PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ANÁLISE DE ÁUDIO DE VOZ PARA IDENTIFICAÇÃO DO EMISSOR UTILIZANDO TÉCNICAS DE PROCESSAMENTO DE SINAIS E REDES NEURAIS ARTIFICIAIS

MARIA REGINA SANTOS DE DEUS

GOIÂNIA
2021

MARIA REGINA SANTOS DE DEUS

ANÁLISE DE ÁUDIO DE VOZ PARA IDENTIFICAÇÃO DO EMISSOR UTILIZANDO TÉCNICAS DE PROCESSAMENTO DE SINAIS E REDES NEURAIS ARTIFICIAIS

Monografia de conclusão de curso apresentada ao curso de Ciência da Computação, da Escola Politécnica da Pontifícia Universidade Católica de Goiás, como requisito parcial à conclusão do curso.
Orientador(a):
Prof. Max Gontijo de Oliveira

GOIÂNIA
2021

MARIA REGINA SANTOS DE DEUS

ANÁLISE DE ÁUDIO DE VOZ PARA IDENTIFICAÇÃO DO EMISSOR UTILIZANDO TÉCNICAS DE PROCESSAMENTO DE SINAIS E REDES NEURAIS ARTIFICIAIS

Este Trabalho de Conclusão de Curso julgado adequado para obtenção o título de Bacharel em Ciência da Computação, e aprovado em sua forma final pela Escola Politécnica, da Pontifícia Universidade Católica de Goiás, em ____/___/_____.

_____
Coordenador(a) de Trabalho de Conclusão de Curso

Banca examinadora:

_____
Orientador: Me. Max Gontijo de Oliveira

_____
Profa. Ma. Lucília Ribeiro

_____
Prof. Me. Fernando Gonçalves Abadia

GOIÂNIA
2021

Para meus pais, Antônia e Salvino,
minha eterna gratidão
por sempre me apoiarem

AGRADECIMENTOS

Aos meus pais Antônia e Salvino, por me ensinarem a sonhar, por sempre me apoiarem e acreditarem em mim, mesmo quando eu mesma não acreditava.

Aos meus irmãos, Salvino, Lindormar (*in memorian*), Luismar e Lucas, pelo amor, suporte, proteção e por me aguentarem.

Ao meu orientador, Me. Max Gontijo, minha gratidão pela dedicação, incentivo e compreensão.

Agradeço esta universidade, em especial à Escola Politécnica pela oportunidade de realizar este trabalho.

A todos os professores, por compartilhar seus conhecimentos, sem eles nada disso seria possível.

Em especial agradeço a amizade dos professores Alexandre, Lucília e Max.

Aos meus colegas pelo auxílio no decorrer do curso.

Agradeço as amizades que se formaram no decorrer do curso.

Finalmente, agradeço a todos que direta ou indiretamente contribuíram para minha formação e construção deste trabalho.

RESUMO

Este projeto visa realizar a identificação do emissor de uma amostra de voz, caso existam amostras cadastrados em uma base de dados. Essas amostras serão usadas para treinamento da rede neural *multilayer perceptron* (MLP) que realizará a identificação do emissor do sinal de voz. Para isso é necessário utilizar métodos de manipulação de sinais e extração de características da voz, sendo utilizados neste trabalho o método de extração de características *Mel-Frequency Cepstral Coefficents* (MFCC). Em posse das características é possível identificar os padrões do sinal de voz do emissor, utilizando neste trabalho o pacote *Python scikit-learn* com ferramentas de *machine learning*. Os resultados obtidos nos experimentos mostram que o algoritmo é eficiente para a identificação, quando as variáveis de configuração estão devidamente calibradas. Porém apresenta limitações quando é considerado a possibilidade de manipular os sinais de voz ou até mesmo quando consideramos a possibilidade de uma inteligência artificial imita o sinal de voz original.

Palavras-chave: Processamento de sinais. Identificação de emissor. Redes neurais. MPL. Máquinas de Comitê

ABSTRACT

This project aims to identify the sender of a voice sample, if there are samples registered in a database. These samples will be used for training the neural network multilayer perceptron (MLP) that will perform the identification of the voice signal emitter. For that, it is necessary to use methods of signal manipulation and voice characteristics extraction, being used in this work the Mel-Frequency Cepstral Coefficents (MFCC) characteristic extraction method. In possession of the characteristics, it is possible to identify the patterns of the voice signal of the sender, using in this work the Python scikit-learn package with machine learning tools. The results obtained in the experiments show that the algorithm is efficient for identification, when the configuration variables are properly calibrated. However, it has limitations when considering the possibility of manipulating the voice signals or even when considering the possibility that an artificial intelligence mimics the original voice signal

Keywords: Signal processing. Issuer identification. Neural networks. MPL. Committee Machines

LISTA DE FIGURAS

LISTA DE SIGLAS

DCT - Transformada Discreta de Cosseno

DFT - Transformada Discreta de Fourier

FFT - Transformada Rápida de Fourier

MFCC - Mel Frequency Cepstral Coefficients

SVM - Support Vector Machine

SUMÁRIO

1. INTRODUÇÃO

Estudos de processamento de sinais de voz começaram em 1950 quando os laboratórios Bell criaram um sistema capaz de identificar números de 0 a 9. Em 2000 já era possível entender as palavras ditas e transcrever para textos (HASSAN et al., 2019).

O processo de identificação de um emissor através da fala pode ser usado tanto para a ciência forense como para a autenticação (TIRUMALA et al, 2017). Para esse último existem outras formas de autenticação biométricas, como por exemplo, impressão digital, reconhecimento facial e reconhecimento de íris. Porém, uma vantagem da biometria usando a voz garante sobre às outras é uma maior acessibilidade aos seus usuários.

Para o uso em ciências forense, a identificação seria possível caso a voz do emissor estivesse cadastrada em um banco de dados e fosse comparada uma amostra obtida por algum meio, como por exemplo, chamadas telefônicas.

Celulares em poder de detentos é um problema recorrente nas penitenciárias brasileiras. Notícias sobre apreensão de celulares em presídios são comuns por todo território nacional. No início do ano (2021), mais de 1400 celulares apreendidos em presídios seriam doados a estudantes da rede pública (DIAS, 2021). Recentemente (11/2021) um detento foi descoberto compartilhando sua rotina em uma rede social. Depois desse episódio ele foi transferido para uma prisão de segurança máxima (G1, 2021).

Pensando em um cenário em que, assim como as impressões digitais, uma base de dados com amostras de voz das pessoas fosse construída, seria possível identificar a pessoa fora da prisão com quem um detento se comunica ou o detento com o qual uma pessoa grampeada fora da cadeia se comunica.

Com base nisso, este trabalho visa responder a seguinte questão de pesquisa: é possível identificar o emissor de um áudio tendo amostras de sua voz previamente cadastradas em uma base de dados?

Assim, o objetivo geral deste trabalho é identificar um emissor de um sinal de voz e classificar as características do sinal. Para isso será necessário: realizar a extração das características dos sinais, e a classificação das características obtidas para identificação.

Este trabalho é organizado da seguinte maneira: introdução que apresenta o

problema, objetivo e estrutura do trabalho.

Capítulo 2, a Fundamentação teórica, onde são detalhados os conceitos que foram utilizados durante o desenvolvimento do trabalho, além dos trabalhos relacionados.

Capítulo 3, Método, que apresenta as metodologias de pesquisa utilizadas e as tecnologias e das especificações técnicas utilizadas.

Capítulo 4, o Desenvolvimento, é detalhado os passos para montar a base de dados e extração de características dos sinais de voz.

