INSTITUTO FEDERAL DE SANTA CATARINA

HENRIQUE HILLESHEIN

Utilização de Técnicas de Aprendizagem de Máquina em Biometria de Voz

São José - SC

dezembro/2018

UTILIZAÇÃO DE TÉCNICAS DE APRENDIZAGEM DE MÁQUINA EM BIOMETRIA DE VOZ

São José - SC

dezembro/2018

LISTA DE ILUSTRAÇÕES

Figura 1 – Características Biométricas 13
Figura 2 – Impressão Digital 14
Figura 3 – Íris 14
Figura 4 – Sinal de Voz 15
Figura 5 – Quantização 19
Figura 6 – Relação entre frequência em Hz e mel 21
Figura 7 – Diagrama de funcionamento do MFFC 22
Figura 8 – Banco de filtros MFCC 23
Figura 9 – Separador Hiperplano SVM 24
Figura 10 – Modelagem do sistema utilizando GMM 28
Figura 11 – Modelagem do sistema 28

LISTA DE TABELAS

Tabela 1 – Cronograma das atividades previstas 28

LISTA DE ABREVIATURAS E SIGLAS

LPC *Linear predictive coding* 10

MFCC *Mel frequency cepstral coefficients* 10

PLP *Perceptual Linear Coding* 10

ELSDSR *English Language Speech Database for Speaker Recognition* 27

PIN Número de Identificação Pessoal 16

ID Identidade 16

CPF Cadastro de Pessoas Físicas 18

FFT *Fast Fourier Transform* 22

SVM *Support Vector Machines* 23

GMM *Gaussian Mixture Model* 10

EM *Expectation-Maximization* 26

UBM Modelo de *Background* Universal 27

DTW *Dynamic Time Warping* 9

API Interface de Programação de Aplicativos 27

SUMÁRIO

1 INTRODUÇÃO . . . 9
1.1 Objetivo Geral . . . 10
1.2 Objetivos Específicos . . . 10

2 FUNDAMENTAÇÃO TEÓRICA . . . 13
2.1 Biometria . . . 13
2.1.1 Impressão Digital . . . 13
2.1.2 Íris . . . 14
2.1.3 Voz . . . 15
2.2 Conceitos de Reconhecimento de Falante . . . 16
2.2.1 Ramos do Reconhecimento de Falante . . . 16
2.2.1.1 Verificação . . . 16
2.2.1.2 Identificação . . . 16
2.2.1.3 Classificação . . . 17
2.2.1.4 Segmentação . . . 17
2.2.1.5 Detecção . . . 17
2.2.1.6 *Tracking* . . . 17
2.2.2 Modalidades . . . 17
2.2.2.1 Dependente de Texto . . . 17
2.2.2.2 Independente de Texto . . . 17
2.2.2.3 *Text-Prompted*(Textos Aleatórios) . . . 18
2.2.2.4 *Knowledge-Based*(Base de Conhecimento) . . . 18
2.3 Extração de Caraterísticas do Sinal de Voz . . . 18
2.3.1 Sinal de Voz . . . 18
2.3.1.1 Amostragem . . . 18
2.3.1.2 Quantização . . . 19
2.3.1.3 Ruído . . . 20
2.3.2 Técnicas de Extração de Características do Sinal de Voz e MFCC . . . 20
2.3.2.1 MFCC . . . 21
2.3.2.1.1 Funcionamento do MFCC . . . 21
2.4 Aprendizado de Máquina . . . 23
2.4.1 Algoritmos de aprendizado de Máquina . . . 23
2.4.1.1 *Support Vector Machines* . . . 23
2.4.1.2 *Naive Bayes* . . . 24
2.4.1.3 GMM . . . 25

3 PROPOSTA . . . 27

REFERÊNCIAS . . . 29

1 INTRODUÇÃO

A importância da segurança e da identificação de pessoas é algo crescente na sociedade atual, fazendo surgir vários métodos para que se consiga suprir essa necessidade. Uma forma de atender essa necessidade é utilizando a biometria para identificar uma pessoa pelas suas características únicas. A biometria faz medidas das características fisiológicas ou comportamentais de um indivíduo para identificá-lo, podendo ser, por exemplo, as características da face de uma pessoa(característica fisiológica) ou a maneira que uma pessoa anda(comportamental).

Existem tipos de biometria que são frequentemente usadas para segurança, como a digital, íris e voz. Digital e íris se enquadram como uma biometria de característica fisiológica, onde se trata de algo imutável. No caso da voz pode-se fazer biometria das características fisiológicas e também do comportamento, possibilitando métodos alternativos, pois a fala oferece bastante flexibilidade e diferentes níveis para realizar o reconhecimento. Por exemplo, o sistema pode forçar o usuário para que fale de uma maneira particular e diferente para cada tentativa(FAUNDEZ-ZANUY; MONTE-MORENO, 2005). Uma outra vantagem da biometria de voz é que ela pode ser usada à distância, diferente das outras biometrias, um exemplo é quando se deseja autenticar ou identificar alguém via telefone.

A voz de cada pessoa possui características únicas igual a impressão digital, devido a formação orgânica da laringe, cordas vocais e todo o sistema responsável pela voz. Além das diferenças fisiológicas, também existem diferenças na maneira de falar de cada pessoa como sotaque, ritmo, entonação, padrão de pronúncia, escolha de vocabulário e assim adiante (KINNUNEN; HAIZHOU, 2010). Baseando-se nessas características é possível reconhecer um indivíduo, ou seja, reconhecer o falante.

O reconhecimento do falante é comumente utilizado no ramo da verificação ou identificação, mas também pode ser utilizado em outros ramos, como classificação, detecção, segmentação e etc. O objetivo da verificação é descobrir se a pessoa que está falando é realmente quem ela alega ser, nesse caso o sistema deve estar preparado para possíveis tentativas de enganação. No caso da identificação o objetivo é distinguir uma pessoa entre um grupo de pessoas previamente registradas (BOLES; RAD, 2017). Atualmente o ramo da verificação é o mais popular para reconhecimento de falante devido sua importância em segurança e controle de acesso(BEIGI, 2011), por esse motivo será o ramo utilizado nesse trabalho.

O sistema de reconhecimento possui as modalidades dependente de texto, independente de texto ou uma variação entre elas. No caso da modalidade dependente de texto, a pessoa deve falar uma palavra ou um conjunto de palavras que foram cadastradas previamente pelo falante no sistema para que seja feito o reconhecimento. Por outro lado, na modalidade independente de texto, o sistema deve identificar a pessoa que está falando independentemente do que está sendo falado, ou seja, o sistema aprende exclusivamente a voz da pessoa e não a combinação da voz com o texto(FAUNDEZ-ZANUY; MONTE-MORENO, 2005). Neste trabalho, será utilizado a modalidade independente de texto, pois é o mais versátil e pode ser utilizado em todos os ramos de reconhecimento de falante(BEIGI, 2011).

Com a recente popularidade de aprendizagem de máquina e *deep learning*(aprendizagem profunda), aplicações de reconhecimento de falante estão ficando mais poderosas. O treinamento da máquina tem se tornado mais rápido e o número de falantes que podem ser qualificados se tornou maior (BOLES; RAD, 2017). Isso faz com que a aprendizagem de máquina se torne uma ferramenta de predição muito atraente para reconhecimento de voz, mas não é a única, existem outras abordagens como a utilização de *Dynamic*

Time Warping (DTW)(MUDA; BEGAM; ELAMVAZUTHI, 2010).

As técnicas de aprendizagem de máquina são algoritmos computacionais que identificam padrões e modelos através da análise de um conjunto de dados(AYODELE, 2010). Tais padrões e modelos podem ser utilizados para fazer predições em diversas aplicações como o reconhecimento de falante.

Para utilizar os algoritmos de aprendizagem de máquina, é necessário que a informação da voz passe por um pré-processamento. Para isso, é utilizada uma técnica para se obter características do sinal de voz. Existem diversas técnicas que podem ser usadas para extrair essas características, como *Linear predictive coding* (LPC), *Mel frequency cepstral coefficients* (MFCC) e *Perceptual Linear Coding* (PLP)(FAUNDEZ-ZANUY; MONTE-MORENO, 2005). Colocando uma informação pré-processada como entrada de uma máquina que já terminou seu processo de aprendizado, se obtém como saída a predição que o falante é quem alega ser ou se obtém como saída quem é o falante, dependendo se o interesse da aplicação é verificação ou identificação. A técnica MFCC tem se mostrado bastante eficiente para extração das características do sinal de voz(RAMGIRE; JAGDALE, 2016), por esse motivo será a técnica utilizada neste trabalho.

