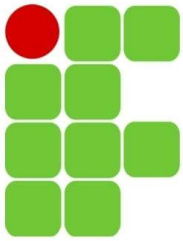


INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELECOMUNICAÇÕES

JANSER JAMES BEZERRA DE OLIVEIRA

RECONHECIMENTO DE VOZ PARA AUTENTICAÇÃO E VOTAÇÃO EM URNAS ELETRÔNICAS

Fortaleza, Ceará

2022

JANSER JAMES BEZERRA DE OLIVEIRA

RECONHECIMENTO DE VOZ PARA AUTENTICAÇÃO E VOTAÇÃO EM URNAS ELETRÔNICAS

Dissertação submetida à Coordenação de Pós-Graduação em Engenharia de Telecomunicações do Instituto Federal de Educação, Ciência e Tecnologia do Ceará, como requisito parcial à obtenção do grau de Mestre em Engenharia de Telecomunicações. Área de concentração: Sistemas de Telecomunicações

Orientador: Prof. Francisco José Alves de Aquino

Fortaleza, Ceará

2022

Dados Internacionais de Catalogação na Publicação
Instituto Federal do Ceará - IFCE
Sistema de Bibliotecas - SIBI
Ficha catalográfica elaborada pelo SIBI/IFCE, com os dados fornecidos pelo(a) autor(a)

---

r , Janser James Bezerra de Oliveira.
Reconhecimento de voz para autenticação e votação em urnas eletrônicas / Janser James Bezerra de Oliveira . - 2022.
112 f. : il.

Dissertação (Mestrado) - Instituto Federal do Ceará, Mestrado em Engenharia de Telecomunicações, Campus Fortaleza, 2022.
Orientação: Prof. Dr. Francisco José Alves de Aquino.

1. Eleições brasileiras. 2. Urnas eletrônicas. 3. Biometria da voz de eleitores. 4. Votação pelo reconhecimento de palavras isoladas. 5. Extração dos coeficientes MFCCs. I. Titulo.

CDD 621.382

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ
Av. 13 de Maio, 2081, - Bairro Benfica – CEP 60040-531 – Fortaleza – CE – www.ifce.edu.br

AVALIAÇÃO

Processo: 23256.015900/2022-79

Interessado: Janser James Bezerra de Oliveira

JANSER JAMES BEZERRA DE OLIVEIRA

RECONHECIMENTO DE VOZ PARA AUTENTICAÇÃO E VOTAÇÃO EM URNAS ELETRÔNICAS

Dissertação apresentada ao curso de Pós-Graduação em Engenharia de Telecomunicações do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) – *Campus Fortaleza*, como requisito parcial para obtenção do Título de Mestre em Engenharia de Telecomunicações.

Área de concentração: Sistemas de Telecomunicações

Aprovado em 12/12/2022

BANCA EXAMINADORA

Prof. Dr. **Francisco José Alves de Aquino** - orientador
Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE

Prof. Dr. **Auzuir Ripardo de Alexandria**
Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE

Prof. Dr. **Francisco Nivando Bezerra**
Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE

Prof. Dr. **John Herbert da Silva Felix**
Universidade da Integração Internacional da Lusofonia Afro-Brasileira - UNILAB

AGRADECIMENTOS

Meus agradecimentos à minha mãe, senhora Maria Helgenar, minha esposa, Maria Solange, e ao meu filho, recentemente nascido, José Aquiles.

Aos que partiram deste mundo, mas que ainda guardo deles eternas lembranças no meu coração: minha avó Helena, meu pai Jader e minha irmã Jadelgenar. Um dia hei de revê-los.

Meus sinceros agradecimentos aos colegas, que, mesmo em época de pandemia, forneceram áudios de treinamento de algumas locuções usadas neste trabalho para teste de reconhecimento de locutor.

Ao meu orientador, Professor Doutor Francisco José, por toda a paciência e perseverança, que só os grandes mestres possuem para com os alunos.

Meu agradecimento institucional ao IFCE, pelo apoio material à realização deste trabalho.

E a Deus, por acreditar que nós, seres humanos, podemos evoluir e construir um mundo cada vez melhor.

Janser James Bezerra de Oliveira

RESUMO

As eleições brasileiras são realizadas por meio de urnas eletrônicas comandadas pelo pressionamento de teclas de computador convencionais e usa atualmente a biometria das impressões digitais como uma forma de evitar fraudes, impedindo que alguma pessoa possa votar no lugar de outra. Entretanto, para eleitores que possuem as impressões digitais desfiguradas, ou que não possuem braços ou estejam momentaneamente impedidos de usá-los, este controle de possíveis fraudes fica prejudicado. Propõe-se neste trabalho a extração de coeficientes *mel cepstrais (MFCCs)* dos áudios de treinamento dos eleitores como dados de entrada para a implementação de um algoritmo de reconhecimento de locutor e de palavras isoladas, para que o eleitor seja reconhecido e, em seguida, possa votar usando unicamente sua voz, sem ferir o sigilo do voto e sem que haja contato físico entre pessoa e máquina. Para chegar a este intento, foram traçadas duas estratégias. A primeira está ligada ao fator psicoacústico e foi implementada pela escolha de palavras fáceis de pronunciar, foneticamente distintas entre si, escolhidas por meio do resultado das correlações entre os vetores de características extraídos dos áudios de treinamento, a fim de reduzir a taxa de erro do algoritmo proposto, e que possam ser representadas não apenas por uma sequência de letras, mas por figuras sugestivas. A segunda estratégia foi criada em razão do sigilo necessário que o processo eleitoral demanda e se consubstanciou na mudança proposital das palavras que normalmente seriam pronunciadas para invocar os comandos básicos da urna eletrônica, de modo que outras palavras mais convenientes em termos de sigilo possam substituir as anteriores, passando a ter relação unicamente posicional com os respectivos comandos. Essa relação posicional, acessível apenas ao eleitor, é expressa através da impressão de correspondências permutadas entre comandos e palavras, após cada ação de comando, no momento do voto. Os resultados das correlações entre os vetores de características extraídos dos áudios de treinamento mostram que, em ambiente controlado, ao se escolher 12 palavras foneticamente distintas entre si para comandar a urna eletrônica, tem-se um ganho na taxa de acerto de 88,68% para 97,18% quando se extrai os coeficientes MFCCs estáticos e dinâmicos. Ao se escolher apenas 6 palavras foneticamente distintas, extraindo-se somente os coeficientes MFCCs estáticos, há um ganho na taxa de acerto de 78,1% para 98,1%, o que demonstra a eficácia da estratégia. Ao se acrescentar nesta última estratégia a extração dos coeficientes MFCCs dinâmicos, obtem-se um ganho na taxa de acerto de 98,1% para apenas 99,95%, não se justificando o aumento do custo computacional.

Palavras-chave: Eleições brasileiras. Urnas eletrônicas. Biometria da voz dos eleitores. Votação pelo reconhecimento de palavras isoladas. Extração dos coeficientes MFCCs.

ABSTRACT

Brazilian elections are carried out through electronic ballot box controlled by pressing conventional computer keys and currently use fingerprint biometrics as a way to prevent fraud, preventing one person from voting in place of another. However, for voters who have disfigured fingerprints, or who do not have arms or are momentarily unable to use them, this control of possible fraud is impaired. It is proposed in this work the extraction of mel-frequency cepstral coefficients (MFCCs) from the voters' training audios as input data for the implementation of a speaker recognition algorithm and isolated words, so that the voter is recognized and then can vote using only your voice, without violating the confidentiality of the vote and without physical contact between person and machine. To achieve this goal, two strategies were designed. The first is linked to the psychoacoustic factor and was implemented by choosing words that are easy to pronounce, phonetically distinct from each other, chosen through the result of the correlations between the vectors of characteristics extracted from the training audios, in order to reduce the error rate of the proposed algorithm, and that can be represented not only by a sequence of letters, but by suggestive figures. The second strategy was created due to the necessary secrecy that the electoral process demands and consisted of the purposeful change of the words that would normally be pronounced to invoke the basic commands of the electronic ballot box, so that other words more convenient in terms of secrecy can replace the previous ones, starting to have a solely positional relationship with the respective commands. This positional relationship, accessible only to the voter, is expressed through the printing of exchanged correspondences between commands and words, after each command action, at the time of voting. The results of the correlations between the feature vectors extracted from the training audios show that, in a controlled environment, when choosing 12 words phonetically different from each other to command the electronic urn, there is a gain in the hit rate of 88.68% to 97.18% when extracting the static and dynamic MFCCs coefficients. When choosing only 6 phonetically distinct words, extracting only the static MFCCs coefficients, there is a gain in the hit rate from 78.1% to 98.1%, which demonstrates the effectiveness of the strategy. By adding the extraction of dynamic MFCCs coefficients to this last strategy, a gain in the hit rate from 98.1% to only 99.95% is obtained, not justifying the increase in computational cost.

Keywords: Brazilian elections. Electronic ballot box. Voter voice biometrics. Voting for single word recognition. Extraction of MFCCs coefficients.

CONTEÚDO

CAPÍTULO 1

CAPÍTULO 2

CAPÍTULO 3

CAPÍTULO 4

CAPÍTULO 5

CAPÍTULO 6

Lista de figuras

Lista de tabelas

Lista de quadros

Lista de abreviaturas

AFIS – Automated Fingerprint Identification System
ANN – Artificial Neural Network
AP – Áudio de Prova
AT – Áudio de Treinamento
CPU – Central Processing Unit
DCT – Discrete Cosine Transform
DFT – Discrete Fourier Transform
DTW – Dynamic Time Warping
EIH – Ensemble Interval Histogram
ERB – Equivalent Retangular Bandwith
FFT – Fast Fourier Transform
FIR – Finite Impulse Response
GMM – Gaussian Mixture Model
HMM – Hidden Markov Model
IAL – Identificação Automática de Locutor
IDFT – Inverse Discrete Fourier Transform
LC – Limiar de Correlação
LS-SVM – Least Square Support Vector Machine
MC – Máxima Correlação
MFCC – Mel Frequency Cepstral Coefficients
PAC – Pedro Álvares Cabral
PCM – Pulse Code Modulation
PDSV – Processamento Digital de Sinais de Voz
RAL – Reconhecimento Automático de Locutor
RPI – Reconhecimento de Palavras Isoladas
SVM – Support Vector Machine
TSE – Tribunal Superior Eleitoral
TRE/CE – Tribunal Regional Eleitoral do Ceará
VAL – Verificação Automática de Locutor
VC – Vetor de Características
ZCPA – Zero-Crossing with Peak Amplitude

CAPÍTULO 1

1. INTRODUÇÃO

Neste capítulo, é feita uma análise sobre a questão do cadastro biométrico de eleitores no processo eleitoral brasileiro usando o processamento digital das impressões digitais, apontando avanços em impedir ou dificultar tentativas de fraudes eleitorais.

Ao mesmo tempo, aponta-se dificuldades ou impossibilidade em captar as impressões digitais de parcela do eleitorado, sugerindo-se a implementação de uma forma alternativa de reconhecimento biométrico por meio da voz, para que os eleitores com problemas de reconhecimento biométrico das impressões digitais possam ser reconhecidos e votar sem tocar na urna eletrônica.

1.1 JUSTIFICATIVA

O presente estudo faz-se necessário em virtude de que a única forma de biometria adotada atualmente pela Justiça Eleitoral Brasileira para impedir multiplicidade de inscrições eleitorais é por meio da captação e processamento das imagens das impressões digitais dos eleitores.

O referido método é eficiente e resolve a maioria das situações. Entretanto há eleitores que têm problemas de reconhecimento biométrico usando como dados suas impressões digitais no dia da eleição. Trabalhadores braçais e da indústria química que mantêm contato direto com produtos degradantes, pessoas idosas, pessoas deficientes em decorrência da ausência de dedos, pessoas portadoras de sudorese das mãos são exemplos de eleitores que possuem suas impressões digitais muito difíceis ou impossíveis de serem captadas.

Outra razão para que se justifique o presente estudo reside no fato de que o eleitor necessariamente precisa usar teclas da urna eletrônica para votar. Parafraseando Ynoguti (1999), existe uma questão social envolvida com interfaces via voz: a dos deficientes físicos. Um eleitor que não tenha braços ou mãos só poderia votar com o auxílio direto de outra

pessoa, quebrando o sigilo do voto, já que a urna eletrônica é comandada por teclas que precisam ser pressionadas fisicamente pelos dedos para serem acionadas.

Quando se opta, por exemplo, pelo método do reconhecimento facial, poder-se-ia supor que ele poderia resolver completamente o problema da ausência de impressões digitais de alguns eleitores. Entretanto, é preciso se averiguar a que custo de memória, de processamento e de recursos financeiros essa escolha seria viável.

O que se propõe neste trabalho é aperfeiçoar o sistema, contemplando os eleitores que apresentarem problemas com o método do reconhecimento das impressões digitais, a um custo menor possível, sem afetar a eficiência necessária para tornar o mecanismo transparente e seguro, objetivando acrescentar um método biométrico alternativo e de fácil implementação, sem modificar drasticamente o que já foi feito e conquistado em termos de transparência pela Justiça Eleitoral.

Acrescente-se o fato de que as eleições, principalmente em municípios pequenos, em especial para prefeitos e vereadores, vêm se tornando cada vez mais competitivas, a ponto de candidatos deixarem de assumir ou assumir cargos em razão de pequenas diferenças após a totalização dos votos.

Existe uma crescente demanda da sociedade para requerer à Justiça Eleitoral urnas eletrônicas com o intuito de realizar eleições não oficiais, como para conselhos tutelares por exemplo. Nada impede, por exemplo, que sindicatos, associações e universidades façam o mesmo.

Isso significa que não se pode tolerar fraudes, principalmente quando um eleitor correr um sério risco de se passar por outro, em virtude da ausência de um processo biométrico alternativo e seguro, como este projeto, em que se propõe o uso da voz como parâmetro biométrico.

1.2 A PANDEMIA DE COVID-19 E O PRESENTE TRABALHO

Segundo Fiorillo (2020), em superfícies plásticas, como é o caso do teclado da urna eletrônica, há estudos que demonstram que o coronavírus pode sobreviver e ser transmissível pelo período de dois a nove dias. No caso de superfícies de vidro, como é o caso do leitor biométrico de impressões digitais, o tempo de sobrevida é de quatro a cinco dias, segundo os mesmos estudos.