Capítulo 4, Considerações Finais, traz a conclusão e sugestões para trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

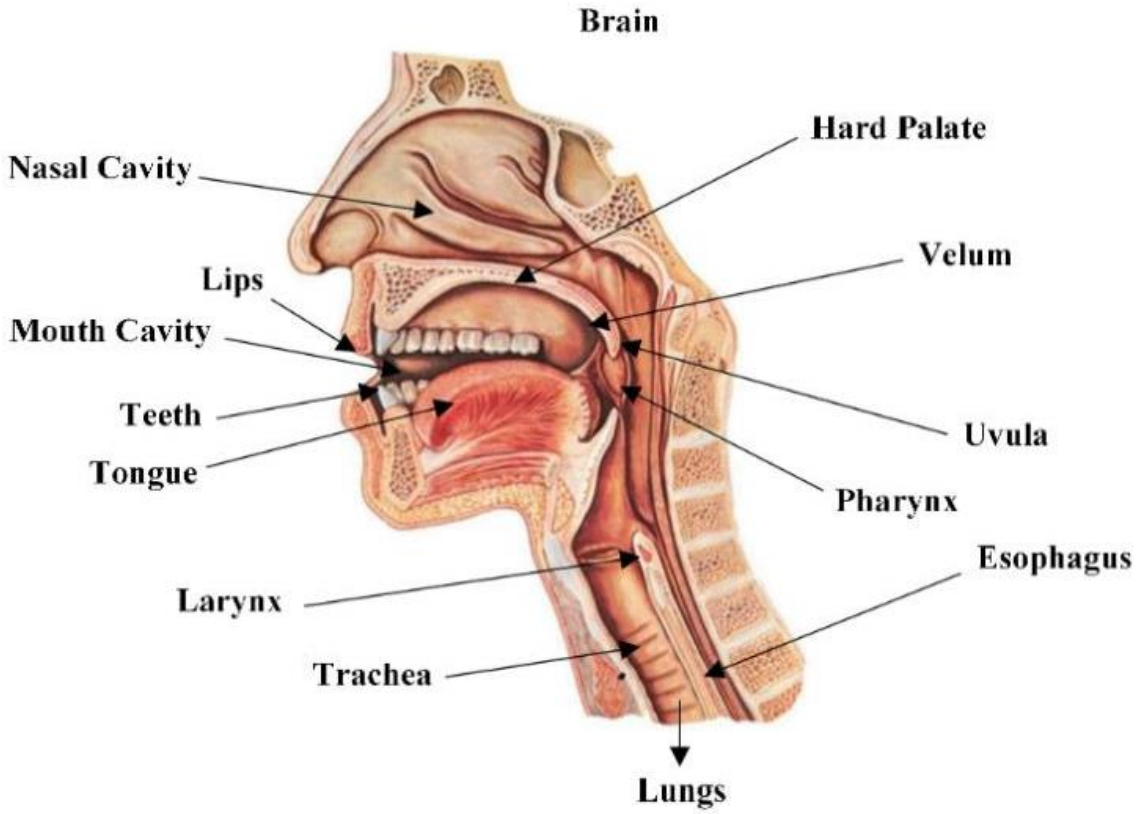
Tendo em vista que o objeto base do estudo deste trabalho são segmentos de áudio em formato digital, faz-se necessário ter um entendimento sobre a natureza do som, mais especificamente, da voz. Além disso, uma vez que a proposta do trabalho passa pela caracterização desses segmentos, compreender algumas técnicas para esse propósito também se faz necessário.

Essa seção irá apresentar os principais conceitos que formaram a base para que o trabalho pudesse ser desenvolvido.

2.1 Emissão e percepção da voz

Cada pessoa possui uma anatomia única, o que faz com que sua voz seja única, tornando possível a realização da identificação do emissor pela voz (HASSAN et al., 2019).

Figura 1 – Anatomia vocal



Fonte: Hassan et al. (2019)

Segundo Hassan et al. (2019):

> “A produção da fala começa nos pulmões. O ar é empurrado dos pulmões para a traqueia. No topo da traqueia existe a laringe ou caixa de ressonância. A abertura e fechamento da laringe é controlada por nosso cérebro.”

Dentro da laringe existem as cordas vocais. Quando o ar passa por essas cordas, faz com que elas vibrem produzindo a voz.

A voz é recebida pelo ouvinte na forma de onda sonora. A onda é transmitida até o tímpano que vibrará na mesma frequência (MACHADO, 2016). As vibrações serão transmitidas ao ouvido médio e interno, chegando à cóclea que atuará como filtro de frequências. Finalmente as informações serão enviadas ao cérebro, onde serão interpretadas (MACHADO, 2016 *apud* HUNG et al., 2001).

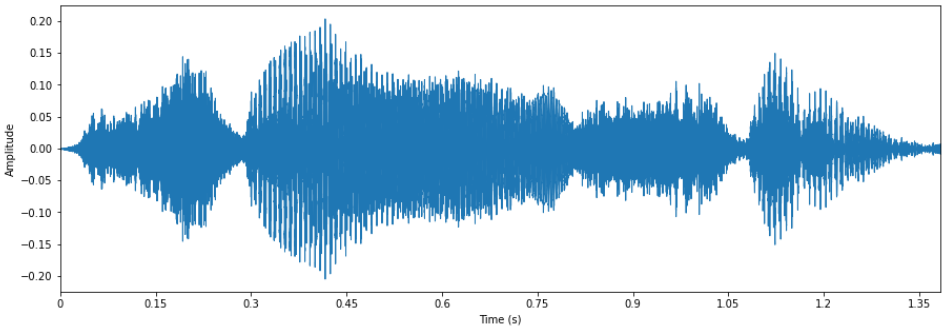
Uma onda de áudio só será percebida por um ser humano caso sua frequência esteja entre 20 Hz e 20.000 Hz, que são as frequências perceptíveis para os ouvidos humanos (MÉNDEZ *at al.*, 1994).

Uma vez que a voz humana é emitida na forma de uma onda, o estudo de sua natureza pode ser realizado analisando-a como um sinal. Dessa forma, caracterizar detalhes que tornem possível a identificação do emissor passa pelo entendimento da voz como um sinal.

2.2 Sinais

Sinais são matematicamente representados por funções com uma ou mais variáveis independentes, que contêm informações sobre o comportamento ou natureza de um fenômeno. Um exemplo de sinal de áudio pode ser visto na Figura 2 o sinal representa a palavra “Ciência”. Sinais podem ser divididos em sinais contínuos ou sinais discretos (OPPENHEIM e SCHAFER, 2012).

Figura 2 – Sinal de voz da palavra ciência

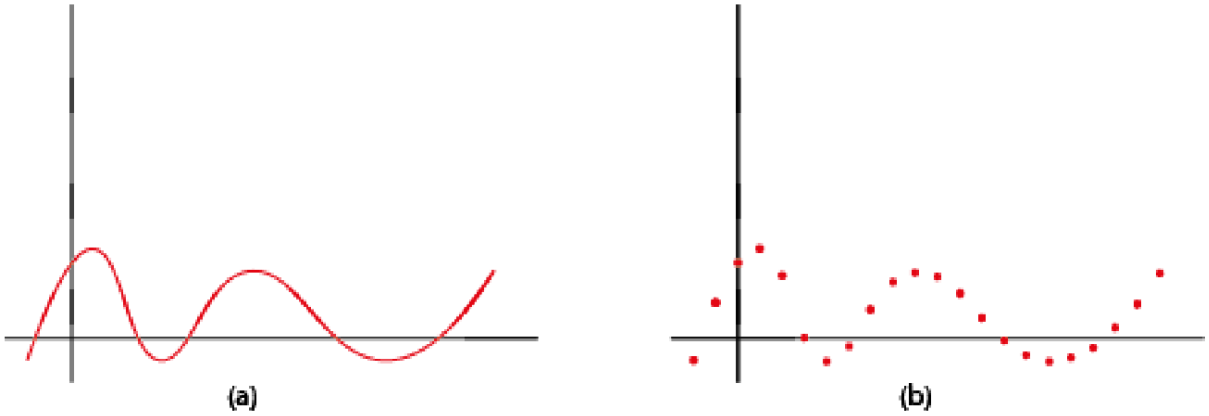


Fonte: Autoria própria (2021)

Sinais contínuos possuem variável independente contínua. Um exemplo é um sinal de fala, em função do tempo (Figura 3(a)). A representação matemática de um sinal contínuo no domínio do tempo será representada por $x(t)$ onde t representa a variável independente de tempo contínuo (OPPENHEIM e SCHAFER, 2012).