Realizando uma busca por sistemas existentes e artigos científicos relacionados ao tema deste trabalho, encontramos ferramentas disponíveis para o desenvolvimento de sistemas de reconhecimento de falante, como Spear(KHOURY; SHAFEY; MARCEL, 2014) e Alize(BONASTRE; WILS; MEIGNIER, 2005) e também aplicações prontas como *Speaker Recognition API*, da Microsoft Azure, e o *Ok Google*, do sistema operacional *Android*. Entretanto, estas ferramentas e aplicações, em algumas situações, são pagas, executadas na nuvem, ou pouco flexíveis para implementações de verificação de falante com modalidade independente de texto. Com relação aos artigos científicos, pode-ser encontrar diversas técnicas que utilizam aprendizado de máquina para o reconhecimento de falante(BAGUL; SHASTRI, 2013; REYNOLDS, 1995; WAN; CAMPBELL, 2000), cada um com foco em uma respectiva aplicação.

É importante destacar que este trabalho não tem a pretensão de propor ferramentas para competir com as citadas anteriormente, nem propor técnicas distintas das encontradas na literatura, mas sim apresentar um introdutório para sistemas de reconhecimento de falante utilizando algumas técnicas de aprendizagem de máquina, destacando aspectos importantes no processo de implementação e disponibilizando códigos de exemplos que auxiliem futuros trabalhos de implementação neste tema. Para tal, utilizaremos a linguagem de programação Python com a biblioteca *toolbox Scikit-learn* e, por ser um dos mais utilizados na literatura, daremos maior ênfase na utilização do algoritmo *Gaussian Mixture Model* (GMM).

1.1 Objetivo Geral

- Testar diversas técnicas de aprendizagem de máquina e ferramentas para reconhecimento de falante.

1.2 Objetivos Específicos

- Estudar as características de um sinal de voz
- Apresentar ferramentas e bibliotecas disponíveis para o desenvolvimento de aplicações de biometria de voz
- Definir bancos de dados a serem utilizados
- Realizar o pré-processamento do sinal de voz

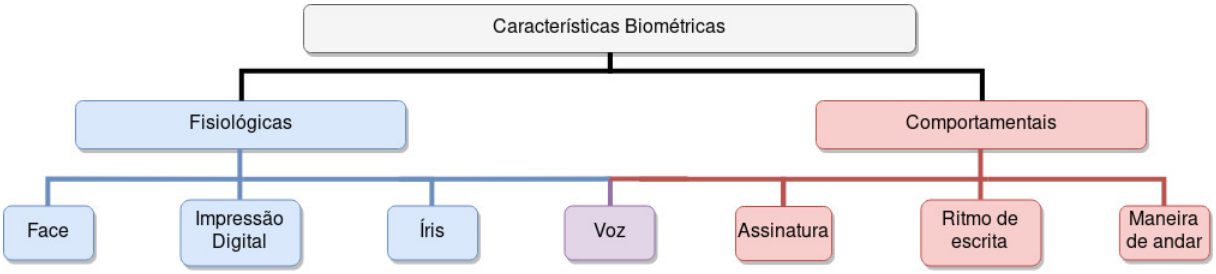
- Implementar técnicas de aprendizagem de máquina para a biometria de voz

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Biometria

A biometria abrange qualquer tipo de medição de características fisiológicas e/ou comportamentais do corpo humano, ela normalmente é associada ao reconhecimento de pessoas, pois é confiável e conveniente para o uso. A informação biométrica é uma informação pessoal que é distinta de outras informações por ser permanentemente ligada ao ser humano e na maioria das vezes resistente à alterações (BERNECKER, 2006). A biometria pode ser o tamanho de um membro do corpo, o jeito que uma pessoa se alimenta, distância entre os olhos ou qualquer outra característica fisiológica e/ou comportamental que pode ser medida. A Figura 1 mostra alguns exemplos de biometria comportamental, fisiológica e ambas(voz).

Figura 1 – Características Biométricas



Fonte: Criado no site <draw.io>

Na área de segurança os tipos de biometrias mais importantes são aquelas que podem distinguir um indivíduo de todos os outros, como a impressão digital, íris, retina, face, voz, etc. É inerente que a biometria é mais confiável e mais competente do que técnicas que usam base de conhecimento e as baseadas em *token*(usuário e senha) na diferenciação entre pessoas autorizadas e impostores, porque as caraterísticas biométricas utilizadas para esse propósito são únicas(HONG, 1998).

2.1.1 Impressão Digital

A impressão digital é o desenho formado pelas papilas dos dedos, ela é formada pelo acumulo de células mortas e cornificadas que se desfazem constantemente como escama da superfície exposta (HONG, 1998). Um exemplo do desenho formado é a Figura 2. A impressão digital provavelmente é o tipo de biometria que a sociedade no geral é mais acostumada, sendo muito usada para documentos de identificação pessoal, controle de acesso, registro de entrada e saída, etc.

Existem muitos tipos de *scanners* de impressão digital no mercado, incluindo óptico, eletrônico de estado sólido e ultrassom. Esses *scanners* começaram a ser produzidos em massa anos atrás e possuem um custo baixo (BEIGI, 2011). Por estar bastante consolidada, e também devido ao custo, a impressão digital é largamente utilizada.

Mesmo já estando bastante avançada, a impressão digital possui problemas com fraudes, pois pode ser facilmente replicada utilizando técnicas avançadas de molde com látex. Outro problema é a legibilidade da digital, onde muitas pessoas que trabalham com serviços braçais não possuem digital legível(BEIGI, 2011).

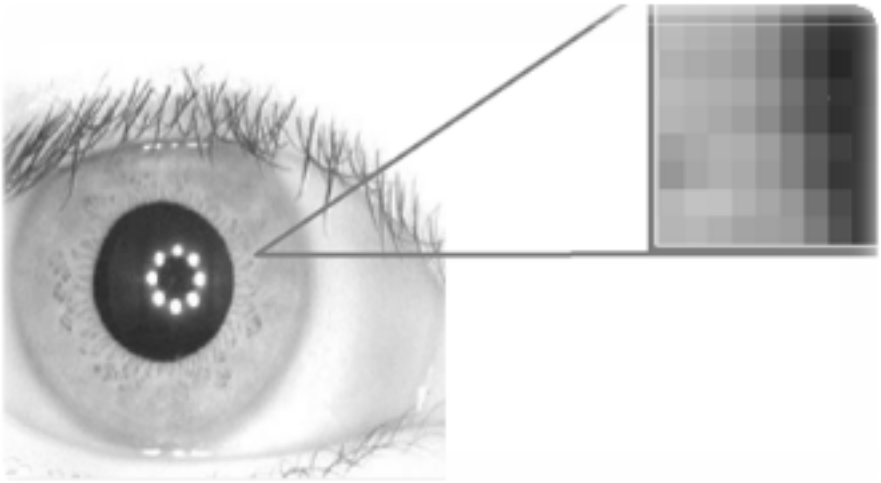
Figura 2 – Impressão Digital

Fonte: (HONG, 1998)

2.1.2 Íris

A íris do olho carrega uma grande quantidade de informação, o suficiente para distinguir uma pessoa. Se trata de algo tão único que inclusive as duas íris de uma pessoa possuem padrões diferentes (BEIGI, 2011). A Figura 3 demonstra a informação contida em uma pequena área da íris.

Figura 3 – Íris



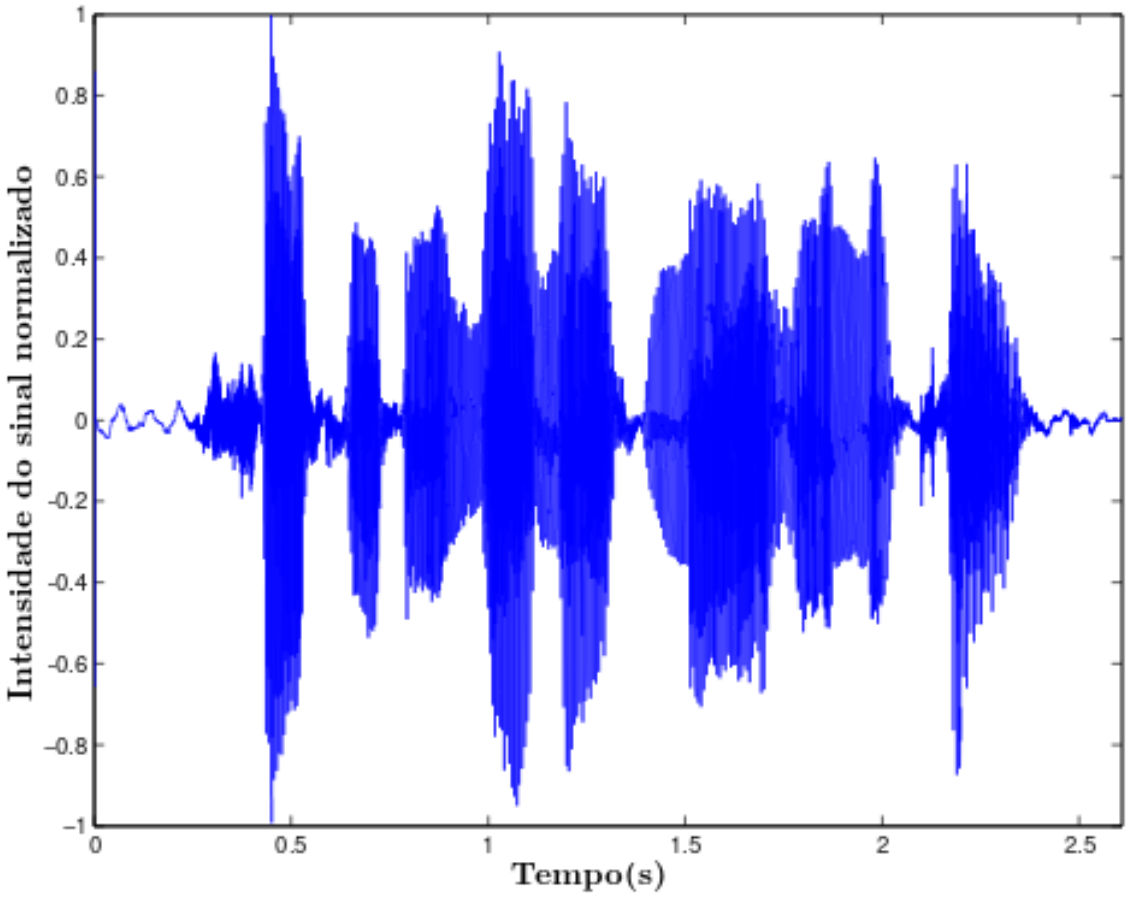
Fonte: (ADAMOVIC; MILOSAVLJEVIC, 2013)