Destarte, se há mecanismos automáticos e de baixo custo para que se evite o contato entre pessoa e máquina em uma pandemia, esse mecanismo deve ser incentivado.

Vale enfatizar também que uma possível proliferação de doenças transmitidas pelo contato não se restringe apenas à Covid-19. Surtos de meningite e hanseníase podem ser combatidos evitando contato entre usuários e máquinas usadas pela população em geral, tais como caixas eletrônicos, menus de opções em restaurantes, lojas e cinemas.

De fato, com o surgimento da pandemia de Covid-19, o reconhecimento de pessoas por meio da leitura de dados de impressões digitais vem sendo muito questionado por autoridades. Para se ter uma ideia desta preocupação, nas eleições de 2020, no Brasil, os eleitores foram desobrigados pelo Tribunal Superior Eleitoral (TSE) de serem reconhecidos por meio das impressões digitais no dia da eleição, conforme Res. TSE 23625/2020, em razão de recomendações expedidas por entidades sanitárias de renome no país, tais como Hospital Sírio Libanês, Fundação Fiocruz e Hospital Albert Einstein (TRIBUNAL SUPERIOR ELEITORAL, 2021). O colhimento das impressões digitais fora temporariamente suspenso no âmbito dos Cartórios Eleitorais, conforme Res. TSE nº 23616/2020.

O escopo deste trabalho não tem a pretensão de substituir o parecer de autoridades de saúde pública e de cientistas da área de infectologia, que devem avaliar, mediante estudos científicos, se realmente o contato físico entre pessoa e máquina, numa pandemia, frente à realidade social e cultural do Brasil, pode aumentar o risco de contaminação.

O que se está afirmando é que o reconhecimento de pessoas pela voz, que evita o contato direto do usuário com qualquer aparelho leitor de impressões digitais ou teclas de computadores, pode ser uma opção viável, pelo menos em momentos de pandemia, a fim de evitar ou diminuir a difusão de doenças, quando do acesso e manipulação de sistemas computacionais disponibilizados para a população em geral.

1.3 OBJETIVOS

O objetivo geral deste trabalho é propor uma solução que possibilite ao eleitor ser reconhecido pela voz e, em seguida, vote por meio da pronúncia e reconhecimento de palavras isoladas, sem tocar em qualquer parte da urna eletrônica, garantindo-se o sigilo do voto por meio da permutação entre as palavras escolhidas para invocar os comandos da urna eletrônica.

O primeiro objetivo específico é mostrar que a extração dos coeficientes *mel cepstrais* dos ATs (áudios de treinamento) é suficiente para que as taxas de acerto do sistema sejam consideradas satisfatórias.

O segundo objetivo específico é mostrar que, ao se reduzir o número de padrões a serem reconhecidos, escolhendo-se palavras foneticamente distintas, em um sistema de RPI (reconhecimento de palavras isoladas), é possível reduzir o custo computacional sem perda considerável de desempenho.

1.4 MATERIAIS E MÉTODOS

Foram objetos de estudo e de pesquisa livros e artigos científicos que tratam de acústica, psicoacústica, trato vocal, sistema auditivo e fonador humanos e reconhecimento digital de locutor e de palavras isoladas.

Em termos de simulação e cálculos, foi utilizada a linguagem Scilab (SCILAB, 2022) para captação, processamento e armazenamento dos áudios e seus VCs, bem como para simular uma eleição usando voz.

Para testar o sistema de reconhecimento de locutor, áudios de vozes de voluntários não identificados foram fornecidos.

O equipamento usado foi um *notebook* de marca *acer,* com processador intel CORE i5, com microfone embutido.

O *software* PRAAT (BOERSMA, et al., 2022), muito utilizado no meio acadêmico, foi utilizado para análise e síntese da fala humana.

1.5 A URNA ELETRÔNICA

O projeto de criação da urna eletrônica foi idealizado no ano de 1995, quando pesquisadores do Centro Técnico Aeroespacial (CTA) e o Instituto Nacional de Pesquisas Aeroespaciais (Inpe) foram convocados pelo Tribunal Superior Eleitoral (TSE) para projetar

uma máquina de votar, como assim é denominada a urna eletrônica pelo artigo 152 do Código Eleitoral de 1965.

O TSE estabeleceu algumas características que deveriam ser levadas em conta na feitura do projeto. A máquina deveria ser leve, compacta, fácil de usar e ser capaz de retirar ao máximo a intervenção humana no momento da apuração e totalização dos votos, dando maior transparência ao processo eleitoral perante a sociedade.

Já nas eleições de 1996, 70 mil urnas eletrônicas foram usadas em 57 cidades do Brasil e 32 milhões de eleitores testemunharam a rapidez, transparência e o avanço tecnológico alcançado a serviço da democracia. Na Figura 1 é apresentada uma urna eletrônica brasileira nos dias atuais. Há uma versão mais moderna que usa o teclado do terminal do mesário em tela *touch screen*.

Figura 1: Urna Eletrônica

Fonte: Adaptado de TRIBUNAL REGIONAL ELEITORAL DO RJ (2020)

Conforme é demonstrado na Figura 1, a urna eletrônica é composta de duas partes: a do terminal do mesário, à esquerda; e a do terminal do eleitor, à direita. A CPU está acoplada junto ao terminal do eleitor. Uma informação importante que precisa ser enfatizada, pois interfere na modelagem do algoritmo para reconhecimento de palavras isoladas, é o fato de que existem 13 teclas, no caso do terminal do eleitor, e 12 teclas no caso do terminal do mesário. O que difere os dois terminais um do outro é a presença da tecla branca no terminal do eleitor, que executa o comando *votar em branco.*

Note-se na Figura 1 que este terminal do mesário já vem equipado com o leitor biométrico, dispositivo de entrada que permite a leitura dos dados das impressões digitais dos eleitores para serem reconhecidos no dia da eleição. No modelo de reconhecimento de locutor pela voz proposto neste trabalho, o microfone deve ser posicionado na cabine de votação, não no terminal do mesário.

Após o reconhecimento biométrico pelas impressões digitais, o eleitor é habilitado a se dirigir ao terminal do eleitor e escolher seus candidatos por meio do contato físico entre

eleitor e teclas de computador, que representam comandos (escolha de números: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9), votar em branco ou corrigir/confirmar o voto.

No dia da eleição, a urna eletrônica está rodeada pela cabine de votação, um anteparo de papelão que impede que se tenha conhecimento em quem o eleitor está votando, para que se garanta o sigilo do voto.

Deve ser enfatizado que as informações que são gravadas na urna eletrônica são apenas dados dos eleitores ligados àquela seção eleitoral. Caso se use a voz como parâmetro biométrico para habilitação do eleitor a ir votar, os VCs de treinamento de voz dos eleitores só podem ser inseminados na urna eletrônica após uma varredura em toda à base de dados da Justiça Eleitoral, inclusive para que se calcule graus de semelhança que o eleitor deve atingir para ser habilitado à cabine de votação.

Desde 2014, o Tribunal Superior Eleitoral usa o sistema *AFIS (Automated Fingerprint Identification System)* (TRIBUNAL SUPERIOR ELEITORAL, 2017), com o intuito de averiguar, em nível nacional, se existe qualquer eleitor do país com mais de uma inscrição eleitoral. Ou seja, este sistema compara cada impressão digital do eleitor com as respectivas impressões digitais dos demais eleitores, em busca de coincidências biométricas. O sistema *AFIS* pode comparar, por dia, até 160 mil impressões digitais.

Portanto, quando uma pessoa vai a um Cartório Eleitoral e se alista, cadastrando suas impressões digitais, o sistema *AFIS* varre todo o banco de dados da Justiça Eleitoral, para detectar duplicidades ou multiplicidades de inscrições eleitorais. Caso não ocorra coincidência, diz-se que o requerimento de inscrição do eleitor foi processado e que está tudo regular.

O sistema *AFIS* é de suma importância, pois é através dele que se garante que os dados biométricos extraídos das impressões digitais dos eleitores incluídos na urna eletrônica são únicos e legítimos, quando comparados com o eleitorado do país inteiro. No caso de uso da voz, é preciso adaptar o sistema *AFIS* ou criar um novo sistema para realizar a tarefa análoga a que é feita para as impressões digitais.

1.6 ANTES E DEPOIS DA BIOMETRIA DAS IMPRESSÕES DIGITAIS

Antes do ano de 2007 a base de dados da Justiça Eleitoral não continha fotos, assinaturas digitalizadas e muito menos impressões digitais de qualquer eleitor do país. Os

cidadãos maiores de 16 anos se dirigiam aos Cartórios Eleitorais, respondiam a uma sequência de perguntas realizadas por servidores e estes a registravam no banco de dados (sistema ELO) da Justiça Eleitoral.

No dia da eleição, o eleitor comparecia com seu documento oficial com foto, era identificado no caderno de votação e, se nenhum fiscal de partido impugnasse a sua identidade, o eleitor assinava sua presença no caderno de votação e em seguida era habilitado a ir votar perante a urna eletrônica.

Somente no ano de 2008 (TRIBUNAL SUPERIOR ELEITORAL, 2021), foram testadas as primeiras verificações biométricas de impressões digitais em seções eleitorais nos municípios de São João Batista (SC), Fátima do Sul (MS) e Colorado do Oeste (RO). Essas eleições foram consideradas um sucesso e a biometria das impressões digitais foi estendida paulatinamente a todo o país ao longo dos anos.

A Tabela 1 mostra a quantidade de eleitores que ainda não fizeram sua biometria (na casa dos 38 milhões) até os dias de hoje.

Tabela 1: Quantidade de eleitores com biometria das digitais no país

Eleitores com biometria	Eleitores sem biometria
117.847.134	38.298.158
75,47%	24,53%

Fonte: TRIBUNAL SUPERIOR ELEITORAL (2022)

Antes de 2007, em razão da falta de biometria, um eleitor poderia chegar à sua seção eleitoral e a urna informar que o mesmo já havia votado, após o Presidente de Mesa digitar o número do título do eleitor no terminal do mesário. Diante dessa situação, não se sabia se o erro teria sido do mesário, que teria digitado o título de eleitor abaixo ou acima do verdadeiro eleitor no caderno de votação (a lista de nomes no caderno de votação é em ordem alfabética) ou se o eleitor seria um impostor tentando se passar pelo verdadeiro eleitor.

Sem a biometria, a fraude era uma possibilidade que sempre pairava no ar. Como não havia foto ou biometria, um eleitor poderia sim facilmente se passar por outro, caso não houvesse impugnação dos fiscais de partido.

Com a biometria das impressões digitais, os eleitores continuaram a se submeter aos mesmos procedimentos anteriores, mas agora as suas dez impressões digitais deveriam ser, antes das eleições e perante o Cartório Eleitoral, capturadas e registradas digitalmente para que, no dia da votação, o eleitor pudesse ser identificado biometricamente. A foto do eleitor e sua assinatura também foram digitalizadas para posterior aferição dos mesários e fiscais de partido. Note-se que a foto do eleitor foi digitalizada mas não há reconhecimento facial

eletrônico no dia da votação. O reconhecimento facial é realizado pelos mesários e fiscais de partidos.

Na atualidade, o eleitor se dirige à sua seção eleitoral e o Presidente de Mesa solicita dele documento oficial com foto, localizando seu nome no caderno de votação, digitando no terminal do mesário o seu número do título do eleitor, pedindo-lhe que posicione seu dedo indicador no leitor biométrico para autenticação. Ou seja, antes da aposição da impressão digital do eleitor no leitor biométrico, o eleitor já é discriminado dos demais, mediante ação do Presidente de Mesa.

Nota-se que não há uma comparação entre as impressões digitais de todos os eleitores inseminados na urna, com as do eleitor que está querendo ser habilitado a votar. Se isso ocorresse, haveria maior custo computacional, embora houvesse ganho de tempo no dia da eleição, em razão do mesário não ter que digitar o número do título de eleitor, tendo apenas que conferir se o eleitor encontrado é o mesmo do documento oficial apresentado.

Em regra, tal procedimento de verificação biométrica usando as impressões digitais do eleitor vem funcionando a contento para a maioria dos casos. Entretanto, há um considerável número de situações em que as impressões digitais do eleitor são muito difíceis de serem capturadas em virtude de diversas razões, não comprovadamente estabelecidas por ausência de um maior estudo empírico e mesmo científico até a presente data.

Contudo, há de se supor que as falhas relacionadas com o reconhecimento biométrico por meio de impressões digitais estão ligadas aos seguintes fatores: idade do eleitor, que muitas vezes tem suas impressões digitais desfiguradas pelo próprio envelhecimento natural; o eleitor é trabalhador braçal (muitas vezes da agricultura) ou trabalha em contato direto com produtos químicos e essas circunstâncias podem prejudicar a legibilidade de suas impressões digitais; o eleitor pode ser portador de uma doença chamada de sudorese das mãos, o que faz com que suas mãos estejam sempre com excesso de suor, dificultando a captação de suas impressões digitais. Há também a situação em que esta captura das impressões digitais seria impossível, em razão da ausência de braços ou mãos.

1.7 NÃO RECONHECIMENTO BIOMÉTRICO NO CEARÁ

Segundo dados do TRIBUNAL SUPERIOR ELEITORAL (2018), no 1º e 2º turnos das Eleições 2018, no Estado do Ceará, haviam 6.344.479 eleitores aptos a exercerem o

direito de votar. Desses, 5.092.349 eleitores compareceram às urnas, sendo que 899.243 não haviam realizado biometria das impressões digitais.

A Tabela 2 demostra a quantidade de eleitores do estado do Ceará, no 1º Turno das Eleições 2018, que tiveram suas impressões digitais reconhecidas eletronicamente na 1ª, 2ª, 3ª e 4ª tentativas, bem como os eleitores que não foram reconhecidos em nenhuma das quatro tentativas: 467.691 eleitores.

