UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de São José do Rio Preto

Graduação em Ciência da Computação

Igor Yoshimitsu Ide

Verificação Espectral De Locutores Na Modalidade *Text-Dependent*

São José do Rio Preto
2021

Igor Yoshimitsu Ide

Verificação Espectral De Locutores Na Modalidade *Text-Dependent*

Orientador: Prof Dr Rodrigo Capobianco Guido

.

Banca Examinadora:

Prof Dr Rodrigo Capobianco Guido
Prof Dr Aleardo Manacero Júnior
Prof Dr Rodolfo Ipólito Meneguette

São José do Rio Preto
2021

I19v

Ide, Igor Yoshimitsu
Verificação espectral de locutores na modalidade text-dependent / Igor Yoshimitsu Ide. -- São José do Rio Preto, 2022
30 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (Unesp), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientador: Rodrigo Capobianco Guido

1. Inteligência artificial. 2. Taxa de Cruzamentos por zero. 3. Processamento de linguagem natural. 4. Voz. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Resumo

O desafio para o desenvolvimento de sistemas de verificação de locutores é extrair representações de fala robustas, considerando os mais diversos fatores que podem afetar a geração de sinais de fala, tais como a presença de ruído ambiente ou o estado de saúde do locutor. Desse modo, considerando uma parcela da base de dados TIMIT adaptada para fins de verificação de locutores, o autor deste trabalho desenvolveu um sistema capaz de verificar indivíduos pela voz com acurácia plena, considerando um classificador linear e assumindo dez amostras de voz contaminadas por ruído para cada um dos 40 locutores matriculados. A etapa de extração de características foi desenvolvida com base nos conceitos de energia e taxa de cruzamentos por zero, associados ao princípio de funcionamento do ouvido humano modelado pela escala Bark e à Transformada Discreta de Fourier. Em vista dos resultados, foi possível concluir que as características experimentadas em associação com os conceitos e ferramentas matemáticas utilizados permitiram levar a bom termo a proposta inicial.

Abstract

The main challenge to develop speaker verification systems is to extract robust speech representations, considering the diversity of issues which may affect those signals, such as the presence of ambient noise and possible speaker's health problems. Thus, by using a branch of TIMIT dataset adapted to speaker verification, the author of this work developed a full-accuracy system to verify speakers, assuming a set of ten voices per speaker, contamined with different levels of noise, for the fourty speakers enrolled. The feature extraction stage was developed based on the concepts of energy and zero crossing rate, in addition to the Bark scale and the Discrete Fourier Transform. In view of the results, it is possible to state that the features adopted in association with the concepts and mathematical tools used allowed for the initial proposal to be successfully completed.

Lista de Acrônimos

MFCC *Escala Mel de frequência*

UBM *Universal Background Model*

GMM *Gaussian Mixture Mode*

CDBN *Convolutional Deep Belief Network*

HMM *Modelos Ocultos de Markov*

LPC *Codificação Preditiva Linear*

OLS *Ordinary Least Square ou do inglês Mínimos Quadrados Ordinários*

TIMIT *Base de dados para reconhecimento de fala*

MIT *Massachusetts Institute of Technology*

ELSDSR *English Language Speech Database for Apeaker Recognition*

WAVE *WAVE Form audio format*

RIFF *Resource Interchange File Format*

ZCR *Taxa de Cruzamento de Zero ou do inglês Zero Crossing Rate)*

Sumário

Capítulo 1

Introdução

Técnicas de identificação de indivíduos por voz têm se tornado cada vez mais comuns, principalmente em aplicações de controle de acesso, substituindo tradicionais sistemas baseados em senhas [1]. O potencial crescente dos computadores, assim como das *graphical processing units* (GPUs), constitui um forte fator motivador para que os sistemas de reconhecimento de padrões, inclusive aqueles baseados em voz, estejam presentes em aplicações diversas [2][3]. Notavelmente, o referido tema tem merecido considerável atenção da comunidade científica, conforme é possível notar mediante uma busca realizada na base científica do *Web of Science*.

Claramente, estratégias específicas das áreas de Inteligência Artificial e Processamento de Sinais têm oferecido um grande número de possibilidades para a implementação de sistemas de reconhecimento de padrões em sinais de voz. Técnicas do tipo *deep learning* [1], em associação com extratores de características que permitam a obtenção de informações no domínio conjunto tempo-frequência [1][2][3], têm possibilitado resultados promissores.

Diante do exposto, este trabalho concentra-se no estudo de conceitos, no projeto e na implementação em linguagem C/C++ de uma estratégia computacional para verificação de locutores, na modalidade *text-dependent*, assumindo a existência de dez níveis diferentes de ruídos nos sinais de voz, oriundos da base de dados TIMIT, de cada um dos 40 locutores matriculados no experimento.

Inicialmente, para a realização deste trabalho, foi realizado um levantamento bibliográfico dos trabalhos correlatos e, a partir deles, das técnicas utilizadas como base para o desenvolvimento, conforme registrado no Capítulo 2. Prosseguindo, os experimentos foram realizados, de acordo com a descrição constante no Capítulo 3, e os testes foram executados, colhendo-se os resultados registrados no Capítulo 4 para, então, permitir a obtenção de interessantes conclusões disponíveis no Capítulo 5, o qual antecede as referências bibliográficas.

Capítulo 2

Revisão Bibliográfica

Neste capitulo são apresentados trabalhos correlatos ao tema em questão, assim como conceitos pertinentes que serviram como base para o desenvolvimento da metodologia utilizada, preparando, assim, o leitor para melhor compreender o conteúdo dos capítulos vindouros.

2.1 Trabalhos Relacionados

A área de verificação de locutores vem sendo bastante estudada, devido ao avanço tecnológico alcançado nos últimos anos, permitindo o efetivo uso do sinais de voz nos sistemas de autenticação biométrica. Assim, os parágrafos seguintes reúnem os principais trabalhos da área que serviram de base para o desenvolvimento desta monografia.

A estratégia usada para verificação de locutores constante no artigo [4] é similar àquela usada neste trabalho, tendo como maior diferencial a modalidade de reconhecimento que não é *text-dependent*. Nela, são utilizados os algoritmos A3 e B3 também usados neste trabalho; em contrapartida o classificador utilizado foi baseado em distâncias euclidianas. Percebeu-se que, utilizando somente o método B3, a acurácia não superou 52,5%, mesmo alterando a resolução do algoritmos de extração de características. Com o a adição do método A3, entretanto, a acurácia alcançou 90%.

Assim como o artigo citado anteriormente, a proposta do artigo [5] possui como diferencial o uso dos modelos ocultos de Markov simplificados. Especificamente, foram utilizados 200 sinais acústicos de 40 locutores e, para o classificador, além da distância euclidiana, foi usada a distribuição Gaussiana. Pode-se perceber que a referida distribuição probabilística constituiu o método mais promissor. Usando a distância euclidiana e os métodos A3 e B3, obteve-se uma acurácia de 40%. Adicionando a derivada de primeira ordem dos vetores de características obtidos com A3 e B3 resultou em 25% de acurácia. Refinando ainda mais a resolução dos vetores de características, alcançou-se 39.3% de acurácia e, por fim, além dos métodos A3 e B3 foram acrescentadas as derivadas de primeira e segunda ordem em associação com a distância euclidiana, produzindo uma acurácia de 61.8% .