A identificação via íris é feita via posicionamento do olho em frente a uma câmera, a qual utilizará luz para localizar a íris, isolando outras partes do olho por fotografia digital. Após localizar a íris é feita análise de segmentos da íris para identificar a pessoa(ADAMOVIC; MILOSAVLJEVIC, 2013). Essa questão do posicionamento onde a pessoa precisa manter-se parada em frente e perto da câmera durante um tempo, é uma desvantagem do uso e algo bastante invasivo. A íris possui outros problemas, algumas doenças oculares não permitem a análise correta e uma foto de alta qualidade permite fraudar muitos sistemas(BEIGI, 2011).

2.1.3 Voz

A Figura 1 mostra que a biometria da voz pode ser feita tanto baseada no comportamento quanto na fisiologia da voz de uma pessoa. A biometria fisiológica é devido a formação do nariz, boca, laringe, cordas vocais e o restante da anatomia responsável pela voz, fazendo com que a voz de uma pessoa seja única(KINNUNEN; HAIZHOU, 2010), tendo informação suficiente para que se consiga usar em sistemas de segurança como uma forma de autenticação. Já a biometria comportamental da voz fica focada no ritmo de fala, sotaque, idioma, etc.

Figura 4 – Sinal de Voz



Fonte: Adaptado de (BEIGI, 2011)

A Figura 4 mostra um sinal de voz gerado por uma pessoa, esse sinal de voz foi moldado pelos órgãos responsáveis pela voz, deixando suas características nele, ou seja, informação que pode ser usada para reconhecer uma pessoa.

Diferente de muitas biometrias, a biometria da voz é pouco invasiva, pois é possível reconhecer uma pessoa simplesmente conversando com ela, tendo um dispositivo que esteja captando o áudio e executando essa aplicação. Um exemplo seria conversando com um gerente de um banco, onde teria um microfone captando a conversa. Neste caso, o sistema reconhece a voz do gerente, então pode separar a voz do gerente com a do cliente, permitindo a autenticação do cliente simplesmente pela conversa, não sendo necessário o cliente passar informações pessoais para que consiga fazer os trâmites desejados e caso apareça um impostor, o sistema avisará o gerente(BEIGI, 2011).

A utilização da biometria de voz também possui suas desvantagens, pois é necessário que a pessoa seja capaz de falar. Outras situações que podem apresentar problemas é quando o usuário está com algum problema na voz, seja por nariz obstruído por causa de uma doença ou por a pessoa estar rouca.

Neste trabalho, focaremos na biometria de voz para reconhecimento de pessoa, normalmente chamado de reconhecimento de falante.

2.2 Conceitos de Reconhecimento de Falante

Reconhecimento de falante é um termo genérico usado para qualquer procedimento que envolve o reconhecimento da identidade de uma pessoa baseada em sua voz(BEIGI, 2011).

Uma menção importante a fazer é a diferença entre o reconhecimento de fala para o reconhecimento de falante. Em reconhecimento de falante é utilizado as características biométricas para que se reconheça um individuo. Em reconhecimento de fala, é desejado reconhecer o que está sendo falado em um áudio, o único foco é a identificação das palavras e números presentes no áudio, ou seja, a informação desejada no áudio é diferente entre os dois sistemas de reconhecimento. Esse trabalho é focado nas características utilizadas para reconhecimento de falante.

Um sistema de reconhecimento de falante tenta modelar as características do trato vocal de uma pessoa. Isso pode ser feito via modelo matemático do sistema fisiológico da produção da fala humana ou simplesmente um modelo estatístico com saída similar às características da fala humana. Uma vez que o modelo é gerado e associado ao indivíduo, novas instâncias de fala podem ser comparadas com o modelo da pessoa alvo, tendo como contraste o modelo de outras pessoas para saber a probabilidade de a instância pertencer a mesma pessoa que gerou o modelo alvo. Esta é a metodologia fundamental para todas as aplicações de reconhecimento de falante(BEIGI, 2011).

É digno de constar que o reconhecimento de falante utiliza a única biometria que pode ser facilmente testada remotamente através da infraestrutura existente, a rede de telefonia. Isso faz o reconhecimento de falante bastante valioso e incomparável em muitas aplicações do mundo real(BEIGI, 2011).

2.2.1 Ramos do Reconhecimento de Falante

O reconhecimento de falante tem muitos ramos que são diretamente ou indiretamente relacionados. No geral, são 6 ramos principais, no qual são verificação, identificação, classificação, segmentação, detecção e *tracking*. Os ramos podem ser utilizados em conjunto. Atualmente a verificação é o ramo mais popular pela sua importância em segurança e controle de acesso(BEIGI, 2011). Neste trabalho é utilizado o ramo de verificação.

2.2.1.1 Verificação

Em uma aplicação genérica de verificação, a pessoa que está sendo verificada primeiro deve se identificar, normalmente por um método que não envolve a voz. A apresentação do Identidade (ID) pode ser feita, por exemplo, fornecendo o Número de Identificação Pessoal (PIN) ou o usuário. Fornecida a identificação, o sistema recupera o modelo do trato vocal que está relacionada àquela identificação para que seja feita então a verificação. Ou seja, a voz da pessoa será comparada com o modelo para confirmar(autenticar) que a pessoa realmente é o indivíduo da identidade passada. Esse modelo é chamado de *target speaker model*(modelo de falante alvo)(BEIGI, 2011).

2.2.1.2 Identificação

Existem dois diferentes tipos de identificação de falante, o *closed-set*(Conjunto fechado) e *open-set*(conjunto aberto). Em conjunto fechado a voz é comparada com o modelo de falantes cadastrados e retorna o ID do falante que possui a voz mais próxima da voz que foi colocada como entrada, ou seja,

independente da pessoa que falou, mesmo que ela não esteja cadastrada, sempre será retornado um ID(BEIGI, 2011).

A identificação de conjunto aberto pode ser vista como uma combinação do conjunto fechado com a verificação. Neste caso, é feita identificação do falante igual como é feito com o conjunto fechado, onde é identificado o ID da pessoa mais próxima daquela voz, depois é feito o teste da voz no modelo de verificação relacionado à esse ID. O ID é somente retornado se a pessoa passar no teste da verificação(BEIGI, 2011). A voz é utilizada para localizar o ID da pessoa, depois disso o sistema funciona da mesma forma que seria se o usuário tivesse fornecido um PIN como foi comentado na subseção 2.2.1.1.

2.2.1.3 Classificação

O objetivo de classificação é um pouco mais vago. A ideia desse tipo de sistema é classificar o falante, podendo ser a idade ou o gênero da pessoa por exemplo(BEIGI, 2011).

2.2.1.4 Segmentação

Segmentação é utilizada para separar as partes contidas no áudio. Podendo ser música, falantes, ruídos ou qualquer outro som presente no áudio(BEIGI, 2011). Uma aplicação em reconhecimento de falante seria uma conferência, onde é necessário separar a voz de todos as pessoas presentes no áudio. Depois da separação pode ser feita identificação e verificação de cada falante presente.

2.2.1.5 Detecção

Se trata do ato de detectar um ou mais específicos falantes em um áudio. Portanto ele é fundamentalmente a junção de segmentação com identificação e/ou verificação(BEIGI, 2011).

2.2.1.6 *Tracking*

Possui uma diferença sútil em relação à detecção onde a ideia é retornar em que momento a pessoa falou e quanto foi a duração da fala, ou seja, marcar(localizar) em quais momentos e qual o período que um indivíduo está presente no áudio(BEIGI, 2011).

