UNIVERSIDADE DO ESTADO DO AMAZONAS – UEA

GRADUAÇÃO EM ENGENHARIA ELÉTRICA

NÍVEL GRADUAÇÃO

GUSTAVO DE AQUINO E AQUINO

CONTROLE DE ACESSO BASEADO EM RECONHECIMENTO DE VOZ

MANAUS

05/12

GUSTAVO DE AQUINO E AQUINO

CONTROLE DE ACESSO BASEADO EM RECONHECIMENTO DE VOZ

Projeto de Pesquisa desenvolvido durante a disciplina de Trabalho de Conclusão de Curso I e apresentado à banca avaliadora do Curso de Engenharia Elétrica da Escola Superior de Tecnologia da Universidade do Estado do Amazonas, como pré-requisito para a obtenção do título de Engenheiro Eletricista.

Orientador: Wheidima Carneiro de Melo

Coorientador: Jozias Parente de Oliveira

MANAUS

05/12

Universidade do Estado do Amazonas – UEA

Escola Superior de Tecnologia - EST

Reitor:

Cleinaldo de Almeida Costa

Vice-Reitor:

Cleto Cavalcante de Souza Leal

Diretor da Escola Superior de Tecnologia:

Roberto Higino Pereira da Silva

Coordenador do Curso de Engenharia Elétrica:

Ingrid Sammyne Gadelha Figueiredo

Banca Avaliadora composta por:

Prof. Jozias Parente de Oliveira *(Co-orientador)*

Prof. Walfredo Da Costa Lucena Filho *Data da defesa: 05 / 12/ 2018.*

Prof. Jose Ruben Sicchar Vilchez

CIP – Catalogação na Publicação

Aquino, Gustavo de Aquino e

Controle de acesso baseado em reconhecimento de voz /Gustavo Aquino; [orientado por] Wheidima Carneiro de Melo. [co-orientado por] Jozias Parente de Oliveira. – Manaus: 2018.

71 f. p.: il.

Trabalho de Conclusão de Curso (Graduação em Engenharia Elétrica). Universidade do Estado do Amazonas, 2018.

1. Processamento Digital de Sinais. 2. Mel Frequency Coefficients.
3. Reconhecimento de voz.
I. Carneiro de Melo, Wheidima. II Jozias Parente de Oliveira.

GUSTAVO DE AQUINO E AQUINO

CONTROLE DE ACESSO BASEADO EM RECONHECIMENTO DE VOZ

Pesquisa desenvolvida durante a disciplina de Trabalho de Conclusão de Curso II e apresentada à banca avaliadora do Curso de Engenharia Elétrica da Escola Superior de Tecnologia da Universidade do Estado do Amazonas, como pré-requisito para a obtenção do título de Engenheiro Eletricista.

Aprovado em 05 de dezembro de 2018

BANCA EXAMINADORA

Jozias Parente de Oliveira – Universidade do Estado do Amazonas

---

Walfredo da Costa Lucena Filho -Universidade do Estado do Amazonas

---

Jose Ruben Sicchar Vilchez – Universidade do Estado do Amazonas

Área de concentração: Processamento Digital de Sinais

DEDICATÓRIA

Dedico este trabalho aos meus familiares, a todos que contribuíram para a minha formação acadêmica e aqueles que se interessam por processamento de áudio.

AGRADECIMENTOS

A Deus, por me dar forças para conseguir concluir este trabalho;

Ao meu pai Lazaro Barros e minha mãe Hermina, que me proporcionaram o apoio necessário durante todo o curso;

Ao professor, Wheidima Melo, orientador deste trabalho, por me apresentar formas de como realizar a minha ideia inicial;

Ao professor, Jozias Parente, devido a todo o seu auxílio e disponibilidade de sanar minhas dúvidas;

A todos os meus professores da graduação, que de alguma forma contribuíram para aumentar meus conhecimentos.

EPÍGRAFE

"A educação tem raízes amargas, mas os seus frutos são doces". Aristóteles.

RESUMO

Um sistema de reconhecimento de voz é projetado para identificar o usuário a partir da sua fala. Por meio do uso de técnicas modernas de Processamento Digital de Sinais, aplicadas no aplicativo MATLAB, a voz do administrador pode ser autenticada. A ideia básica por trás disso é converter a forma de onda de fala em um tipo de representação paramétrica, para que as características possam ser extraídas e analisadas. Existe uma ampla gama de possibilidades para representar parametricamente o sinal de fala, para o sistema de reconhecimento da voz, como o método *Mel Frequency Cepstrum Coeffients* (MFCC). O sinal de voz de entrada é gravado e por meio dos procedimentos do MFCC, as características podem ser extraídas e armazenadas em um banco de dados. No sistema biométrico deste projeto o usuário profere um trecho de voz, no caso, seu nome, na fase de treinamento, de modo a treinar o programa. Na fase de testes, o trecho é novamente falado pelo usuário, a fim de alcançar o reconhecimento, caso haja a correspondência desejada. Os testes, realizados com 20 usuários, apontaram que o sistema implementado alcançou níveis de precisão bastante satisfatórios de 95%, reconhecendo com sucesso a voz dos usuários cadastrados e rejeitando a voz de outros indivíduos. O uso da voz como forma de autenticação de segurança se mostrou possível através dos métodos citados neste documento.

Palavras-chave: Biometria vocal - Mel Frequency Cepstral Coefficients (MFCC) - Reconhecimento de voz.

ABSTRACT

A voice recognition system is designed to identify the user from your speech. Through the use of modern Digital Signal Processing techniques, applied in the MATLAB, the user's voice can be authenticated. The basic idea behind this is to convert the speech waveform into a type of parametric representation, so these characteristics can be extracted and analyzed. There is a wide range of possibilities for parametrically representing the speech signal, for the voice recognition system, such as the Mel Frequency cepstrum coeffients (MFCC) method. The input voice signal is recorded and through the MFCC procedures, the characteristics can be extracted and stored in a database. In the biometric system of this project, the user utters a voice snippet, in this case, his name, in the training phase, in order to train the program. In the testing phase, the snippet is again spoken by the user in order to achieve recognition, if there is the desired correspondence. The tests, performed with 20 users, pointed out that the implemented system achieved very satisfactory levels of accuracy of 95%, successfully recognizing the voice of the registered users and rejecting the voice of other individuals. The use of voice as a form of security authentication proved possible through the methods cited in this document.

Keywords: Vocal Biometrics - Mel Frequency Cepstral Coefficients (MFCC) – Voice Recognition.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

O uso da tecnologia a favor da segurança é cada vez mais comum, sendo em áreas comerciais ou residenciais, afim de elevar o nível de proteção da propriedade. Infelizmente, todos os sistemas de segurança apresentam falhas, as senhas podem ser *hackeadas* e os cartões de acesso podem ser duplicados. Baseado nisso, toda uma nova tecnologia tem surgido para aumentar o nível de segurança do usuário.

A tecnologia biométrica se destaca, pois utiliza características do usuário como senha. Isso a torna mais conveniente para o usuário pois nenhuma senha precisa ser lembrada ou quaisquer objetos precisam ser transportados. Os padrões biológicos, são únicos de cada pessoa, como por exemplo da digital, da face e da voz (RASHID et al, 2008). Dentre esses, o reconhecimento de locutor apresenta algumas vantagens intrínsecas que o destaca dos demais sistemas biométricos. A voz é o meio mais natural de comunicação, é possível fazer a aquisição do sinal sem desconforto, não há necessidade de contato físico com o usuário e, além disso, a voz fornece não só características físicas como também comportamentais (VACCA, 2007). Portanto, sistemas de reconhecimento de voz estão sendo cada vez mais estudados.

A tecnologia de autenticação vocal se tornou possível em meados dos anos 70, e nos anos 90 finalmente consolidou-se comercialmente, possibilitando sua utilização no mercado atual através de uma excelente relação custo/benefício. De um ponto de vista tecnológico, esses sistemas são conhecidos como *Automatic Speech Recognition* (ASR) e se dividem em dependentes de texto ou independentes de texto de uma palavra (CAMPBELL, 1997). Esses sistemas codificam a voz do usuário para analisar características específicas da fala. O estudo em questão irá focar em sistemas DVI, utilizando o método de extração de características *Mel Frequency Cepstrum Coefficients* (MFCC), que é recomendado para reconhecimento vocal por modelar a voz de acordo com as características auditivas humanas (MUDA, 2010).

O projeto em questão utilizou basicamente duas ferramentas, o MATLAB e o ARDUINO. A primeira serviu para a parte da criação de algoritmos de reconhecimento vocal, enquanto a segunda terá o propósito de ser usada para demonstrar uma aplicação prática do projeto, como enviar um sinal de autenticação para controlar um dispositivo.

A problemática de criação desse sistema é extrair os padrões vocais e quantifica-los de uma forma matemática, para que esses possam ser comparados e medidos. Diferente dos outros sistemas biométricos, o de biometria vocal não trabalha com algo puramente físico,

como o caso da biometria por digital. A hipótese então é que possa ser criado um sistema de controle de acesso baseado em reconhecimento de voz.

Portanto, mais especificamente, o objetivo deste trabalho é cadastrar usuários, extrair suas características vocais, através do método MFCC, e armazena-las em um banco de dados, para no final verificar se a precisão do sistema criado é suficiente para o uso em aplicações na área de segurança.

No Capitulo 1.1 são mostradas as características biológicas da voz, para dar um embasamento de como é feita a produção vocal e os motivos de porque a voz é única. Além de mostrar caraterísticas intrínsecas para trabalhar com sinais vocais e a importância do espectrograma.

No Capítulo 1.2 são abordados os conceitos de processamentos de sinais e sinais e sistemas, que formam a base teórica para os procedimentos que serão abordados no capítulo seguinte.

O Capítulo 1.3 apresenta ao leitor o conceito de sistemas ASR, seus diferentes tipos, possíveis fatores de erros, bem como formas para ser implementado utilizando MFCC.

Já o Capítulo 2 trata a respeito dos métodos e procedimentos, utilizados para o desenvolvimento e validação da pesquisa.

O Capítulo 3 mostra como o sistema foi implementado, como foi formado o banco de dados, além de comentar passo a passo os algoritmos no MATLAB e como ocorre a integração entre eles.

Então no Capítulo 4 são discutidos os resultados e como foram feitos os testes no sistema implementado, mostrando também o que cada teste significa.

Por fim, no Capitulo 5 é apresentada a conclusão do trabalho.

1. **REFERENCIAL TEÓRICO**

1.1 A VOZ HUMANA

O sinal de fala possui uma série de especificidades que o torna único, devido a razões biológicas na formação da fala há muitos fatores físicos que influenciam para isso. O discurso falado é uma série de fonemas aleatórios, porém cada pessoa os fala de forma única. Há faixas de frequências específicas da voz. Uma análise espectral, através de um espectrograma, revela o enorme conteúdo harmônico presente na fala, o processo de saber identificar e analisar essas características pode ser útil para uma serie de aplicações.

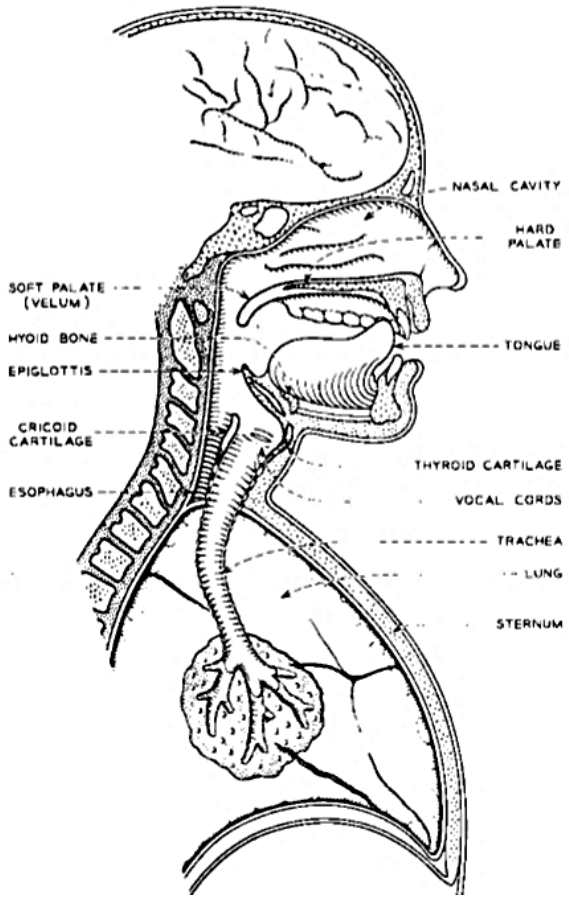
Nesta primeira parte será feito um estudo dos fatores biológicos da voz, bem como as características para se trabalhar com um sinal de fala. Será também analisado os espectrogramas e sua importância para análise vocal.

1.1.1. **Formação biológica da voz**

A voz humana é única, devido aos seus aspectos físicos e psicológicos. O trato vocal é uma distinção física importante, ele é geralmente considerado como o órgão responsável pela produção da fala e se encontra acima das dobras vocais (CAMPBELL, 1997). Como mostrado na Figura 1, isso inclui o seguinte:

- Faringe da laringe (sob a epiglote);
- Faringe oral (por trás da língua, entre a epiglote e velum);
- Cavidade oral (para a frente do velum e limitada pelos lábios, língua e palato);
- Faringe nasal (acima do velum, extremidade traseira da cavidade nasal);
- Cavidade nasal (acima do palato e estendendo-se da faringe às narinas).

Figura 1: Sistema vocal humano.



Fonte: Campbell, 1997.

Um trato vocal masculino adulto é de aproximadamente 17 cm de comprimento. As pregas vocais (anteriormente conhecidas como cordas vocais) são mostradas na Fig. 1. A laringe é composta das pregas vocais, o topo da cartilagem cricóide, as cartilagens aritenóide, e a cartilagem da tireóide. As pregas vocais são esticadas entre a cartilagem da tireóide e as cartilagens aritenóide. A área entre as pregas vocais é conhecida como glottis. As cordas vocais são formadas por dois pares de músculos esticados transversalmente ao fluxo de ar, e localizam-se entre a traquéia e a laringe.

Conforme a onda acústica passa através do trato vocal, o seu conteúdo de frequência (espectro) é alterado pelas ressonâncias do trato vocal. As ressonâncias vocais são chamadas de formadores. Assim, a forma do trato vocal pode ser estimada a partir da forma espectral (por exemplo, a localização do interlocutor e a inclinação espectral) do sinal de voz.

Sistemas de autenticação vocal normalmente usam recursos derivados apenas do trato vocal. Como visto na Fig. 1, o mecanismo vocal humano é impulsionado por uma fonte de

excitação, que também contém informações dependentes do falante. A excitação é gerada pelo fluxo de ar nos pulmões, realizada pela traqueia através das pregas vocais. A excitação pode ser caracterizada como fonação, sussurrarão, fricção, compressão, vibração, ou uma combinação destes.

1.1.2. **Características da voz**

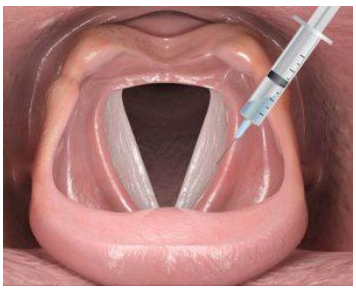
Agora que já foi mostrada a formação biológica do sinal de voz é importante ressaltar as características específicas intrínsecas desse sinal, tanto no domínio do tempo quando no da frequência.

A análise no domínio do tempo mostra que em um discurso são faladas várias sílabas por segundo. A pronúncia dessas sílabas é separada por intervalos de tempos que apresentam uma variação irregular e aleatória. A combinação desses sons, aqui chamados sílabas, é governada pelas regras da linguagem. O estudo das regras de formação das palavras e suas implicações na comunicação humana é conhecido como linguística, e o estudo e classificação dos sons vocálicos recebe a denominação de fonética (RABINER; SCHAFER, 1978).

Os sons produzidos pela fala humana podem ser classificados de duas formas, vocálicos ou fricativos.

Sons frictativos ocorrem quando o ar é forçado pelos pulmões, através das cordas vocais, em direção à boca ou nariz, por onde o som escapa. Devido as vibrações que ocorrem nas cordas vocais, Figura 2, a voz atinge frequências de 50 a 1000Hz, resultando em sopros 26 periódicos de ar injetado na traqueia (SMITH, 1997).

Figura 2: Representação das cordas vocais.

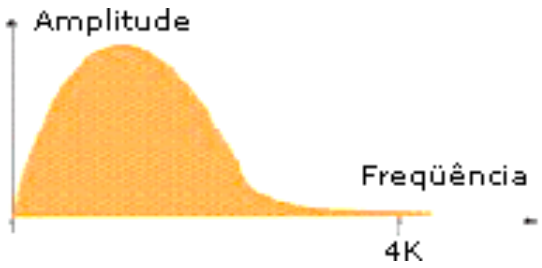


Fonte: (http://marcoseferin.com.br/2016/12/05/cordas_vocais/)

O som vocálico, é como o próprio nome sugere um som que surge ao se pronunciar vogais (a, e, i, o, u). Em contrapartida, sons fricativos ou não-vocálicos originam-se quando ocorre a constrição de algum ponto do trato vocal, geralmente em direção à boca, e o ar é forçado através da constrição a uma velocidade suficientemente grande para produzir turbulência, criando uma fonte de ruído que excita o trato vocal (RABINER; SCHAFER, 1978). Sons fricativos são aqueles cuja pronúncia inclui: /ch/, /f/, /s/, /v/, /x/, e /z/.

As recomendações G.132 e G.151 do ITU-T indicam a banda atribuída ao sinal de voz de 300 Hz a 3KHz, porém todo o processo de análise do sinal de voz, é feito levando em conta a faixa de frequência de 1 Hz a 4 kHz (DÍGITRO, 2003), como mostrado na Fig.3.