Já sinais discretos possuem variáveis independentes que podem assumir apenas valores discretos (Figura 3(b)). Serão representados por $x[n]$, onde n representa a variável independente de tempo discreto. Assume que $x[n]$ tem valor indefinido quando n não é um inteiro (OPPENHEIM e SCHAFER, 2012).

Figura 3. Sinais contínuo e discreto



Fonte: Autoria própria (2021)

Sinais discretos podem ser obtidos de diversas formas. A maneira mais comumente usada é através da amostragem de sinais contínuos.

Usualmente a conversão de sinal contínuo para discreto é realizado por um conversor analógico-digital (A/D), multiplicando o sinal contínuo por uma função impulso unitário. Para obter a relação no domínio da frequência é usado o teorema da amostragem de Nyquist, que diz que a taxa de amostragem mínima deve ser, no mínimo, o dobro da frequência. Caso essa condição não seja atendida, tem-se uma distorção chamada de *aliasing*, onde as amostras se sobrepõem e quando somadas resultarão em uma frequência mais baixa. Para ignorar essas frequências pode ser aplicado um filtro passa-baixa. (OPPENHEIM; SCHAFER, 2012).

2.3 Filtros de frequências

São filtros que selecionará as frequências que respeitam certos parâmetros. O filtro passa-baixa, por exemplo, permite frequências inferiores a um valor limiar; o filtro passa-alta permite frequências superiores a um valor limiar; o filtro passa-banda (também chamado de passa-faixa) permite determinada faixa de frequências e atenua os valores não permitidos (OPPENHEIM; SCHAFER, 2012).

2.4 Ruídos

Ruídos são sinais indesejados somados ao sinal original. Podem ser causados por interferências sonoras no ambiente, como conversas paralelas, barulho de veículos na rua, barulhos de animais etc. Além disso, ruídos podem ainda acontecer devido a alguma perda de dados ao transportar ou na conversão entre formatos (HANECHE et al., 2020).

Para atenuar os ruídos e melhorar o sinal original existem filtros que podem ser aplicados ao áudio.

Métodos mais avançados para redução de ruídos, constantes no áudio, consiste em apresentar uma amostra do ruído, obter e minimizar suas frequências do sinal principal. Essas amostras podem ser encontradas em pausas nas falas ou no início ou fim de cada áudio.

2.5 Mel-Frequency Cepstral Coefficients (MFCC)

Diferenças fisiológicas e comportamentais no sistema de produção de fala humana permitem que um sinal de voz tenha características próprias. O principal fator que determina essas diferenças é a forma do trato vocal (MOLLA; HIROSE, 2004).

Desenvolvido por Davis e Mermestein em 1980, o Mel-Frequency Cepstral Coefficients (MFCC) tem como objetivo extrair informações de sinais de falas com base em suas sintática e duração (DAVIS; MERMELSTEIN, 1980).

Segundo TIRUMALA et al (2017), MFCC é o método mais popular e bem-sucedido para a extração de características. Sendo dividido em:

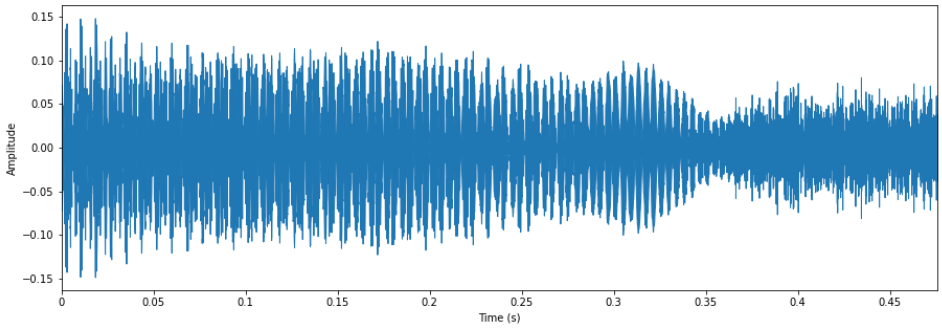
- Enquadramento de sinal e Janelamento

- Transformada rápida de Fourier (do inglês *Fast Fourier Transform, FFT*)
- Cálculo dos filtros MEL
- Transformação Discreta dos Cossenos

2.6.1 Enquadramento de Sinal e Janelamento

Um sinal de voz está sempre variando no decorrer do tempo, desta forma, é necessário dividi-lo igualmente em pequenas amostras de forma que a variação quase não exista. Essa segmentação pode causar descontinuidade do sinal nas extremidades de cada amostra (MACHADO, 2016), como podemos ver na Figura 4.

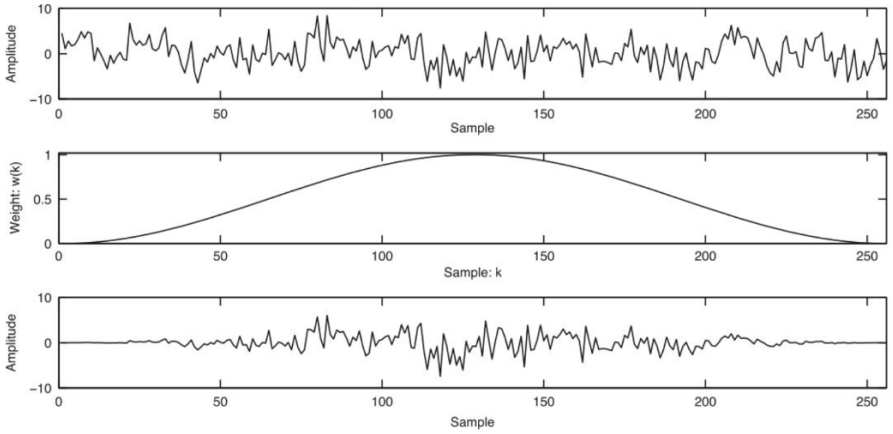
Figura 4 – Sinal com descontinuidades



Fonte: Autoria própria (2021)

Para evitar que esta descontinuidade na amostra cause ruídos e impacte o resultado, multiplicamos a mesma por uma função de janelamento que limita a duração do sinal, suavizando suas bordas, conforme Figura 5.

Figura 5 – Aplicação da função de janelamento



Fonte: (VERLADO, 2020)

Neste trabalho será usado o janelamento de *Hamming.* Sendo definido pela equação a seguir (MACHADO, 2016):

$$w(n) = 0.54 - 0.56\ cos\ \frac{2\pi n}{N-1} \qquad 0\ \leq\ n \leq N-1 \tag{1}$$

Onde n é a posição atual que está sendo percorrida na amostra e N seu comprimento.

2.6.2 Transformada Rápida de Fourier

Inicialmente o trabalho desenvolvido pelo francês Jean Baptiste Joseph Fourier, publicado no ano de 1822, ajudou outros pesquisadores a desenvolver e finalizar a série e a transformada de Fourier. Seu trabalho impacta áreas como matemática, ciência e engenharia (OPPENHEIM e SCHAFER, 2012).

Em 1965 Cooley e Tukey desenvolvem a Transformada Rápida de Fourier, do inglês *Fast Fourier Transform* (FFT). Trata-se de um algoritmo eficiente, baseado na Transformada Discreta de Fourier (do inglês *Discrete Transform Fourier,* DFT). Essa transformada pode ser usada para converter amostras de uma série (obtidas do janelamento) no domínio do tempo para o domínio da frequência (MACHADO, 2016). O resultado desta transformada é chamado de espectro. Seus gráficos mostram a amplitude e a fase em relação à frequência (PAULA FILHO, 2000). A transformada pode ser definida por:

$$x_k = \ \sum_{n=0}^{N-1} x[n]\ e^{-i2\pi Kn/N} \qquad 0\ \leq\ n \leq N-1 \tag{2}$$

Sendo $x[n]$ é o sinal discreto e a frequência angular representada por $2\pi/N$.