2.2.2 Modalidades

Os sistemas de reconhecimento de falante podem ser implementados usando diferentes modalidades, no qual são ligadas ao uso linguístico, contexto e outros meios. No entanto, no senso prático essas modalidades são relevantes somente para verificação de falante(BEIGI, 2011). A seguir, é apresentado de forma sucinta as modalidades mais relevantes para verificação de falante.

2.2.2.1 Dependente de Texto

Nessa modalidade a pessoa deve falar palavras ou frases que foram usadas no cadastramento dela no sistema. O sistema aprende e gera um modelo do indivíduo levando em consideração a biometria da voz da pessoa e também o que está sendo falado. Este sistema só faz sentido para o ramo da verificação(BEIGI, 2011).

2.2.2.2 Independente de Texto

O sistema de reconhecimento de falante independente de texto é o mais versátil de todas as modalidades. Ele é também a única modalidade viável que permite ser usada para todos os ramos de reconhecimento de falante. O sistema deve reconhecer dependendo somente das características do trato

vocal do falante e não faz suposição sobre o texto da fala. Como ele é um sistema que não depende de texto, pode ser facilmente fraudado, pois qualquer gravação de boa qualidade consegue enganar o sistema(BEIGI, 2011).

2.2.2.3 *Text-Prompted*(Textos Aleatórios)

Se trata de um sistema que mostra para o falante uma frase aleatória e força para que o falante fale aquela frase específica no momento do teste. Essa modalidade foi criada para combater impostores. Se o falante não puder antecipar a frase solicitada, não poderá se preparar para que consiga enganar o sistema. Este sistema só faz sentido para o ramo da verificação(BEIGI, 2011).

Existem duas abordagens principais para esse sistema. A primeira utiliza a modalidade dependente de texto para gerar randomicamente frases e criar um modelo dependente de texto para a frase gerada. Esse modelo irá levar em consideração o conteúdo da fala e a biometria da voz. A segunda abordagem, é a utilização de um sistema de reconhecimento de falante independente de texto em conjunto com um sistema de reconhecimento de fala. Neste caso, é necessário que o texto presente na fala corresponda com a frase solicitada pelo sistema de reconhecimento de fala e que o sistema de reconhecimento de falante verifique a biometria da voz(BEIGI, 2011).

2.2.2.4 *Knowledge-Based*(Base de Conhecimento)

É uma modalidade que faz a combinação da modalidade independente de texto, ou *Text-Prompted*, com sistema de reconhecimento de fala. Neste caso, o usuário deve responder algumas perguntas que ele deveria saber, como número de Cadastro de Pessoas Físicas (CPF), nome da mãe e outras. O sistema de reconhecimento de fala verifica se a resposta está correta e o sistema de reconhecimento de falante verifica a biometria. Este sistema tem uma importância maior para verificação, mas também pode ser usada em outros ramos como detecção, onde pode-se procurar áudios de um falante falando sobre determinados tópicos(BEIGI, 2011)

2.3 Extração de Caraterísticas do Sinal de Voz

Um sinal de voz possui uma grande quantidade de informação, mas é necessário conseguir extrair a parte importante da informação para que se tenha um sistema eficiente e confiável. Em reconhecimento de falante o importante é obter a informação pessoal que chamamos de biometria, tendo-se então um foco maior na fonética. Em reconhecimento de fala por outro lado, seria importante obter a informação que permita identificar fonemas ou qualquer outra informação que ajude identificar o conteúdo(palavras) do áudio. Antes de apresentar as características do sinal de voz, apresentaremos os os conceitos básicos de conversão analógica-digital para o sinal de voz.

2.3.1 Sinal de Voz

A extração das características é feita em um sinal de voz já existente, então um fator bastante importante é a qualidade do áudio. Existem fatores que influenciam na qualidade do áudio, como frequência de amostragem, *jitter*, erro de truncamento, sobreposição e ruído.

A seguir, é apresentado de forma sucinta alguns fatores que influenciam na qualidade do áudio.

2.3.1.1 Amostragem

A voz gerada por uma pessoa é um sinal analógico, então é necessário fazer uma amostragem do sinal de voz para que se possa fazer operações computacionais em cima do sinal. Ao se fazer a

amostragem de um sinal, deve se considerar a maior frequência do sinal, pois é necessário que a frequência de amostragem seja no mínimo duas vezes maior do que a maior frequência do sinal para que não ocorra sobreposição do sinal(LATHI, 2009).

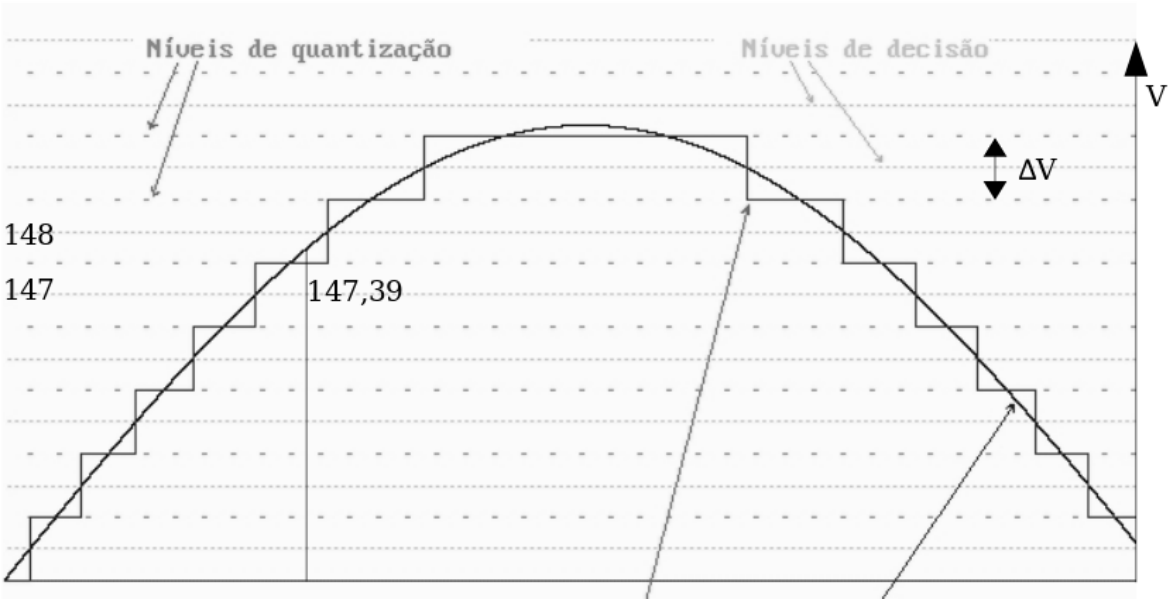
Devido à situações técnicas, a telefonia analógica utiliza uma frequência de amostragem de 8 kHz, mas são áudios de baixa qualidade, pois nesses casos o importante é a inteligibilidade e não a fidelidade da voz(LATHI, 2009). O motivo da baixa qualidade é a informação perdida, algumas frequências do sinal de voz necessitam de uma frequência de amostragem maior para que não ocorra sobreposição.

O aumento da taxa de amostragem melhora a qualidade do áudio, mas em contrapartida precisa de um aumento de processamento, uma taxa de transmissão maior quando se trata de uma aplicação em tempo real, além de um espaço maior de armazenamento.

2.3.1.2 Quantização

Depois da amostragem o sinal ainda possui infinitas possibilidades de amplitude e para torna-lo digital é necessário discretizar a amplitude do sinal. A técnica utilizada para isso é a quantização, onde o valor de amplitude do sinal passa a ter quantidade finita de valores. O número de valores possíveis é 2^b, onde b é a quantidade de bits usados na quantização(LATHI, 2009).

Figura 5 – Quantização



Fonte: http://www.dee.ufrn.br/pcm.pdf (Acesso em 08/04/2018)

A Figura 5 é um exemplo de quantização de um sinal analógico contínuo no tempo, pode-se perceber que a amplitude deixa de ser contínua e passa a ser discreta. É possível observar nessa figura os níveis de quantização e os níveis de decisão. O nível de quantização é um valor que pode ser representado em bits e o nível de decisão é o intervalo de tensão ao qual o sinal pertence. Nesta figura aparecem os níveis de quantização de 147v e 148v e uma amostra com valor de 147,39 que será representada pelo valor 147v, pois é a representação que irá possuir o menor erro de quantização(-0,39v). .

Aumentando o número de bits é possível reduzir a distância entre os níveis de quantização e como consequência diminuir o erro de quantização e melhorar a qualidade de áudio. Da mesma forma que a taxa de amostragem, quanto mais bits, mais o áudio é custoso em rede de transmissão, em armazenamento e processamento.

2.3.1.3 Ruído

O ruído pode ser gerado devido à aspectos físicos ou do ambiente onde o áudio é capturado. O ruído gerado por aspecto físico pode ser em função da qualidade do microfone que foi utilizado para a captura do áudio ou pelo meio no o qual o sinal de áudio é transmitido. O ruído gerado pelo ambiente seria quando o próprio ambiente possui som de fundo, como pessoas próximas conversando ou qualquer outro tipo de som(BEIGI, 2011).