Figura 3: Conteúdo espectral da voz



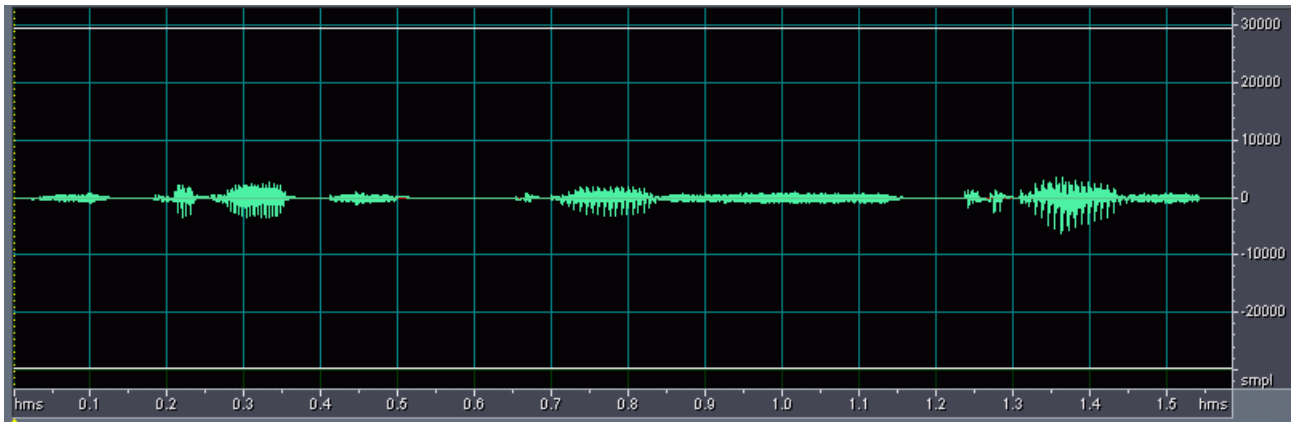
Fonte: Dígitro, 2003.

1.1.3. **Espectrogramas**

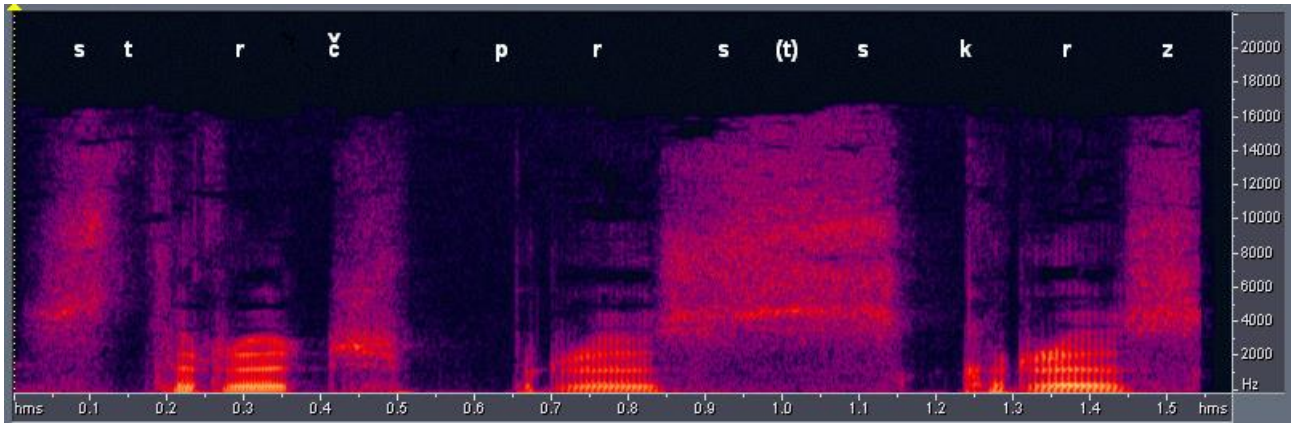
Um sinal de voz humana é complexo de ser analisado por inteiro, sua análise se torna mais simples quando feita no domínio da frequência através da transformada Fourier, que quebra o sinal de voz em uma combinação de sons mais simples, harmonicamente relacionados, ditos assim por possuir frequências múltiplas de uma frequência fundamental.

O estudo de um sinal por meio dos componentes harmônicos mais simples é dito análise espectral, os gráficos espectrais, ou espectrogramas, são úteis para expor a quantidade de informação que está presente em um trecho sonoro, Fig.4. Os espectrogramas expõem dinamicamente a quantidade de energia através do tempo. Seu gráfico simula um espaço de três dimensões entre amplitude, frequência e tempo em um plano bidimensional, de frequência e tempo para isso ele usa diferentes cores, variando do violeta ao vermelho do espectro visível.

Figura 4 (a): Gráfico de um sinal de voz através da distribuição das amplitudes pelo tempo.



(b): Espectrograma do mesmo trecho sonoro, no qual as cores mais próximas do violeta, ou seja, mais fortes, representam uma maior amplitude.



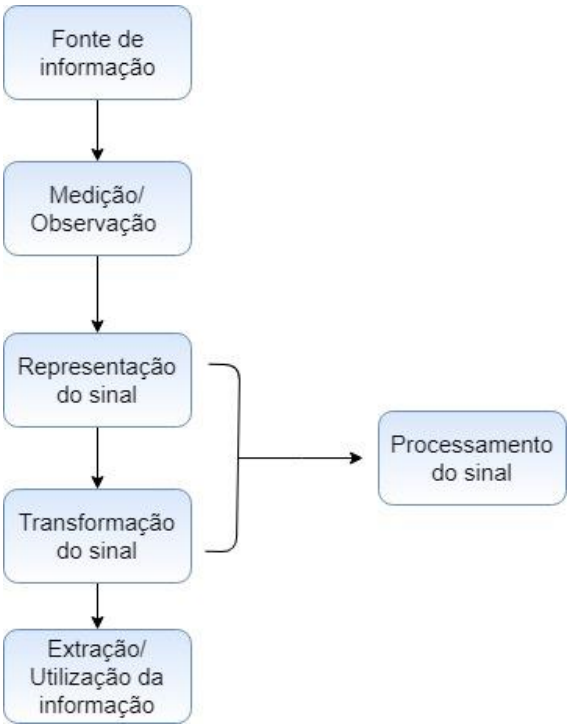
Fonte : (https://commons.wikimedia.org/wiki/File:StrcPrstSkrzKrk.png)

1.2 PROCESSAMENTO DE SINAIS

Uma das formas mais populares de caracterizar a fala é em termos de um sinal ou forma de onda acústica. A figura 5 demonstra que a informação contida em um sinal de fala pode ser extraída por ouvintes humanos ou computadores. O processamento digital desempenha um papel fundamental no processamento da linguagem falada. Neste tópico são descritos os fundamentos de processamento digital de sinais: sinais e sistemas digitais, transformadas discretas e contínuas no domínio da frequência, filtros digitais e a relação entre sinais analógicos e digitais.

O foco desta parte é o desenvolvimento de métodos no domínio da frequência computados por meio da transformada de Fourier. A representação desses sinais no domínio da frequência é especialmente útil pois a estrutura de um fonema é geralmente única (HUANG, 2001).

Figura 5: Importância do processamento de sinais no processamento da fala



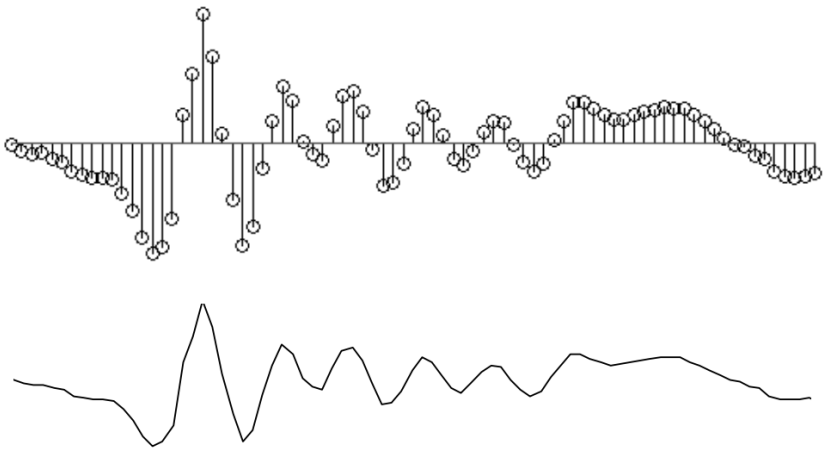
Fonte: Huang, 2001

1.2.1 **Sinais digitais**

Para processamento do sinal de fala, é conveniente representa-las matematicamente como uma função de uma variável contínua t, que representa o tempo. Define-se um sinal analógico $x_a(t)$ como uma função variando continuamente no tempo. Se um sinal x for amostrado com um período de amostragem T (sendo o tempo representado por t=n.T) pode-se definir um sinal discreto no tempo como $x[n] = x_a(nT)$, Fig.6, também conhecido como sinal digital (HUANG, 2001). É comum o uso de parênteses para sinais analógicos e colchetes para sinais digitais. Além disso, pode-se definir a frequência F_s como sendo o inverso o período.

Mais adiante será mostrado que em determinada circunstâncias um sinal analógico pode ser perfeitamente representado por um sinal digital.

Figura 6: Representação de sinais analógicos, e digitais, respectivamente.



Fonte: Huang, 2001

O termo Processamento Digital de sinais (PDS) se refere aos métodos de manipulação das sequências de $x[n]$ em um computador digital. O termo PDS também pode se referir a um processador digital de sinais.

É importante começar a análise dos sinais por ondas senoidais, visto que Fourier demonstrou por meio de suas séries que estas podem ser usadas para representar quase todos os tipos de sinais.

1.2.2 **Sinais senoidais**

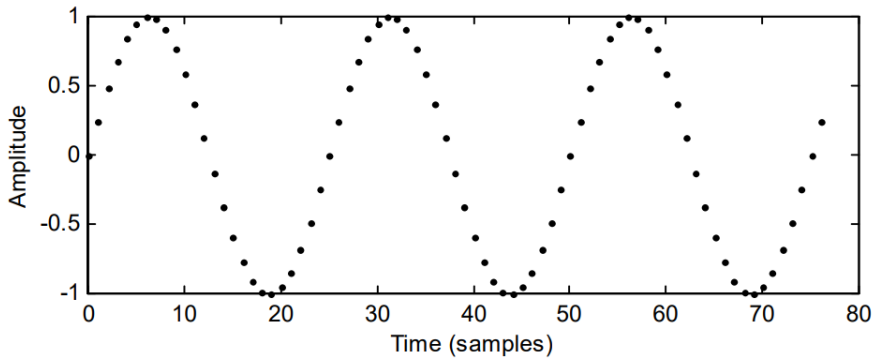
Um dos mais importantes sinais na engenharia elétrica, devido ao uso da corrente alternada nos sistemas elétricos, Fig. 7. Conhecido como onda senoidal ou senoide esta pode ser definida como mostrado abaixo:

$$x_0[n] = A_0 \cos(\omega_0 n + \phi_0) \quad (1)$$

Onde A_0 é a amplitude, ω_0 é a frequência angular e ϕ_0 a fase. O ângulo em funções trigonométricas é expresso em radianos, então a função ω_0 é usada para normalizar a frequência $\omega_0 = 2\pi f_0$ estando f_0 no intervalo de 0 a 1.

As senoides possuem várias formas de serem manipuladas matematicamente, é comum o uso de fasores e números complexos para manipula-las, além da trigonometria. A soma de senoides de diferentes frequências continua sendo uma senoide.

Figura 7: Representação de uma senoide



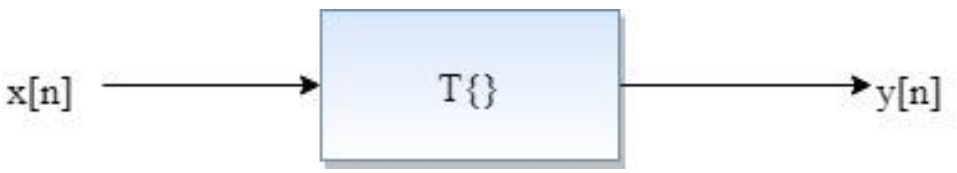
Fonte: Huang, 2001

1.2.3 **Sistemas digitais**

Um sistema digital, Fig.8, transforma uma entrada de sinal $x[n]$ em uma saída $y[n]$:

$$x[n] = T\{x[n]\} \tag{2}$$

Figura 8: Diagrama em blocos de um sistema digital



Fonte: Huang, 2001

O estudo de sistemas digitais é feito de acordo com seu tipo, se são ou não lineares e invariantes ou invariantes no tempo.

Em geral um sistema digital T é definido como sendo linear se e somente se:

$$T\{a_1x_1[n] + a_2x_2[n]\} = a_1T\{x_1[n]\} + a_2x_2T\{x_2[n]\} \quad (3)$$

Para quaisquer valores de a_1, a_2, x_1 e x_2.

Um sistema invariante no tempo é um sinal cujo um atraso na entrada acarreta no mesmo atraso na saída:

$$y[n - n_0] = T\{y[n - n_0]\} \quad (4)$$

Os sistemas lineares e invariantes no tempo ou LTI (Linear Time-Invariant) são especiais pois podem ser completamente caracterizados pelo sinal $h[n]$, que é conhecido como resposta ao impulso do sistema. Eles são descritos por:

$$y[n] = \sum_{k=-\infty}^{\infty} x[k]h[n - k] = x[n] * h[n] \quad (5)$$

Onde * é definido como um operador de convolução. Este operador tem propriedades comutativa, associativa e distributiva.

1.2.4 **Transformada de fourier**

É importante saber o que é a saída de um sistema LTI com resposta ao impulso $h[n]$ quando a entrada é uma exponencial complexa. Substituindo $x[n] = e^{j\omega_0 n}$ na equação 5 e usando a propriedade comutativa da convolução tem-se:

$$y[n] = \sum_{k=-\infty}^{\infty} h[k]e^{j\omega_0(n-k)} = e^{j\omega_0 n}\sum_{k=-\infty}^{\infty} h[k]e^{-j\omega_0 k} = e^{j\omega_0 n}H(e^{j\omega_0}) \quad (6)$$

Que é outra exponencial complexa de mesma frequência e amplitude multiplicada pela quantidade complexa $H(e^{j\omega_0})$ dada por:

$$H\left(e^{j\omega_0}\right) = \sum_{n=-\infty}^{\infty} h[n]e^{-j\omega n} \quad (7)$$

Dessa forma é dito que a saída de um sistema LTI, neste caso, é um autovetor, sendo sua entrada um autovalor.

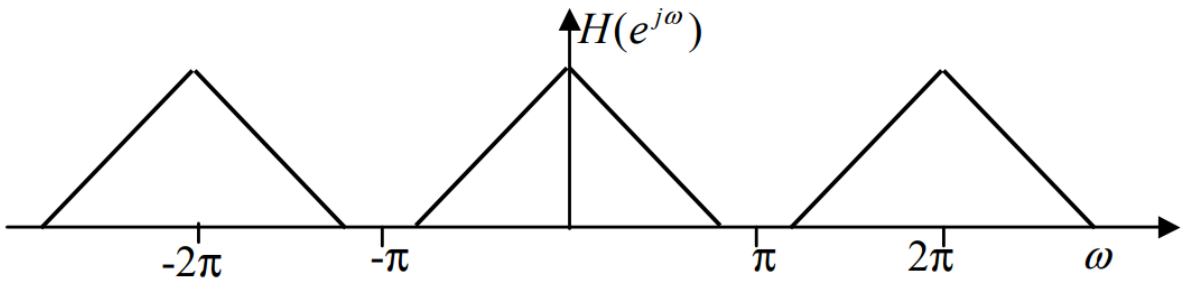
O termo $H(e^{j\omega_0})$ é definido como transformada de Fourier discreta no tempo de $h[n]$. Sendo $H(e^{j\omega_0})$ uma função periódica de ω com o período 2π, e como toda função periódica é necessário apenas um período para descreve-la completamente, tipicamente $-\pi < \omega < \pi$, como mostrado na Figura 9.

$H(e^{j\omega_0})$ é uma função complexa que como tal possui uma parte real, parte imaginária, amplitude e fase. Portanto se a entrada de um sistema LTI é uma senoide como na equação 1, a saída vai ser:

$$y[n] = A_0\left|H\left(e^{j\omega_0}\right)\right| \cos\left(\omega_0 n + \phi_0 + arg\left\{H\left(e^{j\omega_0}\right)\right\}\right) \tag{8}$$

Dessa forma se $\left|H\left(e^{j\omega_0}\right)\right| > 1$, o sistema LTI vai amplificar aquela frequência, e da mesma forma ele vai filtra-la se $\left|H\left(e^{j\omega_0}\right)\right| < 1$. Por essa razão que esses sistemas também são chamados de filtros. A transformada de Fourier $H(e^{j\omega_0})$ de um filtro $h[n]$ é chamada de resposta em frequência ou função de transferência do sistema.

Figura 9: $H\left(e^{j\omega_0}\right)$ é uma função periódica de ω.



Fonte: Huang, 2001

A transformada de Fourier admite inversa, dada por:

$$h[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} H\left(e^{j\omega}\right)e^{j\omega n}d\omega \tag{9}$$

Substituindo a Eq. 7 na Eq. 9, obtém-se:

$$h[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\sum_{-\infty}^{\infty} h[m]e^{-j\omega n}\right)e^{j\omega n}d\omega$$

$$= \sum_{-\infty}^{\infty} h[m]\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{j\omega(n-m)}d\omega = \sum_{m=-\infty}^{\infty} h[m]\delta[n-m] \quad (10)$$

Desde que o sistema seja LTI, podemos fazer a manipulação acima. Dessa forma, uma condição suficiente para a existência da transformada de Fourier é:

$$\sum_{-\infty}^{\infty}|h[n]| < \infty \quad (10)$$

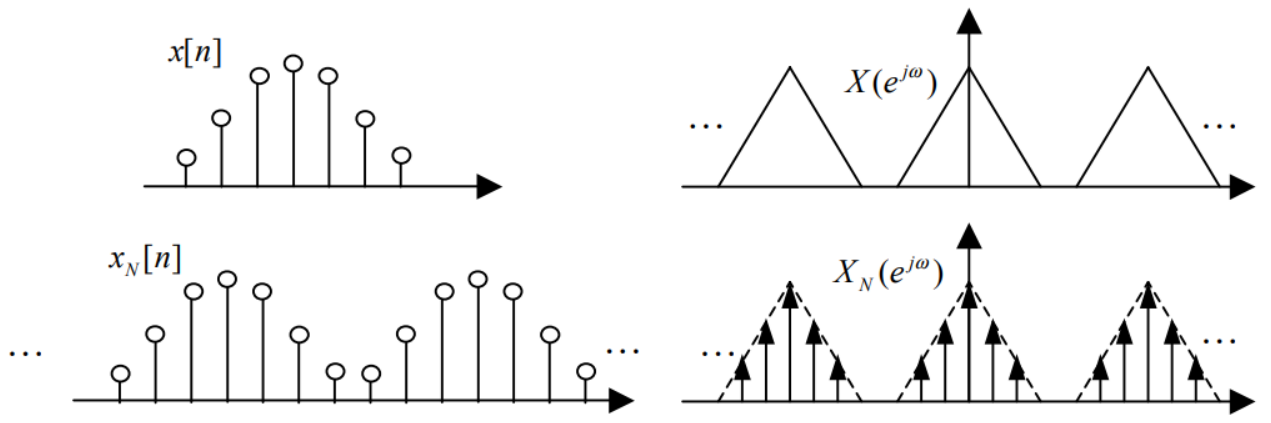
É importante definir a Transformada discreta de Fourier, ou *Discrete Fourier Transform* (DFT). Se um sinal é periódico com período N então:

$$x_N[n] = x_N[n+N] \quad (11)$$

O que significa que o sinal é unicamente representado por N amostras consecutivas. Os conceitos apresentados acima servem para a DFT. Uma relação entre a transformada de Fourier discreta e contínua é mostrada na Fig. 10, na qual o primeiro é um sinal discreto e sua transformada é contínua e periódica. O segundo, um sinal periódico e sua transformada é discreta e periódica.