O resultado da

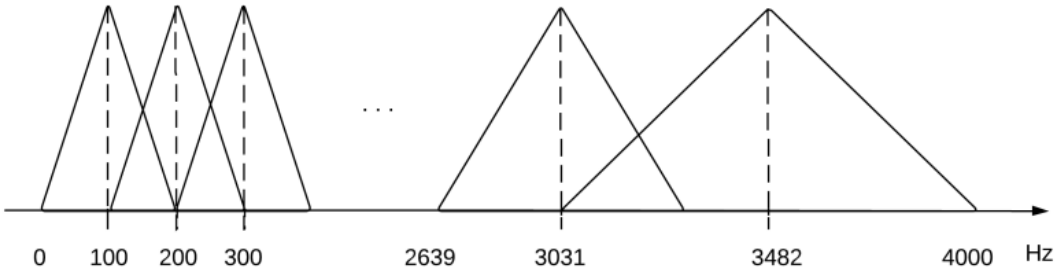
2.6.3 Cálculo dos Filtros MEL

Foi desenvolvido em 1940 por Stevens e Volkmann uma escala chamada de escala Mel, com o objetivo de criar uma relação entre a frequência de uma onda e como ela era interpretada pelo aparelho auditivo humano. Para frequências de até 1000 Hz a relação é linear. Já para frequências acima de 1000 Hz, a relação pode ser representada por (THIAGO, 2017):

$$Fmel = \frac{1000}{log\ (2)}[1 + \frac{F_{Hz}}{1000}] \tag{3}$$

A escala também pode ser obtida através da implementação de filtros passa-banda, por exemplo a Figura 6.

Figura 6 – Banco de filtros na escala *mel*



Fonte: (MARTINS; YNOGUTI, 2014)

2.6.4 Transformada Discreta dos Cossenos

A Transformada Discreta de Cosseno (do inglês Discrete Cosine Transform, DCT), foi introduzida em 1974 por Ahmed, Natarajan e Rao sendo proposta em 1983 por Wang e Hunt. Podendo ser usada em análises de espectro de frequência ou em compressão de dados (ZHOU; CHEN, 2009).

A Transformada Discreta de Cosseno irá transformar os coeficientes do domínio da frequência para o domínio do tempo (MACHADO, 2016). Sendo definida por:

$$C_n = \sum_{k=1}^{K} log\ S_k\ \ cos\ [n(k - \frac{1}{2})\frac{\pi}{k}] \quad (4)$$

2.7 Redes Neurais

É um sistema computacional que tenta simular a forma que o cérebro realiza uma tarefa. Uma rede neural adquire seu conhecimento através de um processo de aprendizado. Ela é formada por unidades de processamento de informações e assim como o cérebro, essas unidades são chamadas de neurônios. Uma rede neural com um único neurônio é a forma mais simples de uma rede neural sendo chamada de *perceptron* (HAYKIN, 2001).

Uma rede neural pode ter diferentes estruturas, são elas redes com camada única, múltiplas camadas e redes recorrentes. A camada de uma rede neural é composta por uma série de neurônios. Em uma rede com camada única haverá

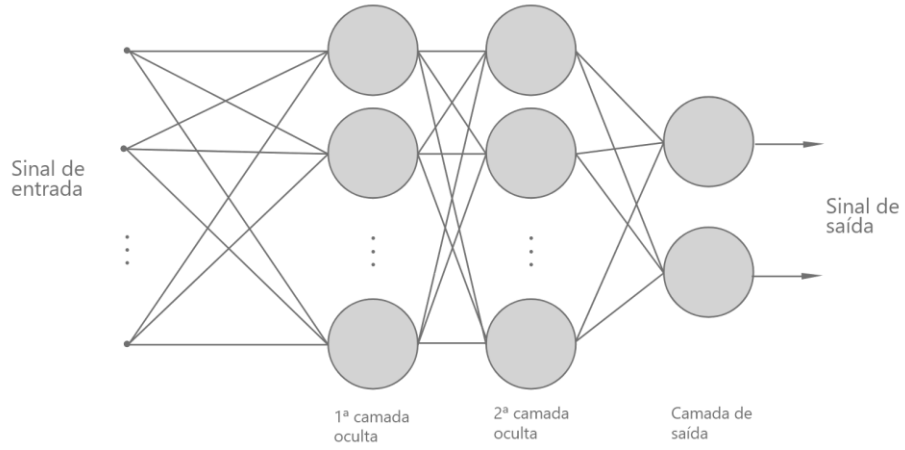
apenas uma camada. Enquanto em uma rede com múltiplas camadas é composta por uma ou mais camadas ocultas que permite uma perspectiva global (HAYKIN, 2001).

Algumas redes neurais são do tipo classificadores, que têm como objetivo extrair características dos exemplos de treinamento. Geralmente, um exemplo é representado por uma tupla de valores (HAYKIN, 2001).

2.7.1 *Perceptrons de múltiplas camadas*

Os *perceptrons* de múltiplas camadas (MLP, *multilayer perceptron*), se trata de redes com uma ou mais camadas ocultas (Figura 7). Possui treinamento supervisionado através do algoritmo de retropropagação de erro (HAYKIN, 2001).

Figura 7 – MLP



Fonte: Adaptado de Haykin (2001)

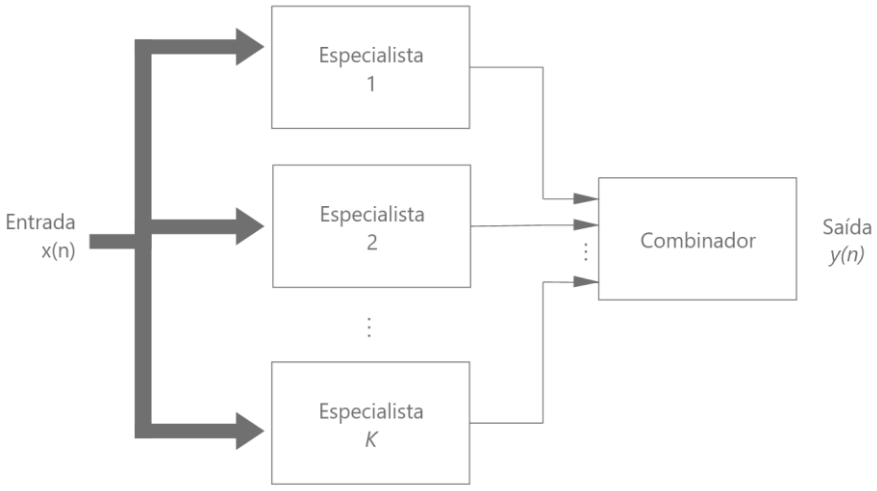
O aprendizado de retropropagação consiste em duas etapas, sendo a primeira a propagação, onde os valores de entrada são aplicados aos nós sensoriais e seus valores se propagam camada por camada. No segundo passo consiste na retropopagação quando os pesos sinápticos são ajustados de acordo com a regra de correção de erro. O erro é calculado subtraindo a reposta encontrada pela rede e a reposta desejada. Este processo de aprendizado é chamado de aprendizado por retropropagação (HAYKIN, 2001).

2.8 Máquina de Comitê

Conforme Haykin (2001) são máquinas que abordam o princípio de dividir e conquistar. Trata de um conjunto de redes neurais intituladas de *especialistas*, a

máquina de comitê funde o conhecimento adquirido pelos especialistas para encontrar uma decisão global, conforme Figura 8. Podem ser classificadas em duas categorias: Estrutura estática e estrutura dinâmica. Neste trabalho será utilizado a estrutura estática.

Figura 8 – Máquina de comitê



Fonte: Adaptado de Haykin (2001)

Ainda conforme Haykin (2001), na estrutura estática os especialistas não são combinados por um mecanismo que envolva o sinal de entrada. Um método da categoria é a *média de ensemble,* onde as redes são treinadas para obter mínimos locais e seus resultados são combinados para aumentar o desempenho global.

2.9 Trabalhos relacionados

Hassan et al. (2019) propôs um sistema classificador *Support Vector Machine* (SVM). Utilizando 120 amostras, extraindo duas características, sendo elas o máximo gradiente do tom e o máximo gradiente do *cepstrum*. Inicialmente é realizada a identificação do gênero e em seguida a identificação do interlocutor. A solução apresentada alcançou acurácia de 83.33%.