2.3.2 Técnicas de Extração de Características do Sinal de Voz e MFCC

As características do sinal de voz podem ser usadas tanto para reconhecimento de falante quanto para reconhecimento de fala. Essas características compõem as biometrias comportamentais e fisiológicas, mas também possuem características que permitem reconhecer a fala(BEIGI, 2011). Este trabalho tem foco em características de biometria fisiológicas.

Em um sinal de voz existe uma grande porção de informação fortemente correlacionada ou que não é essencial para o entendimento da mensagem, um exemplo de informação não essencial é o silêncio que reduz o desempenho do sistema e pode causar resultados inesperados(BEIGI, 2011). Desta forma, é conveniente retirar trechos de silêncio do sinal de voz.

Uma boa parte da informação presente na voz é relacionada com as características vocais individuais do falante, portanto o trabalho do reconhecimento de fala é tentar filtrar essa informação irrelevante do sinal e interpretar a mensagem que está sendo transmitida. Os sistemas de reconhecimento de falante, por outro lado, buscam utilizar toda a informação contida no sinal que o torna único(BEIGI, 2011).

As características obtidas do sinal de voz devem ser o suficiente para conseguir um reconhecimento de boa qualidade, mas em grande quantidade gera a necessidade de mais processamento, o que pode ser um problema para reconhecimentos em tempo real(KINNUNEN; HAIZHOU, 2010).

Uma forma de se adquirir características do sinal de voz é utilizando uma técnica de características espectrais de curto prazo(*short-term spectral features*). Características espectrais de curto prazo, se trata da computação de quadros entre 20 e 30 milissegundos de duração. Essas características normalmente são descritores do envelope espectral de curto prazo(*short-term spectral envelope*), os quais são acusticamente correlacionados com o timbre, como as propriedades de ressonância do trato vocal(KINNUNEN; HAIZHOU, 2010).

O motivo da separação do áudio em pequenos quadros é devido a variação constante no sinal de voz. Separando em pequenos quadros se consegue uma variação do sinal e de suas características de forma relativamente estacionária(HASAN et al., 2004), facilitando a extração das características. A seguir, uma breve explicação de duas características importantes para reconhecimento de falante, a ressonância do trato vocal e o timbre.

A ressonância do trato vocal se trata das frequências fundamentais geradas pela vibração das cordas vocais em uma certa duração de tempo, ela fica em função do as características do fluxo de ar, formato do trato vocal, formato e tensão das cordas vocais no momento observado(BEIGI, 2011), tendo uma grande importância para características espectrais de curto prazo, pois a frequência fundamental varia de pessoa para pessoa.

O timbre é o conteúdo harmônico que está relacionado com a localização de formantes e suas características(BEIGI, 2011), tendo uma relação direta à ressonância do trato vocal. Pode-se perceber a importância do timbre em reconhecimento de falante com a telefonia analógica, onde são filtradas as

frequências acima de 3,4 kHz. Neste caso, não há dificuldade de compreender o que está sendo falado, mas ocorre a dificuldade para identificar a pessoa que está falando, pois uma grande parte das harmônicas da voz foram filtradas.

Existem diversas técnicas de extração de características fisiológicas do sinal de voz como o MFCC, LPC(MARKOV; NAKAGAWA, 1999) e PLP(MOHAMED; HAMIDIA; AMROUCHE, 2013), mas devido ao tempo limitado, focaremos apenas no MFCC, o qual é bastante utilizado no tipo de aplicação proposto neste trabalho.

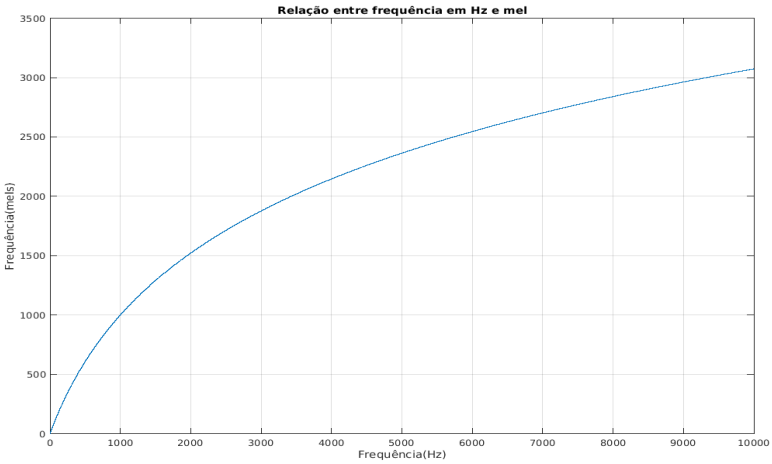
2.3.2.1 MFCC

MFCC é uma técnica aplicada em *short-term spectral features* que se baseia na audição humana para adquirir características de um sinal de áudio(ABDULLA, 2002), a representação da resposta em frequência da audição humana é feita utilizando a frequência mel.

Para o entendimento do MFCC é primeiro necessário entender o que é frequência em escala mel. A audição humana não possui a mesma resposta para todas as frequências, tendo uma percepção linear até 1000 Hz e uma percepção logarítmica à partir de 1000 Hz(BEIGI, 2011). Existem várias equações para conversão da frequência em Hz para frequência na escala mel que tentam se aproximar da percepção humana, uma delas é a equação 2.1, que foi retirada de (O'SHAUGHNESSY, 1987).

$$mel(f) = 2595 * \log_{10}(1 + f/700) \tag{2.1}$$

Figura 6 – Relação entre frequência em Hz e mel



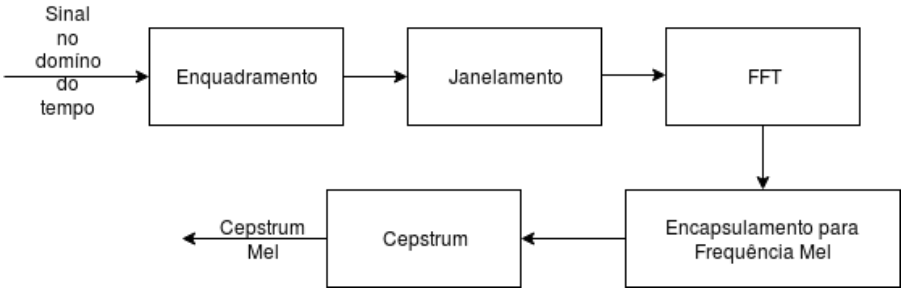
Fonte: Feito no Matlab

Aplicando a fórmula da equação 2.1 obteve-se o gráfico da Figura 6, onde percebe-se a curva logarítmica e o ponto onde o valor de mel é igual a frequência em Hz(1000 Hz). Este ponto é a referência entre as duas escalas.

2.3.2.1.1 Funcionamento do MFCC

O funcionamento do MFCC pode ser demonstrado pelo diagrama da Figura 7. O sistema de extração de característica tem como entrada um sinal de áudio no domínio do tempo e tem como saída os coeficientes *cepstrum* mel.

Figura 7 – Diagrama de funcionamento do MFFC



Fonte: Adapatado de (HASAN et al., 2004)

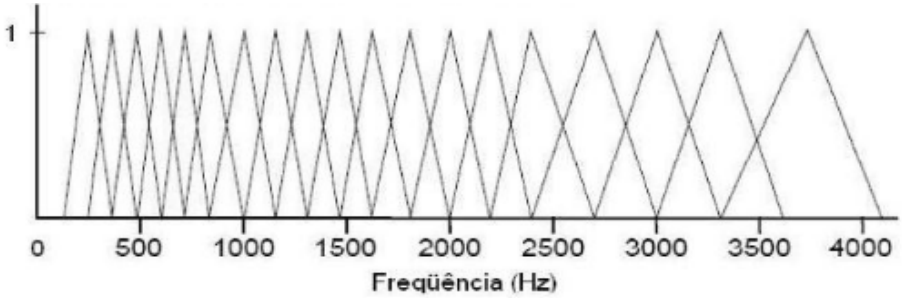
O enquadramento tem como entrada um sinal de voz no domínio do tempo e separa em quadros de, por exemplo, 20 ms. Um sinal de voz não pode simplesmente ser recortado abruptamente em quadros de 20 ms, pois ocorreria descontinuidade no sinal. Desta forma, é feito um janelamento do sinal maior que o tempo de quadro, fazendo uma sobreposição do sinal entre os quadros e mantendo a continuidade do mesmo. Assim, o sinal será atenuado de acordo com a resposta em frequência do filtro do janelamento(HASAN et al., 2004). O tipo de janelamento normalmente utilizado é o de Hamming(KINNUNEN; HAIZHOU, 2010). Os processos nos próximos blocos são realizados para cada quadro separadamente. Após o enquadramento e o janelamento, é aplicada uma *Fast Fourier Transform* (FFT) no sinal para que se obter o sinal no domínio da frequência.