:

Figura 10: Relação entre sinais finitos e suas transformadas de Fourier.



Fonte: Huang, 2001

A DFT de um sinal analógico é definida como:

$$X(\Omega) = \int_{-\infty}^{\infty} x(t)e^{j\Omega t}dt \quad (12)$$

Com a transformada inversa sendo:

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\Omega)e^{j\Omega t}\,dt \quad (13)$$

A transformada de Fourier comum requer um grande custo computacional (geralmente N^2 interações, sendo N o tamanho do vetor) então há uma família de algoritmos para calcular a DFT de forma mais rápida, que são chamados *Fast Fourier Transforms* (FFT). Dentre os algoritmos existentes de FFT o mais comum é o radix-2, que consegue reduzir as operações para $Nlog_2N$.

1.2.5 **Transformada Z**

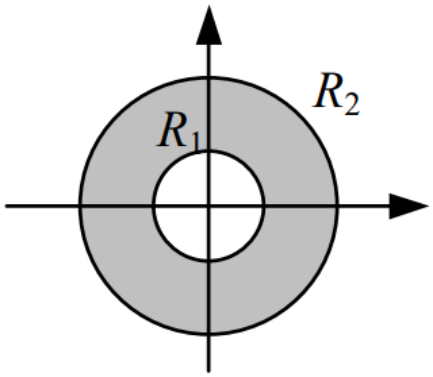
A transformada z é a generalização da transformada de Fourier. A transformada z de um sinal $h[n]$ é definida como:

$$H(z) = \sum_{n=-\infty}^{\infty} h[n]z^{-n} \quad (14)$$

Onde z é uma variável complexa. De fato, a transformada de Fourier de $h[n]$ é igual a sua transformada z dada por $z = e^{j\omega}$. A transformada de Fourier é mais usada para plotar respostas em frequência dos filtros. Enquanto a transformada z é utilizada para analisar características mais matemáticas do filtro, dadas por funções na forma polinomial, além de a mesma também trabalhar com filtros instáveis, que não possuem transformada de Fourier.

Uma condição suficiente para a existência da transformada z é que a soma da Eq. (10) seja finita. Isso é verdadeiro para uma região de convergência, ou *Region Of Convergence* (ROC), no plano complexo $R_1 < |z| < R_2$ como indicado na Fig.11.

Figura 11: Região de convergência de uma transformada z em um plano complexo



Fonte: Huang, 2001.

Isso significa que a transformada z inclui um circulo unitário em sua região de convergência. Essa transformada pode ser usada para qualquer sinal, e não somente para filtros com resposta ao impulso. Ela também admite uma inversa dada por:

$$h[n] = \frac{1}{2\pi j} \oint H(z) z^{n-1} dz \tag{16}$$

Onde a sua integral é circular ao longo do contorno da ROC. Se sua integral for calculada ao longo do circulo unitário a transformada admite inversa.

É comum o uso de tabelas de alguns pares de transformada z e tabelas demonstrando as propriedades para simplificar a solução dessas transformadas, tal como mostrado nas tabelas 1 e 2.

Tabela 1: Alguns pares de transformada Z.

Sequência	Transformada	RDC
1. $\delta[n]$	1	Todo z
2. $u[n]$	$\frac{1}{1-z^{-1}}$	$\lvert z\rvert > 1$
3. $-u[-n-1]$	$\frac{1}{1-z^{-1}}$	$\lvert z\rvert < 1$
4. $\delta[n-m]$	z^{-m}	Todo z exceto 0 (se $m>0$) ou ∞ (se $m<0$)
5. $a^n u[n]$	$\frac{1}{1-az^{-1}}$	$\lvert z\rvert > \lvert a\rvert$
6. $-a^n u[-n-1]$	$\frac{1}{1-az^{-1}}$	$\lvert z\rvert < \lvert a\rvert$
7. $na^n u[n]$	$\frac{az^{-1}}{(1-az^{-1})^2}$	$\lvert z\rvert > \lvert a\rvert$
8. $-na^n u[-n-1]$	$\frac{az^{-1}}{(1-az^{-1})^2}$	$\lvert z\rvert < \lvert a\rvert$
9. $\cos(\omega_0 n)u[n]$	$\frac{1-\cos(\omega_0)z^{-1}}{1-2\cos(\omega_0)z^{-1}+z^{-2}}$	$\lvert z\rvert > 1$
10. $\operatorname{sen}(\omega_0 n)u[n]$	$\frac{\operatorname{sen}(\omega_0)z^{-1}}{1-2\cos(\omega_0)z^{-1}+z^{-2}}$	$\lvert z\rvert > 1$
11. $r^n\cos(\omega_0 n)u[n]$	$\frac{1-r\cos(\omega_0)z^{-1}}{1-2r\cos(\omega_0)z^{-1}+r^2z^{-2}}$	$\lvert z\rvert > r$
12. $r^n\operatorname{sen}(\omega_0 n)u[n]$	$\frac{r\operatorname{sen}(\omega_0)z^{-1}}{1-2r\cos(\omega_0)z^{-1}+r^2z^{-2}}$	$\lvert z\rvert > r$
13. $\begin{cases} a^n, & 0 \le n \le N-1, \\ 0, & \text{caso contrário} \end{cases}$	$\frac{1-a^N z^{-N}}{1-az^{-1}}$	$\lvert z\rvert > 0$

Fonte: Oppenheim, 2013

Tabela 2: Algumas propriedades da transformada z

Propriedade Número	Seção de referência	Sequência	Transformada	RDC
		$x[n]$	$X(z)$	R_x
		$x_1[n]$	$X_1(z)$	R_{x_1}
		$x_2[n]$	$X_2(z)$	R_{x_2}
1	3.4.1	$ax_1[n] + bx_2[n]$	$aX_1(z) + bX_2(z)$	Contém $R_{x_1} \cap R_{x_2}$
2	3.4.2	$x[n - n_0]$	$z^{-n_0}X(z)$	R_x, exceto pela possível adição ou exclusão da origem ou ∞
3	3.4.3	$z_0^n x[n]$	$X(z/z_0)$	$\|z_0\|R_x$
4	3.4.4	$nx[n]$	$-z\frac{dX(z)}{dz}$	R_x
5	3.4.5	$x^*[n]$	$X^*(z^*)$	R_x
6		$\mathcal{R}e\{x[n]\}$	$\frac{1}{2}[X(z) + X^*(z^*)]$	Contém R_x
7		$\mathcal{I}m\{x[n]\}$	$\frac{1}{2j}[X(z) - X^*(z^*)]$	Contém R_x
8	3.4.6	$x^*[-n]$	$X^*(1/z^*)$	$1/R_x$
9	3.4.7	$x_1[n] * x_2[n]$	$X_1(z)X_2(z)$	Contém $R_{x_1} \cap R_{x_2}$

Fonte: Oppenheim, 2013

1.2.6 **Transformada cosseno discreto**

A transformada Cosseno Discreto, ou *Discrete Cosine Transform* (DCT) é amplamente usada em processamento da fala. Ela possui muitas definições. A DCT-II $C[k]$ de um sinal $x[k]$ é definida por:

$$C[k] = \sum_{n=0}^{N-1} x[n]\cos(\pi k(n + 1/2\,)/N) \quad para\ 0 \leq k < N \tag{18}$$

Que admite inversa dada por:

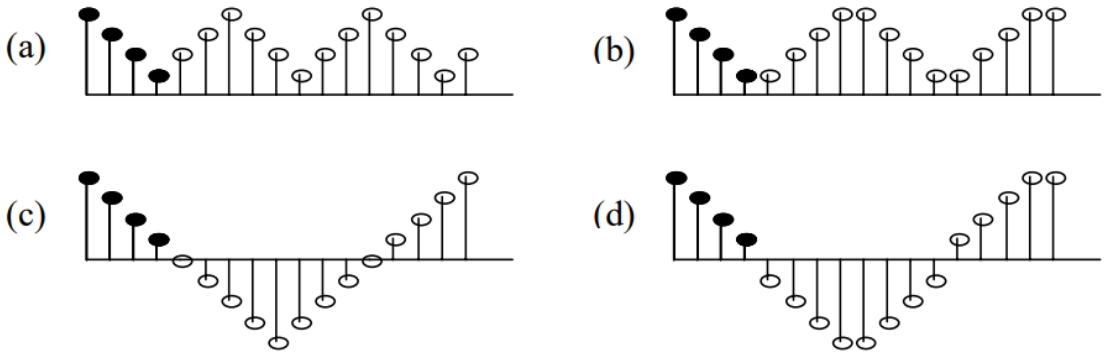
$$x[n] = \frac{1}{N}\{C[0] + 2\sum_{k=1}^{N-1} C[k]\cos(\ \pi k(n + 1/2)/N)\} \tag{19}$$

A DCT-II pode ser derivada da DFT assumindo $x[n]$ é uma sequência real periódica com período $2N$ e com simetria uniforme $x[n] = x[2N - 1 - n]$. Há oito formas diferentes para estender uma sequência de N pontos e torná-la periódica e par, de forma que possa ser recuperada. É fácil provar que $X[k]$ e $C[k]$ estão relacionadas por:

$$X[k] = 2e^{j\pi k/N}C[k] \quad para\ 0 \leq k < N \tag{20}$$

A DCT-II é mais usada devido a sua compactação de energia, que resulta em seus coeficientes mais concentrados em mais baixos índices que a DCT comum. Essa propriedade permite aproximar o sinal com poucos coeficientes. Da Eq. (20) é possível ver que a DCT-II (Fig. 12) de uma sequência real pode ser calculada com um tamanho 2N da FFT de uma sequência real par, que por sua vez pode ser calculado com um comprimento $N/2$ de uma FFT complexa mais alguns cálculos adicionais. As figuras de (a) a (d) representam correspondem consecutivamente a DCT-I a DCT-IV.

Figura 12: Quatro formas de prolongar uma sequência de quatro pontos $x[n]$ tornando-a periódica e simétrica.



Fonte: Huang, 2001

1.2.7 **O Teorema da amostragem**

Vamos definir

$$x_p(t) = x(t)p(t) \tag{21}$$

Onde $p(t)$ é uma versão amostrada de $x(t)$, onde:

$$p(t) = \sum_{n=-\infty}^{\infty} x(t)\delta(t - nT) \tag{22}$$

Onde $\delta(t)$ é um impulso de t. Portanto, $x_p(t)$ pode ser expressado como:

$$x_p(t) = \sum_{n=-\infty}^{\infty} x(t)\delta(t - nT) = \sum_{n=-\infty}^{\infty} x(nT)\delta(t - nT) = \sum_{n=-\infty}^{\infty} x[n]\delta(t - nT) \tag{23}$$

$x_p(t)$ pode ser especificado exclusivamente dado o sinal $x[n]$.

Usando a propriedade de modulação da transformada de Fourier para sinais analógicos, obtém-se

$$X_p(\Omega) = \frac{1}{2\pi} X(\Omega) * P(\Omega) \tag{24}$$

A transformada de um trem de impulsos é dada por:

$$P(\Omega) = \frac{2\pi}{T} \sum_{k=-\infty}^{\infty} \delta(\Omega - k\Omega_s) \tag{25}$$

Onde $\Omega_s = 2\pi F_s$ e $F_s = 1/T$, então

$$X_p(\Omega) = \frac{1}{T} \sum_{k=-\infty}^{\infty} X(\Omega - \mathrm{k}\Omega_s) \tag{26}$$

Através da Figura 13 é possível ver que se

$$X(\Omega) = 0 \; para \; |\Omega| > \frac{\Omega_s}{2} \tag{27}$$

Então $X(\Omega)$ pode ser completamente recuperado através de $X_p(\Omega)$, como segue

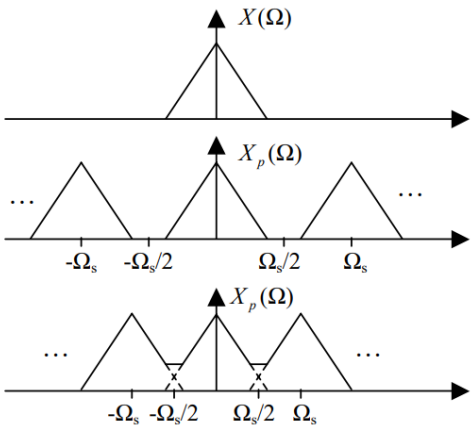
$$X(\Omega) = R_{\Omega_s}(\Omega) X_p(\Omega) \tag{28}$$

Onde

$$R_{\Omega_s}(\Omega) = \begin{cases} 1 & |\Omega| < \frac{\Omega_s}{2} \\ 0 & caso\ contrário \end{cases} \tag{29}$$

É um filtro passa baixa ideal. É possível perceber que se a Eq. 27 não é seguida, então o efeito *aliasing* irá ocorrer e $X(\Omega)$ não poderá ser recuperada através de $X_p(\Omega)$. Uma vez que, em geral, não se sabe ao certo se a condição da Eq. 27 é atendida, o sinal analógico é filtrado com um filtro passa baixa ideal, dado pela Eq. 29, que é chamado filtro anti-aliasing, para a amostragem. Limitar a largura de banda do sinal analógico é o preço a se pagar para manipula-lo digitalmente.

Figura 13: $X(\Omega)$ e $X_p(\Omega)$ para o caso sem e com aliasing



Fonte: Huang, 2001

A transformada inversa de Fourier da equação 29, calculada através da equação 14, é dada por

$$r_T(t) = sinc\left(\frac{t}{T}\right) = \frac{sen\left(\frac{\pi t}{T}\right)}{\frac{\pi t}{T}} \tag{30}$$

Usando a propriedade da convolução, obtêm-se

$$x(t) = r_T(t) * x_p(t) = r_T(t) * \sum_{k=-\infty}^{\infty} x[k]\delta(t - kT) = \sum_{k=-\infty}^{\infty} x[k] r_T(t - kT) \tag{31}$$

O teorema da amostragem afirma que um sinal contínuo no tempo $x(t)$ pode ser recuperado, a partir de suas amostras $x[n]$, usando as equações 30 e 31. A frequência angular $\Omega_s = 2\pi F_s$ é expressa em termos da frequência de amostragem $F_S.T = 1/F_s$ é o período de amostragem, e $F_s/2$ a frequência de Nyquist. A equação 31 se refere a interpolação de banda limitada pois $x(t)$ é reconstruído por interpolação de $x[n]$ usando funções sinc, que são de banda limitada.

1.3 SISTEMAS DE RECONHECIMENTO VOCAL

As pessoas fazem um reconhecimento vocal, ao saber a identidade do falante pelo seu timbre de voz, o que intuitivamente, leva a pensar que isso pode ser forma mais metódica com uso computacional.

Aqui será feito um estudo sobre os tipos existentes de sistemas de reconhecimento de voz, bem como formas de construção de algoritmos para estes sistemas.

1.3.1 **Contexto histórico**

Os sistemas automáticos de reconhecimento vocal, ou *Automatic Recognition Speaker* (ASR), podem parecer algo recente, porém eles são estudados e implementados desde o final do século XX. Tantos sistemas dependentes de texto como os independentes de texto. Seu desenvolvimento aconteceu em indústrias, laboratórios nacionais e universidades. Entre os que pesquisaram e conceberam várias gerações de sistemas de reconhecimento de alto-falante estão AT&T (e seus derivados); Bolt, Beranek e Newman; o Instituto Dalle Molle para Perceptual Artificial Intelligence (Suíça); ITT Instituto de tecnologia Lincoln laboratórios de Massachusetts; Universidade Nacional de Tsing Hua (Formosa); Universidade de Nagoya (Japão); Nippon Telegraph e telefone (Japão); Instituto Politécnico de Rensselaer; Universidade Rutgers; e Texas Instruments (TI). A maioria da pesquisa de ASV (Sistemas de verificação altomática do falante) é dirigida para áreas telefonia. Os laboratórios nacionais Sandia, o Instituto Nacional de normas e tecnologia e a Agência Nacional de segurança realizaram avaliações de sistemas de reconhecimento de falante.

A Tabela 3 mostra o avanço cronológico em sistemas de reconhecimento de voz. Os seguintes termos são usados para definir as colunas na Tabela 3: "Fonte" refere-se a uma citação nas referências, "Org" é a empresa ou escola onde o trabalho foi feito, "Características" é o método de extração utilizado (por exemplo, cepstrum), "Entrada" é o tipo de discurso de entrada (laboratório, qualidade, ou telefone), "Texto" indica se um texto-dependente ou modo de operação independente de texto é usado, "Método" é o coração do processo de correspondência de padrões, "Pop" é o tamanho da população do teste (número de pessoas), e "Erro" é a porcentagem de erro igual para sistemas de verificação do falante (v), ou a porcentagem de erro de reconhecimento para os sistemas de identificação de falante(i), dada a duração especificada do discurso de teste em segundos, a porcentagem e o tempo são unidos por “@”. Estes dados são apresentados para dar uma vista geral simplificada da pesquisa passada do orador-reconhecimento.