Machado (2016) apresentou um sistema para reconhecimento de voz utilizando o método MFCC para a extração de características e quantização vetorial para classificação e reconhecimento de padrões. O trabalho propõe reconhecer palavras ou comandos e a identificação do emissor. Foi utilizado 144 amostras de 8 pessoas. Dos testes realizados em 20 comandos, foram identificado o emissor de maneira correta em 18 deles.

3 MÉTODO

Para o desenvolvimento deste trabalho foram utilizados os procedimentos metodológicos pesquisa bibliográfica e experimental.

A pesquisa bibliográfica realizada na fase inicial do trabalho, consiste em buscar pelas soluções que estão sendo apresentadas pela sociedade acadêmica. Isso nos permite conhecer as limitações e definições do problema e solução que estamos propondo (WAZLAWICK, 2009).

A pesquisa experimental consiste em um trabalho onde o pesquisador realiza alterações no ambiente pesquisado e observa se suas alterações causam as respostas esperadas (WAZLAWICK, 2009).

Os experimentos foram realizados utilizando uma máquina com a seguinte configuração:

1. Processador AMD *Ryzen* 7 3700X 3.59 GHz
2. 16 GB de memória 2800MHz
3. *Nvidia GeForce GTX* 1050 Ti

Os sistemas e softwares utilizados para o desenvolvimento foram:

1. Sistema Operacional (SO) *Windows* 10 *Home* versão 20H2
2. Linguagem de programação *Python* 3.9
3. Sistema Gerenciador de Banco de Dados (SGBD) *PostgreSQL* 9.6
4. *Software* para a manipulação de áudio *Adobe Audition* 14.4
5. Ambiente de desenvolvimento integrado (IDE do inglês *Integrated Development Environment*) *PyCharm 2021.2*
6. Versionamento do código realizado utilizando o Git

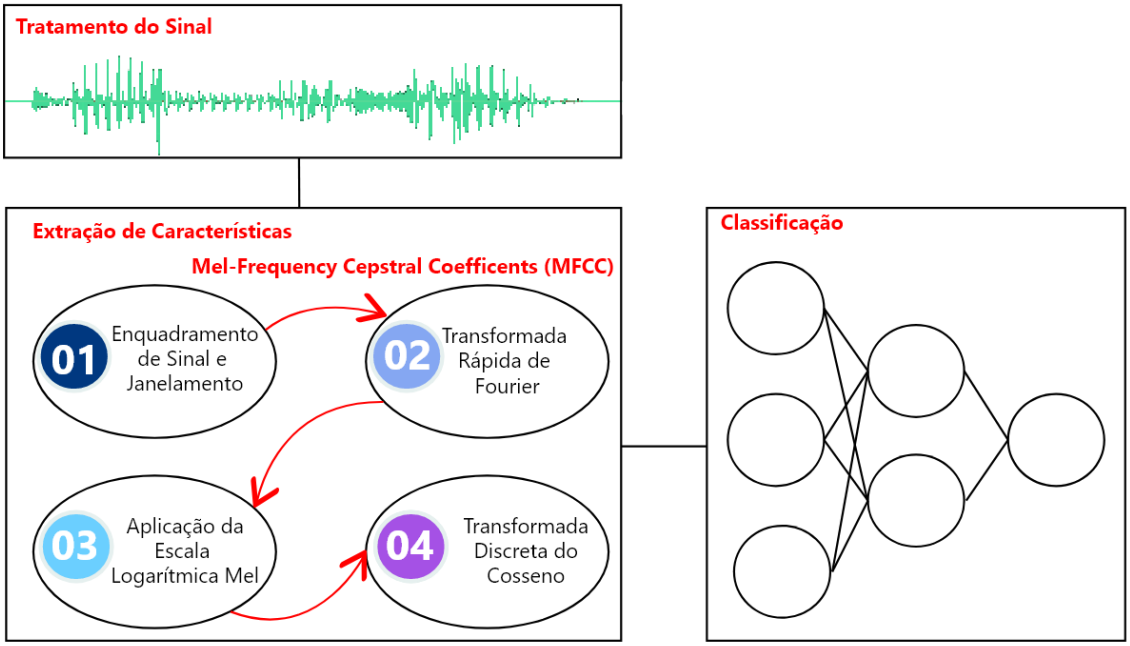
Inicialmente foi realizada a escolha da base de dados que seria utilizada nos experimentos.

O segundo passo foi a escolha e tratamento dos áudios que seriam utilizados.

Por último, implementar a extração de características do sinal de voz e a identificação do emissor.

Esses passos podem ser vistos na Figura 9 e são mais detalhados no capítulo 4.

Figura 9 – Fluxo reconhecimento do emissor



Fonte: Autoria própria (2021)

4. DESENVOLVIMENTO

No decorrer deste capítulo serão relatados os passos realizados no desenvolvimento deste trabalho.

4.1 Base de dados

A base de dados escolhida é fornecida pelo *Mozilla Common Voice*. Uma iniciativa de código aberto que tem como objetivo construir e disponibilizar uma base de dados rica em idiomas e variedade de áudios. A base é construída através da contribuição do público e atualmente possui dados em 76 idiomas (MOZILLA, 2021).

Os arquivos são disponibilizados no formato *MP3,* além dos arquivos de áudio existe também arquivos *CSV* com as informações de cada áudio, como o identificador do emissor, e o texto dito pela pessoa. Para este trabalho utilizaremos apenas a base no idioma português que possui o total de 94.265 amostras.

Além dos arquivos de áudio, são fornecidos também arquivos CSV (*comma separated values*) com dados adicionais sobre os áudios, esses arquivos foram carregados em uma base de dados *PostgreSQL*. Neste trabalho, utilizaremos três desses dados, são eles, o identificador único da pessoa que gerou o áudio, o nome do arquivo final e a descrição do texto.

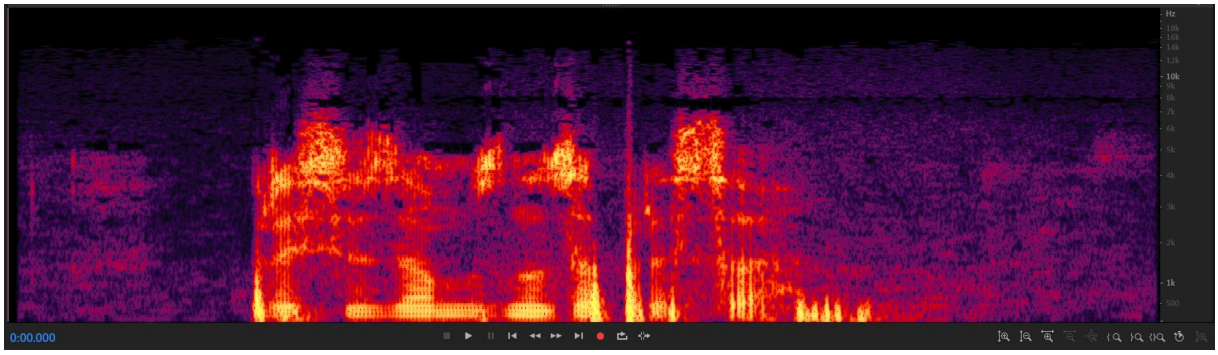
Para obter maior precisão na análise, alguns parâmetros foram definidos na escolha dos áudios que são utilizados. Pôde-se observar que existe uma variação na velocidade de fala, dessa forma, um áudio de menor duração pode ter mais texto que um áudio de maior duração. Por conta disso, foram realizadas consultas e selecionados os áudios em que seu texto possua mais de 70 caracteres e que cada pessoa tenha pelo menos 10 amostras.

Após a aplicação desses filtros a base de dados válida possui 930 amostras geradas por 42 pessoas.

4.2 Tratamento dos sinais de áudio

Cada sinal de áudio foi analisado utilizando o *software Adobe Audition*. Nele vemos duas representações do áudio, a primeira em seu formato de onda, e a segunda seu espectrograma. Neste último podemos perceber visualmente melhor os ruídos presentes no sinal, com é possível ver na Figura 10.