Tabela 2: Tentativas de reconhecimento (1º Turno das Eleições 2018 – Ceará)

1ª tentativa	2ª tentativa	3ª tentativa	4ª tentativa	Não reconhecidos (11,1 %) do Total	Total
2.508.391	750.456	311.735	154.833	467.691	4.193.106

Fonte: setor de TI do TRE-CE

Ou seja, segundo a Tabela 2, 11,1% dos eleitores que haviam realizado biometria das impressões digitais e foram votar no 1º turno da Eleições de 2028, não foram reconhecidos pela urna eletrônica no dia da eleição.

Na Tabela 3 são demonstradas a quantidade de tentativas de reconhecimento no segundo turno das Eleções 2018.

Tabela 3: Tentativas de reconhecimento (2º Turno das Eleições 2018 – Ceará)

1ª tentativa	2ª tentativa	3ª tentativa	4ª tentativa	Não reconhecidos (10,9%) do Total	Total
2.481.926	748.487	302.199	150.307	451.808	4.134.727

Fonte: setor de TI do TRE-CE

Ou seja, segundo a Tabela 3, 10,9% dos eleitores que realizaram biometria de suas impressões digitais no Estado do Ceará nas Eleições de 2018 não foram reconhecidos em nenhuma das 4 tentativas.

Ambas as Tabelas 2 e 3 mostram que a biometria dos eleitores usando impressões digitais não conseguiu fazer com que todos os eleitores pudessem ser autenticados. Aliás, nenhum método biométrico conseguiu tal proeza até a presente data.

Os dados das Tabelas 2 e 3 são estatísticas divulgadas pelo Tribunal Regional Eleitoral do Ceará (TRE/CE). Nas eleições de 2020, como os eleitores foram desobrigados de se autenticarem biometricamente usando as impressões digitais em razão da pandemia de Covid-19, estes dados não foram catalogados.

1.8 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação é estruturada em 6 capítulos, como seguem: no **Capítulo 2**, são discutidos aspectos fundamentais sobre áudios de treinamento (ATs) e de prova (APs), vetores de características (VCs) de treinamento e de prova, operação de correlação, Reconhecimento Automático de Locutor (RAL) e propõe um algoritmo para autenticação do eleitor mediante sua voz; no **Capítulo 3**, é feita uma exposição sobre o algoritmo proposto neste trabalho para que o eleitor vote usando sua voz sem quebra de sigilo do voto, com uso de palavras estrategicamente escolhidas para invocar as funções da urna eletrônica, correspondentes às suas teclas; no **Capítulo 4**, é feita uma análise da técnica de extração dos coeficientes *MFCCs,* usada nestre trabalho para fins de reconhecimento de locutor e de palavras isoladas; o **Capítulo 5** demonstra os resultados dos testes em computador, usando a linguagem Scilab; e, por fim, o **Capítulo 6**, no qual são demonstradas as contribuições deste trabalho, propondo novos estudos e avanços, tendo como base a ideia de comandar máquinas usando voz.

CAPÍTULO 2

2. RECONHECIMENTO BIOMÉTRICO DO ELEITOR PELA VOZ

Neste capítulo, é apresentado um estudo sobre os principais aspectos que envolvem o reconhecimento do eleitor pela voz, antes de habilitá-lo a ir votar, fazendo importantes considerações sobre a escolha da locução que deve ser pronunciada e processada para se realizar testes de reconhecimento no dia da eleição.

Algumas convenções e definições são estabelecidas, tais como vetores de características (VCs), áudios de treinamento (ATs), Áudios de Prova (APs), correlação estatística, máxima correlação (MC), limiar de correlação (LC) e reconhecimento automático de locutor (RAL).

Um algoritmo de reconhecimento de locutor é proposto, estabelecendo um leque de possibilidades para implentação do sistema.

2.1 ÁUDIOS DE TREINAMENTO E ÁUDIOS DE PROVA

Para solucionar o problema suscitado neste trabalho, é preciso estabelecer a diferença entre dois tipos de áudios, levando em conta o momento da sua gravação e o objetivo que cada um se destina.

Os áudios de treinamento (ATs) são as amostras dos sinais que reproduzem as frases pré-estabelecidas. No caso das eleições, eles devem ser captados e gravadas junto ao Cartório Eleitoral.

Os áudios de prova (APs) são os decorrentes da gravação das locuções pronunciadas no dia da eleição pelos eleitores, para verificar se o eleitor é ou não é autêntico. Geralmente se adota um número máximo de tentativas de reconhecimento no dia da eleição, como, por exemplo, 4 tentativas, sob pena de não aceitar o locutor como legítimo.

Com relação à taxa de amostragem das gravações dos áudios, recomenda-se a leitura do Apêndice II deste trabalho.

2.2 OS VETORES DE CARACTERÍSTICAS, AS CORRELAÇÕES E OS LIMIARES

Não é tecnicamente possível comparar diretamente ATs com o objetivo de encontrar semelhança entre os mesmos. É preciso processar digitalmente estes áudios, comprimir e extrair informações dos mesmos e alocá-las sequencialmente em um vetor de características (VCs).

Sem se preocupar como se calcula os dados dispostos nesses VCs, os valores das correlações e dos limiares (LC), algo que é feito na Seção 4.8.5, a presente Seção se limita a definí-los, analisando alguns aspectos relevantes para se entender o funcionamento do sistema proposto.

Uma das operações matemáticas mais usadas para estabelecer semelhanças entre os VCs é a correlação estatística, adotada neste trabalho.

Como os VCs são extraídos do processamento dos ATs e APs, pode-se também denominá-los respectivamente de VCs de treinamento e VCs de prova.

A correlação estatística é um valor numérico (conhecido como coeficiente de *Pearson*) que estabelece um nível de semelhança entre dois vetores de mesmo tamanho. No caso, estes valores numéricos podem variar de menos um (-1) até mais um (+1), de sorte que quanto mais próximo de mais um (+1) o resultado da correlação entre dois vetores X e Y, mais perecidos um com o outro esses vetores serão. Por outro lado, quanto mais próximo de zero (0) o resultado da correlação entre dois vetores X e Y, mais diferentes um do outro eles serão.

É importante, para fins de elaboração do algoritmo de reconhecimento de locutor, diferenciar dois tipos de correlação. A correlação calculada entre VCs extraídos de áudios pronunciados pelo mesmo locutor, aqui denominada de autocorrelação. E a correlação calculada entre VCs extraídos de áudios pronunciados entre locutores diferentes, denominada neste trabalho de correlação cruzada.

Tanto na fase de treinamento quanto na fase de prova, caso os VCs tenham sido extraídos de maneira correta, espera-se que as autocorrelações sejam consideravelmente maiores que as correlações cruzadas. Para que o sistema esteja funcionando bem, as correlações cruzadas podem servir para medir diferenças e as autocorrelações podem servir para medir semelhanças.

Na fase de treinamento, tem-se a certeza de quais são os áudios do locutor e quais os áudios que não são do locutor e isso leva à certeza de 100% de quais correlações entre os VCs respectivos são do tipo correlação cruzada e quais são do tipo autocorrelação. Em razão desta

certeza, é que nesta fase é que se calcula o limiar de correlação (LC), usando os VCs. Este LC é o valor mínimo de correlação que se deve obter na fase de prova (no caso deste trabalho, no dia da Eleição) para que o locutor seja presumidamente reconhecido como legítimo.

Na fase de prova, não se tem certeza alguma se a correlação é do tipo cruzada ou do tipo autocorrelação. Todas as correlações são feitas entre o VC de prova e todos os VCs de treinamento gravados na base e, baseado no valor máximo entre essas correlações, é que existe uma presunção de que a correlação é do tipo cruzada ou do tipo autocorrelação.

Se a máxima correlação (MC) atingir o LC na fase de prova, há uma presunção probabilística de que o locutor foi reconhecido (autocorrelação). Do contrário, se a MC não atingir o LC na fase de prova, há uma presunção probabilística de que o locutor não foi reconhecido (correlação cruzada). Entretanto, se essa presunção estiver errada, ocorre um erro. Há dois tipos de erros neste caso: o falso positivo, quando o sistema reconhece uma locução ou um locutor ilegítimo; falso negativo, quando o sistema não reconhece uma locução ou um locutor legítimo.

2.3 RECONHECIMENTO AUTOMÁTICO DE LOCUTOR (RAL)

O RAL é o primeiro passo antes de habilitar o eleitor a votar, pois este precisa ter sua voz reconhecida digitalmente. Segundo Oliveira (2001), o RAL tem o propósito de controlar ou restringir o acesso a redes, computadores, bases de dados, bem como restringir a disponibilização de informações confidenciais para pessoas não autorizadas, dentre várias outras aplicações.

Segundo Paranaguá (1997), o RAL engloba as maneiras computacionais de se identificar automaticamente pessoas por meio de características vocais levando em conta a ausência ou a presença de prévia identificação numérica para que o locutor reivindicante seja ou não habilitado a ter acesso a determinado sistema ou informações.

O estudo do RAL foi iniciado há mais de 35 anos e obteve força com a introdução e aplicação dos Modelos Ocultos de Markov, em inglês *Hidden Markov Models* (*HMMs*) (PARANAGUÁ, 1997).

Com relação aos tipos e quantidades de informações fornecidas ao sistema para reconhecer o locutor, o RAL pode ser dividido em Verificação Automática de Locutor (VAL) e Identificação Automática de Locutor (IAL) (CAMPBELL, 1997).

A IAL objetiva identificar o autor de uma dada locução pronunciada, com base em uma amostra de sua voz, diferenciando-a de um conjunto de possíveis locutores previamente treinados e cadastrados em toda a base de consulta do sistema. Na IAL não ocorre reivindicação de autenticidade da identidade, pois o sistema é que deve ir em busca desta identidade. O sistema faz comparações com os VCs de toda a base e seleciona a de maior verossimilhança. Essa IAL poderá ser com rejeição, se o grau de verossimilhança tiver que atingir um valor mínimo de limiar previamente calculado para que o locutor seja considerado autêntico. E a IAL poderá ser sem rejeição, cuidando o sistema apenas de fornecer o maior grau de semelhança, cabendo a decisão de aceitar o locutor como legítimo a uma pessoa.

Por outro lado, a VAL é o processo de aceitar ou rejeitar a legitimidade de um locutor pela voz, anteriormente discriminado por uma identificação prévia que não tenha sido a voz. Conforme Campbell (1997), a VAL faz-se uso da máquina para verificar a identidade da voz de uma pessoa que a reivindica e, independentemente do número de locutores cadastrados, toma-se uma decisão de aceitar ou não aceitar o pretenso locutor baseado em um limiar mínimo de aceitação.

A inclusão de mais locutores no sistema não indica que o tempo de processamento necessário para realizar as comparações será aumentada, na medida em que só haverá comparações com as características do locutor previamente informado à máquina. Destarte, na VAL, a probabilidade de ocorrência de erros de verificação mantém-se a mesma para cada locutor, mesmo depois da inclusão de um ou mais locutores (PETRY, 2002). Ou seja, na VAL a identidade é previamente fornecida e confirmada pela voz. Na IAL a voz é fornecida ao sistema para que este busque a identidade do locutor.

Extraídas as características da voz do locutor, há duas formas gerais tradicionalmente conhecidas para comparar estas características e reconhecer locutores: os métodos estatísticos e os métodos determinísticos.

Os métodos estatísticos são baseados em medidas de verossimilhança ou probabilidade condicional da observação do padrão extraído, em que as técnicas da função densidade de probabilidade e os *Hidden Markov Models* (*HMMs*) (TISBY, 1991) são seus maiores exemplos. Nestes métodos, os padrões de treinamento extraídos são considerados cópias imperfeitas uns dos outros.

Já nas técnicas determinísticas, os padrões extraídos são tidos como cópias perfeitas, havendo a necessidade de ocorrer um alinhamento dos testes de prova para se fazer o reconhecimento almejado. São exemplos de métodos determinísticos as *Artificial Neural*

Networks (*ANNs*) (FARRELL, et al., 1994), os *Gaussian Mixture Models* (*GMMs*) (REYNOLDS, et al*.,* 1995) e o *Dynamic Time Warping* (*DTW*) (CAMPBELL, 1997).

Tendo os recursos computacionais suficientes (geralmente exigindo-se muita capacidade de processamento e memória), todas essas técnicas podem ter ótimas performances. Por exemplo, em termos de eficiência, as técnicas que usam *HMMs* podem chegar a quase 99% de acerto, mas com altíssimas demandas computacionais (PEACOCKE, et al*.,* 1990).

O que se tem feito na prática é realizar pesquisas e experimentos para se encontrar métodos que reduzam a complexidade computacional, baseados em estratégias já existentes, mas customizando-as e adaptando-as aos problemas específicos. É o caso deste trabalho.

Um exemplo de redução de complexidade computacional é mencionado em DAN (2008), onde, baseado no *Least Square Support Vector Machine* (*LS-SVM*), transformou-se um problema de programação quadrática, do já conhecido *Support Vector Machine* (SVM), num problema de programação linear, reduzindo a complexidade computacional.

Outras pesquisas e experimentos visam ao aprimoramento da eficiência dos métodos de reconhecimento de locutores em ambientes ruidosos, como em Wang (2007) e Shao (2006).

Com relação às escolhas das locuções, o RAL geralmente é dividida em RAL dependente de texto e RAL independente de texto. Na RAL dependente de texto, as locuções são previamente escolhidas pelo locutor ou por um agente externo (governo ou empresa). Na RAL independente de texto o locutor pode falar qualquer locução para ser reconhecido.

O RAL do tipo IAL e independente de texto é a que exige maior custo computacional, razão pela qual não se aconselha usá-la para identificação de locutor em eleições, por causa da escassez de recursos inerente à urna eletrônica. O RAL do tipo VAL dependente de texto é sem dúvida a estratégia mais adequada e foi a forma usada para a simulação descrita neste trabalho.

A IAL, no caso das eleições brasileiras, pode ser adotada, com a vantagem de economia de tempo no dia da eleição, pois o mesário não perderia tempo em digitar o número do título de eleitor no terminal do mesário. Apenas iria conferir se o nome fornecido pela urna eletrônica é o mesmo nome do documento oficial apresentado pelo eleitor. Entretanto, há um aumento do custo computacional, o que deve ser avaliado pela Justiça Eleitoral.