Tabela 3: Histórico dos sistemas ASR

Fonte	**Org**	**Características**	**Método**	**Entrada**	**Texto**	**Pop**	**Erro**
Atal, 1974	AT&T	Cepstrum	Pontuação de padrões	Lab	Dependente	10	i: 2%@0.5s v: 2%@1s
Markel and Davis 1979	STI	LP	Long Term Statistcs	Lab	Independente	17	i: 2%@39s
Furui 1981	AT&T	Cepstrum Normalizado	Pontuação de padrões	Telefonia	Dependente	10	v: 0.2%@3s
Schwartz, et al 1982	BBN	LAR	Nonparametricpdf	Telefonia	Independente	21	i:2.5%@2s
Li e Wrench 1983	ITT	LP cepstrum	Pontuação de padrão	Lab	Independente	11	i: 21%@3s v:4%@10s
Doddington 1985	TI	Filter-bank	DTW	Lab	Dependente	200	V:0.8%@6s
Soong, et a. 1985	AT&T	LP	VQ. Distorção média de verossimilhança	Telefone	10 dígitos isolados	100	i: 5%@1.5s i: 1.5%@3.5s
Higgins e Wohlford, 1986	ITT	Cepstrum	DTW Pontuação de verossimilhança	Lab	Independente	11	v: 10%@2.5s v: 4.5%@10s
Attili, et al. 1988	RPI	Cepstrum LP Autocorrelação	Estatísticas projetadas de longo termo	Lab	Dependente	90	v: 1%@3s
Higgns, et al. 1991	ITT	LAR, LP Cepstrum	DTW pontuação de verossimilhança	Escritório	Dependente	186	v: 1.7%@10s
Tishby. 1991	AT&T	LP	HMM (AR mix)	Telefonia	10 dígitos isolados	100	v: 2.8%@1.5s v: 0.8%@3.5s

Reynolds 1995; Reynolds e Carlson 1995	MIT-LL	Mel Cepstrum	HMM (GMM)	Escritório	Dependente	138	i: 0.8%@10s
Che e Lin. 1995	Rutgers	Cepstrum	HMM	Escritório	Dependente	138	i: 0.56%@2.5s i: 0.14%@10s v:0.62%@2.5s
Colombi, et al. 1996	AFIT	Cep, Eng dCep, ddCep	HMM monofone	Escritório	Dependente	138	i:0.22%@10s v:0.28%@10s
Reynolds. 1996	MIT-LL	Mel Cepstrum e Mel dCepstrum	HMM (GMM)	Telefonia	Independente	416	v:11%16%@3s v:6%8%@10s v:3%5%@30s

Fonte: Campbell, 1997

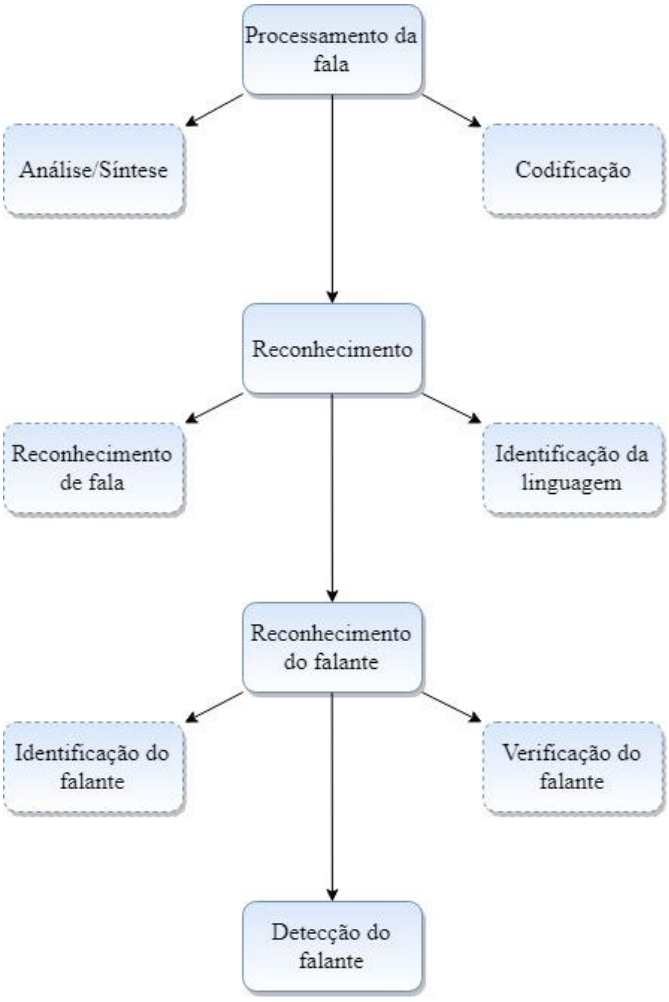
Essa tabela demonstra a flexibilidade de implementação dos sistemas de reconhecimento vocal, dentre o qual há empresas dos mais variados ramos que utilizam desse sistema. É possível também observar que há variados métodos de extração de características, dentre os quais é destacado de Reynolds, de 1996. Esse método utiliza os coeficientes cepstrais coletados com o uso da escala Mel, o mesmo método que foi implementado neste trabalho

1.3.2 **Tipos de sistemas ASR**

Conforme mostrado na Tabela 3 o processamento de fala é um campo diverso e com muitas aplicações. A Fig. 14 mostra algumas dessas áreas e como o reconhecimento de falante se relaciona com cada uma delas. O reconhecimento de voz abrange verificação e identificação (CAMPBEL, 1997). O *Automatic Speaker Verification* (ASV) ou verificação automática do

falante é o uso de uma máquina para verificação da identidade reivindicada, através de um trecho de sua voz. A literatura abunda com termos diferentes para verificação do falante, incluindo a verificação da voz, a autenticação do falante, a autenticação da voz, a autenticação do falador, e a verificação do falador. No *Automatic Speaker Identification* (ASI), identificação automática do falante, não há uma reivindicação de identidade a priori, e o sistema decide quem é a pessoa, o grupo que a pessoa é um membro, ou (no caso de conjunto aberto) quando a pessoa é desconhecida.

Figura 14: Áreas que o reconhecimento de fala abrange



Fonte: próprio autor

ASV e ASI são provavelmente os métodos mais naturais e econômicos para resolver os problemas de uso não autorizado de sistemas de computador e controle de acesso multinível. Com o avanço da rede de telefonia móvel e microfones empacotados com computadores, o

custo de um sistema de reconhecimento de voz pode ser apenas o de software. Os sistemas biométricos reconhecem automaticamente uma pessoa usando traços distintivos. O reconhecimento de falante é uma área biométrica. Sua voz, como outras características biométricas, não pode ser esquecida ou extraviada, ao contrário dos métodos baseados em memorização (por exemplo, senha) ou baseados em posse (por exemplo, chave).

1.3.3 **Fatores de erros**

É preciso identificar possíveis fatores de erro para esses sistemas. A tabela 4 relaciona alguns dos fatores humanos e ambientais que contribuem para esses erros. Esses fatores geralmente estão fora do escopo de algoritmos ou são melhor corrigidos por meios diferentes de medidas (por exemplo, o uso de melhores microfones). Estes fatores são importantes, porque não importa o quão bom um algoritmo de reconhecimento de falante é, se o erro humano (por exemplo, interpretando mal ou não sabendo o significado), em última análise, pode limitar o seu desempenho.

Tabela 4: Fontes de verificação de erro

Frases solicitadas com erro ou mal interpretadas
Estados emocionais extremos (por exemplo, multipercurso e ruído)
Placas de microfone com variação de tempo (intra ou intercessão)
Sala acústica deficiente ou inconsistente (por exemplo, percurso múltiplos e ruído)
Incompatibilidade do canal (por exemplo, usar diferentes microfones para inscrição e verificação)
Envelhecimento (o trato vocal pode se afastar daquele gravado no modelo)
Doenças (por exemplo, resfriados podem alterar o trato vocal)

Fonte: Campbell, 1997

1.3.4 **Algoritmos de reconhecimento vocal**

Essa é a parte chave do problema, construir um algoritmo que transforme o conteúdo vocal em um modelo matemático, afim de que isso possa ser usado para distinguir as vozes dos usuários. Isso requer uma estrutura que faça manipulações no sinal de entrada, após o mesmo ser gravado por um microfone, em diferentes níveis, tal como, Filtragem, Enquadramento, Janelamento, análise de *Mel* Cepstrum, Transformada de Fourier, entre outros.

As etapas de um algoritmo de reconhecimento vocal, de acordo com (MUDA, 2010), sempre envolvem as fases de treino e teste, Figura 15. Na fase de treino cada usuário deve promover amostras da sua voz para construir um modelo de referência. Enquanto isso, na fase de teste, é feita uma nova entrada de voz que é então comparada com o modelo de referência gravado na fase anterior, e a partir disso uma decisão é tomada.

Figura 15: Diagrama em blocos de um algoritmo de reconhecimento vocal

Fonte: Muda, 2010

De maneira simples, existe uma serie de formas para extrair as características do sinal de voz para sistemas dependentes de texto, que é o caso a ser adotado neste documento, tal como, Hidden Markov Model (HMM), Perceptual Linear Prediction (PLP), Mel Frequency Coefficients (MFCC) entre outros. Dentre esses a que foi utilizada nesta pesquisa foi o MFCC, que é recomendado o mais recomendado por modelar a voz de acordo com as características auditivas humanas (MUDA, 2010). Cada método de extração tem seus procedimentos para obter no final parâmetros matemáticos (vetores de características) que podem ser usados para

construir um modelo de referência. No próximo tópico será abordado mais detalhadamente o MFCC.

Essa sequência de vetores características é comparada aos modelos do falante por correspondência de padrões. Isso resulta em uma pontuação de correspondência para cada vetor ou sequência de vetores. A pontuação de correspondência mede a semelhança dos vetores de recurso de entrada computada para modelos de voz reivindicada ou padrões vetoriais de recurso para o orador reivindicado. Por último, uma decisão é tomada para aceitar ou rejeitar o requerente de acordo com a Pontuação obtida.

1.3.5 **Extração de características – MFCC**

A primeira etapa em qualquer sistema automático de reconhecimento de voz é a extrair as características, ou seja, identificar os componentes do sinal de áudio que são uteis para capturar o conteúdo linguístico e descartar as outras coisas que transportam informações desnecessárias como ruído de fundo, emoção etc.

No processamento da fala o MFCC é amplamente utilizado devido a sua precisão. Ele foi introduzido por (DAVIS, 1980) na década de 1980, e têm se destacado desde então. Antes da introdução de MFCCs, coeficientes de previsão linear (LPCs) e prognósticos lineares *Cepstral* coeficientes (LPCCs) foram o principal tipo de recurso para o reconhecimento automático de fala.

Os Coeficientes de Cepstrum Mel-Frequency (MFCC) são uma representação definida como cepstrum de um sinal de curto prazo em janelas derivadas da FFT desse sinal. A diferença do mel cepstrum é que é utilizada uma escala de frequência não linear, que se aproxima do comportamento do sistema auditivo humano. Davis e Mermelstein mostraram que a utilização do MFCC pode ser benéfica para o reconhecimento de fala. Os procedimentos matemáticos realizados no sinal de entrada são mostrados a seguir.

Dado a DFT do sinal de entrada:

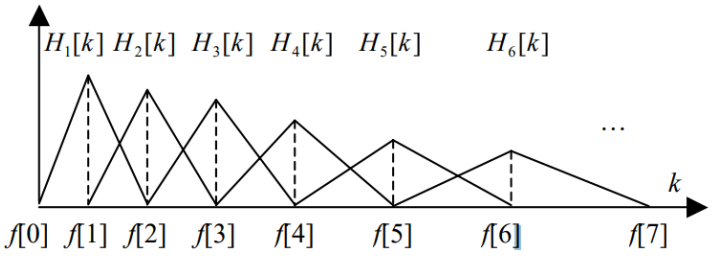
$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-\frac{j2\pi nk}{N}} \ , \ 0 \leq k < N \tag{32}$$

Nós definimos um banco de filtros com filtros M (m = 1,2, M), onde o filtro m é um filtro triangular dado por:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{2(k-f[m-1])}{(f[m]-f[m-1])} & f[m-1] \le k \le f[m] \\ \frac{2(f[m+1]-k)}{(f[m+1]-f[m])} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1] \end{cases} \quad (33)$$

Esses filtros calculam o espectro médio em torno de cada frequência central com o aumento largura de banda, e elas são exibidas na Figura 16.

Figura 16: Filtros triangulares usados no cálculo dos coeficientes mel cepstrum.



Fonte: Huang ,2001

Alternativamente, os filtros podem ser escolhidos como:

$$H'[k] = \begin{cases} 0 & k < f[m-1] \\ \frac{(k-f[m-1])}{(f[m]-f[m-1])} & f[m-1] \le k \le f[m] \\ \frac{(f[m+1]-k)}{(f[m+1]-f[m])} & f[m] \le k \le f[m+1] \\ 0 & k > f[m+1] \end{cases} \quad (34)$$

Que satisfaz a condição:

$$\sum_{m=0}^{M-1} H'[k] = 1 \quad (35)$$

O mel-cepstrum calculado com $H_m[k]$ ou $H'_m[k]$ será diferente por um vetor constante para todas as entradas, então a escolha torna-se sem importância quando usada em sistema de reconhecimento de fala que treinou com os mesmos filtros.

Define-se f_1 e f_h como as menores e mais altas frequências do banco de filtros em Hz. Essas vão ser as frequências que o filtro vão ser iniciadas (f[0] da Figura 16) e a frequência final (f[7] da Figura 16), Fs frequência de amostragem em Hz, M o número de filtros e N o tamanho da FFT. O os pontos de fronteira f[m] estão uniformemente espaçados na escala mel:

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_1) + \frac{m(B(f_h) - Bf(_1)}{M+1}\right) \quad (36)$$

onde a escala de mel B é dada pela Eq. (2.6) e B^{-1} é o seu inverso

$$B^{-1}(b) = 700(\exp((b/1125) - 1) \quad (37)$$

Em seguida, calculamos a energia de log na saída de cada filtro como:

$$S[m] = \ln[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]] \quad (38)$$

A mel frequência cepstrum é então a transformada de cosseno discreta das saídas do filtro M:

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos(\pi n(m + 1/2)/M) \quad 0 \leq n < M \quad (39)$$

onde M varia para diferentes implementações de 24 a 40. Para reconhecimento de fala, tipicamente apenas os primeiros 13 coeficientes cepstrum são usados. É importante notar que o MFCC representação não é mais uma transformação homomórfica. Seria se a ordem do somatório e logaritmos na Eq. 38 fossem revertidos:

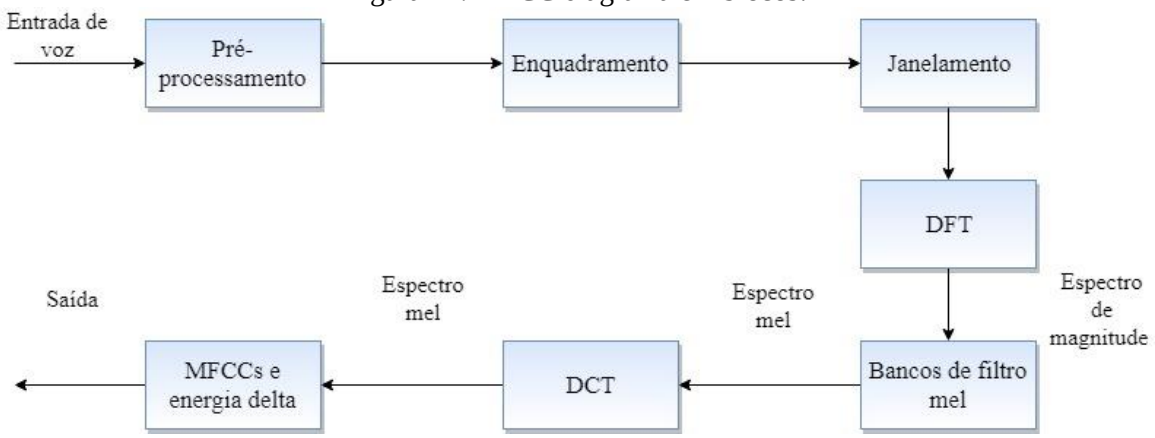
$$S[m] = \sum_{k=0}^{N-1} \ln(|X_a[k]|^2 H_m[k]) \quad 0 \leq m < M \quad (40)$$

Na prática, no entanto, a representação MFCC é aproximadamente homomórfica para filtros que têm uma função de transferência suave. A vantagem da representação MFCC usando 38 em vez de 40 é que as energias de filtro são mais robustas ao ruído e à estimativa espectral aos erros. Esse algoritmo tem sido usado extensivamente como um vetor de recursos para reconhecimento de fala sistemas.

1.3.6 **Algoritmo MFCC**

Como já visto, a diferença do MFCC para os modelos lineares é que esse utiliza a escala Mel de frequência, uma escala logarítmica, esta escala faz com que as características extraídas combinem mais com a percepção auditiva humana. O diagrama em blocos do método é baseado no descrito por MUDA (2010), mostrado na Figura 17.

Figura 17: MFCC diagrama em blocos.



Fonte: Muda, 2010

Essas etapas podem ser melhor explicadas a seguir:

a) Pré-processamento: esta etapa processa a passagem do sinal através de um filtro que enfatiza frequências mais elevadas, geralmente é dado pela equação 41. Este processo irá aumentar a energia do sinal em maior frequência, um valor padrão para a filtragem é $\alpha = 0{,}97$;

$$y(t) = x(t) - \alpha x(t-1) \tag{41}$$

Fazer o pré-processamento é útil pois: (1) equilibra o espectro de frequências, já que as frequências altas geralmente possuem magnitudes menores em comparação com freqüências mais baixas, (2) evitam problemas numéricos durante a operação de transformada de Fourier e (3) também podem melhorar o sinal Relação de Ruído.

b) Enquadramento: Aqui, é necessário dividir o sinal em quadros de tempo curto. A lógica por trás dessa etapa é que as freqüências de um sinal mudam com o tempo, então, na maioria dos casos, não faz sentido fazer a transformada de Fourier ao longo de todo o sinal, perdendo os contornos de frequência do sinal ao longo do tempo. Para evitar isso, é possível assumir com segurança que as freqüências de um sinal são estacionárias durante um período muito curto de tempo. Portanto, fazendo uma transformada de Fourier sobre este curto período de tempo, obtêm-se uma boa aproximação dos contornos de frequência do sinal pela concatenação de quadros adjacentes.