Figura 10 – Espectrograma

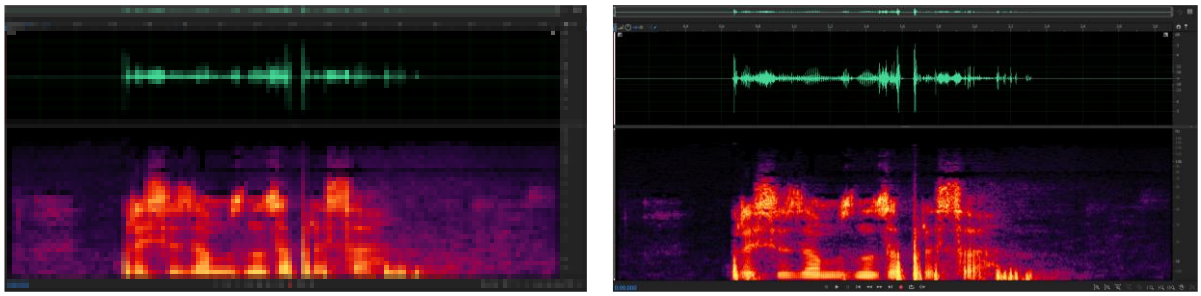


Fonte: Autoria própria (2021)

O *software* possui sua própria ferramenta de redução de ruídos, primeiramente apresentamos um corte do áudio com exemplo do ruído e após realização da Transformada Rápida de Fourier para a identificação dessas frequências, elas são minimizadas por todo o sinal. Além dessa opção, o software permite que os ruídos sejam minimizados manualmente em seu espectrograma, sendo utilizado para minimizar picos de ruídos.

Na Figura 11 é possível ver um exemplo do sinal de áudio e seu espectrograma antes (11.a) e depois (11.b) da redução de ruídos no *Adobe Audition*.

Figura 11 – Redução de ruídos



(4.a) Sinal antes da redução de ruídos (4.b) Sinal depois da redução de ruídos

Fonte: Autoria própria (2021)

Após a redução de ruídos, os intervalos de silêncio presentes nos sinais foram removidos.

4.3 Extração de características

Os arquivos fornecidos estão em formato MP3 (MPEG *Layer* 3), porém a biblioteca utilizada para a extração de características não permite este formato.

Dessa forma, o primeiro passo foi converter para o formato permitido WAV (*Waveform Audio File Format*). Para realizar essa conversão foi utilizado o pacote *pydub*.

O algoritmo da extração de características pode ser visto na Figura 12.

Figura 12 – Algoritmo extração de características

```
def getCaracteristicas(mfcc):
    caracteristicas = [mean(mfcc), std(mfcc), min(mfcc), max(mfcc)]
    return caracteristicas

def geraVetorMfcc(arquivo):
    y, sr = librosa.load(arquivo.path)
    mfcc = librosa.feature.mfcc(y=y, sr=sr, n_fft=n_fft, hop_length=hop_length,
n_mfcc=n_mfcc)
    vet = []
    for i in range(n_mfcc):
        caracteristicas = getCaracteristicas(mfcc[i])
        vet = concatenate(vet, caracteristicas)
    return vet

def gerarBase():
    arquivos = buscaArquivos()
    for arquivo in arquivos:
        escreveCsv(geraVetorMfcc(arquivo))
```

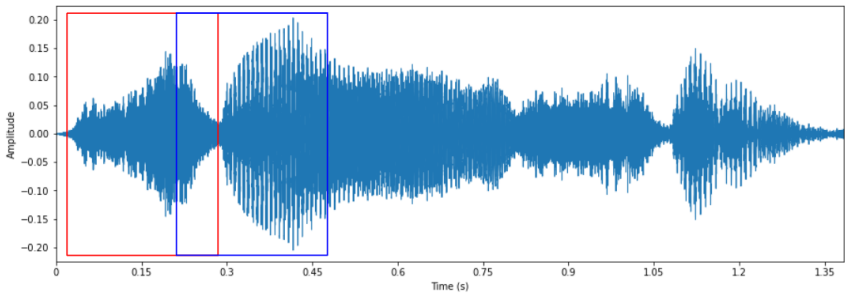
Fonte: Autoria própria (2021)

A função *buscaArquivos* é a responsável por devolver um vetor de objetos com os dados de cada arquivo válido. Esses dados são o identificador único da pessoa, nome e caminho em que o arquivo se encontra.

Para cada arquivo é então chamada a função *geraVetorMfcc,* onde o arquivo será carregado e suas características extraídas. Para a extração de características dos áudios foi utilizado o pacote *python librosa* na versão 0.8.1 (MCFEE et al., 2015). Trata-se de um conjunto de ferramentas que abstrai as operações realizadas para o processamento dos sinais. Dessa forma, apenas três parâmetros são usados para configurar a extração do MFCC, são eles o tamanho do janelamento (*n_fft*), o tamanho do salto entre as janelas (*hop_length*) e o número de filtros *Mel* (*n_mfcc*).

O salto entre as janelas trata do quanto uma janela irá sobrepor a anterior (Figura 13). Isso é recomendado para evitar perdas de dados. O tamanho do salto é configurado pelo parâmetro *hop_length.*

Figura 13 – Sinal com janelas sobrepostas



Fonte: Autoria própria (2021)

Inicialmente o arquivo de áudio é carregado e obtemos duas variáveis (*y, sr*) onde y é um vetor de números complexos representando a série temporal do sinal e *sr* a taxa de amostragem daquele sinal.

Para cada sinal de áudio será gerado n vetores, sendo n o número de filtros *Mel*. Como os vetores podem ter tamanhos distintos em áudios distintos, usá-los em algoritmos de aprendizagem de máquina poderia ser complicado ou mesmo inviável, uma vez que as tuplas não teriam a mesma quantidade de atributos. Para contornar isso e normalizar os valores em vetores de tamanho fixo, no lugar de usar os valores dos vetores, para cada vetor foram utilizados quatro valores estatísticos básicos sobre as características: valor médio, valor mediano, valor mínimo e valor máximo.

A implementação dessa extração pode ser vista na função getCaracteristicas. Essas características foram escolhidas por serem parâmetros comuns. O conjunto desses valores é então escrito em um novo arquivo CSV e este será usado para o treinamento das redes.

A rede neural escolhida para teste foi a *MLP.* Os testes foram realizados em uma Máquina de Comitê, composta por especialistas *MLP.* Foram implementadas utilizando o pacote *Python scikit-learn* versão 0.24.2, que é um conjunto de ferramentas de código aberto para *machine learning*. (Pedregosa et al., 2011)

4.4 Treinamento

Os algoritmos de treinamento foram selecionados como a intenção de realizar uma comparação entre eles.

O MLP foi implementado utilizando três camadas ocultas com 200, 150 e 100 neurônios respectivamente, com no máximo 250 iterações, para a tolerância de erro foi mantida o valor padrão do pacote *scikit-learn* de 0,0001. 30% da base de dados

foi utilizado para testes.

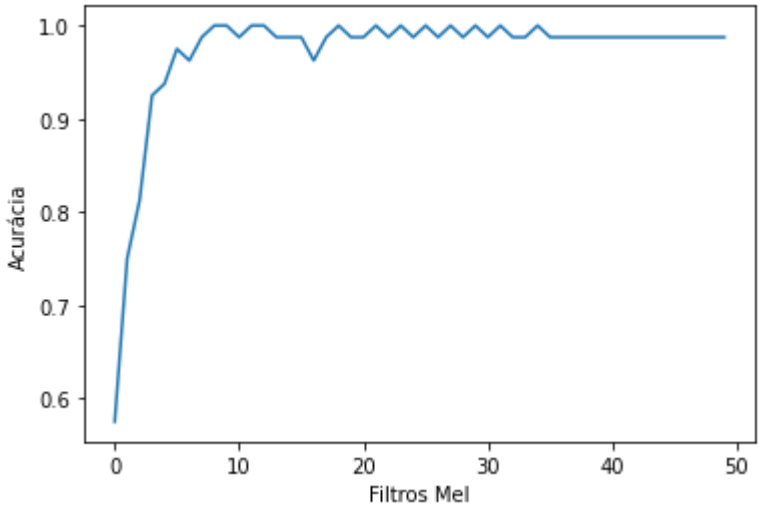
A máquina de comitê foi desenvolvida utilizando 20 especialistas, sendo cada um deles um *Perceptron* de Múltiplas Camadas.