O encapsulamento para frequência mel é a parte particular do MFCC. Esse bloco faz a conversão do espectro de frequência linear(Hz) para o espectro de frequência logarítmica(mel). Uma abordagem bastante utilizada é a utilização de banco de filtros triangulares(HASAN et al., 2004; TIWARI, 2010; KINNUNEN; HAIZHOU, 2010). Nessa abordagem, é utilizado um filtro para cada componente da frequência mel desejada, onde a quantidade de filtros varia de acordo com a implementação. Normalmente são usados 20 filtros, considerando que o primeiro coeficiente é a média da energia do sinal, ele não é utilizado por possuir pouca informação específica do falante(HASAN et al., 2004).

Os filtros triangulares são de banda passante, na escala linear(Hz) possuem um espaçamento e largura de banda determinados por uma constante de intervalo de frequência mel. O banco de filtro na escala mel é uma série de X filtros triangulares de banda passante de mesmo espaçamento e largura de banda que foram designados para simular o filtro que ocorre no sistema auditório humano(TIWARI, 2010). Um exemplo desse banco de filtros em escala linear pode ser visto na Figura 8, pode-se notar nessa imagem que até 1000 Hz a distância entre os filtros é a mesma e depois começa a aumentar, pois é a parte que a audição humana tem sua percepção logarítmica. O banco de filtros é aplicado no sinal em escala linear, então é obtido como saída o sinal filtrado em escala linear, mas de acordo com as componentes mel. Neste caso, é feito a conversão para escala logarítmica utilizando, por exemplo, a equação 2.1, deixando o sinal menos sensível a pequenas variações de amplitudes(S.D., 2013).

O *cepstrum* se trata da conversão de um espectro de frequência em escala logarítmica para o domínio do tempo. A representação cepstral do espectro da fala fornece uma boa representação do sinal para análise de quadro, devido aos coeficientes de espectro mel serem números reais. Além disso, pode-se converte-los para o domínio do tempo utilizando a DCT(*discrete cosine transform*)(TIWARI, 2010). Os coeficientes de saída desse bloco são usados como entrada dos modelos de aprendizado de máquina estudados neste trabalho.

Figura 8 – Banco de filtros MFCC



Fonte: (COURAS, 2017)

2.4 Aprendizado de Máquina

Na internet pode-se encontrar diversas definições interessantes sobre aprendizagem de máquina, dentre elas podemos citar a de Google Cloud (2017), "*Machine learning is functionality that helps software perform a task without explicit programming or rule*" e dos autores Bali et al. (2016), "*The field of study interested in the development of computer algorithms to transform data into intelligent action is known as machine learning*". Com estas definições podemos dizer que as técnicas de aprendizagem de máquina podem ser utilizadas para ensinar um computador a tomar decisões futuras baseadas nas características de um determinado conjunto de dados disponível. As técnicas de aprendizado de máquina podem ser classificadas como supervisionadas, não supervisionadas e por reforço.

As técnicas de aprendizagem supervisionada são aquelas em que o processo de treinamento é realizado com um conjunto de dados que possui uma ou mais variáveis independentes (entradas) e uma ou mais variáveis dependentes (saídas), em que tanto as entradas quanto as saídas são conhecidas. Estas técnicas podem ser utilizadas para resolver problemas de regressão, em que se deseja prever dados numéricos, ou em problemas de classificação, em que se deseja prever categorias(AYODELE, 2010).

No aprendizado não supervisionado, o processo de geração de modelo é realizado com um conjunto de dados em que apenas as entradas são fornecidas. Estes métodos podem ser utilizados para problemas em que deseja identificar padrões úteis nos dados e para segmentação de grupos com semelhanças, por exemplo(RUSSELL; NORVIG, 2003).

As técnicas de aprendizado por reforço não serão abordadas neste trabalho. De forma sucinta, essas técnicas funcionam com um processo de treinamento realizado a partir de recompensas nas interações com o ambiente. Exemplos de utilização podem incluir otimização jogos, entre outros.

Neste trabalho, o sistema pode ser encarado como um classificador, pois é feito reconhecimento de falante para verificação, onde a saída deste sistema classifica se a voz de entrada corresponde ou não corresponde com a voz que está sendo verificada. Levando isso em consideração, na sequência apresentaremos algumas técnicas de aprendizagem de máquina que poderão ser testadas neste trabalho.

2.4.1 Algoritmos de aprendizado de Máquina

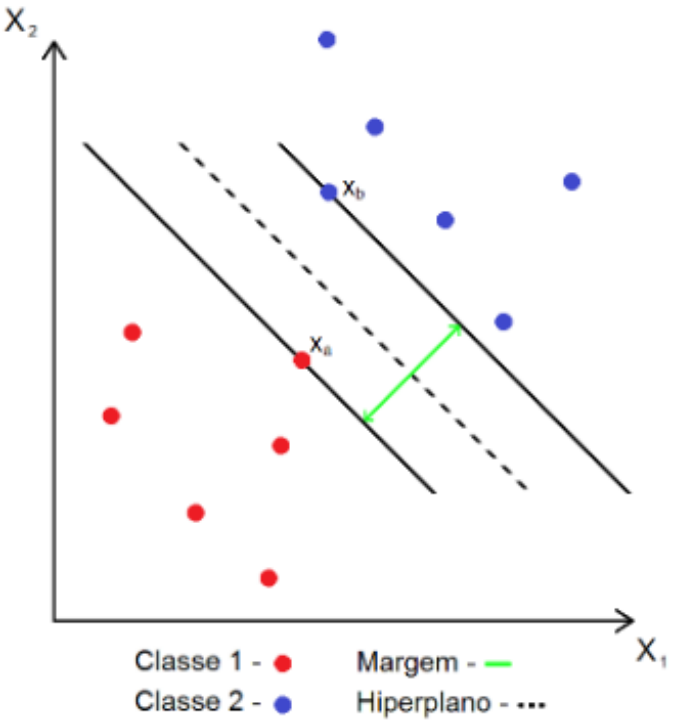
2.4.1.1 *Support Vector Machines*

O *Support Vector Machines* (SVM) é um classificador que modela limites de decisão entre classes através de hiperplano(KINNUNEN; HAIZHOU, 2010). Por exemplo, Pensando em plano cartesiano, o separador de hiperplano separa o plano em dois, onde um lado do plano pertence a uma classe e o outro pertence à outra classe, como pode ser visto na Figura 9. Se existirem múltiplas soluções de limite de decisão, então é necessário tentar localizar o que gere a menor quantidade de erros de generalização. O

SVM aborda esse problema através do conceito de margem para qual é definido ser à menor distância entre os limites de decisão e qualquer outra amostra. O limite de decisão escolhido deve ser aquele que possui a margem maximizada(BISHOP, 2006).

Os *support vectors*(vetores de suporte) corresponde aos vetores que ficam na margem máxima do hiperplano. Na Figura 9 seriam os vetores X_a e X_b. Após feito o treinamento o restante dos vetores podem ser descartados, pois eles não fazem diferença para o limite de decisão(BISHOP, 2006).

Figura 9 – Separador Hiperplano SVM



Fonte: (OLIVEIRA, 2017)

Em verificação de falante, uma classe consiste dos vetores de treinamento do falante alvo e a outra classe consiste dos vetores de falantes impostores. Usando esses vetores rotulados por suas classes, o SVM localiza um separador de hiperplano que maximiza a margem de separação entre as duas classes(KINNUNEN; HAIZHOU, 2010).

A Figura 9 também mostra a separação feita quando se trata de duas classes que são linearmente separáveis no plano original de entrada(X_1 e X_2), mas muitas vezes não é o caso. Nestas situações, são usadas as funções *kernel* que permitem a computação dos produtos internos de dois vetores no espaço de característica do *kernel*, o qual possui espaço de dimensão maior do que do espaço de entrada. Em um espaço de dimensão maior, é mais fácil de separar as duas classes com um hiperplano. De uma forma intuitiva, um hiperplano linear no espaço de característica de *kernel* corresponde em limite de decisão não linear no espaço original de entrada, um exemplo seria o espaço do MFCC(KINNUNEN; HAIZHOU, 2010).