Valores típicos recomendam usar 25 ms como tamanho do quadro e um tamanho de passo de 10 ms (com sobreposição de 15ms). Como exemplo, enquadrar o sinal com os valores padrões mostrados acima significa que o comprimento do quadro para um sinal de 16kHz é de 0,025 *

16000 = 400 amostras. A etapa do passo do quadro é geralmente algo como 10 ms (160 amostras, neste caso), o que permite alguma sobreposição aos quadros. O primeiro quadro de amostra de 400 começa na amostra 0, o próximo quadro de amostra de 400 começa na amostra 160 etc. até que o final do arquivo de fala seja atingido. Se o arquivo de fala não se dividir em um número par de quadros, então são colocados zeros para que ele seja compatível.

c) Janelamento: A janela de Hamming é utilizada como forma de janela, considerando o bloco seguinte na corrente de processamento da extração da característica e integra todas as linhas de frequência as mais próximas. A equação da janela de Hamming por:

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right), 0 < n < N - 1 \quad (42)$$

Onde N é o número de amostras em cada quadro;

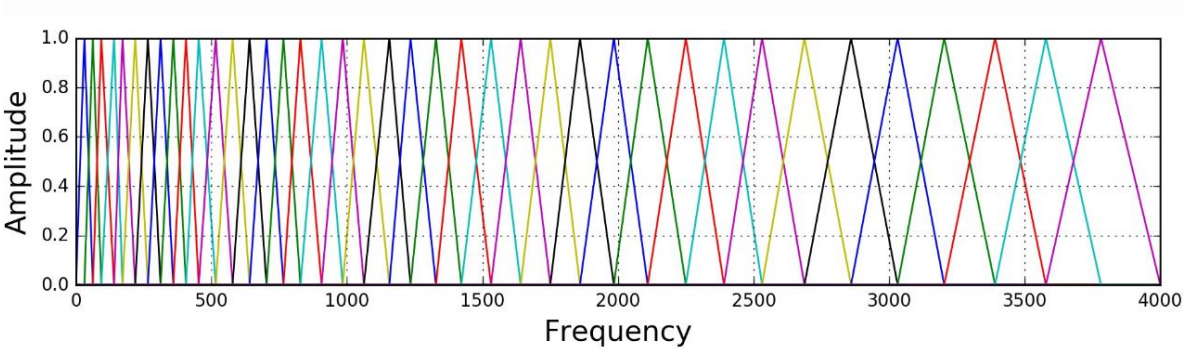
Existem várias razões pelas quais se faz necessário aplicar uma função de janela aos quadros, a principal é para compensar a suposição feita pela FFT de que os dados são infinitos e reduzir o vazamento espectral.

d) Transformada Rápida de Fourier (FFT): A transformação rápida de Fourier (FFT) usada para calcular a transformação de Fourier discreta, convertendo o sinal de domínio do tempo para o domínio de frequência. O tempo de cálculo FFT é 10 vezes inferior a um DCT clássico. Ela serve também para que seja extraído o espectro e para que depois se possa converter para a escala Mel de frequência.

Geralmente, é realizada uma FFT de 512 pontos, em cada quadro, e mantem-se apenas os primeiros 257 coeficientes.

e) Banco de filtro Mel: Nessa etapa é feita a aplicação de filtros triangulares, normalmente 40 filtros, espaçados não linearmente e sim em uma escala Mel ao espectro de energia para extrair bandas de frequência. A escala Mel tem como objetivo imitar a percepção não-linear do ouvido humano, sendo mais discriminativa nas freqüências mais baixas e menos discriminativa nas freqüências mais altas.

Figura 18: Bancos de filtros Mel



Fonte: (https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html)

A figura 18 mostra um conjunto de filtros triangulares que são usados para calcular uma soma ponderada de componentes espectrais de filtro para que a saída do processo se aproxime de uma escala de mel. A resposta de frequência de magnitude de cada filtro é triangular em forma e igual a unidade na frequência central e diminui linearmente para zero na frequência central de dois filtros adjacentes.

f) Transformada Cosseno Discreto (DCT): O resultado da multiplicação do sinal com os bancos de filtro mostrados na etapa anterior é altamente correlacionado, o que pode ser problemático em alguns algoritmos de aprendizado de máquina. Então, é conveniente aplicar a Transformada de Cosseno Discreta (DCT) para descolar os coeficientes do banco de filtros e gerar uma representação comprimida dos mesmos.

O resultado da conversão é chamado de MFCC. O primeiro coeficiente do vetor Cmel, denotado por c0, pode carregar muita informação do meio de transmissão (DELLER, 2000). normalmente, para reconhecimento automático de fala (ASR), os coeficientes cepstrais resultantes 2-13 são retidos e o restante é descartado. Esta redução é feita por meio de uma propriedade da DCT conhecida como compactação da energia, concentrando os valores mais significativos nos primeiros termos do vetor, e descartando os últimos, melhorando assim a eficiência computacional.

g) Energia delta e espectro delta: Também conhecidos como coeficientes diferenciais e de aceleração. A ideia principal da extração de atributos é captar as mudanças temporais bruscas presentes no espectro. Devido a isto, utilizam-se além dos coeficientes extraídos até agora,

chamados coeficientes “estáticos”, os coeficientes delta e de aceleração, chamados coeficientes “dinâmicos”, que capturam essas mudanças e incorporam informação relativa à transição dos coeficientes estáticos entre quadros vizinhos. O vetor de característica do MFCC descreve apenas o envelope espectral de potência de um único quadro, mas a fala também possui informações dinâmicas, ou seja, as trajetórias dos coeficientes do MFCC ao longo do tempo, isso é conhecido como delta. Calcular as trajetórias do MFCC e anexá-las ao vetor de característica original pode aumentar bastante o desempenho do ASR.

Para calcular os coeficientes delta, é usada a seguinte relação:

$$d_t = \sum_{n=1}^{N} \frac{n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (43)$$

onde d_té um coeficiente delta, do quadro $t$computado em termos dos coeficientes estáticos c_{t+N}para c_{t-N}. Um valor típico Né 2.

1.3.7 **Sistema para controle de acesso**

Juntando as etapas de construção de um algoritmo de reconhecimento de locutor e a extração de características utilizando o método MFCC é possível pensar em um sistema de reconhecimento vocal, que utilize essas ferramentas para uma aplicação voltada para controle de acesso.

De maneira resumida, começando a partir da fase de treino, a voz é gravada por um microfone e são extraídas características que são armazenadas em um banco de dados, essas características são comparadas com as que serão extraídas na fase de teste. A comparação gera uma pontuação e a partir desse valor é decidido aceitar ou rejeitar o usuário. O ARDUINO faz a comunicação do sistema entre o algoritmo de reconhecimento do locutor e as ferramentas de controle, como controle de um LED.

1.3.8 **Erro quadrático médio**

Na estatística, o conceito de erro quadrático médio é um critério importante que é utilizado para medir o desempenho de um estimador. O erro quadrático médio, abreviado como

MSE (Mean Squared Error), é um modelo de classificação de baixa complexidade que se baseia no conceito de distância Euclidiana.

Ele é dado por:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2 \quad (44)$$

Onde Xi e Yi são as amostras que estão sendo comparadas, podemos ainda colocar uma delas, por exemplo o Xi, como sendo a média das amostras de um modelo de observação. Esse método de classificação ajuda a comparar uma dada amostra e ver o quanto ela se distanciou do modelo de referência.

Esse modelo apresenta alguns pontos negativos, como por exemplo desprezar os valores negativos de um determinado dado, além de elevar eventuais discrepâncias, já que é um erro quadrático. Há outras formas de classificação utilizadas em modelos de reconhecimento de fala, porém que demandam um grau de complexidade muito elevado, e que por isso fogem do objetivo desse trabalho, como os modelos estatísticos de Hidden Markov.

2. **METODOLOGIA**

O Trabalho discutido foi uma Pesquisa aplicada, a qual tem como objetivo a realização de pesquisa exploratória sobre o material bibliográfico adquirido a respeito do assunto. É utilizado o procedimento técnico de pesquisa bibliográfica e a pesquisa experimental. O método de abordagem utilizado foi o hipotético-dedutivo e o método de procedimento de elaboração foi monográfico. Para coleta de dados foi utilizada documentação indireta, com auxílio de documentos primários e secundários, e a análise e interpretação de seus dados foi quantitativa.

Serão feitas pesquisas bibliográficas na área de Comunicações de dispositivos com o MATLAB, processamento digital de sinais, sinais e sistemas, e linguística, tal como consta na Tabela 2.1. As pesquisas bibliográficas serão realizadas para coletar dados sobre o estado da arte e as técnicas mais recentes, que sejam viáveis para a elaboração do projeto. Os dados servirão também de embasamento para verificação da precisão, qualidade, e outros critérios, uma vez que o sistema esteja pronto.

Tabela 5: Pesquisas bibliográficas

Comunicações de dispositivos com o MATLAB.	Desenvolvimento de algoritmos para comunicação entre Arduino e MATLAB.
Processamento digital de sinais	Conceitos básicos, Transformadas de Fourier, DFT, DCT FFT.
Sinais e Sistemas	Classificação de sinais, Sistemas digitais, Sinais Senoidais.
Linguística	Fonética acústica, sintaxe e semântica.

Fonte: Produzido pelo autor

Para o desenvolvimento prático deste trabalho foi utilizado um notebook VAIO com processador Intel ® CoreTM i7-6500U e 2.59GHz. Um arduino MEGA. Um microfone Shure Beta51A. Foram utilizados 95 arquivos de áudio mono de 20 usuários em formato .wav, com taxa de amostragem de 44100Hz, 16 bits por amostra e com um canal. 15 desses voluntários eram cadastrados no sistema e para isso gravaram a sua voz 5 vezes e os outros 5 eram impostores que gravaram sua voz apenas uma vez, forjando um nome para tentar obter acesso. Utilizou-se, também, o software MATLAB R2018a e o programa para gravação Audacity (2.3.0).

A partir dos conhecimentos adquiridos iniciou-se o desenvolvimento do algoritmo de reconhecimento vocal. Inicialmente foi feita a aquisição de arquivos de áudio a partir de usuários voluntários. Cada usuário gravou um trecho de voz, com o seu nome, quatro vezes. Para que um modelo de referência fosse criado. Logo após, iniciou-se o desenvolvimento do algoritmo, no Matlab, em new script. Esse algoritmo tinha por intuito extrair as características, utilizando o método MFCC, de voz de cada usuário e cadastra-las em um banco de dados.

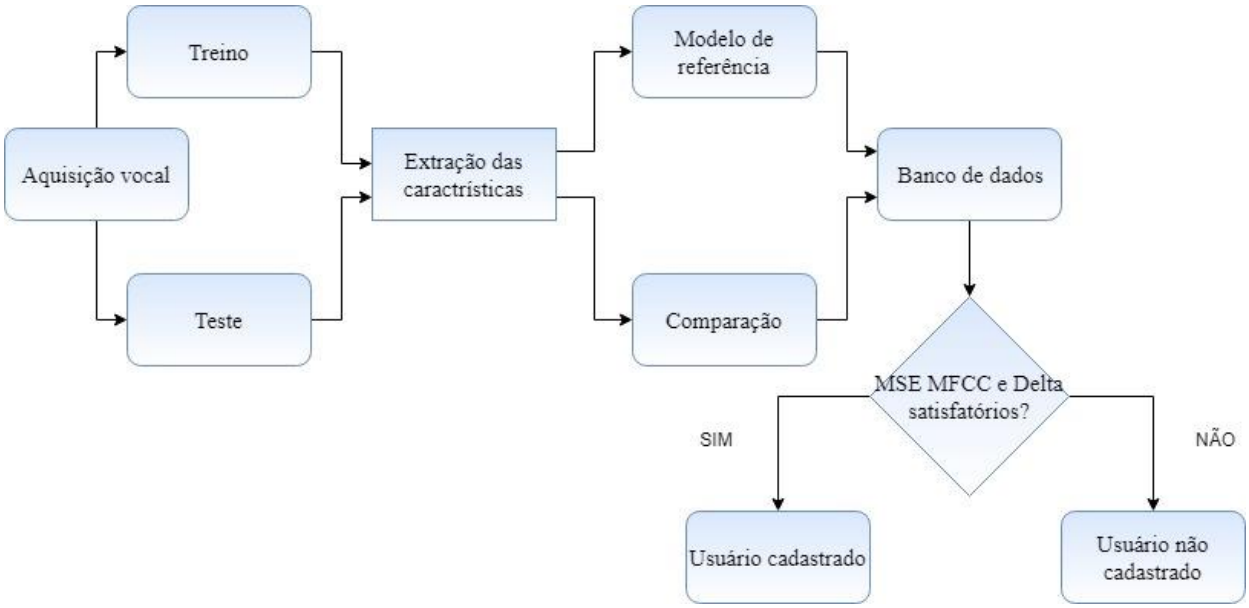
Os sistemas de reconhecimento vocal apresentam basicamente duas fases principais, a de treinamento e a de teste. Na etapa de treinamento, são extraídas características da voz que serão cadastradas em um banco de dados. Na etapa de teste os padrões armazenados serão novamente extraídos para serem então comparados com os do banco de dados.

A parte final era integrar os algoritmos construídos no MATLAB com o Arduino, mostrando uma implementação de um sistema de controle de acesso utilizando reconhecimento de voz. A comparação dos parâmetros de vocais codificados nas fases de treino e testes gerou uma pontuação. A métrica de avalição, é feita utilizando o MSE, conforme a figura 19, que

realiza a pontuação e valida ou rejeita o acesso do usuário ao sistema baseados nos padrões biomédicos vocais.

Os critérios de precisão do sistema serão testados de duas formas, com o usuário cadastrado proferindo a seu nome e um usuário não cadastrado falando o nome de outro usuário cadastrado.

Figura 19: Metodologia aplicada



Fonte: próprio autor

3. IMPLEMENTAÇÃO

3.1 Gravação dos arquivos de voz

Os áudios foram gravados com o uso da ferramenta Audacity (2.3.0), Fig.19. O Audacity é um software livre de edição digital de áudio disponível principalmente nas plataformas: Windows, Linux e Mac e ainda em outros Sistemas Operacionais. Ele é próprio para gravações de voz, e, portanto, executa essa tarefa melhor que o MATLAB.

Figura 19: Logo marca do programa Audacity

Fonte: https://www.audacityteam.org

No programa as gravações foram realizadas com um Project Rate de 44100 Hz, um canal (mono) e o formato de amostragem de 16 bits. Cada usuário gravou a sua voz 6 vezes. Obtendo o total de 90 amostras de aúdio, já 15 que usuários foram cadastrados. Dentre as seis gravações, cinco foram usadas para formar o modelo de referência e uma delas serviu para testes. Cada gravação durou 2 segundos e foi convertida no formato wav. Os arquivos de voz foram salvos com o nome do falante e sua respectiva gravação, conforme mostrado na Fig 20 abaixo.

Figura 20: Gravação da voz de um dos usuários

Fonte: próprio autor

Para a gravação foram utilizados um microfone Shure Beta 58A, uma interface de áudio ZOOM G2NU, e um computador VAIO com processador Intel ® CoreTM i7-6500U e 2.59GHz. As gravações foram realizadas em uma mesma sala, contendo o mínimo de ruídos possível, buscando diminuir os fatores de erro.

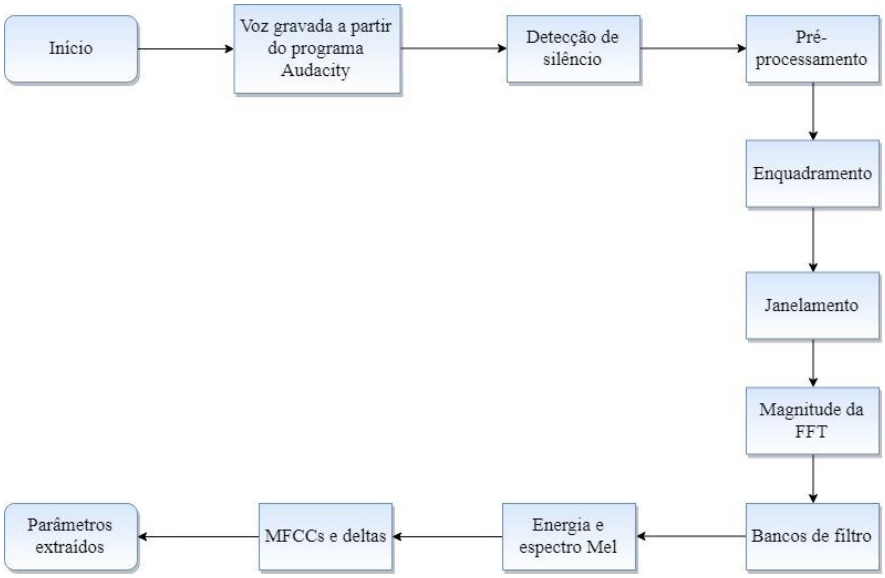
A gravação por meio desse software fornece mais confiança do que aquela executada diretamente pelo MATLAB. Tendo em vista que fornece uma série de dados como volume do microfone, que pode ser ajustado para um ponto que causa uma melhor relação possível das amplitudes da amostra sem defasagem, além de já ser mostrado o gráfico de voz no momento da gravação, identificando instantaneamente algum possível ruído ou problema no áudio gravado.

3.2 **Desenvolvimento dos algoritmos**

3.2.1 **Método de extração de características**

O reconhecimento de voz é dividido em duas fases que são fase de treino e fase de teste, porém, para ambas as fases é necessário um algoritmo que parametrize a voz extraia as suas características. Isso pode ser feito com o MFCC, o diagrama implementado foi baseado em (MUDA, 2010) e (DAVIS,1980) com algumas modificações realizadas pelo próprio autor deste documento. Um diagrama em blocos com cada fase do processo implementado é mostrado na Figura 21. Que cita os dados de projeto, colocando como exemplo os procedimentos feitos para extração das características vocais do usuário Igor soares.