4.5 Experimentos

No Gráfico 1 é apresentado o resultado do experimento 1, neste teste foi utilizado 10 amostras por pessoas, o janelamento de tamanho 2048 e salto de 512. O eixo das abscissas representa a variação dos filtros *Mel* e o eixo das ordenadas a acurácia alcançada.

Este experimento busca verificar a qualidade das classificações quando alterado o número de filtros Mel.

Gráfico 1 – Variação da acurácia em relação ao filtro *Mel*

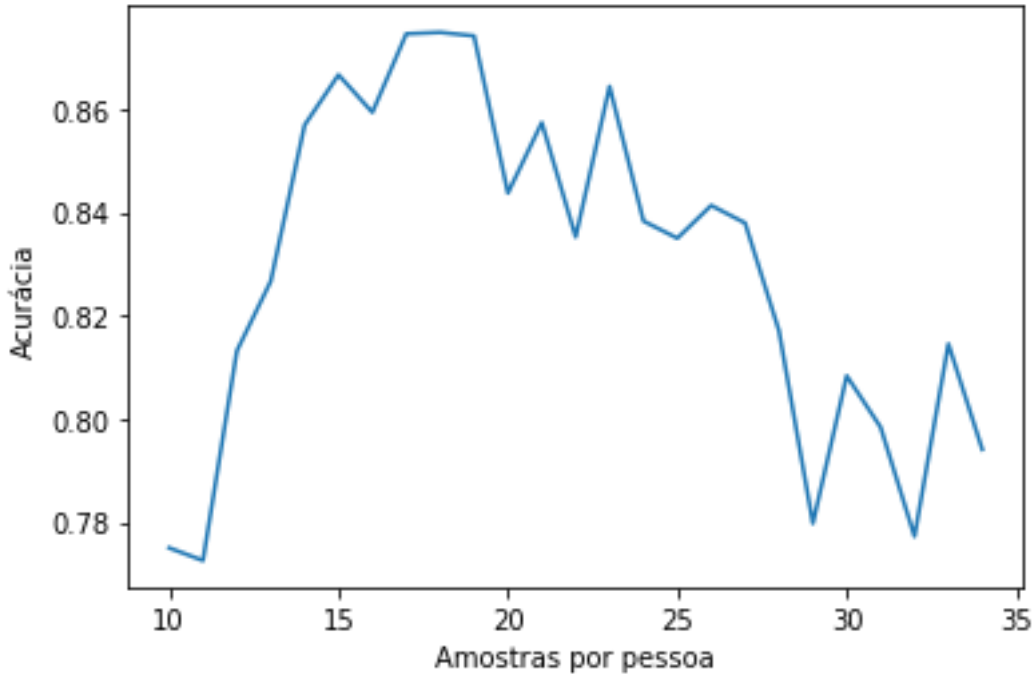


Fonte: Autoria própria (2021)

Como é possível ver na Figura 6, para o filtro 1 a frequência central é 100Hz, dessa forma temos um intervalo pequeno das frequências que serão usadas como treinamento. Para o treinamento realizado com mais de 10 filtros onde a frequência central é maior que 1000Hz um maior escopo e uma melhor acurácia. É possível perceber também que para valores superiores à 10, temos uma estabilização dos resultados, de modo que passa a não valer mais a pena adicionar mais filtros, uma vez que o custo de treinamento contina a crescer e o benefício do aumento da acurácia não.

O Gráfico 2 foi gerado utilizando 2 filtros Mel

Gráfico 2 – Acurácia com número de filtros igual a 2



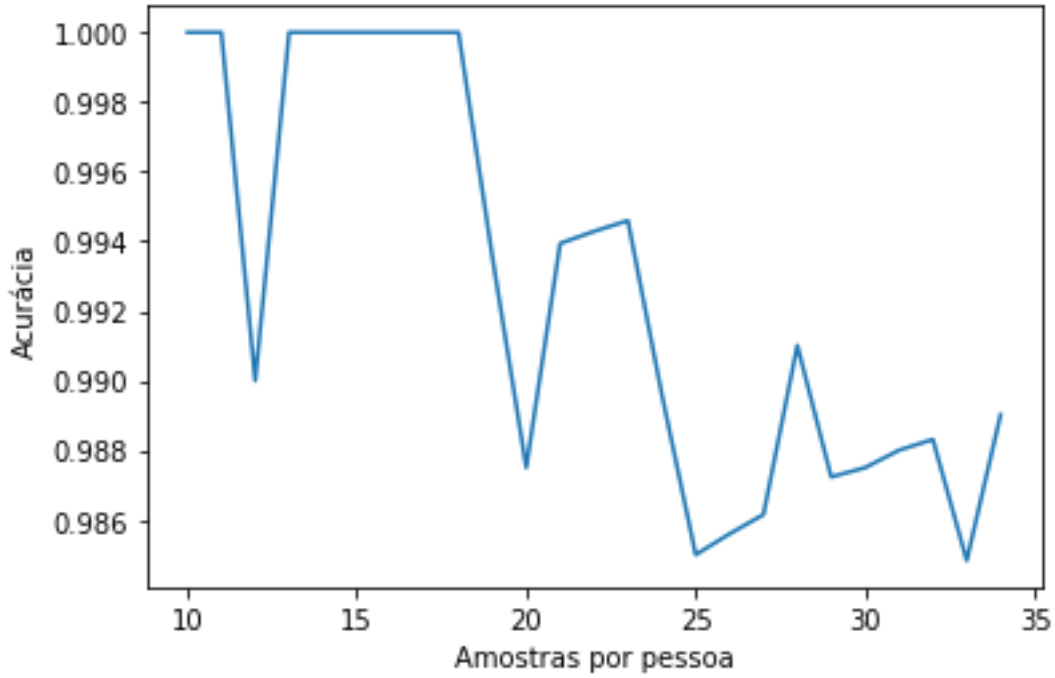
Fonte: Autoria própria (2021)

É possível notar a melhora dos resultados conforme o aumento do número de amostras por pessoas.

Foi realizado mais um experimento, dessa vez fixando o número de filtros *Mel* alterando o número de amostras por pessoas. O resultado pode ser visto no Gráfico 3. Para este o número de filtros Mel foi 25, o tamanho das janelas continuou com 2048 e o salto com 512.

Neste experimento, busca observar o comportamento dos resultados quando estes foram gerados em um cenário onde todas as variáveis estão calibradas corretamente.

Gráfico 3 – Variação da acurácia em relação ao número de amostras



Fonte: Autoria própria (2021)

O gráfico mostra os valores obtidos em 25 testes, é possível notar que para todos a acurácia obtida foi superior a 0.9. A pequena oscilação apresentada é meramente resultado das características randômicas do treinamento.

O código implementado para a realização deste trabalho está disponível em um repositório Git (DEUS, 2021).

6. CONSIDERAÇÕES FINAIS

Este trabalho tem como objetivo geral extrair as características do sinal de voz e identificar o emissor de um áudio, caso ele tenha amostras de sua voz cadastradas em uma base de dados.

Ao final do método MFCC obtemos as características dos sinais de voz. Observando a acurácia dos experimentos realizados, notamos que o sistema foi capaz de realizar a identificação de maneira eficiente em todos os cenários em que as variáveis foram devidamente calibradas. Porém é possível notar algumas limitações para a aplicação em cenários reais.

Inicialmente a ideia surgiu do problema que há nas penitenciárias, onde seus detentos têm acesso à celulares e se comunicam com o exterior através dele. Para a implementação do sistema em uma penitenciária exigiria que todos os detentos fossem catalogados, além de que uma estrutura para o armazenamento dos dados.

Além disso, também é possível usar softwares que distorcem a voz original, de forma que o algoritmo para a extração de características implementado não funcionaria.