2.4 A ESCOLHA DO TEXTO

Partindo da decisão de que nesta simulação o reconhecimento de locutor é do tipo VAL dependente de texto, sendo a forma mais adequada para o reconhecimento de locutor em eleições, em razão de um menor custo computacional, resta saber quem pode ou deve escolher esta locução: o eleitor ou a própria Justiça Eleitoral.

Caso a decisão de escolha do texto seja do eleitor, a Justiça Eleitoral poderia estabelecer por exemplo que o texto seria os 3 primeiros nomes do eleitor, ou seu mês e dia de nascimento ou até mesmo uma frase escolhida livremente por ele (uma senha vocal).

Caso a decisão de escolha seja da Justiça Eleitoral, o texto poderia ser padronizado para todos os eleitores do país, o que exigiria um maior custo computacional para classificação dos locutores, tanto a nível de base de dados da Justiça Eleitoral como na base da própria urna eletrônica.

Se o texto é único para identificação de qualquer eleitor e o sistema de classificação é capaz de distinguir todos os eleitores, esse mesmo sistema, por uma questão lógica, seria capaz também de identificar locutores por meio de textos diferentes, escolhidos pelo próprio eleitor ou pela Justiça Eleitoral.

Por uma questão de sigilo dos locutores que forneceram seus áudios voluntariamente, os testes feitos em computador para identificação de locutor foram realizados usando a frase PEDRO ÁLVARES CABRAL (doravante referida como PAC), por quatro razões:

1 – é uma expressão por todos conhecida (quem nunca ouviu a pergunta: quem descobriu o Brasil?);

2 – se refere a um personagem histórico ligado ao descobrimento do Brasil e que já se encontra sepultado há mais de quinhentos anos, o que dá um certo clima de imparcialidade política, o que é bom para as eleições;

3 – é uma frase fácil de pronunciar, sem qualquer dificuldade em sua articulação e já naturalmente treinada por todos os brasileiros desde as primeiras aulas de História do Brasil no ensino fundamental. Mesmo quem nunca frequentou a escola sabe o nome do personagem que descobriu o Brasil;

4 – é uma expressão sem fonemas nasalizados, o que descartaria os possíveis problemas com reconhecimento de voz de eleitores com doenças respiratórias como gripe, que incham as fossas nasais, modificando as frequências formantes do sinal sonoro, o que poderia dificultar o reconhecimento da voz;

Não fossem as doenças respiratórias que inflamam os tratos nasais das pessoas, os sinais nasalizados seriam os mais indicados para reconhecimento de locutor, em virtude de não haver articulação nenhuma nas cavidades nasais para se produzir esses sons.

Importante mencionar que esta possível modificação das frequências formantes do som causada por problemas respiratórios é apenas uma hipótese, que, salvo melhor juízo, não fora averiguada empiricamente até o presente momento, não se tendo uma posição definitiva sobre se essa mudança seria suficiente para provocar erro no reconhecimento do locutor (ver Seção 6.2).

2.4.1 Usando o nome do eleitor

Partindo-se do princípio de que o nome completo de uma pessoa é reconhecidamente e provavelmente a frase vocal mais pronunciada por ela ao longo de sua vida, tem-se desta forma uma padronização vocal muito bem estabelecida, pois a pergunta “Qual é o seu nome?” é sem dúvida uma das primeiras perguntas que um ser humano ouve desde a sua tenra idade e nos primeiros anos de sua vida escolar e se perpetua ao longo de sua existência, principalmente quando se dirige aos órgãos públicos e empresas para realizar cadastros os mais diversos.

Destarte, há uma tendência natural para que esta pronúncia do nome completo seja, ao longo do tempo, padronizada e acomodada nas cordas vocais e estruturas ressonantes (boca, pescoço, faringe, laringe, fossas nasais) do indivíduo.

Entretanto, escolher o nome do eleitor como parâmetro de reconhecimento biométrico de locutor pode ter dois inconvenientes: primeiro, a questão da divulgação do nome para pessoas não autorizadas; segundo, o fato de que alguém pode tentar se passar pelo eleitor tentando imitar sua voz, em razão de ter ouvido sua pronúncia em reconhecimento anterior.

2.4.2 Usando uma senha vocal com RAL

Uma forma de resolver o proplema do sigilo do nome do eleitor é possibilitar que ele escolha uma frase, obviamente diferente de seu nome e de qualquer dado sensível.

Entretanto, o problema da revelação da senha no primeiro acesso não seria resolvido, pois alguém que por ventura tenha ouvido o eleitor pronunciar a senha vocal poderá tentar imitar a voz do eleitor pronunciando a mesma frase em um posterior acesso.

Ademais, como só exitem eleições oficiais a cada dois anos, corre-se um sério risco de possível esquecimento desta senha vocal por parte de grande parte dos eleitores no dia da eleição.

Este provável esquecimento levaria uma grande massa de eleitores a consultarem sua senha vocal por meio da *internet,* no dia da eleição*,* gerando congestionamento na rede, atraso na eleição e corrida aos Cartório Eleitorais do país para saber qual seria sua senha vocal.

Neste trabalho, há uma outra proposta de senha vocal (Seção 3.3.2), que usa permutação e reconhecimento de palavras isoladas, muito mais robusta e segura.

2.5 POSSÍVEIS ERROS

Na circunstância em que o eleitor tente se autenticar usando voz em todas as tentativas permitidas e não consiga atingir o LC, é presumido a existência de cinco situações:

1 – o eleitor estaria pronunciando as locuções de forma muito diferente das que ele havia pronunciado no dia que forneceu suas locuções de treinamento no Cartório Eleitoral;

2 – o eleitor estaria com problemas nas cordas vocais (portador de câncer, cisto, com rouquidão...etc.)

3 – o ambiente da seção eleitoral estaria contaminado com muito ruído ambiente (o que não é comum, no dia da eleição, estando o ruído provocado provavelmente pelo barulho de um ar-condicionado ou ventilador);

4 – o *hardware* da urna eletrônica estaria com muitos ruídos internos que se somariam às gravações das locuções de prova;

5 – o eleitor estaria tentando se passar por outro (tentativa de fraude);

Nos casos dos ítens 1, 2, 3 e 4, uma solução seria optar pelo procedimento descrito em Seção 5.7, quando o mesário ou alguém de confiança do eleitor vota pelo eleitor sem saber em quem o eleitor votou, por meio do acionamento de teclas do terminal do mesário. Este procedimento não deve ser obrigatório, pois seu acionamento deve ser de livre escolha do eleitor que não quer que seu acompanhante saiba em quem ele votou. Destarte, se o eleitor pronunciou a frase e não foi reconhecido nas 4 tentativas, resta-lhe 3 opções: votar pela forma tradicional, pressionando as teclas de plástico da urna, caso possa movimentar suas mãos; autorizar que seu acompanhante vote por ele, com a quebra de sigilo autorizada por lei; ou solicitar ao mesário o procedimento descrito em Seção 5.7.

Caso não ocorra nenhuma das hipóteses dos itens 1, 2, 3 e 4, restando apenas o item 5, em havendo impugnação de algum fiscal de partido questionando a identidade do eleitor em razão do não reconhecimento vocal, a solução é fazer com que o eleitor retorne outro horário próximo ao término da eleição, na tentativa de que o verdadeiro eleitor apareça na seção eleitoral.

Outra solução para o caso apontado no item 4 seria substituir a urna eletrônica. Para o caso do item 3, outra solução seria adotar ou reforçar o tratamento acústico proposto no Apêndice II.

2.6 O ALGORITMO PARA RECONHECIMENTO DE ELEITOR

Descartada a possibilidade de que o eleitor pronuncie qualquer frase ao microfone e possa ser identificado biometricamente, em virtude da escolha do reconhecimento de locutor ser do tipo VAL dependente de texto, passa-se agora para a forma como as características extraídas dos áudios serão comparadas, sem apelar para os classificadores tradicionais (como as Redes Neurais e os Modelos Ocultos de Marcov), que requerem um custo computacional excessivo, para o problema que se quer resolver neste trabalho.

A técnica usada neste trabalho é a *template matching* (que significa correspondência de modelo, em português) usando correlações, onde os padrões gerados por locutor serão extraídos dos ATs e dos APs. Nesta técnica, os VCs extraídos dos ATs e dos APs são comparados diretamente de forma que cada um deles é tido como cópia imperfeita do outro.

É sabido que a técnica *template matching,* não obstante ser simples e requerer poucos recursos computacionais, possui relativamente desempenho menor, quando comparada com a

técnica da Rede Neural, por exemplo. Entretanto, pode-se aumentar consideravelmente este desempenho aplicado ao reconhecimento de voz em urnas eletrônicas usando as seguintes ações:

1 – escolhendo locuções naturalmente treinadas pelo locutor (ano de nascimento, nome do próprio eleitor, locuções fáceis e muito conhecidas como PAC ou senha vocal treinada pelo eleitor);

2 – fazendo um tratamento acústico padronizado, a fim de impedir ruídos no momento da captação dos ATs e dos Aps (ver Apêndice II);

3 – parametrizando a forma de captação dos APs e dos ATs, distância da boca ao microfone, tipos de microfones e espumas acústicas, estabelecendo tempo médio de pronúncia da frase e posicionamento da urna eletrônica no local de votação de forma acusticamente estratégica;

2.6.1 Usando MC e LC

Para que se entenda como a técnica da *template matching* usando correlação foi usada na simulação deste trabalho, faz-se necessário mostrar a forma padronizada com que os ATs e seus respectivos VCs foram nomeados na base de dados da simulação.

Os ATs foram nomeados e salvos usando as letras iniciais da frase PAC, seguida da ordem da locução e o caractere *underline* (_) antes das iniciais do locutor. Exemplo: *pac3_jj* é o nome do arquivo de áudio da frase PAC gravado da terceira pronúncia do locutor *jj*.

Os VCs extraídos dos ATs foram identificados apenas adicionando “v_” (a letra “v” + *underline).* Exemplo: *v_pac7_r* é o nome do arquivo do VC extraído do arquivo de áudio da frase PAC gravado da sétima pronúncia do locutor *r*.

Como, para fins de reconhecimento, o que interessa são os VCs, o Quadro 1 mostra 8 vetores de características dispostos em colunas, extraídos dos 8 ATs fornecidos por cada um dos 9 locutores, representados pelas letras “*jj”, “c”, “d”, “e”, “g”, “k”, “r”, “s”* e *“t”*, que pronunciaram a frase PAC 8 vezes. A próxima linha abaixo indica o limiar mínimo de correlação (LC) de cada locutor (*lc_jj*, no caso do locutor *jj*), previamente calculado antes dos testes, na fase de treinamento.

Quadro 1: Matriz de VCs dos locutores *jj, c, d, e, g, k, r, s, t*

v_pac1_jj	*v_pac1_c*	*v_pac1_d*	*v_pac1_e*	*v_pac1_g*	*v_pac1_k*	*v_pac1_r*	*v_pac1_s*	*v_pac1_t*
v_pac2_jj	*v_pac2_c*	*v_pac2_d*	*v_pac2_e*	*v_pac2_g*	*v_pac2_k*	*v_pac2_r*	*v_pac2_s*	*v_pac2_t*
v_pac3_jj	*v_pac3_c*	*v_pac3_d*	*v_pac3_e*	*v_pac3_g*	*v_pac3_k*	*v_pac3_r*	*v_pac3_s*	*v_pac3_t*
v_pac4_jj	*v_pac4_c*	*v_pac4_d*	*v_pac4_e*	*v_pac4_g*	*v_pac4_k*	*v_pac4_r*	*v_pac4_s*	*v_pac4_t*
v_pac5_jj	*v_pac5_c*	*v_pac5_d*	*v_pac5_e*	*v_pac5_g*	*v_pac5_k*	*v_pac5_r*	*v_pac5_s*	*v_pac5_t*
v_pac6_jj	*v_pac6_c*	*v_pac6_d*	*v_pac6_e*	*v_pac6_g*	*v_pac6_k*	*v_pac6_r*	*v_pac6_s*	*v_pac6_t*
v_pac7_jj	*v_pac7_c*	*v_pac7_d*	*v_pac7_e*	*v_pac7_g*	*v_pac7_k*	*v_pac7_r*	*v_pac7_s*	*v_pac7_t*
v_pac8_jj	*v_pac8_c*	*v_pac8_d*	*v_pac8_e*	*v_pac8_g*	*v_pac8_k*	*v_pac8_r*	*v_pac8_s*	*v_pac8_t*
lc_jj	*lc_c*	*lc_d*	*lc_e*	*lc_g*	*lc_k*	*lc_r*	*lc_s*	*lc_t*

Fonte: o autor (2022)

Importante perceber que, no caso de uma Seção Eleitoral com 400 eleitores, com cada um destes tendo fornecido seus treinamentos representados pelos 8 VCs, haveria não uma matriz 9 x 9 (72 VCs e 9 limiares), como a do Quadro 1, mas uma matriz de 9 x 400 (3.200 VCs e 400 limiares).

O Quadro 2 mostra o resultado das correlações entre os VCs do Quadro 1 com um VCs de prova do locutor *jj,* denominado de *v_pac9_jj* (nona pronúncia da frase PAC do locutor *jj*). Note-se que o Quadro 2 tem a mesma quantidade de linhas e colunas do Quadro 1.