Outro trabalho interessante é aquele documentado na referência [6], que foca na remoção de vulnerabilidades no processo de verificação de locutores. Diferentemente dos trabalhos citados anteriormente, foi utilizado na classificação a escala Mel de frequência (MFCC) e para extração de chaves binárias o *Universal Background Model* (UBM) e o *Gaussian Mixture Model* (GMM), implementados com base na ferramenta Matlab. Uma base de dados própria foi utilizada, sendo conhecida como CEFALA-1. Foram realizadas 43264 simulações, analisando os índices de falsa aceitação e rejeição, afastando a possibilidade de uma falsa aceitação, mantendo um índice de falsa rejeição abaixo dos 5%.

No trabalho [7], assim como no anterior, foi utilizado o Matlab, mas a diferença é que, desta vez, uma interface para cálculo dos MFCCs que serão utilizadas para treinar os modelos ocultos de Markov (HMM) foi usada. Para a verificação dos locutores, foram utilizadas curvas de densidade de probabilidade com as verossimilhanças dos MFCCs de três grupos: um do locutor de interesse, outro de locutores desconhecidos e outro de um locutor com alto grau de parentesco. Como resultado, os locutores desconhecidos não foram reconhecidos, mas o com parentesco foi identificado algumas vezes.

Outro trabalho interessante é aquele documentado na referência [8], que utiliza aprendizado de máquina para identificação de locutores. Nele, são utilizadas as bases de dados TIMIT (sem ruído) e MIT (com ruído) para treinamento não-supervisionado, valendo-se do classificador *Convolutional Deep Belief Network* (CDBN). Dois foram os testes realizados: se esse modelo consegue aprender características importantes para identificação de um locutor de uma base que apresenta ruído e se essas características aprendidas são úteis em outra base, implicando uma transferência de aprendizado. Para os testes foram utilizados 168 locutores com 10 sinais acústicos cada, totalizando 1680 arquivos de áudio. Como resultado, foi possível inferir que os dados aprendidos pela CDBN podem ser utilizados em transferência de aprendizado.

Seguindo o caminho do artigo anterior, o trabalho [9] também utiliza aprendizado de máquina mas, desta vez, com o algoritmo GMM em conjunto com o UBM. Nele, também foi utilizado uma base de dados diferente, a ELSDSR, que possui 22 locutores com dois sinais acústicos para cada um. Para extração de características, foi utilizado o MFCC. Ao final, concluído que, com a base de dados ELSDSR, houve uma acurácia de 99.59%.

Um artigo com um caminho um pouco diferente é o [10], que utiliza codificação preditiva linear (LPC) e *Ordinary Least Square* (OLS) para verificação de locutores na análise forense. Para os testes foram utilizados 26 sinais acústicos. Foi possível notar que, com a exclusão de alguns casos, a acurácia alcançou 100%.

Diferentemente dos trabalhos mencionados, a ideia-chave utilizada neste trabalho é acoplar os métodos A1, A3, B1 e B3 com a caracterização espectral oriunda da escala Bark, a qual aproxima o funcionamento do ouvido humano, assim como a análise de Fourier. Contou-se, como base de dados, com sinais oriundos da base TIMIT contaminados com diferentes níveis de ruído, permitindo, assim, desenvolver

uma estratégia para verificação de locutores robusta a um conjunto de degradações modestas e intensas nos sinais de cada um dos 40 locutores matriculados no sistema desenvolvido.

2.2 A Voz Humana e a Verificação de Locutores

A voz é um elemento nitidamente importante para a comunicação dos seres humanos, mesmo porque o ser humano vive em sociedade e precisa se comunicar. Assim, ao longo dos tempos, os seres humanos desenvolveram um padrão de aparelho vocal que é dividido em três partes [11]:

- o sistema respiratório (pulmão)
- a laringe (pregas vocais)
- os ressonadores (pavilhão faringo-bucal)

Essencialmente, a voz é produzida quando ar é expelido dos pulmões, pressionados pelo diafragma, e atravessa a laringe, onde as pregas vocais vibram aleatoriamente ou quase periodicamente, produzindo os diversos sons que constituem a fala.

Notavelmente, para locuções específicas, o período de vibração das pregas vocais e o formato da estrutura vocal e nasal são diferentes para cada indivíduo, criando diferenças na voz que permitem identificar os locutores. Para tal identificação, ou mesmo verificação como é o caso do presente trabalho, é interessante observar que a anatomia dos tratos vocal e nasal de cada locutor caracteriza a sua voz como resultado de frequências geradas por uma fonte de excitação, ou seja, o sinal pulmonar que atravessou as pregas vocais, e de um sistema de equalização, isto é, o conjunto de ressonadores que ressalta ou atenua frequências específicas conhecidas como formantes e anti-formantes, respectivamente.

Nesse contexto, nota-se que o fonema é a menor unidade sonora de uma língua; por exemplo a palavra "hoje" tem quatro letras mas apenas três fonemas. De fato, a variação fonética entre os falantes é frequentemente um fator negativo quando se lida com a acurácia em sistemas de identificação de locutores. Alterações em um ambiente acústico, fatores técnicos de captura de som (por exemplo, microfone usado) e variação na voz da pessoa por razões biológicas (doença, envelhecimento, entre outros) ou emocionais, além da velocidade da fala e do timbre representam ocorrências indesejáveis que tornam a verificação de locutores um problema consideravelmente trabalhoso [14] [15].

Assim, de modo geral, a verificação biométrica nada mais é do que os conjuntos de passos e regras a serem seguidas para, mediante atributos específicos tal como a voz, individualizar uma pessoa perante a sociedade [12]. Nesse sentido, para que uma característica do ser humano possa ser utilizada como forma de reconhecimento ela deve existir em todas as pessoas (universalidade), deve ser distinta (singularidade), não

pode desaparecer ou mudar com o tempo (invariabilidade ou permanência), pode ser medida (mensurabilidade) e, além disso, considerar se fatores externos afetam a precisão da identificação (desempenho), se os usuários aceitam o sistema (aceitabilidade) e se há técnicas de proteção no sentido de que seja possível confiar no sistema [13].

Rotineiramente, sinais de voz são analisados por *frames* [14], ou seja, intervalos curtos de tempo em torno de 20 ou 30ms. Quando a análise é realizada de modo *off-line*, o formato wave de arquivos, criado pela Microsoft e IBM para armazenamento de sinais acústicos em computadores, é classicamente utilizado pelo fato de conter, além do cabeçalho, as amostras digitalizadas do sinal analógico correspondente [17][5], conforme a Figura 2.1. Em roxo, pode-se observar o *chunck* descritor "RIFF", em verde observa-se o sub-bloco "fmt" que descreve o formato da informação do som e, em laranja, consta o sub-bloco "data" que indica o tamanho e a informação bruta de interesse, ou seja, as amplitudes do sinal digitalizado.