Figura 21: Diagrama em blocos para extração de características

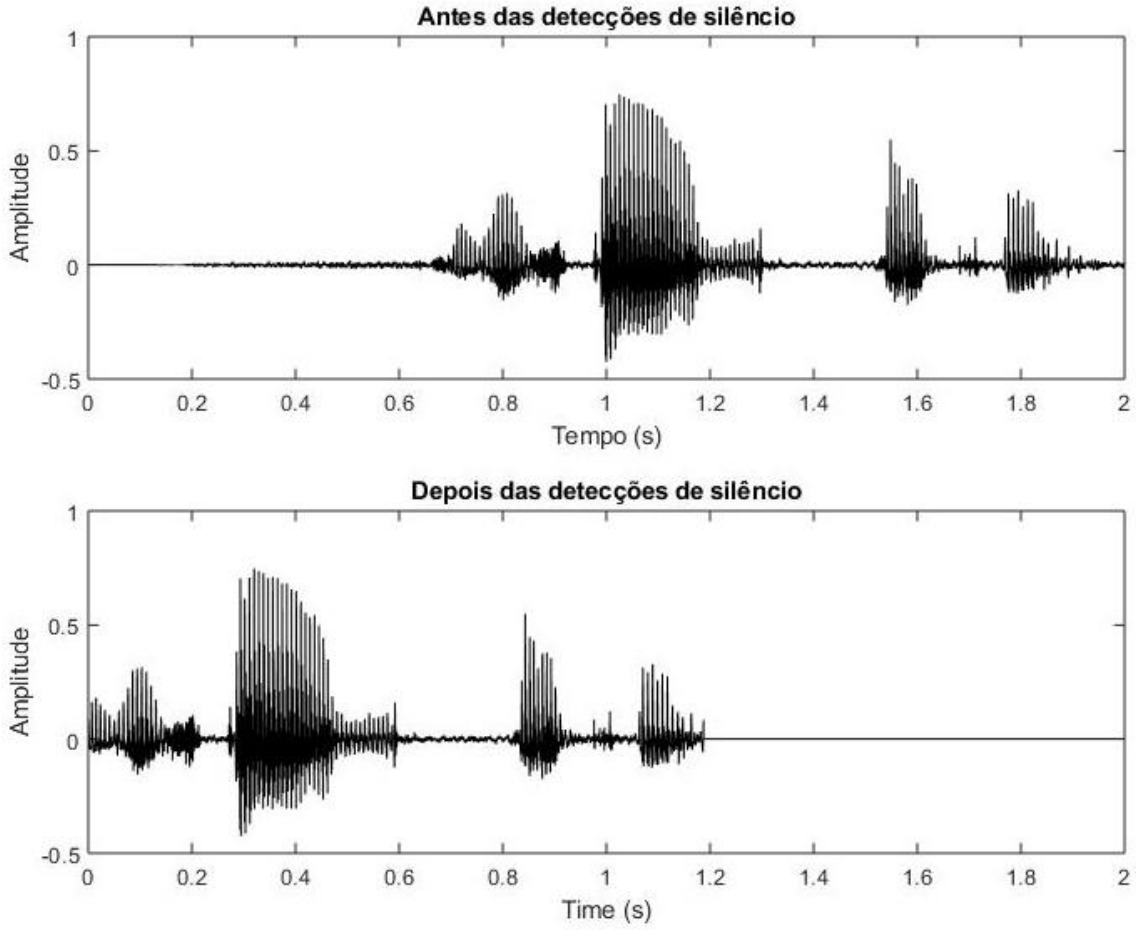


Fonte: Próprio autor

Será mostrado o resultado obtido de um dos usuários cadastrados, para cada etapa, visando uma melhor compreensão. No diagrama a análise do algoritmo será feito a partir da detecção de silêncio, visto que a etapa anterior não é realizada pelo MATLAB.

a) Detecção de silêncio: Esta parte é importante para eliminar ruídos indesejáveis do sinal de voz, prevenindo futuros erros. Nessa parte é feito o truncamento do sinal, para que este comece a partir do trecho em que há informação útil para o reconhecimento vocal. É feita também uma detecção de silêncio no final, já que a gravação de áudio dura 2 segundos e após o usuário falar seu nome ainda há um trecho em que não há presença de fala, essa detecção substitui as amostras de amplitude pequena, abaixo de 0.065, por 0. Após isso o sinal é preenchido com zero, para ter um tamanho padrão de 88000 amostras, já que a frequência de amostragem é 44 kHz. Como pode ser visto na Figura 22.

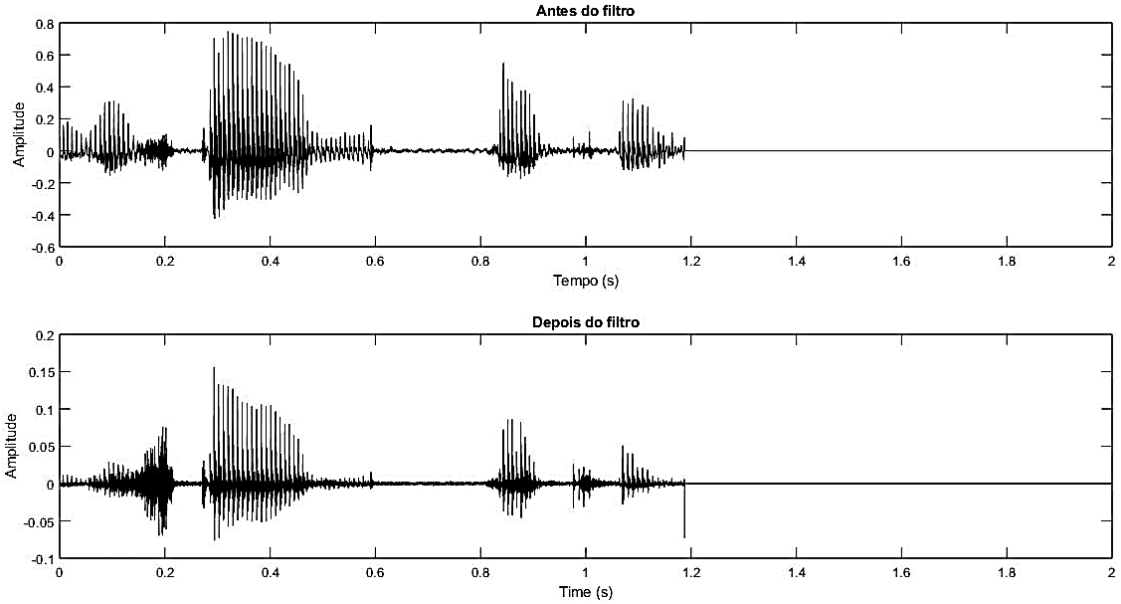
Figura 22: Amostra de áudio antes e depois da detecção de silêncio



Fonte: Próprio autor

b) Pré-processamento: aqui o uso do filtro se dá pelo mesmo critério explicado no tópico 1.3.6. Esse filtro é aplicado com $\alpha = 0.97$, no sinal após a detecção de silêncio, como mostrado na Fig.23. Esse filtro atenua um pouco a amplitude do sinal, mas melhora os resultados no processamento.

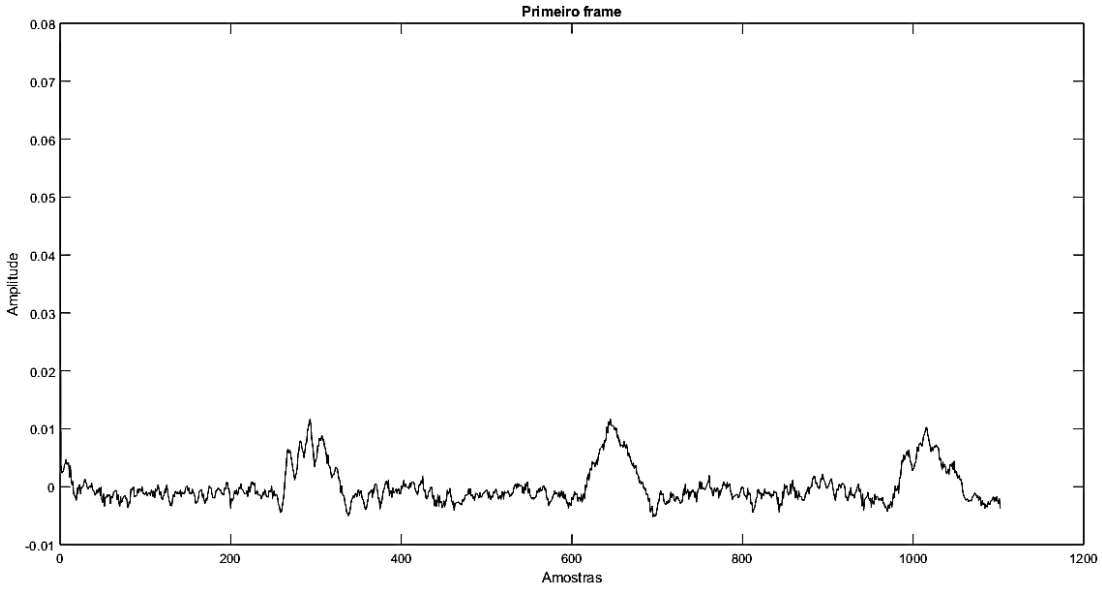
Figura 23: Antes e depois da passagem do filtro



Fonte: Próprio autor

c) Enquadramento: Não faz mais sentido mostrar o sinal como um todo já que ele foi decomposto em vários outros menores, apenas para ilustrar, o primeiro quadro do sinal, Fig.24, no total o sinal foi quebrado em 198 quadros ou *frames*, com largura do frame de 25ms (Ts) e deslocamento de 10ms (Td). O primeiro possui as amostras de 1 até 1103.

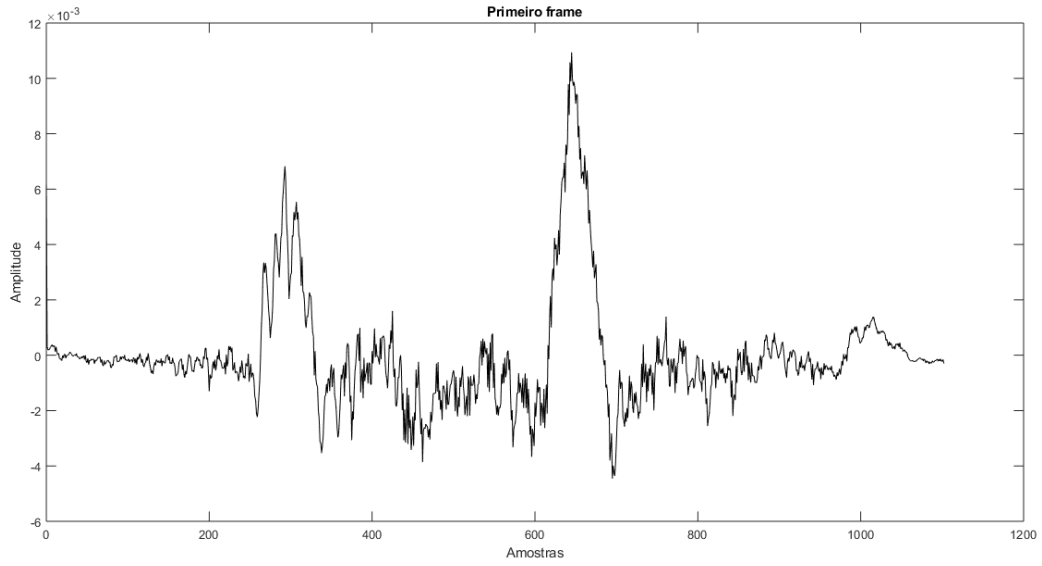
Figura 24: Primeiro frame do sinal em questão



Fonte: Próprio autor

d) Janelamento: Para cada um dos 198 *frames* é passada uma janela hamming, essa janela tem o mesmo número de pontos de cada *frame*, para que possa ser feita uma multiplicação ponto a ponto entre ambas, Fig. 25.

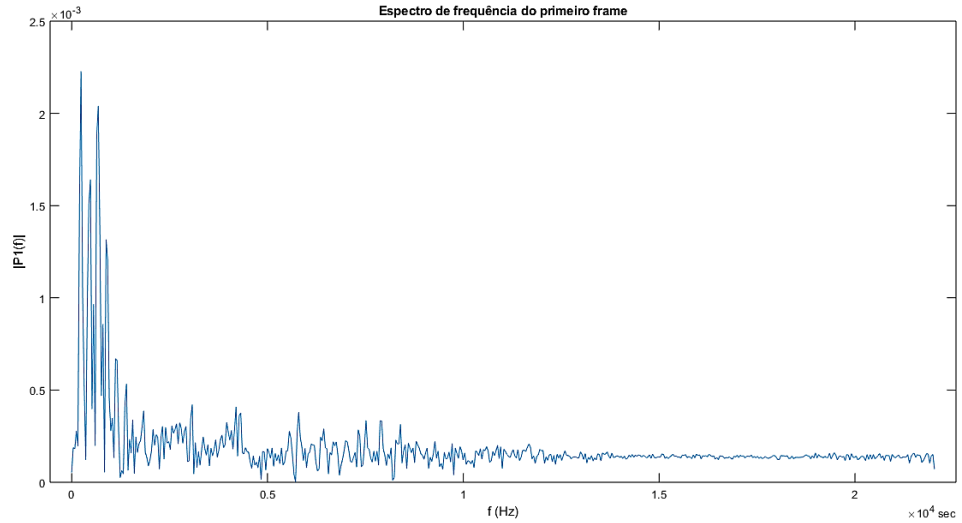
Figura 25: O primeiro *frame* após passar por uma janela hamming.



Fonte: Próprio autor

e) Magnitude da FFT: o sinal então foi passado para o domínio da frequência por meio de uma transformada rápida de Fourier, de 512 pontos, para cada um dos 198 frames, como consta em Fig.26.

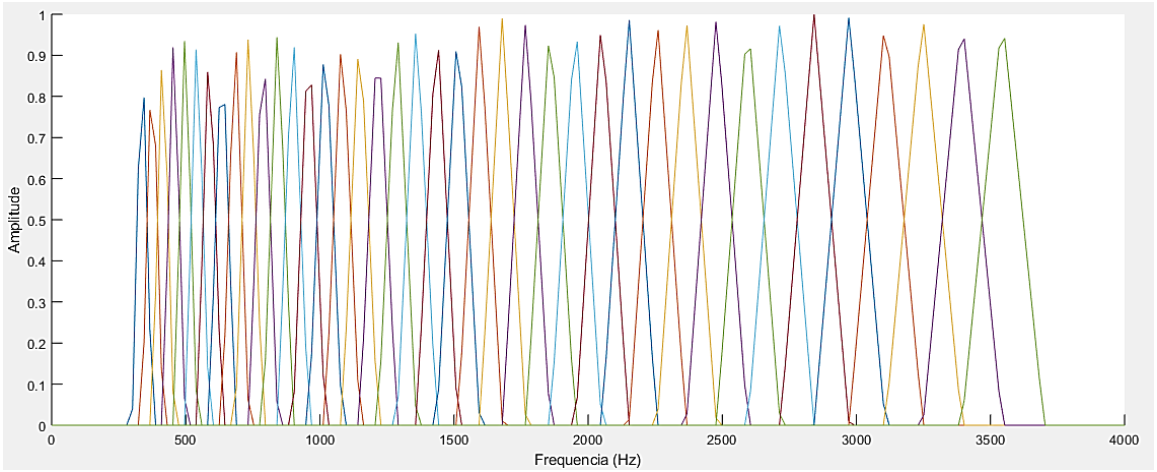
Figura 26: Espetro de frequência para o primeiro frame



Fonte: Próprio autor

f) Bancos de filtro: Foram utilizados 40 bancos de filtro, com uma frequência inferior de 300 e a superior de 3700 Hz. Como mostrado na Figura 27 esses bancos não são espaçados linearmente na frequência em Hertz, porém o espaçamento deles é linear na frequência mel, e isso é tido como base na hora de montar a função desses bancos de filtro.

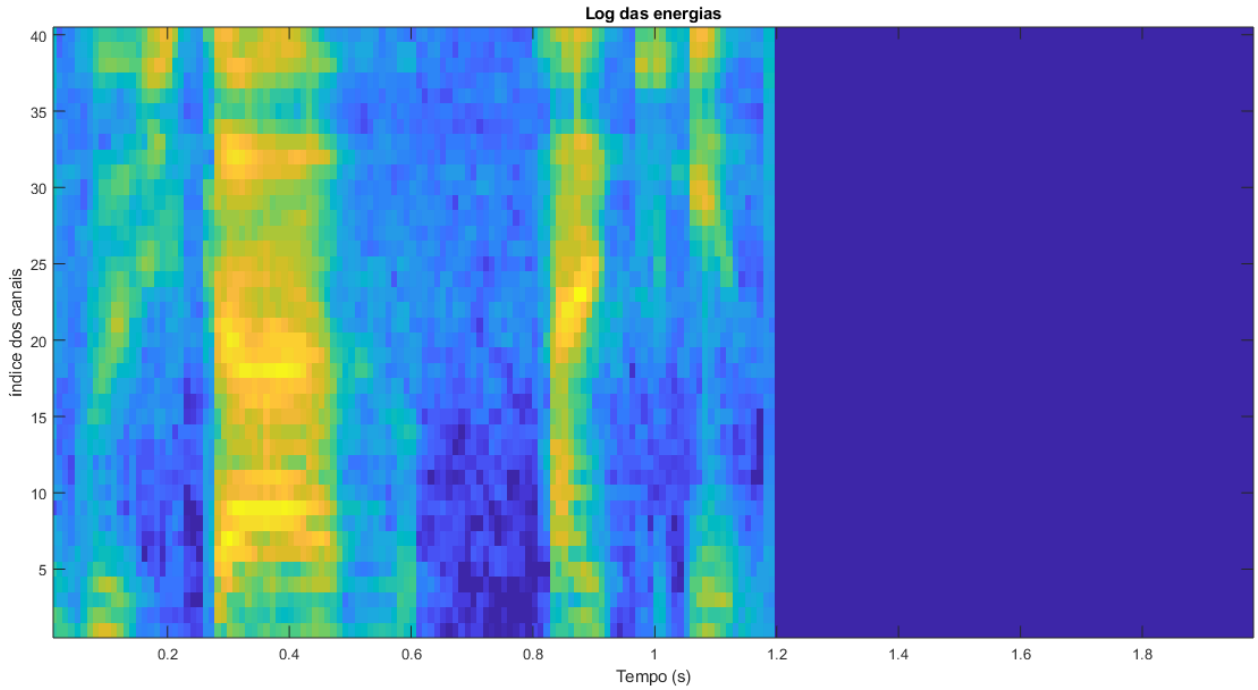
Figura 27: Bancos de filtros



Fonte: própria

g) Energia e espectro mel: Para cada um dos 40 bancos de filtro são calculadas as energias, e isso é mostrado no espectrograma da Fig.28.