Quadro 2: Matriz de correlações entre os VCs acima e o VC de prova *v_pac9_jj*

0,64	**0,44**	**0,33**	**0,37**	**0,25**	**0,40**	**0,28**	**0,46**	**0,38**
0,68	**0,37**	**0,33**	**0,45**	**0,31**	**0,43**	**0,32**	**0,32**	**0,39**
0,65	**0,36**	**0,34**	**0,32**	**0,29**	**0,41**	**0,29**	**0,35**	**0,40**
0,63	**0,39**	**0,40**	**0,45**	**0,32**	**0,29**	**0,33**	**0,36**	**0,42**
0,59	**0,41**	**0,48**	**0,31**	**0,34**	**0,45**	**0,34**	**0,31**	**0,41**
0,67	**0,32**	**0,49**	**0,34**	**0,27**	**0,42**	**0,34**	**0,44**	**0,40**
0,65	**0,36**	**0,40**	**0,37**	**0,29**	**0,48**	**0,32**	**0,42**	**0,42**
0,66	**0,32**	**0,31**	**0,39**	**0,31**	**0,36**	**0,34**	**0,36**	**0,41**
lc_jj	***lc_c***	***lc_d***	***lc_e***	***lc_g***	***lc_k***	***lc_r***	***lc_s***	***lc_t***
0,50	-	-	-	-	-	-	-	-

Fonte: o autor (2022)

Conforme os resultados das correlações dispostas no Quadro 2, como se quer reconhecer o locutor "*jj"* por meio de VAL (havendo necessidade de prévia identificação do locutor mediante identificador numérico), o sistema deverá fazer correlação do VC de prova *v_pac9_jj* apenas com os VCs de treinamento da coluna 1 do Quadro 1 e verificar se a MC atingiu o *lc_jj* (LC) de 0,50 (calculado usando a Eq. 12, Seção 4.8.5).

Como se pode ver no Quadro 2, a MC da coluna 1 foi de 0,68, superior ao LC de 0,50, o que já habilita o locutor "*jj"* como autêntico, usando-se o modelo VAL. No caso em análise,

foram realizadas as demais correlações da matriz, apenas para demonstrar que estas são bem inferiores quando comparadas com as corrrelações da coluna 1, o que demonstra o êxito da estratégia.

Também é possível perceber a diferença de custo computacional entre os modelos VAL e IAL, analisando a quantidade de correlações do Quadro 2. Para se implementar a VAL, na situação do Quadro 2, só há necessidade de 8 correlações, uma operação de busca de valor máximo (MC) e uma verificação de limiar (LC). Para se implentar a IAL, necessita-se de 64 correlações, uma operação de busca de valor máximo (MC) e uma verificação de limiar (LC).

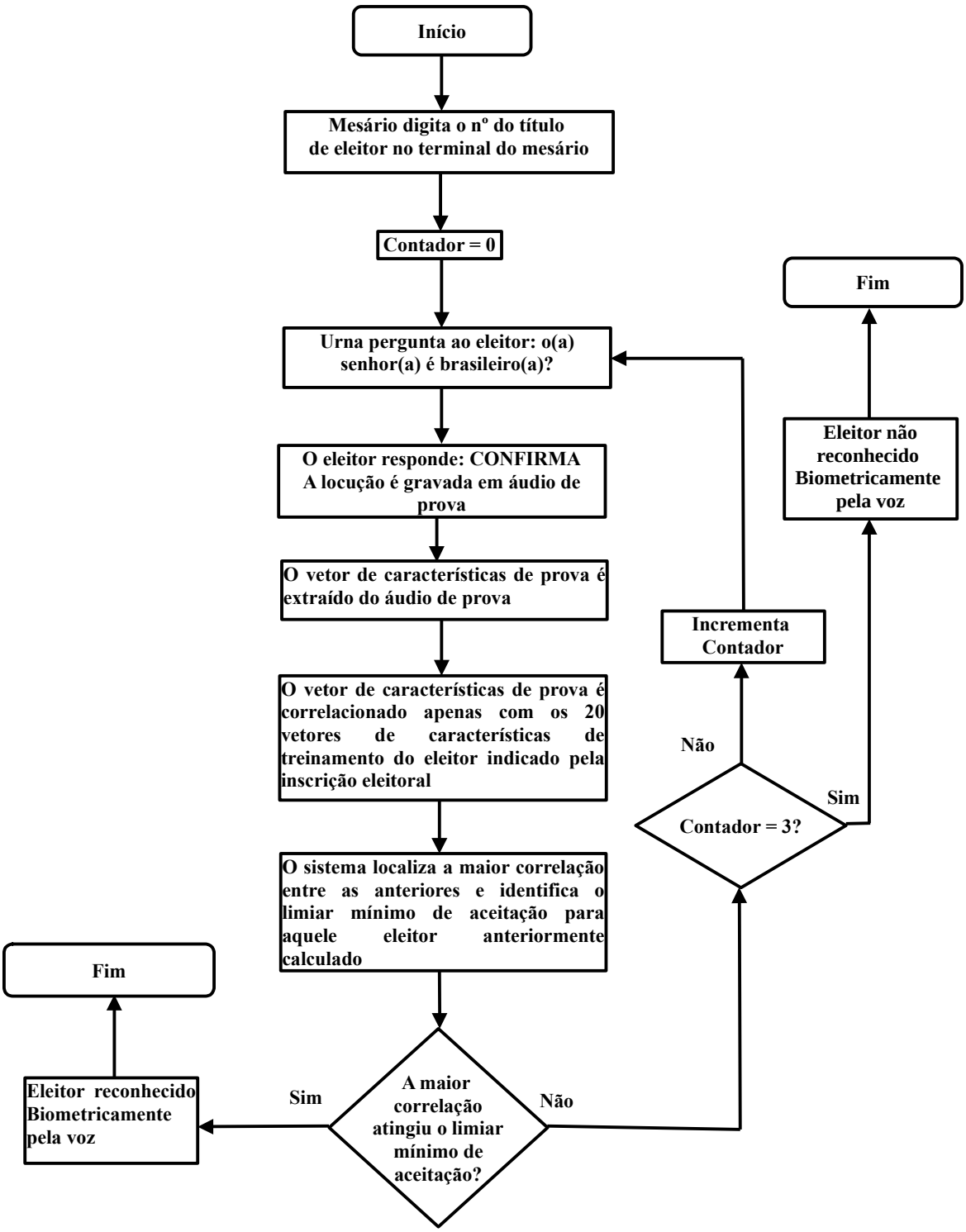
Note-se que, no caso da IAL, quanto maior a quantidade de locutores, maior o custo computacional, pois a tendência natural da ordem da matriz do Quadro 2 é crescer com o aumento de usuários. Por outro lado, o custo computacional quando se adota a VAL não depende da quantidade de usuários cadastrados na base de dados, permanecendo o mesmo custo, caso se aumente indistintamente o número de usuários.

Por uma questão de custo computacional, o algoritmo proposto em Figura 2 para reconhecimento de locutor foi modelado com uso da VAL. Entretanto, o mesmo algoritmo pode ser usado usando a IAL, apenas retirando do fluxograma a parte em que o mesário informa o número do título de eleitor à urna eletrônica e fazendo as correlações com toda a base.

Apesar do aumento de processamento, a vantagem da IAL é ganho de tempo, pois não há necessidade do mesário digitar os 12 números do título de eleitor para identificá-lo previamente. Este apenas pronuncia a frase e é automaticamente identificado pela MC, no caso de IAL do tipo incondicional ou é automaticamente identificado pela MC, caso essa MC atinja um determinado LC, no caso de IAL condicional. Cabe ao mesário apenas averiguar, mediante documento oficial, se o eleitor automaticamente identificado corresponde realmente ao eleitor reinvindicante, autorizando-o a votar ou simplesmente não habilitar o eleitor, caso o nome automaticamente identificado não coincida com o nome indicado no documento oficial apresentado pelo eleitor.

A Figura 2 detalha o fluxograma para reconhecimento biométrico de eleitor por meio de voz usando RAL do tipo VAL, dependente de texto, usando *template matching,* por meio de MC e LC.

Figura 2: Fluxograma proposto para reconhecimento de locutor



Fonte: o autor (2022)

2.7 CONCLUSÃO DO CAPÍTULO

Neste capítulo, é exposto um algoritmo para que o eleitor seja reconhecido pela voz no dia da Eleição, dissecando as possibilidades de se optar pelo próprio nome do eleitor ou não.

São traçadas diferenças entre ATs e APs, bem como o resultado dos seus processamentos, que são os VCs.

É feita uma pequena simulação calculando as correlações entre os VCs extraídos dos áudios da frase PAC do locutor *jj*, autor deste trabalho, e de mais 8 voluntários, identificados por letras, constatando-se que a técnica de *template matching* é viável neste tipo de problema, pelo resultado das correlações.

CAPÍTULO 3

3. RECONHECIMENTO DE PALAVRAS ISOLADAS

Neste capítulo, é discutida uma análise sobre a fase em que o eleitor, após ser reconhecido em termos biométricos pela voz usando RAL, é habilitado a escolher os seus candidatos, ou, em sendo o caso, votar em branco ou nulo, por meio do Reconhecimento de Palavras Isoladas (RPI) usadas para acionamento dos comandos da urna eletrônica.

3.1 AS TECLAS DA URNA ELETRÔNICA E OS SEUS COMANDOS

Há três tipos básicos de comandos a serem acionados na urna eletrônica. São os comandos de escolha, os de confirmação e os de correção. Os comandos de escolha são aqueles que invocam os números de 0 até 9 e o comando *branco*. Os comandos de confirmação e correção são respectivamente *confirmar* e *corrigir*.

Quando se usa teclas de computador para se acionar comandos na urna eletrônica, não existe um perigo considerável de quebra de sigilo, pois só o eleitor tem acesso às teclas que são acionadas através da movimentação do seu dedo indicador.

Como se quer substituir o acionamento das teclas por um acionamento pela voz, é preciso também se estabelecer uma relação entre os comandos acionados por estas teclas e os respectivos sinais vocálicos adotados.

Quando se intenta usar a voz para invocação dos comandos da urna eletrônica, o problema fica bem mais complexo, em razão de possíveis ruídos, da similaridade acústica entre as palavras escolhidas e também em razão do sigilo do voto, pois as pessoas presentes na Seção Eleitoral ouvem a voz do eleitor.

Destarte, um problema a ser enfrentado neste caso seria a questão do sigilo do voto, pois esse relacionamento entre comandos e palavras não pode ser fixo, como ocorre com as teclas da urna atualmente.

O Quadro 3 mostra os 13 comandos básicos da urna eletrônica acionados pelas respectivas teclas sugestivas. Quaisquer outros comandos ou invocação de funções são acionadas mediante combinação entre as teclas.

Quadro 3: Teclas e comandos da Urna Eletrônica (Terminal do Eleitor)

TECLAS	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Confirma**	**Corrige**	**Branco**
COMANDOS	***Escolhe o nº 0***	***Escolhe o nº 1***	***Escolhe o nº 2***	***Escolhe o nº 3***	***Escolhe o nº 4***	***Escolhe o nº 5***	***Escolhe o nº 6***	***Escolhe o nº 7***	***Escolhe o nº 8***	***Escolhe o nº 9***	***Confirma voto***	***Corrige voto***	***Vota em branco***

Fonte: o autor (2022)

Conforme Quadro 3, parece fácil fazer com que o eleitor pronuncie qualquer número em um microfone acoplado à urna eletrônica e o número apareça no terminal, atestando o reconhecimento da fala do eleitor.

Entretanto, o eleitor que quer votar em um candidato de número hipotético 18263 não pode pronunciar ao microfone as locuções UM, OITO, DOIS, SEIS e TRÊS. Se assim o fizer, todos na Seção Eleitoral passam a saber em quem o eleitor está votando e isso quebra o sigilo do voto, o que torna inviável o mecanismo.

Outro problema está relacionado com a similaridade acústica entre as palavras OITO e DOIS, TRÊS e SEIS e CINCO e BRANCO. Mesmo que se use uma Rede Neural para classificação destes padrões, a alta taxa de erro do sistema pode tornar o mecanismo inviável em termos práticos.

Então, como fazer o eleitor usar a voz, sem quebrar o sigilo do voto, com uma menor taxa de erro possível e com um baixo custo computacional? É o que é visto na próxima seção, por meio de um procedimento aqui denominado de correspondência posicional entre comandos e palavras.

3.2 AS PALAVRAS PARA COMANDAR A URNA ELETRÔNICA

Para que o eleitor consiga dar comandos à urna eletrônica mediante locuções vocais, sem que quaisquer ouvintes possam saber em quem o eleitor esteja votando, são escolhidas as

locuções relacionadas no Quadro 4, de maneira que haja permutação entre as palavras no sentido horizontal, após cada decisão de comando automaticamente reconhecida.

Quadro 4: correspondência entre palavras e comandos

URSO	BOLA	PÁSSARO	ARCO-ÍRIS	MAÇÃ	JACARÉ	TIGRE	NAVIO	SERRA	FEIJÃO	ABELHA	CORRIGE
0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*branco*	*corrige*

Fonte: o autor (2022)

Entretanto, não se chegou às locuções do Quadro 4 de forma trivial ou de maneira fácil, mas com vários testes, correlações cruzadas e autocorrelações realizadas, para saber qual o conjunto melhor de locuções que possam representar os comandos da urna eletrônica. Essas locuções foram escolhidas estrategicamente levando-se em conta cinco fatores:

a) O tamanho das palavras, para economizar o máximo de espaço na urna eletrônica (de duas a quatro sílabas);
b) A diferença acústica entre os sinais (quanto menor a semelhança entre palavras distintas, melhor para o bom funcionamento do sistema);
c) O fato delas poderem ser representadas por figuras, para facilitar a votação de eleitores analfabetos;
d) Facilidade de pronúncia;
e) Os resultados das correlações expostas em Tabela 10 (Capítulo 5);

No Quadro 4, sugere-se as 12 locuções, todas escritas em caixa alta, posicionadas em 1ª linha. Logo abaixo, na 2ª linha, posicionam-se os seus respectivos comandos, todos escritos em caixa baixa e em itálico.

A assoçiação entre palavas e comandos ocorre por meio das colunas. Convencionou-se que palavras em fundo azul serão sempre permutadas no sentido horizontal (mudam a posição das colunas), após cada decisão de comando de voz solicitada pelo eleitor. A única palavra que não será permutada é a palavra CORRIGE (fundo branco), que sempre será associada ao comando *corrige* na 12ª coluna.

Por outro lado, os comandos nunca permutam, mesmo após qualquer decisão de comando invocados por voz. Em outras palavras, o que está em fundo branco não permuta e o que está em fundo azul sempre permuta.

Nota-se que as locuções escolhidas de 2 a 4 sílabas são facilmente associadas a figuras sugestivas de sentido. Entretanto, neste exemplo, deixou-se de relacionar as locuções com as respectivas figuras por questões de economia textual.