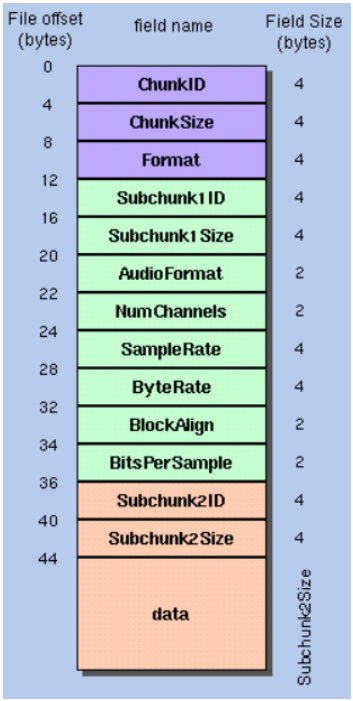

Figura 2.1: caracterização do formato wave. Imagem extraída de [17].

2.3 Pré-enfase

A pré-ênfase é um método simples de pré-processamento dos sinais de voz que busca diminuir o efeito a irradiação nos lábios nos sinais sob análise pois essas informações não contribuem, na extração de características, com tarefas de verificação de locutores [5]. Neste trabalho é utilizado o seguinte filtro passa-banda de primeira ordem para realizar a pré-ênfase, conforme descrito em [6]: $Y[i] = X[i] - 0.95 \cdot X[i-1]$, onde Y e X são os sinais pré-enfatizados e original, respectivamente.

2.4 Zero-crossing Rate (ZCR)

A taxa de cruzamento de zero [18] de um sinal acústico é a taxa de mudança de amplitudes positivas para negativas e vice-versa em um certo período. A ZCR de um sinal $x[\cdot]$ de tamanho N pode ser definida como:

$$\text{ZCR}(x[\cdot]) = \frac{1}{2}\sum_{i=0}^{N-2}\Big(sgn(x_i) - sgn(x_{i+1})\Big) \quad ,$$

onde *sgn*(x_i) é o sinal da amplitude x_i. Se $x_i \geq 0$ então sgn$(x_i) = 1$ e se $x_i \leq 0$ então sgn(x_i) = -1. Neste trabalho, os métodos B1 e B3, descritos em [18], foram usados para extração do ZCR. Particularmente descritos no próximo Capítulo, eles diferem apenas nos trechos dos sinais sob análise que serão inspecionados e como os sinais serão sub-divididos.

2.5 Energia

A energia está presente em toda ação, isso não é uma exceção para a fala. No sinal da fala a energia seria o trabalho do pulmão e do trato vocal do locutor ao passar do tempo para produzir som [3]. O sinal de energia sendo definido como:

$$\text{E(x[.])} = \sum_{i=0}^{M-1}(x_i)^2 \quad ,$$

onde M é o tamanho do sinal $x[\cdot]$. Para capturar essa energia, serão utilizados os métodos A1 e A3 descritos em [3], os quais estão mais detalhados no próximo Capítulo.

2.6 Transformada de Fourier (DFT)

Uma técnica muito usada para análise espectral é a DFT. Com ela pode-se calcular as amplitudes das diversas frequências que compõe um sinal. Particularmente, pode-se encontrar a resposta de frequência de um sistema a partir da resposta de impulso do sistema e vice-versa, permitindo conversões entre os domínios do tempo e da frequência [8][19]. A DFT de um sinal $x[\cdot]$ de tamanho N é dada por

$$\text{DFT}_k(x[\cdot]) = \sum_{i=0}^{N-1} x_n \cdot \Big(cos(\frac{2\pi kn}{N} - j \cdot sin(\frac{2\pi kn}{N})\Big) \quad ,$$

for $k = 0, 1, 2, ...$ e $j = \sqrt{-1}$. Particularmente importante para o caso deste trabalho é o módulo da DFT, conforme mencionado no próximo Capítulo.

2.7 A Base de Dados TIMIT

Amplamente conhecida na área de processamento de sinais de voz, uma das ferramentas mais importantes para este trabalho é a própria base de dados utilizadas, isto

é, a base TIMIT. Desenvolvido em 1993, em conjunto com o Instituto de Tecnologia de Massachusetts (MIT), o *SRI International* (SRI) e a Texas Instruments (TI), contém um conjunto de 6300 sinais de voz de 630 locutores. Cada locutor pronuncia 10 frases, sendo que uma delas, comuns a todos eles, é a sentença selecionada para o desenvolvimento deste trabalho: "*She had your dark suit and greasy wash water all year*".

Capítulo 3

Metodologia

Neste Capítulo são apresentados e explicados todos os métodos e ferramentas utilizados para o desenvolvimento deste trabalho. Particularmente, para o sistema de verificação de locutores construído neste trabalho foi utilizada a sequência de passos exibida na Figura 3.1.

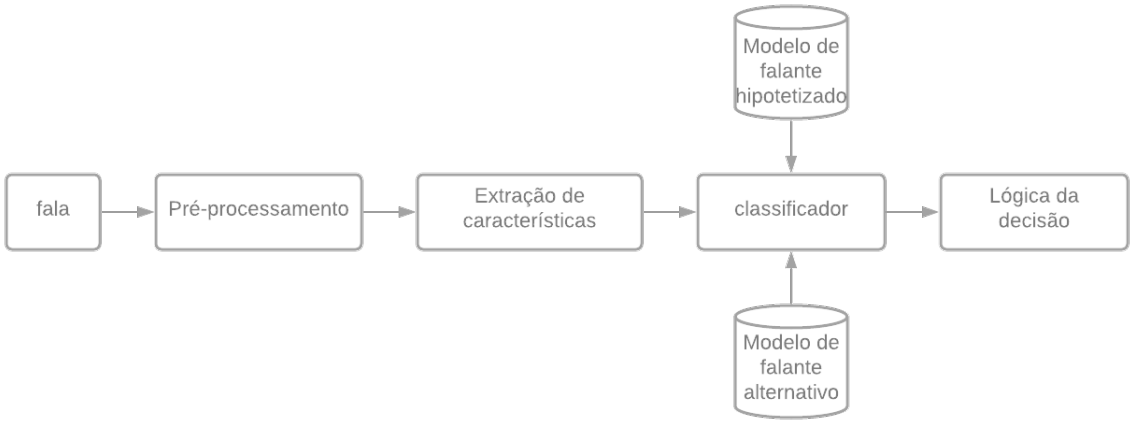

Figura 3.1: sistema de verificação de locutores adotado neste trabalho. Imagem de autoria própria.

Desse modo, os passos executados são os seguintes:

- PASSO 1: Obter os dados brutos do sinal de fala sob análise;
- PASSO 2: Aplicar o pré-processamento.
- PASSO 3: Extrair as características dos sinais de fala.
- PASSO 4: Obter o resultado da análise com o classificador utilizado para verificar o locutor em questão.
- PASSO 5: De acordo com uma lógica pré-estabelecida, uma decisão tomada.