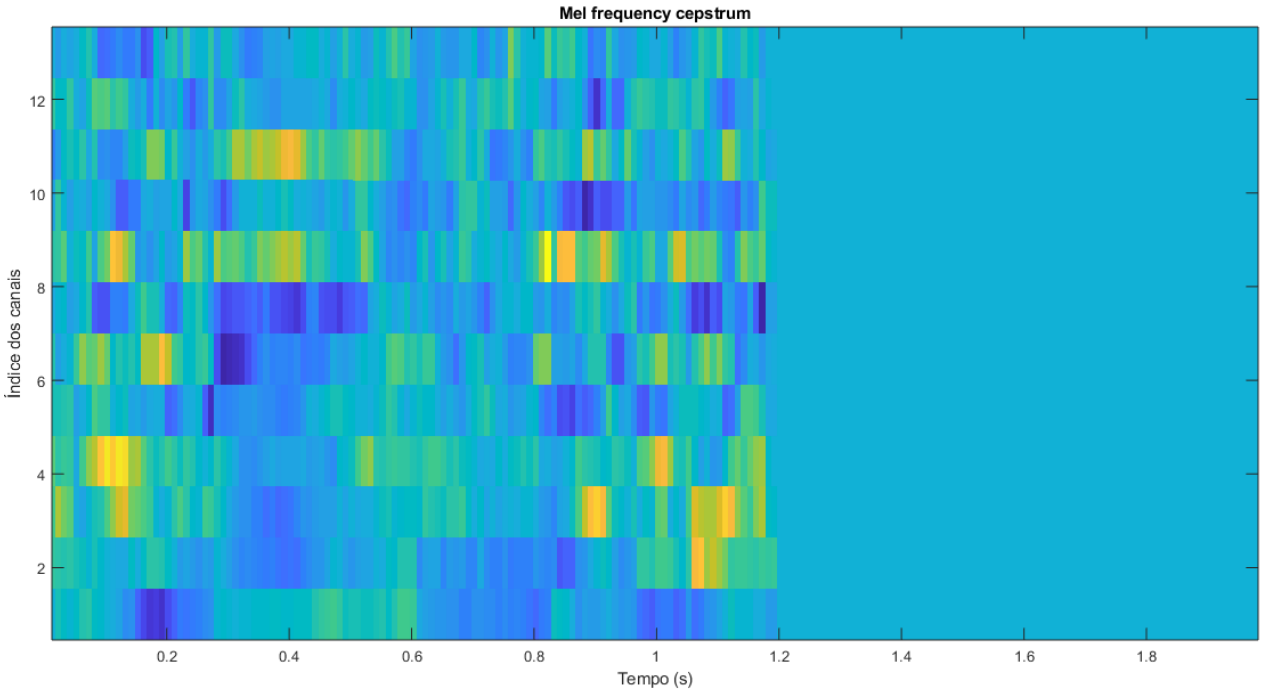
Figura 28: Logaritmo das energias



Fonte: próprio autor

h) MFCCs e deltas: Aqui, o sinal passa por uma DCT, para voltar ao domínio do tempo, Fig 29. Então do vetor de 40 posições, acima só é mantido os coeficientes 2 até 17. Obtendo assim 16 coeficientes cepstrais. Com esses coeficientes extraídos é calculado o parametro delta, que mede a variação do coeficiente com o seguinte.

Figura 29: Os coeficientes MFCCs



Fonte: próprio autor

3.3.2 **Criando o banco de dados**

Nessa parte, as características extraídas, no tópico anterior foram armazenadas em um banco de dados, para formar um modelo de referência. Os dados que foram para o banco de dados foram a matriz com os doze coeficientes do MFCC, o nome do usuário e a frequência de amostragem.

O modelo de referência pode ser construído com o uso da estatística, fazendo a média e desvio padrão dos coeficientes usando isso em uma função gaussiana de probabilidade. Quanto mais amostras forem gravadas mais chances há de que os erros diminuam.

O banco de dados será feito a partir de uma tabela dinâmica, Tabela 6. Essa tabela teve a função de armazenar dados de diferentes tipos, matrizes (no caso da média e desvio padrão dos MFCCs), *strings* (nome do usuário) e inteiros (frequência de amostragem).

Tabela 6: Banco de dados

Usuário	**Média dos MFCCs**	**Média Delta**	**Amostragem (fs)**
Kimberly Mouzinho	[16x198 double]	[16x198 double]	44100 Hz
Gustavo Aquino	[16x198 double]	[16x198 double]	44100 Hz

Roseneth Mouzinho	[16x198 double]	[16x198 double]	44100 Hz
Jacobus de Jager	[16x198 double]	[16x198 double]	44100 Hz
Herik Mouzinho	[16x198 double]	[16x198 double]	44100 Hz
Thiago Patrício	[16x198 double]	[16x198 double]	44100 Hz
Judite Bezerra	[16x198 double]	[16x198 double]	44100 Hz
Italo Santos	[16x198 double]	[16x198 double]	44100 Hz
Everton Cassiano	[16x198 double]	[16x198 double]	44100 Hz
Igor Soares	[16x198 double]	[16x198 double]	44100 Hz
José Sicchar	[16x198 double]	[16x198 double]	44100 Hz
Renan Figueiredo	[16x198 double]	[16x198 double]	44100 Hz
Maria Eduarda	[16x198 double]	[16x198 double]	44100 Hz
Nikolas da Silva	[16x198 double]	[16x198 double]	44100 Hz
Caic Otani	[16x198 double]	[16x198 double]	44100 Hz

Fonte: próprio autor

3.3.3 **Fase de testes**

Construído o modelo de referência, com os padrões de características vocais dos usuários cadastrados em um banco de dados, o sistema já está pronto para a próxima fase. Esta parte é decisiva no projeto, pois é ela quem analisou os dados para garantir o acesso. As características podem ter sido bem extraídas, e os dados podem ter sido bem armazenados, porém, se não houver uma métrica de avaliação correta o objetivo final, controlar o acesso a usuários específicos, não será garantido. Foi utilizado o erro médio quadrático como para a classificação. Com constantes para os coeficientes Cepstrais e os coeficientes delta.

Os coeficientes cepstrais, que no caso são 16 formam uma matriz, de 16x198, esse número (198) é devido a divisão do quadro em frames. Ou seja, cada coeficiente é na verdade um vetor de 198 posições. Para uma melhor precisão deve ser obtida a média dos 198 valores que formam um coeficiente, para cada um dos 16. Essa média foi usada para construir o modelo de referência. Na tabela 7 são mostrados os primeiros Coeficientes Cepstrais de 1 até 11 de dois usuários, colocados com apenas uma casa de precisão, o que é claro, não é suficiente para aplicação de projeto, mas é o suficiente para o entendimento do leitor.

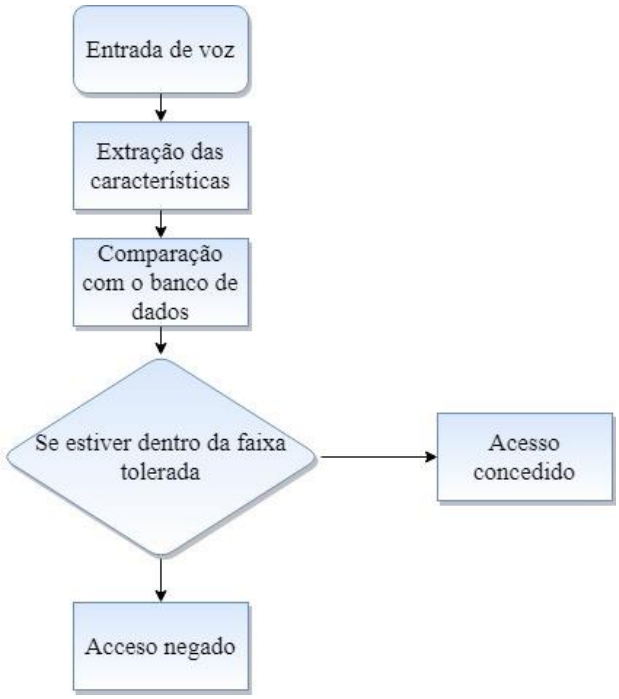
Tabela 7: Coeficientes cepstrais de dois usuários

Nome	CC1	CC2	CC3	CC4	CC5	CC6	CC7	CC8	CC9	CC10	CC11
Maria Eduarda	-0.6	0.2	1.0	-0.5	0.5	-1.6	0.8	0.3	-0.6	-0.5	-0.4
Roseneth Mouzinho	-0.6	0.1	1.2	-0.6	0.7	-1.6	0.8	0.3	-0.5	-0.3	-0.2

Fonte: próprio autor

Na fase de testes é feita uma nova gravação vocal, de um usuário supostamente desconhecido. Após isso, cada amostra contida no banco de dados é analisada, para verificar se o usuário em questão está apto para obter acesso. O teste é feito comparando as amostras cadastradas com essa nova entrada de voz, para cada uma é estipulada uma pontuação. Essa pontuação serve para indicar se o usuário está apto ou inapto para ter acesso ao sistema. O diagrama em blocos, mostrado na Figura 30, ilustra o que foi comentado.

Figura 30: Diagrama em blocos da fase de testes



Fonte: próprio autor

3.3.4 **Integração entre arduino e MATLAB**

Essa é a parte final, ela busca demostrar uma aplicação do sistema, para isso foi utilizado o Arduino. Essa plataforma é uma ferramenta amplamente usada em projetos e desenvolvimentos de protótipo. Sabendo disso várias empresas de software estão criando formas para poder interagir seu produto com as placas da plataforma Arduino. O MATLAB é uma delas, ele já possui um pacote de suporte para a interação com Arduino, permitindo a comunicação através do cabo USB. Com esse pacote é possível controlar o Arduino para leitura/escrita digital/analógica, PWM, comunicação I2C e SPI, controlar servos, fazer plotagens em tempo real de leituras analógicas.

Porém, esse dispositivo possui um poder de processamento e memória limitadas, no caso do Arduino Mega 2560 ele possui uma memória Flash de 256 kb, uma SRAM de 8kb e EEPROM 4kB, conforme mostrado na Tabela 8. Isso significa que a parte do programa principal (Extração de características com MFCC) não poderia ser executada por esse dispositivo. Por essa razão que o MATLAB no computador executa os programas principais e deixa o arduino apenas com o papel de receber o sinal de saída, de permitir ou não o acesso.

Tabela 8: Especificações técnicas do Arduino Mega 2560

Tensão de operação	5V
Tensão de entrada (recomendada)	7-12V
Tensão de entrada (limites)	6-20V
Pinos Digitais I/O	54 (com 14 fornecidos para saída PWM)
Pinos analógicos	16
Corrente DC para pinos I/O	40mA
Corrente DC para pinos de 3.3V	50mA
Memória flash	256 kb com 8kb usados para bootloader
SRAM	8 kb
EEPROM	4 kb
Velocidade do clock	16 MHz

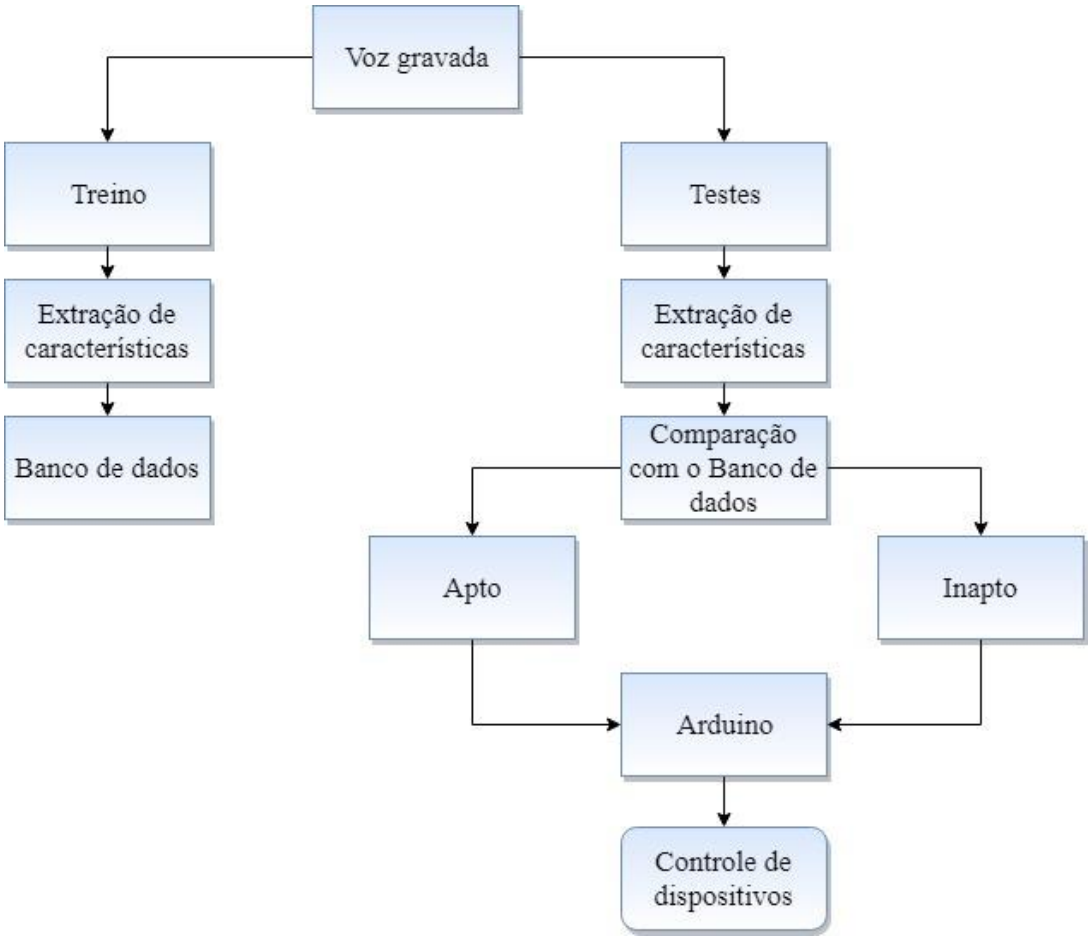
Fonte: (https://www.robotshop.com/media/files/pdf/arduinomega2560datasheet.pdf)

O comando elétrico, através do Arduino, serve para controlar os dispositivos elétricos. Isso abre um leque de opções para a implementação do protótipo desenvolvido nesta pesquisa.

Como por exemplo, isso poderia ser usado para a área de segurança, com o controle de uma fechadura eletrônica.

Essa parte do projeto unifica todos os algoritmos mostrados, para obter o controle de acesso. O diagrama em blocos da Figura 31 mostra como o sistema todo é integrado

Figura 31: Sistema de controle de acesso



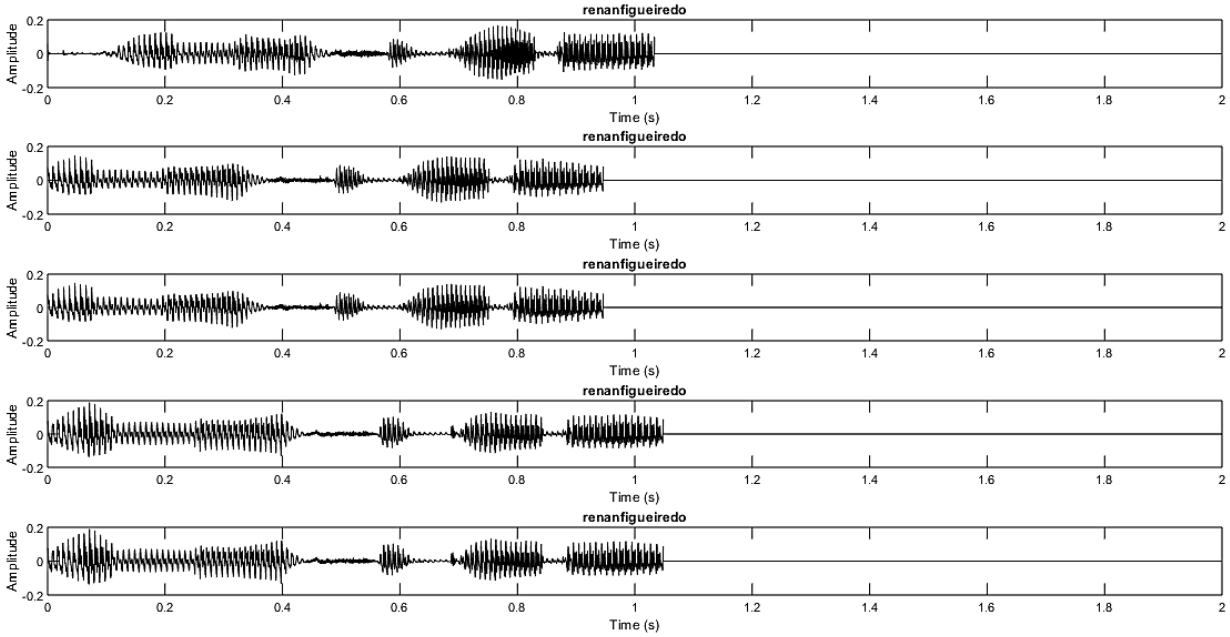
Fonte: próprio autor

Sua implementação no MATLAB é de baixa complexidade, bastando ser instalado a ferramenta no suporte de pacotes. Feito isso o Arduino pode ser chamado como uma variável no programa principal e suas portas podem ser controladas através dos comandos executados no MATLAB.

4. TESTES E RESULTADOS OBTIDOS

As amostras de voz de uma mesma pessoa apresentam diferenças que podem prejudicar o reconhecimento vocal. Essas diferenças normalmente ocorrem pela não adaptação do usuário com a interface de gravação. Além do mais, isso indica que há ruídos no ambiente de gravação, que faz com que as amostras comecem em pontos diferentes e com amplitudes aleatórias. Percebeu-se também que o microfone deve permanecer parado, isso reduz incompatibilidades nos sinais gravados.