O comando *corrige* apenas atua sob o último dígito escolhido, não no número do candidato inteiro. Isso significa que, se o eleitor quiser corrigir o número do candidato por inteiro, deve pronunciar a palavra CORRIGE tantas vezes quanto for o número de dígitos do candidato.

A título de exemplo, com base no Quadro 4, suponha-se que o eleitor esteja votando para presidente da república e resolva votar no candidato cujo número seria *99*. Conforme a Quadro 4, o mesmo deve pronunciar a locução FEIJÃO. Em seguida, uma operação de permutação deve ser realizada e uma nova correspondência é disponibilizada para o eleitor, conforme o Quadro 5.

Quadro 5: correspondência entre palavras e comandos

NAVIO	ABELHA	URSO	ARCO-ÍRIS	PÁSSARO	BOLA	FEIJÃO	JACARÉ	TIGRE	SERRA	MAÇÃ	CORRIGE
0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*branco*	*corrige*

Fonte: o autor (2022)

Note-se que agora o eleitor está em busca do número 9. Diante desta nova correspondência impressa no Quadro 5, qual é a palavra correspondente nesta nova correspondência para se invocar o comando 9? No caso, a palavra SERRA.

Percebe-se que a pronúncia de qualquer palavra (FEIJÃO, por exemplo) pode representar a invocação de qualquer dos comandos *0,1,2,3,4,5,6,7,8,9* e *branco,* dependendo do resultado da permutação aleatória entre as palavras e respectivos comandos da urna eletrônica, após cada ação de comando.

É preciso enfatizar que o voto em branco também deve ser sigiloso. Não há nenhuma razão plausível para que o fiscal de partido, mesário ou cabo eleitoral tenha a certeza se um eleitor votou ou não em branco. Em razão disso é que a escolha do comando *branco* deve ser feita para cada dígito escolhido, ou seja, o eleitor deve escolher o camando *branco* tantas vezes quantas forem a quantidade de dígitos do número do candidato.

De fato, o comando *branco,* que sugere que o eleitor não quer votar em ninguém, aparece tanto no Quadro 4 (relacionado à palavra ABELHA) como no Quadro 5 (relacionado à palavra MAÇÃ) e está posicionalmente relacionado com palavras advindas da operação de permutação.

Se o eleitor escolher um número de candidato inexistente (voto nulo) ou o comando *branco* para todos os dígitos do número do candidato, a urna eletrônica deve pedir a

confirmação de escolha ao eleitor mediante a disponibilização da correspondência do Quadro 11, em Seção 3.3.1.

Suponha-se que o eleitor esteja votando para o cargo de Prefeito (no caso, dois números precisam ser informados à urna eletrônica) e ele queira votar em branco. No caso dos Quadros 4 e 5, o eleitor teria que pronunciar ABELHA e, em seguida, a palavra MAÇÃ, respectivamente. O sigilo do voto estaria preservado, pois quem ouviu o eleitor pronunciar as referidas locuções não teria como saber que o eleitor haveria votado em branco.

Escolhido o número do candidato a prefeito (no caso o número *99*) ou voto em branco (escolhendo o comando *branco* duas vezes por meio da pronúncia das locuções ABELHA e MAÇÃ), resta ao eleitor confirmar seu voto usando sua voz tantas vezes quantas forem a quantidade de dígitos do número do candidato, conforme Seção 5.6 (Uma trava de segurança para confirmar o voto).

Note-se que no caso da votação usando teclas, todas as opções estão disponíveis ao eleitor, bastando que o eleitor movimente seu dedo e escolha uma das 13 opções (*0,1,2,3,4,5,6,7,8,9*, *corrige, branco, confirma*). No caso da votação por voz, até o momento foram disponibilizados apenas os 12 comandos *0,1,2,3,4,5,6,7,8,9, corrige, branco,* excluindo-se o comando *confirma.*

Essa exclusão inicial do comando *confirma* se justifica em razão da necessidade de se aumentar a taxa de acerto do sistema, optando-se por disponibilizá-lo ao usuário somente após a escolha de todos os dígitos do número do candidato ou após a escolha de votar em branco tantas vezes quantas forem a quantidade de dígitos do número do candidato.

Entretanto, ao se disponibilizar ao eleitor a opção de confirmar qualquer decisão, precisa-se também lhe dar a opção de negar esta confirmação. E é justamente aí onde existe a necessidade de modificar um pouco a lógica da votação manual para a votação pela voz. É o que é visto na Seção 3.2.1.

3.2.1 Os comandos de controle e correção

Os comandos *confirma* e *corrige* da urna eletrônica se constituem nos seus dois principais comandos de controle e correção. Sua lógica se resume em confirmar (por meio da execução do comando *confirma*) ou negar (por meio da execução do comando *corrige*) à

pergunta que se está fazendo ao eleitor. É através desses comandos que a urna eletrônica se comunica com o usuário fazendo perguntas por meio de questionamentos do tipo:

"Você quer votar em branco?"
"confirma: para votar em branco"
"corrige: para não votar em branco"

"Você quer votar nulo?"
"confirma: para votar nulo"
"corrige: para não votar nulo"

"Você quer votar no candidato *12345*?"
"confirma: para sim"
"corrige: para não"

Note-se que, em termos funcionais, o comando *corrige* aplicado no contexto acima equivale a reiniciar todo o processo de votação, desde o início. A mudança em relação à lógica da votação por meio de teclas para a de voz é justamente modificar a denominação deste comando *corrige* aplicado a este contexto para o comando *reinicia.* Isto se justifica para diferenciar o comando *corrige* aplicado no contexto da escolha dos números pela voz, conforme Quadro 4 e Quadro 5.

Destarte, tem-se dois comandos de correção aplicados em momentos diferentes no processo de votação pela voz: primeiro, durante a escolha dos números do candidato, tem-se o comando *corrige,* que simplesmente apaga o último dígito escolhido pela voz; segundo, após o eleitor já ter escolhido todos os dígitos que compõem o número do seu candidato, sejam eles brancos ou nulos, tem-se o comando *reinicia,* que começa todo o processo desde o início.

O comando *confirma* pode ser acionado pela própria palavra que tem relação de significado com o nome do comando, (no caso, a palavra CONFIRMA), pois não existe nenhuma quebra de sigilo em saber que o eleitor confirmou alguma coisa sem que se saiba especificamente que coisa foi essa (voto branco, nulo ou escolha de número de candidato), pois sempre há a necessidade, para todos os eleitores, de confirmar alguma decisão no final da votação.

Portanto, se o eleitor quiser confirmar o seu voto impresso na tela da urna, ele teria que pronunciar a palavra CONFIRMA. Por outro lado, se o eleitor não quiser confirmar, teria que

pronunciar qual palavra? A resposta quase que automática seria a palavra REINICIA, mas, como será visto adiante, esta opção é inviável em razão da similaridade acústica entre as palavras CONFIRMA e REINICIA.

Note-se no Quadro 6 as duas palavras separadas para facilitar e entender o problema.

Quadro 6: Comparação fonética entre as palavras CONFIRMA e REINICIA

CON	**FIR**	**MA**
REI	**NICI**	**A**

Fonte: o autor (2022)

Ao se pronunciar os trechos FIR e NICI, na coluna 2 do Quadro 6, constata-se que há uma coincidência posicional da vogal I, bem como no quesito tonicidade, pois as sílabas CI e FIR são as sílabas mais fortes. No caso das sílabas terminais MA e A, ocorre coincidência vocal, com tonicidade fraca.

O que mais diferencia de fato estas duas locuções são as sílabas iniciais CON e REI. Portanto, pode-se dizer, com alguma margem de erro, que as palavras CONFIRMA e REINICIA são 77% semelhantes e 33% diferentes em termos acústicos. Esse fato pode tornar o sistema vulnerável, pois se o sistema confundir a palavra REINICIA com a palavra CONFIRMA na situação em que o eleitor queria reiniciar todo o processo, não há como reverter a situação. O erro neste caso seria fatal, pois o eleitor teria votado em um candidato que não queria votar em virtude da má escolha dos comandos vocálicos por quem elaborou o sistema.

No caso, seria conveniente encontrar uma palavra estrategicamente escolhida para substituir uma das duas. Diante disto, na simulação de votação usando Scilab (SCILAB, 2022), resolveu-se trocar a palavra REINICIA pela palavra PÁSSARO. A razão pode ser aferida pela análise do Quadro 7:

Quadro 7: comparação fonética entre as palavras CONFIRMA e PÁSSARO

CON	**FIR**	**MA**
PÁS	**SA**	**RO**

Fonte: o autor (2022)

Percebe-se do Quadro 7 que, ao se adotar a palavra PÁSSARO em substituição à palavra REINICIA, aumenta-se consideravelmente a distância acústica entre elas, pois não se tem mais coincidências de vogais e nem de tonicidade em mesma posição silábica. Estes fatos fazem com que as correlações cruzadas entre os VCs destas locuções sejam baixas quando comparadas com as autocorrelações entre os VCs, o que é favorável para a performance do modelo.

3.3 DIFERENÇA ENTRE RAL E RPI

Para o acionamento de comandos da urna eletrônica através de RPI, os ATs e seus respectivos VCs são gerados da mesma forma que ocorre no RAL, com certas diferenças que mereçem ser enfatizadas.

A primeira diferença é em relação à duração da pronúncia das palavras. Para a pronúncia dos 3 primeiros nomes do eleitor ou da frase PAC como parâmetro de RAL, foi adotado o tempo de 1,7 segundos. Já para a pronúncia das palavras (de 2 a 4 sílabas) escolhidas para serem usadas como invocação dos respectivos comandos da urna, foi adotado o tempo de 0,8 segundos.

A segunda diferença entre RAL e RPI paira na questão do tipo e quantidade de correlações. Como se está adotando o método VAL, só há necessidade de fazer correlações do VC de prova com os VCs de treinamento extraídos dos áudios de treinamento pronunciados pelo eleitor que reivindica a autenticidade (identificado previamente por identificador numérico). Não há necessidade de se averiguar posição de coluna em que a MC tenha se estabelecido para reconhecimento de locutor, mas apenas se a MC encontrada alcança determinado LC, calculado previamente, na fase de treinamento.

Já para o RPI, além de se fazer correlações entre o VC de prova com todos os VCs extraído dos ATs de todas as locuções de treinamento pronunciadas pelo locutor previamente identificado, detectando-se a MC, precisa-se saber qual posição de coluna essa MC se encontra, para identificação da palavra que foi pronunciada e consequentemente do comando a ser acionado.

Uma terceira diferença entre RAL e RPI está relacionada com o objetivo do LC, que deve ser calculado tanto para reconhecimento de locutor, como para reconhecimento das palavras isoladas para comandar a urna eletrônica.

Para reconhecimento de palavras isoladas, o limiar mínimo de correlação se destina a impedir que o microfone capte qualquer barulho, ruído ou palavra não previamene treinada e o sistema acuse o aceite de comando baseado unicamente na MC identificada na matriz de correlações. Para reconhecimento de locutor, o LC serve para aceitar ou rejeitar o locutor reivindicante, mediante número de tentativas previamente estabelecidas.

Note-se que no RPI, não há rejeição definitiva do reconhecimento da palavra. O locutor tem a possibilidade indeterminada de tentar atingir o limiar e, em o atingindo, corrigir o comando, se for o caso.

Para que o mecanismo da permutação de palavras garanta o sigilo do voto, há a necessidade de trabalhar com dois tipos de correspondências: a pública e a privada, como será visto na Seção 3.3.1 (Algoritmo para RPI do eleitor).

3.3.1 Algoritmo para RPI do eleitor

Reconhecido o eleitor por meio de RAL usando o LC, ele estará habilitado a votar usando palavras pré-estabelecidas, conforme Quadro 8. A partir deste momento, as palavras que serão pronuniadas e reconhecidas se resumem àquelas previamente treinadas e gravadas no Cartório Eleitoral, usadas para escolher comandos da urna eletrônica (podendo ser também acionados por teclas) de forma sonoramente codificada pela correspondência posicional entre palavras e comandos impressas na tela da urna eletrônica.

O Quadro 8 ilustra os 20 VCs de treinamento extraídos da gravação das 20 locuções pronunciadas de cada palavra (ao todo, 240 vetores), pelo mesmo locutor "*jj"*, e representa a correspondência pública, gravada internamete na urna eletrônica, sendo que nesta correspondência não ocorre qualquer espécie de permutação e é invisível para os usuários no momento da votação.

Quadro 8: Relação posicional entre os VCs e as *strings* das palavras

v_f1	***v_a1***	***v_u1***	***v_b1***	***v_p1***	***v_m1***	***v_ai1***	***v_j1***	***v_t1***	***v_n1***	***v_s1***	***v_c1***
v_f2	***v_a2***	***v_u2***	***v_b2***	***v_p2***	***v_m2***	***v_ai2***	***v_j2***	***v_t2***	***v_n2***	***v_s2***	***v_c2***
v_f3	***v_a3***	***v_u3***	***v_b3***	***v_p3***	***v_m3***	***v_ai3***	***v_j3***	***v_t3***	***v_n3***	***v_s3***	***v_c3***
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
v_f20	***v_a20***	***v_u20***	***v_b20***	***v_p20***	***v_m20***	***v_ai20***	***v_j20***	***v_t20***	***v_n20***	***v_s20***	***v_c20***
lc_f	***lc_a***	***lc_u***	***lc_b***	***lc_p***	***lc_m***	***lc_ai***	***lc_j***	***lc_t***	***lc_n***	***lc_s***	***lc_c***
FEIJÃO	**ABELHA**	**URSO**	**BOLA**	**PÁSSARO**	**MAÇÃ**	**ARCO-ÍRIS**	**JACARÉ**	**TIGRE**	**NAVIO**	**SERRA**	**CORRIGE**

Fonte: o autor (2022)

A linha 21 do Quadro 8 representa o limiar de cada palavra. As correlações serão todas entre os VCs extraídos de palavras pronunciadas pelo mesmo eleitor, podendo ser do tipo cruzadas (palavras diferentes) ou autocorrelacionadas (mesmas palavras), sendo estas em

grande parte as de maior valor de correlação e aquelas em grande parte as de menor valor de correlação.