3.1 PASSO 1: Obtenção dos dados brutos

Utilizando 40 sinais acústicos da base de dados TIMIT nos quais a frase "*She had your dark suit in greasy wash water all year*", sendo um para cada um dos 40 locutores selecionados, foram inseridos, para cada sinal, ruídos Gaussianos aleatórios nas seguintes proporções pré-definidas fazendo o uso do aplicativo Audacity: 2,5%; 5%; 7,5%; 10%; 12,5%; 15%; 17,5%; 20%; 22,5%. Assim, cada um dos 40 sinais originais originou outros nove sinais ruidosos, totalizando 10 sinais para cada um dos 40 locutores, implicando $10 \cdot 40 = 400$ sinais dos quais $9 \cdot 40 = 360$ são ruidosos.

3.2 PASSO 2: Pré-processamento

Um pré-processamento foi aplicado aos dados brutos de cada sinal de voz, conforme mencionado no Capítulo anterior. Particularmente, foi aplicado o filtro $Y_i = X_i - 0.95X_{i-1}$ para cada sinal original $X[\cdot]$, originando o respectivo sinal $Y[\cdot]$, visando atenuar os efeitos da irradiação labial do locutor em questão. Adicionalmente, a remoção da média das amplitudes de cada sinal filtrado foi removida visando anular a amplitude da frequência 0 Hz [1]. Para isso, foi calculada a média das amplitudes do sinal sob análise a qual, em seguida, foi subtraída de cada amplitude. Por fim, foi realizada uma normalização das amplitudes, visando anular os efeitos do ganho do microfone no momento em que as gravações foram realizadas. Tais procedimentos encontram-se descritos nos Algoritmos 1, 2 e 3.

Algorithm 1 Filtro passa-banda

Imput : $s[n]$. $n \geq 0$ ▷ Vetor de dados.
for $i := 0, 1, 2, \ldots Tamanho\_Vetor$ **do**
  $s[i] \leftarrow s[i] - 0.95 * s[i-1]$
end for
$Me \leftarrow Me/Tamanho\_Vetor$
for $i := 1, 2, \ldots Tamanho\_Vetor$ **do**
  $s[i] \leftarrow s[i] - Me$
end for

Algorithm 2 Remoção da média

Imput : $s[n]$. $n \geq 0$ ▷ Vetor de dados.
$Me \leftarrow 0$
for $i := 0, 1, 2, \ldots Tamanho\_Vetor$ **do**
  $Me \leftarrow Me + s[i]$
end for
$Me \leftarrow Me/Tamanho\_Vetor$
for $i := 1, 2, \ldots Tamanho\_Vetor$ **do** ▷ subtraindo a média
  $s[i] \leftarrow s[i] - Me$
end for

Algorithm 3 Normalização

```
Imput : s[n].    n ≥ 0                          ▷ Vetor de dados.
Ma ← s[0]                                       ▷ Maior valor.
if Ma < 0 then
    Ma ← Ma ∗ −1
end if
for i := 0, 1, 2, ... Tamanho_Vetor do
    if s[i] < 0 then
        if −s[i] > Ma then
            Ma ← −s[i]
        else if s[i] > Ma then
            Ma ← s[i]
        end if
    end if
end for
for i := 0, 1, 2, ... Tamanho_Vetor do
    s[i] ← s[i]/Ma
end for
```

3.3 PASSO 3: Extração das Características

No processo de extração de características, com base nos conceitos de energia de taxa de cruzamentos por zero, foram utilizados quatro algoritmos denominados como A1 e A3, conforme descrito em [3], além de B1 e B3, conforme consta em [20].

3.3.1 Algoritmo A1

O Algoritmo A1 concentra-se no agrupamento de segmentos do sinal de voz sob análise, gerando informações relevantes para a classificação. Ele utiliza uma janela retangular de comprimento L atravessando o vetor de modo que, para cada posição da janela, a energia do sinal normalizada é calculada[3]. No caso deste trabalho, foi definido que cada posicionamento subsequente se sobrepõe em dois terços das amostras anteriores, conforme ilustrado na Figura 3.2. Na imagem, é possível ver como as janelas $w[\cdot]$ de tamanho L englobam partes do sinal de entrada, produzindo o vetor de característica de acordo com os cálculos de energia destacados nos quadros azuis na parte inferior da imagem.

3.3.2 Algoritmo A3

Em contrapartida ao caso anterior, o Algoritmo A3 determina os comprimentos proporcionais do sinal sob análise que são necessários para atingir níveis pré-definidos de energia, conforme a Figura 3.3. A consequência direta desta abordagem é a caracterização de A3 como sendo ideal para inspecionar a constância em ação da entidade física responsável por gerar o sinal[3]. Particularmente, notam-se na Figura 3.3 as janelas $w[\cdot]$ percorrem o sinal de entrada de acordo com o tamanho L decidido para cada

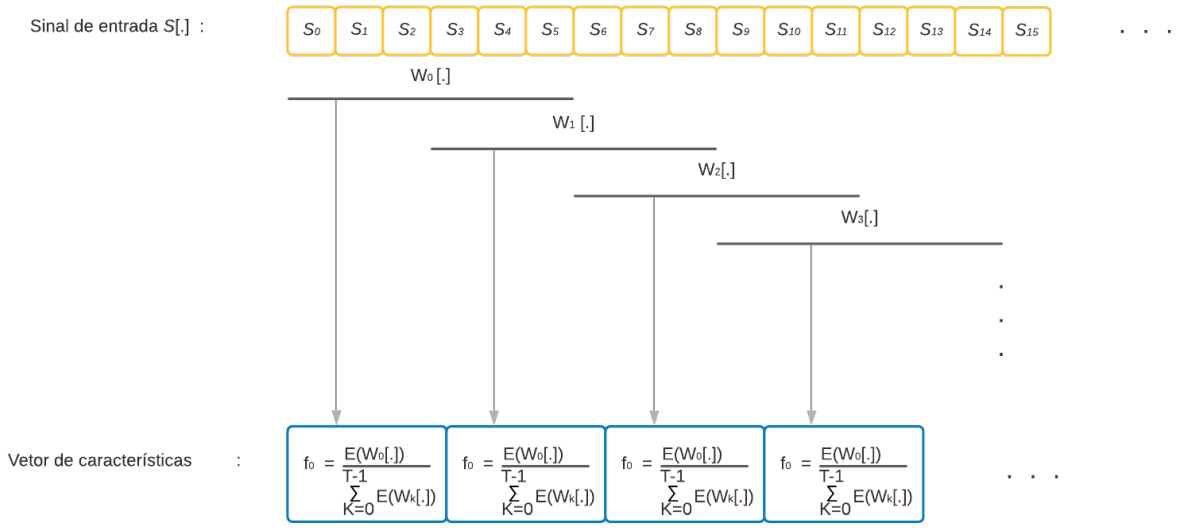

Figura 3.2: exemplo do algoritmo A1. Imagem extraída de [3].

janela, que cobre $i \cdot C\%$ do sinal, para i variando de 0 até $T-1$. Como foi decidido que $C = 4.9$, cada janela cobre $i \cdot 4.9\%$ da energia do sinal de entrada [3].