Figura 32: Amostras de voz de um dos usuários cadastrados no sistema



Fonte: próprio autor

Os conceitos de ASR e ASI apresentados (ver o tópico 1.3.2), podem ser aplicados para testar a precisão do sistema. A precisão é feita aplicando duas constantes, para ambos os tipos. Uma constante para o MSE dos MFCCs e a outra para o MSE dos coeficientes delta. O valor dessas constantes é mais alto para ASI, já que não é requerida uma alegação de identidade. O sistema ASR apenas diz se o usuário está cadastrado no banco de dados ou não. Para o sistema implementado foram definidas as constantes de K1=2 para o MSE de delta e K2=1.8, para o MSE dos MFCCs. Esses valores foram obtidos empiricamente, por meio de testes no sistema. Se o erro do usuário testado com as amostras no banco de dados estiver abaixo de K1 e K2 ele está apto para ter acesso ao sistema. A Tabela 9 mostra um teste no qual o usuário Kimberly

Mouzinho, já cadastrado no sistema, foi identificado corretamente, nessa tabela estão dispostos os nomes dos usuários da forma como eles foram falados na gravação.

Tabela 9: Erros da fase de teste de um dos usuários

Usuário	MSE MFCCs	MSE Delta	Saída
Kimberly Mouzinho	0,82	0,86	Aceito
Gustavo Aquino	7,83	9,82	Rejeitado
Roseneth Mouzinho	9,02	4,74	Rejeitado
Jacobus de Jager	7,01	15,50	Rejeitado
Herik Mouzinho	5,85	2,15	Rejeitado
Thiago Patrício	16,91	22,00	Rejeitado
Judite Bezerra	5,33	7,57	Rejeitado
Italo Santos	8,31	9,07	Rejeitado
Everton Cassiano	11,13	13,06	Rejeitado
Igor Soares	13,20	15,97	Rejeitado
José Sicchar	12,73	23,94	Rejeitado
Renan Figueiredo	8,55	18,54	Rejeitado
Maria Eduarda	7,11	12,25	Rejeitado
Nikolas da Silva	12,76	15,87	Rejeitado
Caic Otani	6,19	5,55	Rejeitado

Fonte: Próprio autor

O usuário Kimberly Mouzinho foi aceito pois foi único que teve os erros de MFCC e delta abaixo dos valores das constantes K1 e K2. Os dados da tabela acima também apontam que o coeficiente delta está relacionado com a palavra falada, por exemplo, nos nomes Kimberly Mouzinho e Herik Mouzinho o erro delta foi baixo pois esses dois nomes são foneticamente parecidos ambos repetem no final (IMOUZINHO).

Testando para os outros usuários obteve-se a precisão de 93,33%. Ou seja, somente um dos 15 usuários cadastrados não foi identificado corretamente pelo sistema. Para isso cada usuário proferiu mais uma amostra da sua voz, obtendo-se ao todo mais 15 trechos vocais, eles foram comparados com as do modelo de referência. A tabela 10 demonstra, os acertos ao se comparar a amostra de voz de um usuário com o seu modelo de referência.

Tabela 10: Erros da fase de teste dos usuários cadastrados

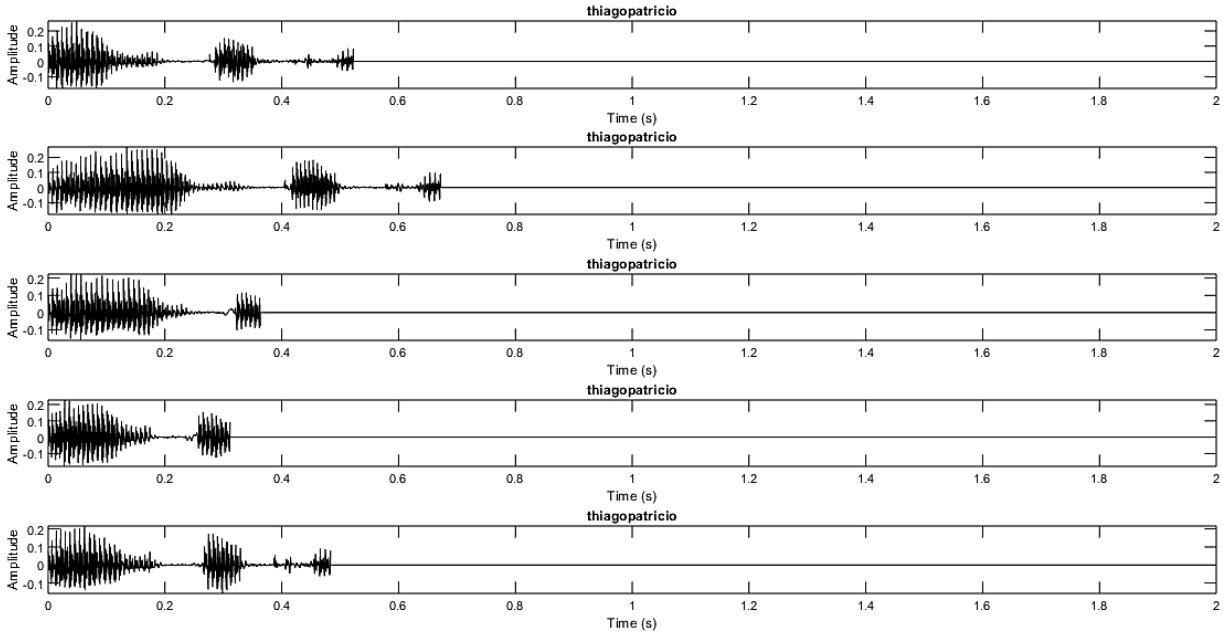
Usuário	MSE MFCCs	MSE Delta	Saída
Kimberly Mouzinho	0.82	0.86	Aceito
Gustavo Aquino	0.60	1.37	Aceito
Roseneth Mouzinho	0.80	0.61	Aceito
Jacobus de Jager	1.14	1.85	Aceito
Herik Mouzinho	0.90	0.81	Aceito
Thiago Patrício	1.16	2.36	Rejeitado

Judite Bezerra	0.39	0.62	Aceito
Italo Santos	0.56	0.77	Aceito
Everton Cassiano	0.24	0.41	Aceito
Igor Soares	0.71	1.38	Aceito
José Sicchar	0.55	1.11	Aceito
Renan Figueiredo	0.40	0.84	Aceito
Maria Eduarda	1.05	0.88	Aceito
Nikolas da Silva	0.57	1.82	Aceito
Caic Otani	0.29	0.58	Aceito

Fonte: próprio autor

Uma análise dos trechos vocais do usuário Thiago patrício, gravadas na fase de treinamento, mostra o porquê essa amostra não foi autenticada. Conforme disposto na Figura 33, que mostra a amostra de voz após a detecção de silêncio final e inicial, os trechos vocais diferem entre si nas 5 amostras, isso provavelmente se deve a algum vício de fala do usuário. Em algumas gravações esse último trecho do áudio reconhecido como ruído, por apresentar uma amplitude baixa e outras a função de detecção de silencio não corta esse áudio pelo mesmo apresentar amplitudes que não podem ser desconsideradas.

Figura 33: Amostras vocais do usuário Thiago patrício



Fonte: próprio autor

Podem também ser feitos testes com impostores, pessoas não cadastradas no sistema tentando obter acesso. Nesses testes os impostores falam o nome de um dos usuários já

cadastrados no sistema. Foram gravadas as vozes de 5 impostores. Para esse caso foi obtido uma precisão de 100%, como mostrado na tabela 11. Os erros mostrados foram os entre as amostras de voz a qual o impostor simulou, falando o nome do usuário cadastrado, em alguns casos o erro delta ficou próximo de 2, porém para todos os impostores o erro pelo MFCC diferiu mais que o dobro da quantidade mínima para o usuário ser aceito por este parâmetro.

Tabela 11: Teste de impostores

Usuário	**MSE MFCCs**	**MSE Delta**	**Saída**
Impostor 1	3.58	2.84	Rejeitado
Impostor 2	4.42	6.62	Rejeitado
Impostor 3	3.24	2.54	Rejeitado
Impostor 4	6.46	6.12	Rejeitado
Impostor 5	7.52	8.54	Rejeitado

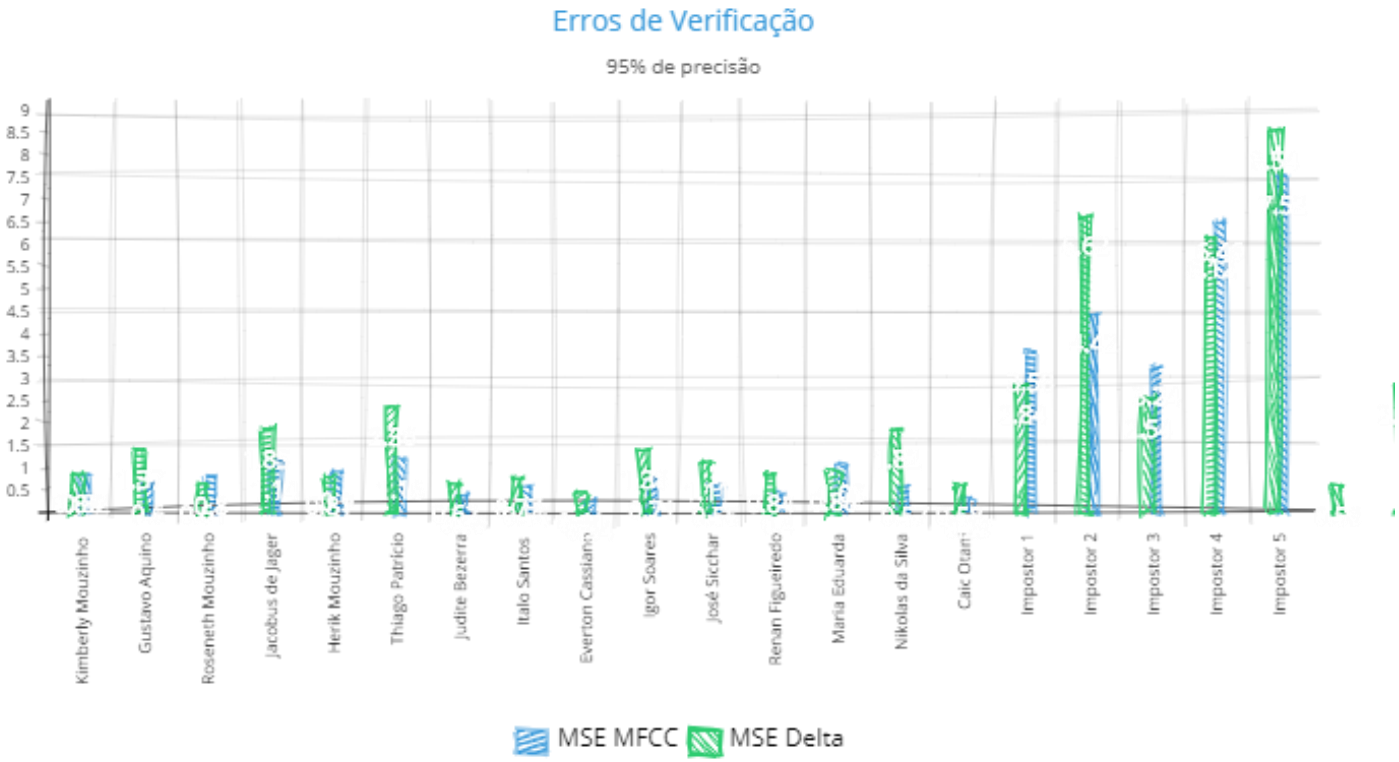
Fonte: próprio autor

Foi realizado outro teste, o de gravar amostras em dois ambientes, um ambiente com o ar condicionado ligado e outros sons de um ambiente ruidoso, e outro com a menor quantidade de ruídos possível. Observou-se então que o erro Delta aumentava, enquanto o erro por MFCC não sofria mudanças significativas. O que significa que o sistema pode ser implementado em ambientes diferentes, desde que na fase de treinamento este não apresente níveis de ruídos elevados. Porém não é recomendado a mudança de ambiente, já que a sala acústica no qual a voz é coletada interfere na forma como a fala é captada pelo microfone

Além disso, devido a função de detecção de silêncio final e inicial, observou-se que a matriz dos coeficientes cepstrais e a matriz delta apresentavam muitos zeros. Então para compor as médias para os testes e treinamento foram desprezados os zeros, e utilizada a média apenas dos valores não nulos, em busca de uma melhor precisão.

De modo geral, o sistema apresentou uma precisão de 95%, considerando 14 acertos de usuários cadastrados e 5 acertos de impostores. Conforme disposto na tabela 14.

Tabela 14: Precisão do sistema



Fonte: próprio autor

5. CONCLUSÃO

A biometria é de fato uma tecnologia inovadora, em específico o reconhecimento de voz, por dessas biometrias ser a mais natural de ser realizada pelo usuário. A pesquisa sobre reconhecimento biométrico de voz, forneceu habilidades que podem ser usadas para todos os outros sistemas ASR. Já que os procedimentos algumas vezes acabam de repetindo. No caso de sistemas de reconhecimento de fala e autenticação do falante os métodos são muito parecidos (HUANG, 2001). Até mesmo para texto dependente e texto independente os alguns procedimentos acabam se repetindo (CAMPBELL,1997).

A extração das características vocais é a parte mais importante no reconhecimento do falante. O método de extração utilizado (MFCC) se mostrou útil. Porém, as pesquisas demonstraram inconsistências de aplicação desse método, alguns autores citam mais procedimentos e outros excluem alguns. Como por exemplo, (SHAH, 1997) não considera importante fazer o enquadramento ou obter os coeficientes delta. Já (MUDA, 2010) e outros autores citam o enquadramento como parte essencial do projeto. Essas inconsistências levaram a fazer um método MFCC com algumas adaptações e sugestões propostas pelo próprio autor deste trabalho. Tal como a função de detecção de silêncio.

As amostras de voz mostraram-se susceptíveis a ruídos externos. Além disso, os testes apostaram que a fase de testes e treino devem ser gravadas no mesmo ambiente, e com o mesmo microfone, com intuito de diminuir erros. Já que as constantes de autenticação, como no caso K1 e K2, variam para cada dispositivo de gravação devido ao som ser captado com amplitudes diferentes.

O desenvolvimento desse trabalho por si já de grande validade para o passar da vida acadêmica para a profissional. Tendo em vista que neste projeto foi necessária muita multidisciplinaridade, por se tratar de algo matemático e bastante complexo. Os conceitos de programação no MATLAB, voltados para a área de processamento de sinais digitais, servirão como base para outras pesquisas. O desenvolvimento dessa pesquisa sanou as dúvidas de conceitos teóricos como por exemplo, bancos de filtro, transformada cosseno discreto, enquadramento do sinal, entre outros.

Para trabalhos futuros deve-se considerar o uso de um método para a parte da construção do modelo de referência, como o modelo estatístico de reconhecimento de padrões proposto por Hilden Markov, em específico o utilizando mistura gaussiana (HMM-GMM). Esse método vem se destacando muito e promete resultados mais satisfatórios como o erro entre 1 a 3 % (REYNOLDS, 1995).

Os objetivos iniciais foram alcançados, foram cadastrados usuários, suas características vocais foram extraídas e armazenadas corretamente e a precisão obtida mostrou que é possível realizar um controle de acesso baseado em reconhecimento de voz. Já que o protótipo biométrico implementado inicialmente, testado com 20 usuários e um total de 95 amostras de áudio, obteve uma precisão de 95%. Com poucas modificações esse sistema poderia ser usado para a área de segurança dispensando o uso de outro método de acesso, como senhas e cartões de acesso.

REFERÊNCIAS BIBLIOGRÁFICAS

Audacity. Software de gravação. Disponível em: https://www.audacityteam.org. Acesso em 22 de novembro de 2018

CAMPBELL JP. Speaker recognition: a tutorial. Proceedings of the IEEE. 1997 Sep; 85(9):1437–62.

DAVIS, S. and P. Mermelstein, "Comparison of Parametric Representations for Monosyllable Word Recognition in Continuously Spoken Sentences," IEEE Trans. on Acoustics, Speech and Signal Processing, 1980, 28(4), pp. 357-366.

Datasheet Arduino mega 2560. Disponível em: https://www.robotshop.com/media/files/pdf/arduinomega2560datasheet.pdf. Acesso em 22 de novembro de 2018

DELLER, J. H. Hansen, and P. J.G, “The relation of pitch to frequency,” IEEE Press, vol. 1, p. 936, 2000.

DÍGITRO, Ensinare. Telefonia Digital: Curso on-line, 2003. Disponível m: http://ensinare.locaweb.com.br. Acesso em 18 novembro de 2018

Cordas vocais. Disponível em: http://marcoseferin.com.br/2016/12/05/cordas_vocais/. Acesso em 18 de novembro de 2018.

Espectrograma vocal. Disponível em https://commons.wikimedia.org/wiki/File:StrcPrstSkrzKrk.png. Acesso em 20 de novembro de 2018.

Fayek. Os coeficientes MFCCs. Disponível em https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html. Acesso em 02 de novembro de 2018.

HUANG, X., Acero, A., Hon, H., 2001. Spoken Language Processing A guide to theory, algorithm, and system development. Prentice Hall, Upper Saddle River, NJ, USA .

MUDA, Begam KM, Elamvazuthi I. Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques. Journal of Computing. 2010; 2(3):138–143

OPPENHEIM, Allan V; Ronald .W.Schafer. Processamento em tempo discreto de sinais. 2013. Pearson Education do Brasil.

RASHID, Ra et al. Security system using biometric technology: design and implementation of Voice Recognition System (VRS). 2008 International Conference on Computer and Communication Engineering.; Kuala Lumpur.p. 898–902.x

RABINER, L.R.; SCHAFER, R.W.; Digital techniques for computer voice response: Implementations and applications. Proceedings of the IEEE, V. 64, Abril 1978

REYNOLDS, D. and B. Carlson, “Text-dependent speaker verification using decoupled and integrated speaker and speech recognizers,” in Proc. EUROSPEECH, Madrid, Spain, 1995, pp.

SHAH, Hairol Nizam Mohd.et al. Biometric Voice Recognition in Security System. Indian Journal of Science and Technology, vol. 7, fev. 2014.

SMITH, S. W. The Scinentist and Engineer’s Guide to Digital Signal Processing. California Technical Publishing. 1997

VACCA, John R. Biometric Technologies And Verification Systems. 1. ed. Estados Unidos da América: Elsevier Inc., 2007.