O eleitor visualizará a correspondência privada, conforme Quadro 9, e pronunciará a palavra correspondente ao comando que deseja. O áudio de prova será então gravado e seu VC será extraído e correlacionado com todos os VCs de treinamento extraídos da pronúncia das 12 palavras, conforme a correspondência pública expressa no Quadro 8.

Quadro 9: Relação posicional entre as *strings* das palavras e os comandos da urna

JACARÉ	URSO	ABELHA	NAVIO	TIGRE	PÁSSARO	SERRA	FEIJÃO	ARCO-ÍRIS	BOLA	MAÇÃ	CORRIGE
0	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***branco***	***corrige***

Fonte: o autor (2022)

A permutação ocorre apenas entre as palavras em células de fundo azul, postas na primeira linha da correspondência privada, representada pelo Quadro 9, visível ao usuário e que deve permutar após cada escolha de dígito do número do candidato em quem o eleitor deseja votar, o que garante o sigilo da votação.

O eleitor só poderá pronunciar ao microfone as locuções da linha 1 do Quadro 9 (correspondência privada), com a finalidade de acionar o comando correspondente na segunda linha, em branco, imediatamente abaixo.

Ao se calcular a permutação simples por meio do fatorial de 12, obtem-se 479.001.600 possibilidades distintas de correspondências privadas possíveis após cada ação de comando, sendo praticamente impossível prever ou adivinhar qualquer resultado de correspondência permutada.

No caso da correspondência exposta no Quadro 9, se o eleitor estiver querendo votar em branco em uma eleição para o cargo de prefeito, ele deve pronunciar a palavra MAÇÃ. O microfone então capta o áudio e o sistema, após a devida amostragem e digitalização do som, extrairá o VC de prova e fará correlação com os 240 VCs de treinamento posicionados na nas primeiras 20 linhas da matriz do Quadro 8, criando uma nova matriz de resultados de correlações de tamanho 20 x 12, cujo valor máximo (MC) deverá ser identificado para se obter sua posição de coluna. Esta posição deverá ser uma das células da coluna 6 do Quadro 8, pois é nesta posição onde se encontram os 20 VCs de treinamento da palavra MAÇÃ.

É claro que esta MC só terá significado se estiver acima ou igual ao LC indicado na linha de limiares, no caso a linha 21. Caso não se atinja nenhum dos limiares da linha 21, o

sistema deve pedir para repetir a operação até que se atinja um dos limiares. Caso, o sistema erre o comando, o eleitor deverá pronunciar a palavra CORRIGE.

Note-se que, conforme Tabela 10, a palavra CORRIGE é a que tem maior probabilidade de acerto, pois se mostra, pelos resultados das correlações catalogadas, a que mais se distingue das demais em termos acústicos, por não ter nenhum falso negativo ou falso positivo calculado na fase de treinamento (momento em que se calcula também o LC).

Dando continuidade ao algoritmo, após calculada a MC e verificado se houve o atingimento do LC, o sistema então captura a *string* correspondente à coluna onde se estabeleceu a MC, no caso a *string* MAÇÃ e verifica na correspondência do Quadro 9, em que coluna esta *string* se encontra e, em a identificando, procura o comando imadiatamente abaixo, ou seja, o comando *branco.*

Como o eleitor está votando para prefeito e este cargo possui dois dígitos, nova tabela de correspondência permutada deverá ser impressa para que o eleitor escolha o segundo dígito, que, no caso, será nenhum, pois o eleitor quer votar em branco. Então nova palavra que corresponda ao comando *branco* deverá ser pronunciada.

Após escolher votar em branco para a eleição de prefeito, a urna eletrônica deverá perguntar se o eleitor realmente quer votar em branco. Nesta situação, a urna eletrônica imprimirá na tela a seguinte pergunta:

“Você quer realmente votar em branco?"

“Se sim responda CONFIRMA”
“Se não responda: PÁSSARO”

Novamente, haverá a necessidade de trabalhar com duas tabelas, uma pública e uma privada. O Quadro 10 ilustra a tabela pública, que não aparece na tela da urna eletrônica, mas é gravada internamente na urna eletrônica, onde devem estar posicionados os VCs.

Quadro 10: Relação posicional entre os VCs e as *strings* das palavras

v_confirma1	***v_passaro1***
v_confirma2	***v_passaro2***
v_confirma3	***v_passaro3***
⋮	⋮
v_confirma20	***v_passaro20***
lc_confirma	***lc_passaro***
CONFIRMA	**PÁSSARO**

Fonte: o autor (2022)

Note-se, no Quadro 10, que a probabilidade de erro neste tipo de reconhecimento de voz é pouco provável, em razão da quantidade de padrões a serem reconhecidos, no caso duas palavras (CONFIRMA e PÁSSARO), bem como em razão da distinção fonética entre as palavras.

O Quadro 11 então é impresso na tela da urna eletrônica, sugerindo ao usuário que só há duas locuções possíveis de pronunciar: CONFIRMA e PÁSSARO. Isso significa que, após a gravação do AP e extração do VC de prova, este só deverá ser correlacionado com os 40 VCs (20 VCs ligados à palavra CONFIRMA e 20 VCs ligados à palavra PÁSSARO).

Quadro 11: Relação posicional entre as *strings* das palavras e os comandos da urna

CONFIRMA	**PÁSSARO**
confirma	***reinicia***

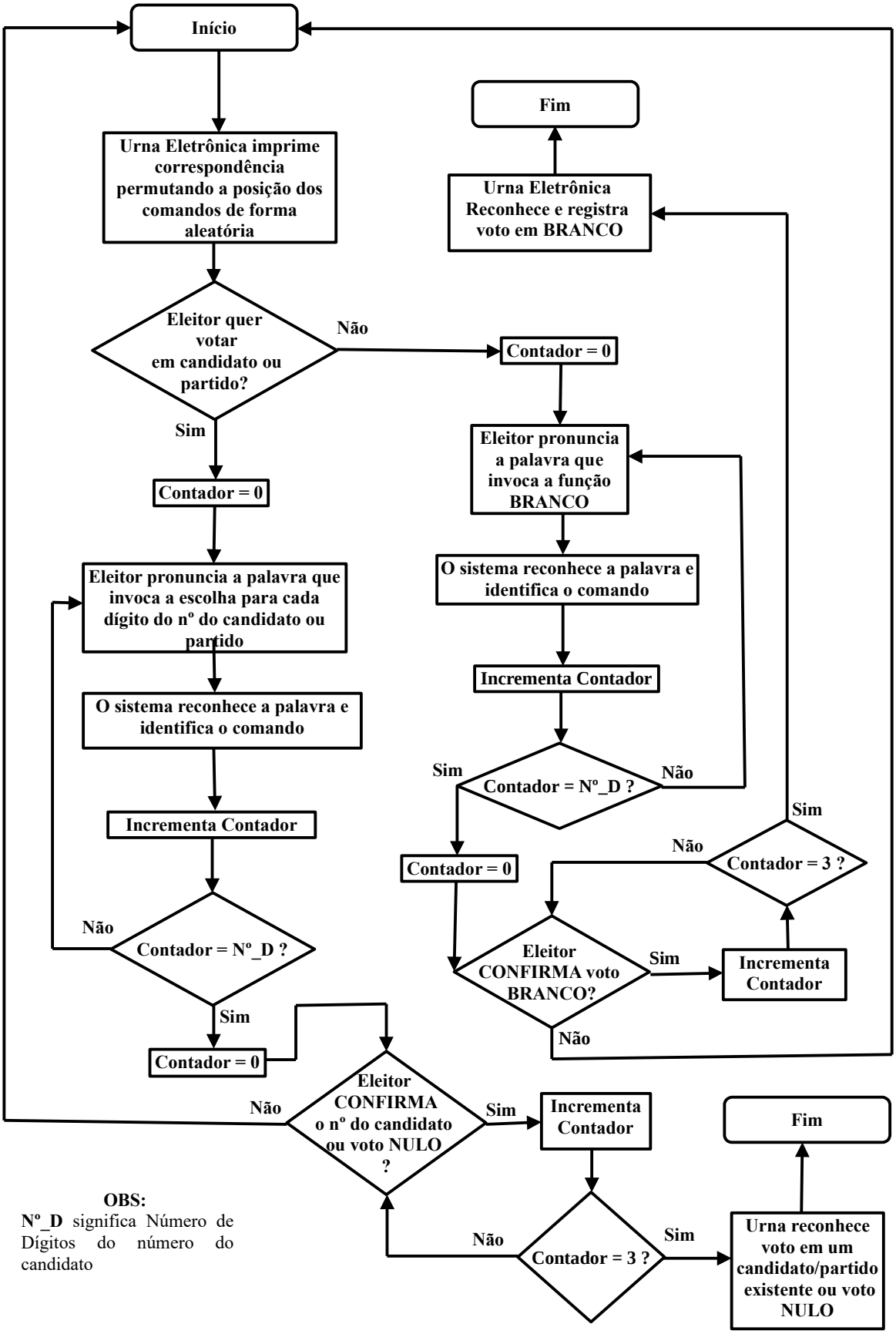
Fonte: o autor (2022)

O comando *reinicia* no caso do Quadro 11 tem um significado diferente do comando *corrige* usado no Quadro 9. Este corrige o último dígito do número do candidato escolhido pelo eleitor. Aquele reinicia a escolha dos dígitos desde o início, limpando tudo que vinha sendo feito anteriormente.

Note-se que não há necessidade de fazer permutação entre as palavras CONFIRMA e PÁSSARO, pois não há perda de sigilo do voto quando se escuta um eleitor pronunciar a locução CONFIRMA, sem saber o que especificamente foi confirmado.

A Figura 3 demostra o fluxograma do algoritmo para se chegar ao número do candidato, por meio de RPI.

Figura 3: Fluxograma para escolhas do número do candidato



Fonte: o autor (2022)

Na Figura 3, Nº_D é uma variável que recebe a quantidade de dígitos que o cargo no qual o eleitor está votando possui. Para Presidente, Governadores, Prefeitos e Partidos são 2 dígitos. Para Senadores são 3 dígitos. Para Deputados Estaduais são 5 dígitos. Para Deputados Federais são 4 dígitos e para Vereadores são 5 dígitos

Da mesma forma que foi realizada para escolha dígitos do número do candidato, a resposta do eleitor deverá ser gravada em formato digital e o sistema deverá extrair seu VC de prova e deverá correlacioná-lo com todos os VCs de treinamento da matriz 20 X 2, do Quadro 10. Esta correspondência do Quadro 10, chamada de correspondência pública, não permuta e é invisível ao usuário.

Uma nova matriz de mesmo tamanho da exposta no Quadro 10 será criada, mas conterá o resultado das correlações respectivas. Se o eleitor falar a palavra CONFIRMA, a MC deverá estar em alguma célula da coluna 1 do Quadro 10. Se o eleitor pronunciar a palavra PÁSSARO, a MC estará em alguma célula da coluna 2 do Quadro 10. Novamente um dos LCs (que estão na linha 21) deve ser atingido (*lc_confirma* ou *lc_passaro*), sob pena de não reconhecimento do comando e solicitação de nova pronúncia de palavra.

3.3.2 Usando uma senha vocal com RPI

Antes da biometria das impressões digitais, ocorriam raramente algumas situações não previstas. Por exemplo: o eleitor chegava à sua seção eleitoral no dia da eleição e, quando o Presidente de Mesa identificava o eleitor, digitando o número do título de eleitor no terminal do mesário, a urna eletrônica informava que o eleitor já havia votado.

Diante disto, pode-se inferir duas situações: ou o eleitor estava tentando se passar por outro; ou o Presidente de Mesa havia digitado o número do título de eleitor errado no terminal do mesário e teria habilitado inconscientemente outro eleitor a ir votar, sendo muito difícil saber a verdade, em razão de ausência de um processo de biometria.

Com o surgimento da biometria das impressões digitais, esses problemas foram reduzidos de forma drástica, mas não foi capaz de incluir aqueles eleitores com impressões digitais desfiguradas ou mesmo com ausência delas.

Nenhum método biométrico é capaz de dar total certeza de reconhecimento, pois as decisões computacionais são baseadas em fatores probabilísticos.

Por outro lado, pode-se dar ao eleitor a oportunidade de fornecer uma senha numérica para acesso no dia da eleição, se houver uma desconfiança que algum eleitor ilegítimo tenha tentado se passar por ele em eleições anteriores.

A Justiça Eleitoral também pode, finalizadas as eleições e baseadas nos arquivos gravados nas urnas eletrônicas (os arquivos de *log*), comunicar ao eleitor, que, apesar dele ter justificado a ausência de seu voto, alguém, não identificado biometricamente, se passou pelo mesmo no dia da eleição, sem que houvesse impugnação por parte dos fiscais de partido.

Neste caso, se o eleitor verdadeiro confirmar que faltou às eleições, com certeza, tem-se uma fraude ou um erro do mesário. Destarte, o eleitor faltante deve se dirigir ao Cartório Eleitoral para criação de uma senha numérica de acesso, para evitar futuras tentativas de fraudes ou erros.

É óbvio que, no dia da eleição, o eleitor deve se lembrar desta senha numérica, sob pena de não poder votar, devendo ser orientado pelo mesário a ir ao Cartório Eleitoral para averiguar qual sua senha foi cadastrada.

É claro também que a senha numérica escolhida pelo eleitor será acionada pela pronúncia das palavras responsáveis pelo acionamento dos comandos, após cada permutação destes, o que garante duas coisas: primeiro que ninguém que escutar a pronúncia das locuções vai saber qual a senha numérica, em razão da permutação; segundo por causa das características biométricas dos sinais vocálicos que só pertencem aquele eleitor. Ou seja, mesmo que o eleitor forneça sua senha numérica para outra pessoa, o sistema deve não aceitar, em razão das diferenças dos dados biométricos vocálicos.

Como as eleições ocorrem a cada dois anos, o recomendável é que essa senha numérica seja bastante conhecida por cada eleitor individualmente: por exemplo, o dia e mês de nascimento do eleitor. Se essa opção fosse adotada, o problema do esquecimento da senha seria contornado, pois dificilmente um cidadão maior de 16 anos não iria se lembrar de seu dia e mês de nascimento.