Algorithm 4 Algoritmo A1

Imput : $s[n]$. $n \geq 0$ ▷ Vetor de dados
Output : $f[m]$. $m \geq 0$ ▷ Vetor de características
$L \leftarrow 256$ ▷ Tamanho da janela
$V \leftarrow 66.67$ ▷ Dois terços de sobreposição entre janelas
$Tamanho\_Novo \leftarrow (((100 * Tamanho\_Vetor) - (L * V))/((100 - V) * L))$
$E \leftarrow 0$
for $k := 0, 1, 2, \ldots Tamanho\_Novo - 1$ **do**
    $f[k] \leftarrow 0$
    for $i := k * (((100 - V)/100) * L)) \ldots k * (((100 - V)/100) * L)) + L - 1$ **do**
        $f[k] \leftarrow f[k] + s[i]^2$
    end for
    $E \leftarrow E + f[k]$
end for
for $i := 0, 1, 2, \ldots Tamanho\_Novo$ **do**
    $f[k] \leftarrow f[k]/E$
end for

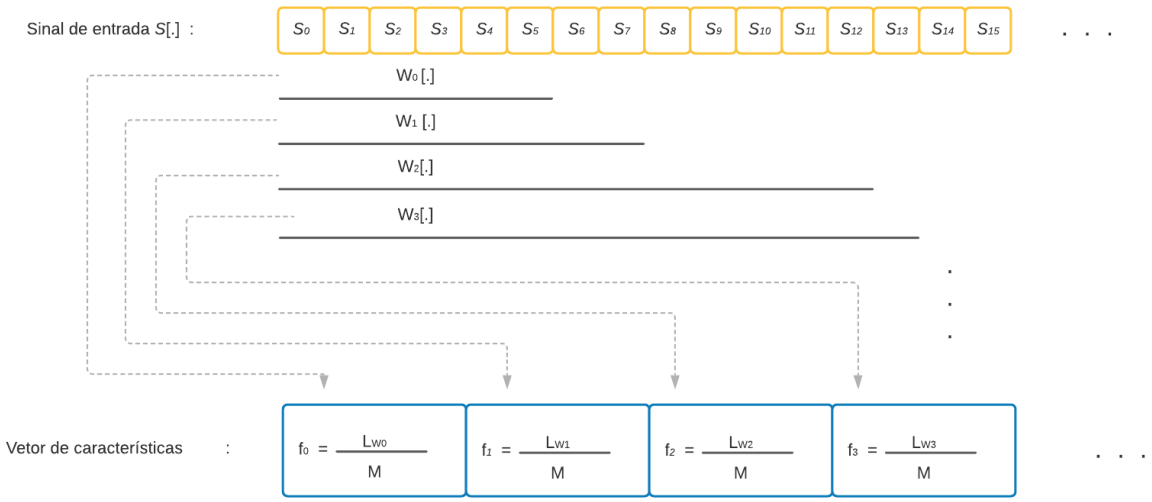

Figura 3.3: exemplo do Algoritmo A3. Imagem extraída de [3].

Algorithm 5 Algoritmo A3

Imput : $s[n]. \quad n \geq 0$ ▷ Vetor de dados.
Output : $f[m]. \quad m \geq 0$
$L \leftarrow 0$
$C \leftarrow 4.9$
$Tamanho\_Novo \leftarrow ((100/C) - ((100/C)) == 0)?(100/C) - 1 : (100/C)$
$z \leftarrow energy(s[0], Tamanho\_Vetor) * (C/100)$
for $k := 0, 1, 2, \ldots Tamanho\_Novo - 1$ **do**
  while $energy(s[0], L) < ((i + 1) * z)$ **do**
    $L \leftarrow L + 1$
  end while
  $f[k] \leftarrow L/Tamanho\_Vetor$
end for

3.3.3 Algoritmo B1

O algoritmo B1, conforme ilustrado na Figura 3.4, consiste em uma janela retangular deslizante de comprimento L atravessando o sinal de modo que, para cada posicionamento, o ZCR sobre essa posição é determinado. Cada posicionamento subsequente se sobrepõe em dois terços o anterior, sendo descartadas as possíveis amostras excedentes no final do sinal, que não são longas o suficiente para serem sobrepostas por uma janela de amostra L [20].

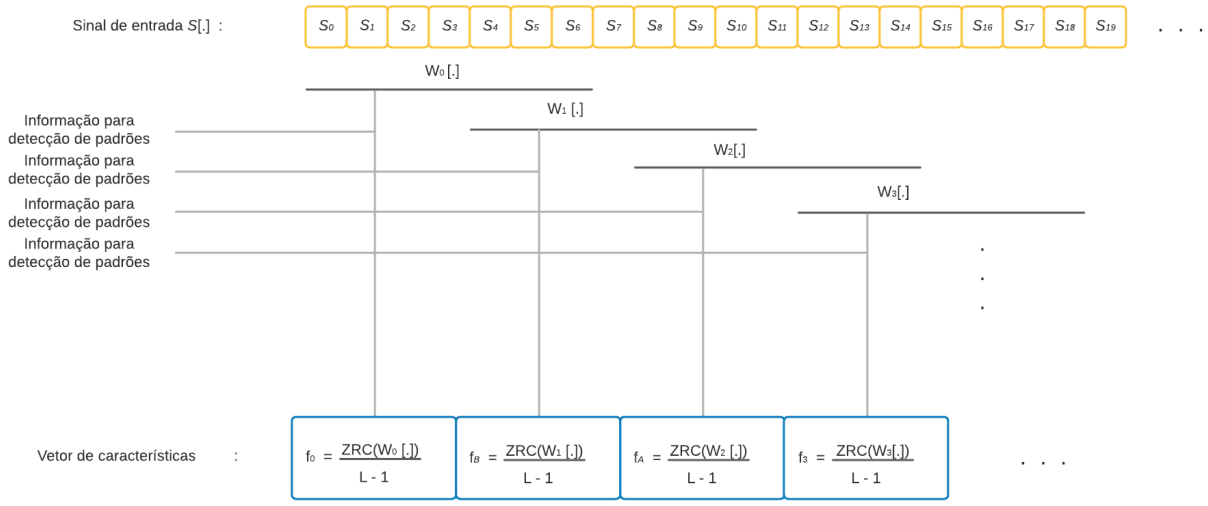

Figura 3.4: exemplo do Algoritmo B1. Imagem extraída de [20].

É possível visualizar, na Figura 3.4 e apenas para fins de ilustração, as janelas $w[\cdot]$ passando pelo sinal de entrada com tamanho $L = 8$, utilizando uma normalização que consiste em dividir cada ZCR por $L - 1$ [20].

Algorithm 6 Algoritmo B1

Imput : $s[n]$. $n \geq 0$ ▷ Vetor de dados.
Output : $f[m]$. $m \geq 0$ ▷ Vetor de características.
$L \leftarrow 256$
$V \leftarrow 66.67$
$Tamanho\_Novo \leftarrow (((100 * Tamanho\_Vetor) - (L * V))/((100 - V) * L))$
for $k := 0, 1, 2, \ldots Tamanho\_Novo - 1$ **do**
    $f[k] \leftarrow 0$
    for $i := k * (((100 - V)/100) * L)) \ldots k * (((100 - V)/100) * L)) + L - 1$ **do**
        $f[k] \leftarrow f[k] + (s[i] * s[i + 1] < 0)?1 : 0$
    end for
    $f[k] \leftarrow f[k]/(L - 1)$
end for

3.3.4 Algoritmo B3

O Algoritmo B3 consiste em determinar os comprimentos proporcionais do sinal sob análise que são necessários para atingir porcentagens predefinidas do ZCR total, conforme também ilustrado na Figura 3.5. Similarmente ao que ocorre com A3 para o conceito de energia, B3 é usado para inspecionar a constância na frequência de trabalho da entidade física responsável por geral o sinal [20].

