza.(35) Archaeologists, nonetheless, continue to look for pieces of this puzzle to further our understanding of the Pyramid Age and the pharaohs that ruled Egypt.(36)

**History of Giza**

Standing at the base of the Great Pyramid, it is hard to imagine that this monument -- which remained the tallest building in the world until early in this century -- was built in just under 30 years.(37) It presides over the plateau of Giza, on the outskirts of Cairo, and is the last survivor of the Seven Wonders of the World.(38) Today, Giza is a suburb of rapidly growing Cairo, the largest city in Africa and the fifth largest in the world.(39)

About 2,550 B.C., King Khufu, the second pharaoh of the fourth dynasty, commissioned the building of his tomb at Giza.(40) Some Egyptologists believe it took 10 years just to build the ramp that leads from the Nile valley floor to the pyramid, and 20 years to construct the pyramid itself.(41) On average, the over two million blocks of stone used to build Khufu's pyramid weigh 2.5 tons, and the heaviest blocks, used as the ceiling of Khufu's burial chamber, weigh in at an estimated 40 to 60 tons.(42)

This question has long been debated, but many Egyptologists agree the stones were hauled up ramps using ropes of papyrus twine.(43) The popular belief is that the gradually sloping ramps, built out of mud, stone, and wood were used as transportation causeways for moving the large stones to their positions up and around the four sides of the pyramids.(44)

Giza, however, is more than just three pyramids and the Sphinx.(45) Each pyramid has a mortuary temple and a valley temple linked by long causeways that were roofed and walled.(46)