2.4.1.2 *Naive Bayes*

Naive Bayes é um modelo em que as características de entrada dos dados de treinamento são consideradas condicionalmente independentes para simplificar a estrutura do modelo. Devido a esta

consideração, pode-se simplesmente aplicar a equação 2.2 adaptada de (SUNNY; S; K, 2013). Com essa equação é calculada a probabilidade de ser o real falante dada as características de entrada. $P(X)$ é a probabilidade marginal das características de entrada, $P(Falante)$ é a probabilidade marginal de ser o real falante e $P(X|Falante)$ é a probabilidade do real falante exibir as características de entrada. No caso de verificação só existe duas classes, então só é necessário aplicar a equação uma vez, pois sabendo a probabilidade de ser o real falante, se sabe também a probabilidade de não ser o real falante e o desejado é localizar a classe que tem a máxima probabilidade para ser feita a classificação.

$$P(Falante|X) = \frac{P(X|Falante) \cdot P(Falante)}{P(X)} \tag{2.2}$$

A assunção de independência condicional no modelo é claramente forte que pode conduzir representações ruins das densidades condicionais das classes. Mesmo se a assunção não é precisamente satisfeita, o modelo pode ainda fornecer uma boa performance na prática porque os limites de decisão podem ser insensíveis para alguns dos detalhes em densidades condicionais das classes(BISHOP, 2006)

2.4.1.3 GMM

Um modelo de mistura finito, é uma abordagem comum em cenários de verificação de falante independente de texto(LIU et al., 2017). Modelos de mistura finitos e seus métodos de estimação de parâmetros típicos podem aproximar uma grande variedade de funções densidade de probabilidade, portanto em situações onde uma única distribuição falha, eles são atrativas soluções para casos onde uma única função é criada. No entanto, de um ponto de vista prático, é frenquentemente comum criar misturas usando um tipo pré-definido de distribuição(BAGUL; SHASTRI, 2013).

Normalmente a distribuição pode ser de qualquer tipo, mas a distribuição gaussiana é uma das mais conhecidas e úteis distribuições em estática, sendo predominante em muitas áreas de aplicações. Por exemplo, em uma análise multivariada a maioria dos procedimentos de inferência existentes tem sido desenvolvidas com assunção de normalidade e de problemas de modelos lineares, o vetor de erro é frequentemente assumido ser normalmente distribuído. Portanto, se não há conhecimento da função densidade de probabilidade do fenômeno, somente um modelo geral pode ser usado. A distribuição gaussiana é um bom candidato devido ao enorme esforço em pesquisa no passado(BAGUL; SHASTRI, 2013).

A ideia fundamental do GMM é a distribuição do vetor de características extraídas da fala de uma pessoa ser modelada por uma densidade GMM(CHEN; WANG; CHI, 1997). Para um vetor de características de dimensão D denotado como $\vec{x}$, a densidade de mistura para o falante s pode ser definida pela equação 2.3 (REYNOLDS, 1995).

$$P(\vec{x}|\lambda_s) = \sum_{i=1}^{M} p_i^s b_i^s(\vec{x}) \tag{2.3}$$

A densidade é uma combinação linear de M componentes de densidades gaussianas unimodais ($b_i^s(\vec{x})$) ponderadas pelo peso(p_i^s). Cada densidade gaussiana unimodal é parametrizada pelo vetor de médias($\vec{\mu}_i^s$) e matriz covariância $\sum_i^s$ como é representado pela equação 2.4 (REYNOLDS, 1995).

$$b_i^s(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\sum_i^s|^{1/2}} \times \exp -\frac{1}{2}(\vec{x} - \vec{\mu}_i^s)(\sum{}_i^s)^{-1}(\vec{x} - \vec{\mu}_i^s) \tag{2.4}$$

A soma de todos os pesos é igual a 1. Os parâmetros do modelo densidade do falante s são denotados como $\lambda_s = \{p_i^s, \vec{\mu}_i^s, \sum{}_i^s\}$ para $i = 1, \cdots, M$(REYNOLDS, 1995).

O método mais conhecido para estimar o modelo que possui máxima probabilidade com GMM (utilizando os dados de treinamento disponíveis) é o algoritmo *Expectation-Maximization* (EM). O algoritmo EM pode ser representado pela equação 2.5, onde T é a sequência dos vetores de características de treinamento e $\vec{x}_t$ é o vetor de treinamento. A ideia básica do algoritmo EM é começar com um modelo inicial e à partir desse modelo estimar um novo modelo, fazendo isso até que se consiga uma convergência da máxima probabilidade(AROON; DHONDE, 2015). A convergência da máxima probabilidade significa que foi localizado o modelo que possui a função densidade de probabilidade(mistura de gaussianas) que mais reflete os dados que foram colocados de entrada, ou seja, é localizada a média e o desvio padrão de cada gaussiana que melhor representa os dados de entrada. O número de iterações necessárias para convergir fica em média de 10 iterações(REYNOLDS, 1995).

$$P(\vec{x}|\lambda_s) = \prod_{i=1}^{T} P(\vec{x}_t|\lambda_s) \tag{2.5}$$

Mais detalhes sobre o algoritmo EM pode ser localizado em Dempster, Laird e Rubin (1977).

3 PROPOSTA

A proposta deste trabalho é ser um introdutório para sistemas de reconhecimento de falante, mostrando aspectos importantes para se montar um sistema que utiliza técnicas de aprendizado de máquina para localizar padrões, descrevendo ferramentas disponíveis e disponibilizando códigos de exemplo que poderão ser usados em trabalhos futuros. O sistema de reconhecimento de falante considerado neste trabalho será baseado em aplicações de verificação usando a modalidade independente de texto.

Serão gerados modelos de falante utilizando diferentes algoritmos de aprendizado de máquina. O algoritmo GMM será o mais analisado, pois tem apresentado bom desempenho em aplicações de verificação de falante em cenários independentes de texto, de acordo com a literatura. Também faremos uma explanação detalhada das ferramentas disponíveis que podem ser utilizadas para esta finalidade. Neste ponto, o objetivo é mostrar ao leitor as vantagens e limitações das ferramentas disponíveis.

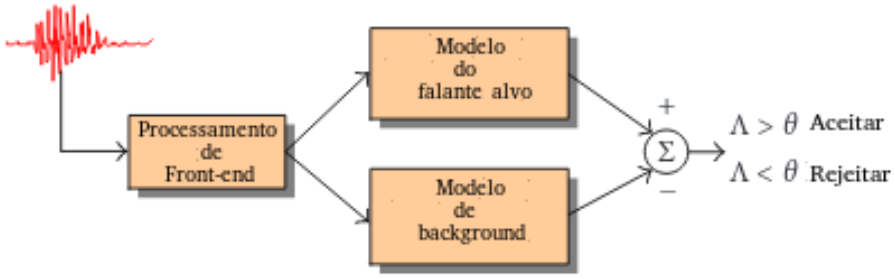
A ferramenta que pretendemos utilizar para o desenvolvimento do sistema proposto neste trabalho é a *toolbox Scikit-learn*, disponível para Python. O *Scikit-learn* uma ferramenta gratuita que possibilita o uso de aprendizado de máquina para não especialistas por meio da linguagem de programação de alto nível(PEDREGOSA et al., 2011). Existem ferramentas gratuitas de desenvolvimento focadas em reconhecimento de falante, por exemplo, *Spear* e *Alize*, mas que não dão muita liberdade no ajuste do algoritmo que gera os modelos. Existem também aplicações prontas como *Speaker Recognition API*, da Microsoft Azure que é paga e o aprendizado de máquina é totalmente abstraído, tendo uma Interface de Programação de Aplicativos (API) que é utilizada tanto para treinamentos quanto para testes. Por desejarmos uma ferramenta gratuita e que também permita uma liberdade maior na alteração de parâmetros do aprendizado de máquina no desenvolvimento, foi escolhido utilizar a *toolbox Scikit-learn* para o desenvolvimento. Para as demais ferramentas citadas será feita uma descrição e também uma análise comparativa com o sistema desenvolvido.

Inicialmente vamos modelar o sistema de acordo com a Figura 10 para o GMM e de acordo com a Figura 11 para os demais algoritmos de aprendizado de máquina. O GMM não é um classificador, portanto será usada uma técnica de modelagem que utiliza um modelo *backgrond* em conjunto com o modelo do falante alvo para fazer a classificação. O modelo de *background* informa a probabilidade do sinal de entrada pertencer a um impostor, podendo ser criada para cada falante cadastrado ou pode ser feito um modelo universal. O Modelo de *Background* Universal (UBM) é a abordagem mais predominante no momento(BIMBOT et al., 2004), onde é criado um único modelo de *background* para todos os falantes. Este modelo é criado da mesma forma que é criado o modelo de falante, sendo que as amostras de áudios usadas são escolhidas de forma aleatória dos dados de treinamento disponíveis. As amostras de voz do UBM serão distribuídas de forma que metade sejam de mulheres e a outra metade de homens, no intuito de tentar evitar que o sistema funcione melhor para um dos gêneros.

Em ambas modelagens, tem-se um sinal de voz como entrada e um processamento de *Front-end*. O processamento de *Front-end* se trata da extração das características do sinal de voz(MFCC) e a retirada de informações inúteis que podem prejudicar no reconhecimento de falante, por exemplo, o silêncio(BIMBOT et al., 2004).