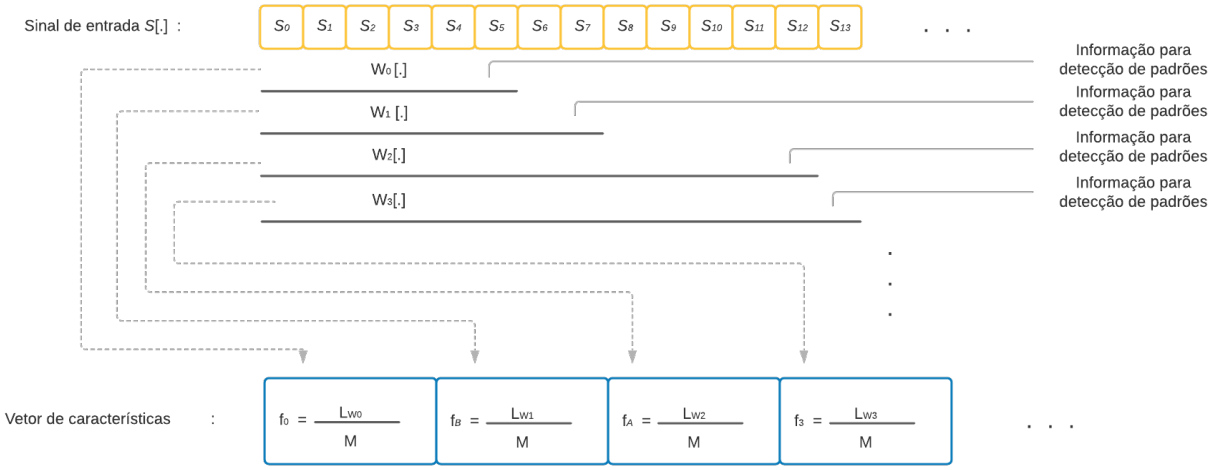

Figura 3.5: exemplo do Algoritmo B3. Imagem extraída de [20].

Na Figura 3.5, é possível observar o comportamento similar ao Algoritmo A3, onde cada janela $w[\cdot]$ cobre $i \cdot 4,9\%$ dos cruzamentos por zero do sinal de entrada [20].

Algorithm 7 Algoritmo B3

Imput : $s[n]$. $n \geq 0$ ▷ Vetor de dados.
Output : $f[m]$. $m \geq 0$ ▷ Vetor de características.
$L \leftarrow 0$
$C \leftarrow 4.9$
$Tamanho\_Novo \leftarrow ((100/C) - ((100/C)) == 0)?(100/C) - 1 : (100/C)$
$z \leftarrow zrc(s[0], Tamanho\_Vetor) * (C/100)$
for $k := 0, 1, 2, \ldots Tamanho\_Novo - 1$ **do**
  while $zrc(s[0], L) < ((k+1) * z)$ **do**
    $L \leftarrow L + 1$
  end while
  $f[k] \leftarrow L/Tamanho\_Vetor$
end for

Notavelmente, para composição final dos vetores de características, foi utilizada uma concatenação dos dados providos pelos algoritmos A1 e B1, assim como A3 e B3. Assim, o vetor de dados brutos $s[\cdot]$ foi submetido aos algoritmos e, ao resultado contendo $64 + 64 + 64 + 64 = 256$ posições, foi aplicada, isoladamente nos quatro blocos de 64 posições, a Transformada Discreta de Fourier, originando um novo vetor $f[\cdot]$ de $64 + 64 + 64 + 64 = 256$ posições. Levando em conta a taxa de amostragem dos sinais acústicos, a resolução espectral do vetor final é de $\frac{8000}{256} = 31.25$ Hz.

Com essa informação em mãos, foi realizado um casamento aproximado do vetor $f[\cdot]$ com a escala de bandas Bark, agrupando-se amostras vizinhas para obter um novo vetor de 22 posições, de acordo com a Tabela 3.1.

banda	faixa exata de frequências	faixa obtida (aprox.) de frequências
1	20 a 100 Hz	20 a 125 Hz
2	100 a 200 Hz	125 a 218.75 Hz
3	200 a 300 Hz	218.75 a 312.5 Hz
4	300 a 400 Hz	312.5 a 406.25 Hz
5	400 a 510 Hz	406.25 a 531.25 Hz
6	510 a 630 Hz	531.25 a 656.25 Hz
7	630 a 770 Hz	656.25 a 781.25 Hz
8	770 a 920 Hz	781.25 a 937.5 Hz
9	920 a 1080 Hz	937.5 a 1031.25 Hz
10	1080 a 1270 Hz	1031.25 a 1281.25 Hz
11	1270 a 1480 Hz	1281.25 a 1500 Hz
12	1480 a 1720 Hz	1500 a 1750 Hz
131	1720 a 2000 Hz	1750 a 2000 Hz
14	2000 a 2320 Hz	2000 a 2343.75 Hz
15	2320 a 2700 Hz	2343.75 a 2718.75 Hz
16	2700 a 3150 Hz	2718.75 a 3156.25 Hz
17	3150 a 3700 Hz	3156.25 a 3718.75 Hz
18	3700 a 4400 Hz	3718.75 a 4406.25 Hz
19	4400 a 5300 Hz	4406.25 a 5312.5 Hz
20	5300 a 6400 Hz	5312.5 a 6406.25 Hz
21	6400 a 7700 Hz	6406.25 a 7718.75 Hz
22	7700 a 9500 Hz	7718.75 a 8000 Hz (*)
23	9500 a 12000 Hz	-
24	12000 a 15500 Hz	-
25	15500 a 20000 Hz	-

Tabela 3.1: bandas Bark e frequências aproximadas correspondentes. Tais aproximações foram obtidas agrupando-se as energias das amostras consecutivas dos vetores de características no domínio de Fourier, considerando que cada amostra possui resolução de 31.25 Hz. (*): limitado pela taxa de amostragem.

Finalmente, para encerrar a etapa de extração de características, foram selecionadas as cinco seguintes características dentre as primeiras $64 + 64 = 128$ posições de $f[\cdot]$, assim como as mesmas cinco características dentre as $64 + 64 = 128$ posições finais de $f[\cdot]$:

- o menor valor dentre os 128 valores de $f[\cdot]$;
- o maior valor dentre os 128 valores de $f[\cdot]$;
- a média dos 128 valores de $f[\cdot]$;
- o desvio-padrão dos 128 valores de $f[\cdot]$;
- a maior diferença, em módulo, entre posições consecutivas dos 128 valores de $f[\cdot]$.