Outro problema é que com o avanço da tecnologia, os chamados *Deepfakes* são comuns, ficando cada vez mais difícil identificar uma voz ou imagem falsa. Dessa forma, as amostras de voz que mimetizam emissores específicos podem ser geradas de modo que os resultados obtidos pela solução apresentadas nesse trabalho fossem equivocadas.

Para a continuação desta pesquisa, sugere-se:

- Desenvolver uma maneira de identificar padrões de fala e não apenas as características do sinal de voz, para que a identificação do emissor seja possível mesmo que a voz esteja distorcida.
- Desenvolver uma forma de identificar vozes geradas por uma inteligência artificial.
- Desenvolver maneiras alternativas de extração de características da voz.

REFERÊNCIAS

DAVIS, Steven B.; MERMELSTEIN, Paul. **Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences**. Readings In Speech Recognition, [S.L.], p. 65-74, 1980. Elsevier. http://dx.doi.org/10.1016/b978-0-08-051584-7.50010-3.

DEUS, Maria Regina. **TCC**. 2021. Disponível em: https://github.com/MariaRegina/TCC. Acesso em: 06 dez. 2021.

DIAS, Flávio**. Mais de 1.400 celulares apreendidos em presídios de MS serão doados para alunos da rede pública**. Disponível em: https://g1.globo.com/ms/mato-grosso-do-sul/noticia/2021/05/10/mais-de-1400-celulares-apreendidos-em-presidios-de-ms-serao-doados-para-alunos-da-rede-publica.ghtml. Acesso em: 09 nov. 2021.

G1; **Preso tiktoker: detento é transferido para presídio de segurança máxima após compartilhar rotina na cadeia em rede social**. Disponível em: https://g1.globo.com/rj/norte-fluminense/noticia/2021/11/02/preso-tiktoker-detento-e-transferido-para-presidio-de-seguranca-maxima-apos-compartilhar-rotina-na-cadeia-em-rede-social.ghtml. Acesso em: 09 nov. 2021.

HANECHE, Houria; BOUDRAA, Bachir; OUAHABI, Abdeldjalil. **A new way to enhance speech signal based on compressed sensing. Measurement,** [S.L.], v. 151, p. 107117, fev. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.measurement.2019.107117.

HASSAN, Bilal; AHMED, Ramsha; LI, Bo; HASSAN, Omar; HASSAN, Taimur. **Autonomous Framework for Person Identification by Analyzing Vocal Sounds and Speech Patterns**. 2019 5Th International Conference On Control, Automation And Robotics (Iccar), [S.L.], v. 5, p. 649-653, abr. 2019. IEEE. http://dx.doi.org/10.1109/iccar.2019.8813463.

HAYKIN, Simon. **Redes Neurais: princípios e práticas**. 2. ed. Porto Alegre: Bookman, 2001. 900 p.

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

MACHADO, Mateus Lichfett. **Implementação de um sistema de reconhecimento automático de voz utilizando as técnicas MFCC e quantização vetorial com atributos dinâmicos, de normalização e detecção de voz ativa**. 2016. 149 f. Dissertação (Mestrado) - Curso de Engenharia Mecânica, Univesidade Federal de Uberlândia, Uberlândia, 2016.

MARTINS, R.; YNOGUTI, C. **Normalização do locutor em sistemas de reconhecimentode fala para usuários crianças**. In: XIII Simpósio Brasileiro Sobre Fatores Humanos em Sistemas Computacionais. [S.l.: s.n.], 2014

MCFEE, BRIAN, RAFFEL, LIANG, ELLIS, MCVICAR, BATTENBERG, NIETO.

librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference, pp. 18-25. 2015.

MÉNDEZ, Antonio M. et al. **Acustica Arquitectonica**. Buenos Aires: Testone Hnos, 1994. 119 p.

MOLLA, M.K.I.; HIROSE, K.. **On the effectiveness of MFCCs and their statistical distribution properties in speaker identification**. 2004 IEEE Symposium On Virtual Environments, Human-Computer Interfaces And Measurement Systems, 2004. (Vcims)., [S.L.], v. 0, n. 0, p. 0-0, jun. 2004. IEEE. http://dx.doi.org/10.1109/vecims.2004.1397204

MOZILLA. **Mozilla Common Voice**. Disponível em: https://commonvoice.mozilla.org/pt. Acesso em: 07 nov. 2021.

OPPENHEIM, Alan V.; SCHAFER, Ronald W.. **Processamento em tempo discreto de sinais**. 3. ed. São Paulo: Pearson Education do Brasil, 2012. 688 p.

PAULA FILHO, Wilson de Pádua. **Multimídia: conceito e aplicações**. [S.l.]: Ltc, 2000. 321 p.

Pedregosa et al.. **Scikit-learn: Machine Learning in Python**, JMLR 12, pp. 2825-2830, 2011.

THIAGO, Ernani Rodrigues de São. **Reconhecimento de Voz utilizando extração de Coeficientes Mel-Cepstrais e Redes Neurais Artificiais**. 2017. 76 f. TCC (Graduação) - Curso de Engenharia de Telecomunicações, Instituto Federal de Santa Catarina, São José, 2017

TIRUMALA, Sreenivas Sremath; SHAHAMIRI, Seyed Reza; GARHWAL, Abhimanyu Singh; WANG, Ruili. **Speaker identification features extraction methods: a systematic review.** Expert Systems With Applications, [S.L.], v. 90, p. 250-271, dez. 2017. Elsevier BV. http://dx.doi.org/10.1016/j.eswa.2017.08.015.

VERLADO, Velario. **How do we extract audio features?** 2020. Disponível em: https://github.com/musikalkemist/AudioSignalProcessingForML/blob/master/6-%20How%20to%20extract%20audio%20features/How%20to%20extract%20audio%20features%20.pdf. Acesso em: 08 jun. 2021.

WAZLAWICK, Raul Sidnei. **Metodologia de Pesquisa para Ciência da Computação**. Rio de Janeiro: Elsevier, 2009. 124 p

ZHOU, Jianqin; CHEN, Ping. **Generalized Discrete Cosine Transform**. 2009 Pacific-Asia Conference On Circuits, Communications And Systems, [S.L.], v. 41, n. 1, p. 135-147, maio 2009. IEEE. http://dx.doi.org/10.1109/paccs.2009.62.

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
GABINETE DO REITOR

Av. Universitária, 1069 ● Setor Universitário
Caixa Postal 86 ● CEP 74605-010
Goiânia ● Goiás ● Brasil
Fone: (62) 3946.1000
www.pucgoias.edu.br ● reitoria@pucgoias.edu.br

RESOLUÇÃO n° 038/2020 – CEPE

ANEXO I

APÊNDICE ao TCC

Termo de autorização de publicação de produção acadêmica

O(A) estudante MARIA REGINA SANTOS DE DEUS do Curso de Ciência da Computação, matrícula 2017.1.0028.0042-4, telefone: 62 99701-8186 e-mail santosdedeusm@gmail.com, na qualidade de titular dos direitos autorais, em consonância com a Lei nº 9.610/98 (Lei dos Direitos do autor), autoriza a Pontifícia Universidade Católica de Goiás (PUC Goiás) a disponibilizar o Trabalho de Conclusão de Curso intitulado Análise de Audio de Voz Para Identificação do Emissor Utilizando Técnicas de Processamento de Sinais e Redes Neurais Artificiais, gratuitamente, sem ressarcimento dos direitos autorais, por 5 (cinco) anos, conforme permissões do documento, em meio eletrônico, na rede mundial de computadores, no formato especificado (Texto (PDF); Imagem (GIF ou JPEG); Som (WAVE, MPEG, AIFF, SND); Vídeo (MPEG, MWV, AVI, QT); outros, específicos da área; para fins de leitura e/ou impressão pela internet, a título de divulgação da produção científica gerada nos cursos de graduação da PUC Goiás.

Goiânia, 14 de dezembro de 2021.

Assinatura do(s) autor(es): Maria Regina Santos de Deus

Nome completo do autor: MARIA REGINA SANTOS DE DEUS

Assinatura do professor-orientador: Max Gontijo

Nome completo do professor-orientador: Me. Max Gontijo de Oliveira