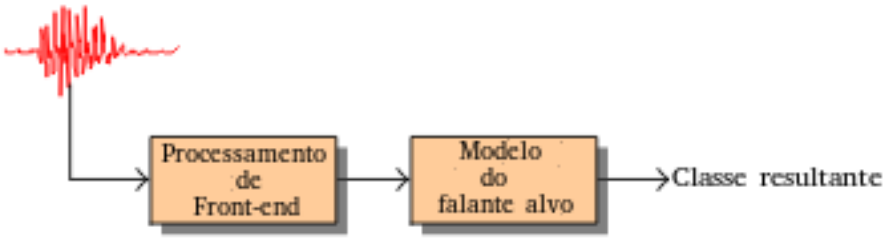
A base de dados utilizada neste trabalho será a base *English Language Speech Database for Speaker Recognition* (ELSDSR), que é um banco de dados dedicado à aplicações de reconhecimento de falante. Esse banco de dados possui a gravação do áudio de pessoas lendo uma única sessão de leitura, bem com possui uma quantidade relativamente extensa de amostras de áudio para aprender as características

Figura 10 – Modelagem do sistema utilizando GMM



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

Figura 11 – Modelagem do sistema



Fonte: Modificado de (BIMBOT et al., 2004)

específicas da fala de uma pessoa.

Para alcançar o objetivo proposto neste trabalho, estamos propondo 6 metas, conforme descrito no cronograma da tabela 1:

- M1 - Organização do banco de dados e desenvolver ou utilizar algum programa que retire os momentos de silêncio do áudio.
- M2 - Fazer a extração das características do sinal de voz utilizando o MFCC.
- M3 - Gerar os modelos de falante utilizando a *toolbox Scikit-learn*.
- M4 - Testar os modelos de falantes, fazer o melhoramento dos modelos e obter os resultados.
- M5 - Analisar ferramentas disponíveis.
- M6 - Elaboração de um documento final.

Tabela 1 – Cronograma das atividades previstas

Meta	Jul	Ago	Set	Out	Nov	Dez
M1	X					
M2	X	X				
M3		X	X			
M4			X	X		
M5				X	X	
M6					X	X

REFERÊNCIAS

ABDULLA, W. Auditory based feature vectors for speech recognition systems. p. 231–236, 01 2002. Citado na página 21.

ADAMOVIC, S.; MILOSAVLJEVIC, M. Information analysis of iris biometrics for the needs of crypto logy key extraction. In: *Serbian Journal of Eletrical Engineering*. [S.l.: s.n.], 2013. v. 10, n. 1, p. 1–12. Citado na página 14.

AROON, A.; DHONDE, S. Speaker recognition system using gaussian mixture model. v. 130, p. 38–40, 11 2015. Citado na página 26.

AYODELE, T. O. Machine learning overview. In: *New Advances in Machine Learning*. [S.l.]: InTech, 2010. Citado 2 vezes nas páginas 10 e 23.

BAGUL, S. G.; SHASTRI, R. K. Text independent speaker recognition system using gmm. In: *2013 International Conference on Human Computer Interactions (ICHCI)*. [S.l.: s.n.], 2013. p. 1–5. Citado 2 vezes nas páginas 10 e 25.

BALI, R. et al. *R: Unleash Machine Learning Techniques*. Packt Publishing, 2016. (Learning path). ISBN 9781787128286. Disponível em: <https://books.google.com.br/books?id=3ZfcDgAAQBAJ>. Citado na página 23.

BEIGI, H. *Fundamentals of Speaker Recognition*. New York, NY, USA: Springer, 2011. Citado 9 vezes nas páginas 9, 13, 14, 15, 16, 17, 18, 20 e 21.

BERNECKER, O. Biometrics: security: An end user perspective. *Information Security Technical Report*, v. 11, n. 3, p. 111–118, 2006. Citado na página 13.

BIMBOT, F. et al. A tutorial on text-independent speaker verification. *EURASIP Journal on Advances in Signal Processing*, Springer, v. 2004, n. 4, p. 101962, 2004. Citado 2 vezes nas páginas 27 e 28.

BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387310738. Citado 2 vezes nas páginas 24 e 25.

BOLES, A.; RAD, P. Voice biometrics: Deep learning-based voiceprint authentication system. In: *2017 12th System of Systems Engineering Conference (SoSE)*. [S.l.: s.n.], 2017. p. 1–6. Citado na página 9.

BONASTRE, J. F.; WILS, F.; MEIGNIER, S. Alize, a free toolkit for speaker recognition. In: *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* [S.l.: s.n.], 2005. v. 1, p. 737–740. ISSN 1520-6149. Citado na página 10.

CHEN, K.; WANG, L.; CHI, H. Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification. *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific, v. 11, n. 03, p. 417–445, 1997. Citado na página 25.

CLOUD, G. *What is machine learning?* 2017. <https://cloud.google.com/what-is-machine-learning/>. (Accessado em 01/07/2018). Citado na página 23.

COURAS, M. de F. K. B. *Classificação de Desvios Vocais Utilizando Características Baseadas no Modelo Linear de Produção da fala*. Dissertação (Mestrado) — Instituto Federal de Educação, Ciência e Tenologia da Paraíba, João Pessoa, fev 2017. Citado na página 23.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, JSTOR, p. 1–38, 1977. Citado na página 26.

FAUNDEZ-ZANUY, M.; MONTE-MORENO, E. State-of-the-art in speaker recognition. *IEEE Aerospace and Electronic Systems Magazine*, v. 20, n. 5, p. 7–12, 2005. Citado 2 vezes nas páginas 9 e 10.

HASAN, M. R. et al. Speaker identification using mel frequency cepstral coefficients. 12 2004. Citado 2 vezes nas páginas 20 e 22.

HONG, L. *Automatic Personal Identification Using Fingerprints*. Tese (Doutorado), East Lansing, MI, USA, 1998. AAI9909316. Citado 2 vezes nas páginas 13 e 14.

KHOURY, E.; SHAFEY, L. E.; MARCEL, S. Spear: An open source toolbox for speaker recognition based on Bob. In: *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. [s.n.], 2014. Disponível em: <http://publications.idiap.ch/downloads/papers/2014/Khoury_ICASSP_2014.pdf>. Citado na página 10.

KINNUNEN, T.; HAIZHOU, L. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, v. 52, n. 1, p. 12–40, 2010. Citado 6 vezes nas páginas 9, 15, 20, 22, 23 e 24.

LATHI, B. P. *Linear Systems and Signals*. 2nd. ed. Oxford, UK: Oxford University Press, 2009. ISBN 0195392566, 9780195392562. Citado na página 19.

LIU, Y. et al. Investigation of frame alignments for gmm-based text-prompted speaker verification. *CoRR*, abs/1710.10436, 2017. Disponível em: <http://arxiv.org/abs/1710.10436>. Citado na página 25.

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 (E)*, Acoustical Society of Japan, v. 20, n. 4, p. 281–291, 1999. Citado na página 21.

MOHAMED, Z.; HAMIDIA, M.; AMROUCHE, A. *Investigation of the Use of PLP Coefficients for Speaker Recognition over IP Network*. 2013. Citado na página 21.

MUDA, L.; BEGAM, M.; ELAMVAZUTHI, I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *CoRR*, abs/1003.4083, 2010. Disponível em: <http://arxiv.org/abs/1003.4083>. Citado na página 10.

OLIVEIRA, G. C. *Localização indoor utilizando a tecnologia LoRaWAN e aprendizado de máquina*. Bacharelado em Engenharia de Telecomunicações — Instituto Federal de Santa Catarina, 2017. Citado na página 24.

O'SHAUGHNESSY, D. *Speech communication: human and machine*. Universities Press (India) Pvt. Limited, 1987. (Addison-Wesley series in electrical engineering: digital signal processing). ISBN 9788173713743. Disponível em: <https://books.google.com.br/books?id=UWmD8aY_448C>. Citado na página 21.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 27.

RAMGIRE, J. B.; JAGDALE, S. M. A survey on speaker recognition with various feature extraction and classification techniques. *International Research Journal of Engineering and Technology*, v. 3, n. 4, p. 709–712, 2016. Citado na página 10.

REYNOLDS, D. A. Speaker identification and verification using gaussian mixture speaker models. *Speech Commun.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 17, n. 1-2, p. 91–108, ago. 1995. ISSN 0167-6393. Disponível em: <http://dx.doi.org/10.1016/0167-6393(95)00009-D>. Citado 3 vezes nas páginas 10, 25 e 26.

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 2. ed. [S.l.]: Pearson Education, 2003. ISBN 0137903952. Citado na página 23.

S.D., B. M. D. Automatic speech and speaker recognition by mfcc, hmm and vector quantization. v. 3, p. 93–98, 07 2013. Citado na página 22.

SUNNY, S.; S, D. P.; K, P. J. Combined feature extraction techniques and naive bayes classifier for speech recognition. p. 155–163, 07 2013. Citado na página 25.

TIWARI, V. Mfcc and its applications in speaker recognition. v. 1, 01 2010. Citado na página 22.

WAN, V.; CAMPBELL, W. M. Support vector machines for speaker verification and identification. In: *Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No.00TH8501)*. [S.l.: s.n.], 2000. v. 2, p. 775–784 vol.2. ISSN 1089-3555. Citado na página 10.