Portanto, cada sinal de voz original foi, ao longo do processo de extração de características, convertido para um vetor com $5 + 5 = 10.$ características. As cinco primeiras correspondem ao conceito de energia refletido pelos Algoritmos A1 e A3. Diferentemente, as cinco últimas correspondem ao conceito de ZCR refletido pelos Algoritmos B1 e B3.

3.4 PASSO 4: Classificação para verificação dos locutores

Depois da extração das características, foi elaborado um sistema de classificação, objetivando verificar os locutores, que funciona da forma como segue. Para cada um dos 40 locutores L_i matriculados no sistema, para $i = 1, 2, ..., 40$, foi criado um verificador linear supervisionado de padrões para o qual o valor de saída constitui uma combinação linear dos dez coeficientes do vetor de características sob análise com coeficientes $x_0, x_1, ..., x_9$. Tais coeficientes foram encontrados durante a etapa de treinamento, a qual consiste na solução do seguintes sistema possível e determinado de equações:

$$\begin{cases} a_0x_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6 + a_7x_7 + a_8x_8 + a_9x_9 &= 1 \\ b_0x_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9 &= 1 \\ c_0x_0 + c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4 + c_5x_5 + c_6x_6 + c_7x_7 + c_8x_8 + c_9x_9 &= 1 \\ d_0x_0 + d_1x_1 + d_2x_2 + d_3x_3 + d_4x_4 + d_5x_5 + d_6x_6 + d_7x_7 + d_8x_8 + d_9x_9 &= 1 \\ e_0x_0 + e_1x_1 + e_2x_2 + e_3x_3 + e_4x_4 + e_5x_5 + e_6x_6 + {}_7\, x_7 + e_8x_8 + e_9x_9 &= 1 \\ p_0x_0 + p_1x_1 + p_2x_2 + p_3x_3 + p_4x_4 + p_5x_5 + p_6x_6 + p_7x_7 + p_8x_8 + p_9x_9 &= -1 \\ q_0x_0 + q_1x_1 + q_2x_2 + q_3x_3 + q_4x_4 + q_5x_5 + q_6x_6 + q_7x_7 + q_8x_8 + q_9x_9 &= -1 \\ r_0x_0 + r_1x_1 + r_2x_2 + r_3x_3 + r_4x_4 + r_5x_5 + r_6x_6 + r_7x_7 + r_8x_8 + r_9x_9 &= -1 \\ s_0x_0 + s_1x_1 + s_2x_2 + s_3x_3 + s_4x_4 + s_5x_5 + s_6x_6 + s_7x_7 + s_8x_8 + s_9x_9 &= -1 \\ t_0x_0 + t_1x_1 + t_2x_2 + t_3x_3 + t_4x_4 + t_5x_5 + t_6x_6 + t_7x_7 + t_8x_8 + t_9x_9 &= -1 \end{cases}$$

onde $a[\cdot]$, $b[\cdot]$, $c[\cdot]$, $d[\cdot]$ e $e[\cdot]$ são os cinco vetores de características do locutor L_i escolhidos aleatoriamente dentre os dez vetores do referido locutor e, ainda, $p[\cdot]$, $q[\cdot]$, $r[\cdot]$, $s[\cdot]$ e $t[\cdot]$ são os cinco vetores de características de outros locutores diferentes de L_i escolhidos aleatoriamente dentre os 39 locutores restantes, formando o *background model*. Notavelmente, as equações que consideram $a[\cdot]$, $b[\cdot]$, $c[\cdot]$, $d[\cdot]$ e $e[\cdot]$ são igualadas ao rótulo 1 e as que consideram $p[\cdot]$, $q[\cdot]$, $r[\cdot]$, $s[\cdot]$ e $t[\cdot]$ são igualadas ao rótulo -1, visando caracterizar o locutor L_i e um locutor que não é L_i, respectivamente.

3.5 PASSO 5: Lógica de Decisão

Uma vez que os 40 sistemas lineares oriundos do passo anterior forem resolvidos, estabelecendo assim o modelo para cada locutor L_i, para $i = 1, 2, ..., 40$, pode-se passar à etapa de testes, a qual faz uso dos 5 vetores restantes de cada locutor, não usados na etapa de treinamento. Portanto, existem $40 \cdot 5 = 200$ vetores de características disponíveis para os testes, sendo 5 de cada um dos 40 locutores. Cada um dos cinco vetores do locutor L_i é, separadamente, combinado linearmente com os coeficientes

do modelo desse locutor e o valor resultante serve de base para a decisão: caso seja menor do que zero, o locutor não é autêntico; contrariamente, caso seja maior ou igual a zero, o locutor é autêntico. O valor 0, usado como fronteira de decisão, foi escolhido por ser a média entre os rótulos utilizados, ou seja, -1 e 1.

Capítulo 4

Testes e Resultados

Neste Capítulo são apresentados os resultados dos testes, realizados conforme descrito ao final do Capítulo anterior. Apenas para fins de ilustração, é possível notar, na Figura 4.1, um exemplo dos 10 vetores de características, cada qual de tamanho 10 conforme descrito no Capítulo anterior, de um mesmo locutor.

Notavelmente, excetuando as amplitudes, todos os vetores da Figura têm formato similar, o que também ocorre com os demais locutores. Portanto, pode-se concluir que, com base no método proposto, os ruídos introduzidos afetam minimamente os vetores de características obtidos com base na estratégia proposta, tornando relevante a combinação dos algoritmos A1, A3, B1e B3 associados ao uso da DFT e da escala Bark.

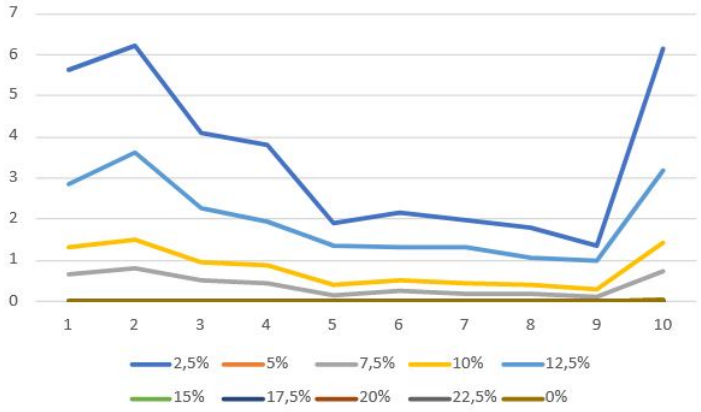

Figura 4.1: exemplo dos 10 vetores de características de um mesmo locutor. Imagem de autoria própria.

Após realizar todos os testes, verificou-se os resultados constantes na Tabela 4.1, a qual contempla a identificação do locutor ($L_1, L2, ..., L_{40}$) e a quantidade de vetores de teste verificados corretamente. É possível perceber que sistema, na expressiva maioria das vezes, verificou corretamente o locutor. As eventuais verificações incorretas corresponderam a algumas àas situações nas quais o ruído adicionado ao sinal original dos locutores ultrapassou o nível de 15%, o que pode ser interpretado como um resultado condizente com as expectativas.

locutor	**NVC** (vetores do mesmo locutor)	**NRC** (vetores de outros locutores)
L_1	5	195
L_2	5	192
L_3	4	192
L_4	5	193
L_5	4	195
L_6	5	193
L_7	5	193
L_8	5	195
L_9	5	195
L_{10}	5	195
L_{11}	5	195
L_{12}	5	195
L_{13}	4	195
L_{14}	4	195
L_{15}	4	193
L_{16}	5	195
L_{17}	5	191
L_{18}	5	193
L_{19}	5	190
L_{20}	5	192
L_{21}	5	190
L_{22}	5	195
L_{23}	5	195
L_{24}	4	194
L_{25}	5	195
L_{26}	5	195
L_{27}	5	195
L_{28}	5	195
L_{29}	5	190
L_{30}	5	195
L_{31}	5	195
L_{32}	5	190
L_{33}	5	195
L_{34}	5	192
L_{35}	5	195
L_{36}	5	195
L_{37}	5	193
L_{38}	5	195
L_{39}	5	192
L_{40}	5	192

Tabela 4.1: resultados do experimento proposto. Legenda: NVC é o número de verificações corretas, ou seja, o número de resultados verdadeiros positivos, que é no máximo 5; NRC é o número de recusas corretas, ou seja, o número de verdadeiros negativos obtidos quando **vetores de outros locutores** foram usados para se passar como o locutor L_i, que é no máximo $39 \cdot 5 = 195$ (pois existem 39 locutores diferentes do locutorL_i em questão, cada um com 5 vetores disponíveis).

Por fim, com base na referida Tabela, é possível calcular, na forma de porcentagens, os seguintes resultados:

- fração de resultados verdadeiros positivos: $\frac{5+5+4+...+5}{40 \cdot 5} = \frac{194}{200} = 0.97 = 97\%$
- fração de resultados falsos positivos: $\frac{6}{40 \cdot 5} = 1 - 0.97 = 0.03 = 3\%$
- fração de resultados verdadeiros negativos: $\frac{195+192+192+...+192}{40 \cdot 39 \cdot 5} = \frac{7745}{7800} \approx 0.99 = 99\%$
- fração de resultados falsos negativos: $1 - 0.99 = 0.01 = 1\%$

Assim, considera-se que os resultados foram significantes e, pelo teor dos trabalhos similares listados no Capítulo 2, condizentes com as expectativas. Nota-se, em especial, a baixíssima taxa de falsos negativos, implicando que raramente o sistema autenticou um locutor não genuíno. Além disso, a expressiva taxa de verdadeiros positivos também permite afirmar que o sistema proposto é eficaz.

Capítulo 5

Conclusões

Neste trabalho, foi apresentada, após a etapa de revisão de literatura, uma estratégia para verificação de locutores na modalidade *text-dependent*. O mecanismo, baseado em duas técnicas de extração de características utilizando o conceito de energia e em duas utilizando o conceito de taxa de cruzamentos por zero, além da análise de Fourier e da escala Bark, permitiu obter resultados significantes, com uma taxa de erros considerada modesta em vista de outros trabalhos do mesmo nível. Assim, o autor deste trabalho considera que a estratégia proposta atendeu aos requisitos e sugere, como trabalhos futuros, o uso de outros classificadores não lineares, tais como as redes neurais perceptron multicamadas e os algoritmos de aprendizado profundo, visando reduzir ainda mais os erros de verificação.

Referências Bibliográficas

[1] John, H.L. Hansen.; Taufiq, H. Speaker Recognition by Machines and Humans A tutorial review. *IEEE Signal Processing Magazine*, v.32, n.6, pp. 74-99, (2015).

[2] Zhizheng, Wua.; Nicholas, E.; Tomi, K.; Junichi, Y.; Federico, A.; Haizhou, L. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, v.66, pp. 130-153, (2015).

[3] Guido, R.C. A tutorial on signal energy and its applications. *Knowledge Based-Systems*. v.106, pp. 248-269, (2015).

[4] Silva, L.F.T. **Um Sistema Para reconhecimento de comandos falados dependente de locutor**. *Universidade Estadual Paulista*, São José do Rio Preto (Trabalho de conclusão de curso) (2017).

[5] Alves, L.O. **Reconhecimento computacional de fala baseado em modelos ocultos de Markov (HMMs) simplificados**. *Universidade Estadual Paulista*, São José do Rio Preto (Trabalho de conclusão de curso) (2018).

[6] Neto, A.F. **Modelo de Autenticação Aplicado a Sistemas de Verificação de Locutor**. *Universidade Federal de Minas Gerais*. Belo Horizonte (Tese de doutorado) (2018).

[7] Silveira, F.; Schueler, C.F.; Cataldo, E. Um programa para verificação de locutor por HMM usando o MATLAB. Trabalho apresentado no *XXXVII CNMAC*, S.J. dos Campos - SP, (2017).

[8] Porpino, T.N. **Identificação de locutores baseada em Aprendizagem Não-Supervisionada de características**. *Universidade Federal de Pernambuco*. Recife (Dissertação de mestrado), (2015).

[9] Hilleshein, H. **Desenvolvimento de um Sistema de Reconhecimento de Locutor Utilizando Aprendizado de Máquina**. *Instituto Federal de Santa catarina*. São José (Trabalho de conclusão de curso), (2018).

[10] Machado, T.J. **Reconhecimento de voz de locutor no contexto forense utilizando mínimos quadrados ordinários**. *Universidade Estadual Paulista*. Ilha Solteira (Tese de doutorado), (2021).

[11] Ramos, B.T.R. **As Seis Canções Trovadorescas de Fructuoso Vianna: aspectos intertextuais e perspectivas interpretativas para voz de contratenor na canção de câmara brasileira**. *Universidade Federal de Minas Gerais*. Belo Horizonte (Dissertação de mestrado), (2013).

[12] Marcondes, J.S. Identificação Pessoal: Documentos Aceitos e Retenção de Documentos. Disponível em: https://gestaodesegurancaprivada.com.br/identificacao-pessoal/ – Acessado em (2021).

[13] Colombo, F.J.; Neto, B.B.; Barros, L.J.R. Um Estudo sobre a Biometria. *Interface Tecnológica*, v. 10, n. 1, pp. 37-44, (2013).

[14] Lopes, G.A.M. **Segmentação de voz em ambientes ruidosos utilizando análise da imagem do espectrograma**. *Universidade Federal de Pernambuco*. Recife (Dissertação de mestrado), (2013).

[15] Reynolds, D.A. Speaker identification and verification using gaussian mixture speaker. *Speech Communication*. v.17, n. 1–2, pp. 91-108, (1995).

[16] http://penta3.ufrgs.br/RNP/cap3/3.2%20Audio/ .Acesso em: Janeiro de (2022).

[17] http://soundfile.sapp.org/doc/WaveFormat/ .Acesso em: Janeiro de (2022).

[18] Giannakopoulos, T.; Pikrakis, A. **Introduction to Audio Analysis**. *Academic Press*, (2014).

[19] Smith, S.W. **Digital Signal Processing: A Practical Guide for Engineers and Scientists**. *Elsevier*, (2002).

[20] Guido, R.C. ZCR-aided neurocomputing: A study with applications. *Knowledge-Based Systems*. n. 105, pp. 248-269, (2016).