Note-se que o dia e mês de nascimento do eleitor sempre será o mesmo, mas, devido à correspondência posicional entre as palavras que invocam os comandos da urna eletrônica, as palavras pronunciadas em eleições consecutivas dificilmente seriam as mesmas, o que torna impossível descobrir a senha numérica escolhida. Quem estivesse ouvindo o eleitor, saberia que a senha seria o mês e dia de nascimento, mas não saberia qual dia e qual mês seria esse. Outra opção seria adotar os três primeiros números do CPF.

A possibilidade de criar esta senha é uma das vantagens de se usar voz em vez de impressões digitais como padrão biométrico.

Informar uma senha numérica usando RPI por meio da pronúncia de palavras permutadas para invocação de comandos, que não têm relação de sentido com as respectivas palavras, como fora anteriormente demonstrado, é provavelmente uma forma de duplo fator de autenticação simultêneo, ou seja, a verificação dos dois padrões (a numérica e a biométrica vocal) ocorrerm de forma simultânea e não consecutiva, como normalmente é feita. Em outras palavras, a biometria vocal é verificada no mesmo instante da verificação da senha numérica.

3.3.3 As teclas da urna eletrônica e o RPI

O fato de se adotar RPI para acionamento de comandos da urna eletrônica, evitando contato físico entre máquina e pessoa, não implica necessariemente que as teclas da urna eletrônica sejam impossibilitadas de serem usadas.

É perfeitamente possível, mediante código, estabelecer uma ordem de prevalência entre comandos ordenados mediante voz e comandos ordenados por teclas, de sorte que esta prevaleça.

Essa medida é extremamente importante para garantir a continuidade da eleição, caso haja problemas de reconhecimento de voz causados, por exemplo, pelos seguintes fatos:

1 - perda de voz do eleitor em razão de alguma doença em suas cordas vocais;

2 - defeito no microfone que capta a voz dos eleitores;

3 - barulhos excessivos e incontroláveis na seção eleitoral;

4 - outros fatores que impeçam a captação e reconhecimento da voz no dia da eleição;

3.4 CONCLUSÃO DO CAPÍTULO

Neste capítulo, analisa-se a permutação posicional de palavras para acionamento de funções da urna eletrônica, com o intuito de fazer com que o eleitor escolha os números dos seus candidatos, vote em branco ou nulo, sem que quaisquer ouvintes saibam as ações reais que o eleitor tenha ordenado à urna eletrônica.

Verifica-se também que, em regra, as palavras escolhidas não podem guardar nenhum significado com seus respectivos comandos, para garantir o sigilo do voto. Exceção à palavra CONFIRMA.

A única relação que deve existir entre as palavras e os comandos da urna eletrônica é de ordem posicional e percebida pela visão do eleitor no momento do voto.

Constata-se que a escolha dessas locuções devem estar relacionadas com as diferenciações acústicas entre as mesmas, de sorte que o resultado das correlações cruzadas entre os VCs entre elas sejam as menores possíveis e que o resultado das autocorrelações sejam as maiores possíveis.

Enfatiza-se também que outros fatores devem ser considerados na escolha das locuções: facilidade de pronúncia e possibilidade de representação por figuras sugestivas, para facilitar a compreensão do mecanismo pelo maior número de pessoas, incluindo os analfabetos.

Chega-se também à conclusão de que usar as locuções CONFIRMA e REINICIA, para confirmar ou negar as ações de comando não se mostra a opção mais conveniente, em virtude da similaridade sonora entre as locuções, razão pela qual sugeriu-se a palavra PÁSSARO para ser usada no lugar da palavra REINICIA.

CAPÍTULO 4

4. EXTRAÇÃO DAS CARACTERÍSTICAS DOS SINAIS

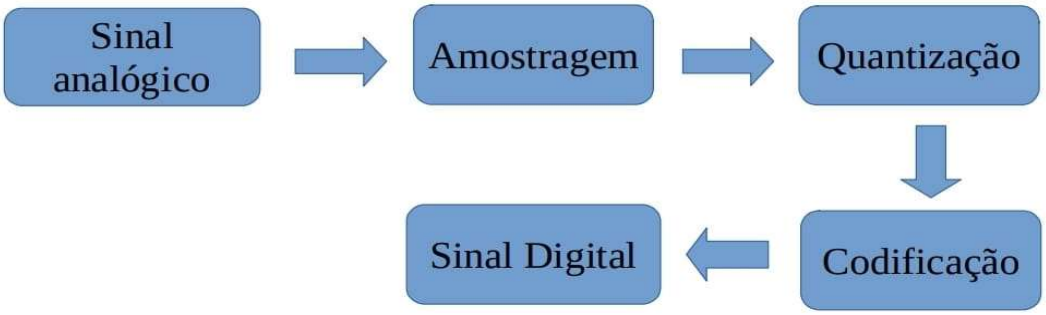
Neste capítulo, serão analisadas as principais técnicas usadas na atualidade para extrair dados da voz humana, de forma que uma atenção especial será dada à técnica *mel cepstral* (as *MFCCs*), adotada nas simulações deste trabalho.

4.1 TEOREMA DA AMOSTRAGEM

Os sinais produzidos pelo trato vocal humano são de características analógicas, ou seja, são contínuos e variantes no tempo, podendo assumir infinitos níveis. Por várias razões, os sinais digitais são muito mais maleáveis e mais fáceis de se trabalhar, razão pela qual o mundo partiu em meados dos anos oitenta para uma tendência à digitalização quase que total dos sistemas eletrônicos.

Para que o sinal de voz humana seja analisado por meio de um computador digital (no caso, a urna eletrônica), faz-se necessário sua transformação em sinal digital por meio de um processo de digitalização, que inclui a amostragem, a quantização e a codificação, como sugere o esquema da Figura 4.

Figura 4: Esquema de digitalização do sinal sonoro



Fonte: o autor (2022)

Segundo Lathi (1989), a amostragem é feita por meio da multiplicação de um sinal analógico por um trem de impulsos de determinada frequência. Esta frequência é conhecida como frequência de amostragem f_s.

Entretanto, pelo teorema de Nyquist (teorema da amostragem), esta frequência de amostragem não pode ser estabelecida em qualquer valor, de sorte que deve estar compreendida no mínimo pelo dobro da frequência máxima do sinal analógico ($f_s > 2f_{max}$), onde f_{max} significa a frequência máxima do sinal analógico. Caso isso não ocorra, o sinal não transmitirá a informação de forma eficiente.

Por outro lado, uma das consequências da amostragem do sinal analógico é o efeito de *aliasing,* que é a superposição de ciclos espectrais repetidos. Uma forma de amenizar o referido problema, seria amostrar o sinal em taxas mais altas.

A solução geralmente adotada é aplicar um filtro passa-baixas com frequência de corte estabelecida pela metade da frequência de amostragem $f_s/2$, para eliminar frequências superiores ao dobro da frequência de amostragem.

No caso do reconhecimento automático de fala, há uma tendência de se optar pela máxima taxa de amostragem para aumentar a precisão dos sistemas (KUMAR, 2022). Entretanto, esse aumento de taxa de amostragem requer custo computacional, algo que requer certo sopesamento, quando se trata de aumentar as funcionalidades da urna eletrônica, hajam vista suas limitações em termos de compartilhamento de recursos e segurança.

Após a amostragem, o sinal deve ser quantizado em uma quantidade fixa de níveis. Estes níveis são pontos do sinal para onde todas as amostras devem ser arredondadas, ou seja, as amostras mais próximas destes níveis devem a eles serem igualadas. Isso traz um nível de erro ao sinal que não pode ultrapassar certos limites no sentido de não prejudicar a qualidade mínima da informação.

Diante do sinal quantizado, parte-se para a codificação em *bits* (zeros ou uns) por meio de algum método, como o *PCM (Pulse-Code Modulation).* Após a codificação, cada nível estará representado por uma sequência de *bits*, o que implica na digitalização total do sinal.

4.2 OS NÍVEIS DAS CORRELAÇÕES E OS LCs

Em um sistema que usa correlação estatística para identificar semelhanças e reconhecer padrões, como é o caso deste trabalho, pode-se identificar dois tipos de correlações: as correlações altas (aqui denominadas de autocorrelações), que conseguem atingir o LC, atestando a presunção do reconhecimento do padrão sob teste; e as correlações baixas (aqui denominadas de correlações cruzadas), que não atingem o LC, atestando o não reconhecimento do padrão sob teste.

É na fase de treinamento que se calcula o LC, envolvendo todos os vetores de características das locuções treinadas, por meio da expresão matemátca descrita na Seção 4.8.5. Nesta fase, já se sabe quais VCs estão associados a determinado padrão de voz pronunciada, algo que não acontece na fase de teste, quando o VC extraído é correlacionado com todos os VCs de treinamento e, por meio do nível de correlação, ocorre uma presunção (não certeza) de que o padrão deve ser reconhecido.

No caso de recohecimento de locutores, interessante observar que, neste trabalho, foi adotada a mesma palavra ou frase para reconhecimento de locutor. Entretanto, deve-se perceber que, se o sistema foi capaz de diferenciar locutores distintos pronunciando as mesmas palavras ou frases, com certeza, deverá reconher com maior robustez locutores distintos pronunciando palavras ou frases distintas.

No caso de reconhecimento de locutores, as correlações entre os VCs extraídos de mesmas palavras ou mesmas frases pronunciadas pelo mesmo locutor (autocorrelações) devem atingir, no mínimo, o LC em sua grande maioria. Por outro lado, as correlações entre os VCs extraídos de mesmas palavras ou mesmas frases pronunciadas por locutores distintos (correlações cruzadas), devem estar abaixo do LC, em sua grande maioria.

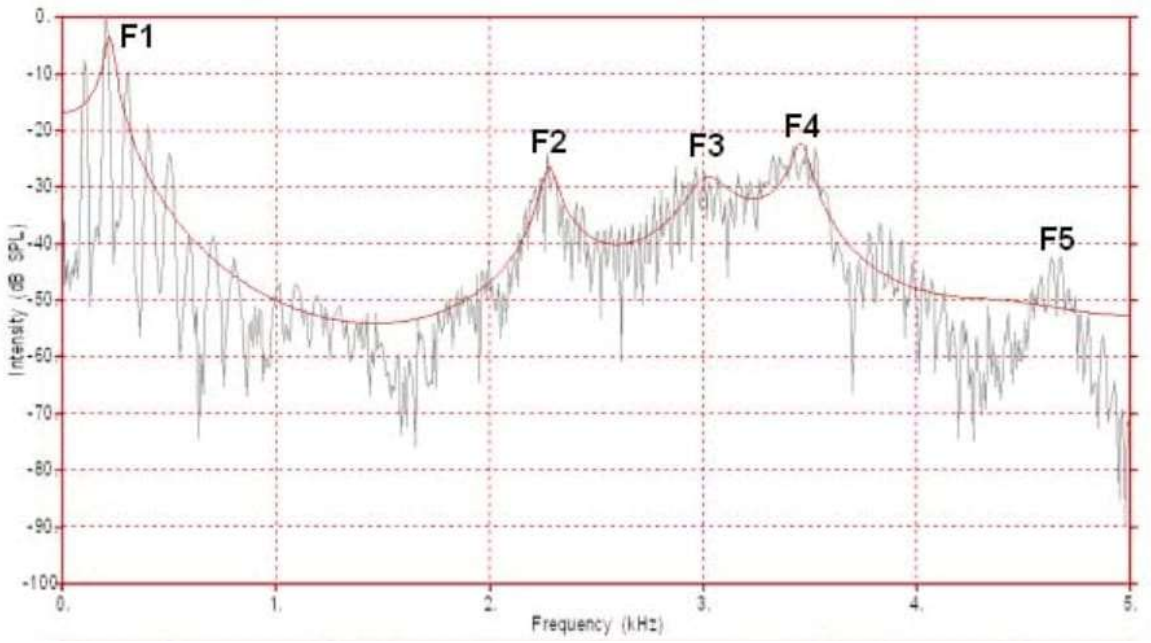
4.3 ESTADO DA ARTE NA EXTRAÇÃO DAS CARACTERÍSTICAS

Segundo Aldarmaki (2022), o primeiro passo para se implementar qualquer sistema de reconhecimento de voz é a extração das principais características do sinal, descartando informações redundantes.

Essa extração pode ser realizada por técnicas de análise de espectro, tais como a Transformada Fourier, métodos de bancos de filtros, análise homomórfica (*mel cepstrum)* e codificação por predição linear (*Linear Predictive Coding* ou *LPC*) (RABINER, 1993).

A técnica *LPC* é comumente usada para códificação de voz, sendo uma das mais antigas e consiste resumidamente em modelar a voz como uma combinação linear das amostras anteriores. A ideia base é representar o sinal de maneira comprimida, sem afetar a legibilidade, por meio da construção da envoltória espectral do sinal em cada trecho estacionário da voz, sendo imprescindível a aplicação da Transformada de Fourier. A Figura 5 procura explicar como funciona a técnica *LPC.*

Figura 5: A envoltória espectral da técnica LPC



Fonte: Adaptado de MANNELL (2020)

Ao se observar a Figura 5, nota-se que, para o reconhecimento de voz de locutores e palavras usando *LPC*, há a necessidade de aplicar a operação de *spectrum* e, em seguida, a de *cepstrum* (analisado na Seção 4.8) sob os coeficientes *LPC*, obtendo-se os *LPCC* (LPC-*cepstrum*).

Outras técnicas surgiram, com a intenção de criar uma forma de caracterizar a voz, sem apelar para a análise espectral de Fourier. As duas técnicas principais desta tendência são a *ZCPA (Zero-Crossing with Peak Amplitude)* (CUADROS, 2007) *e* a *EIH* (*Ensemble Interval Histogram*) (ALMEIDA, et al. 2014).

Tanto a *ZCPA* quanto a *EIH* usam escalas psicoacústicas (*Cochlear filter bank*), de sorte que cada frequência sintonizada (*Filter 1, Filter 2...Filter N*) deve estar relacionada com

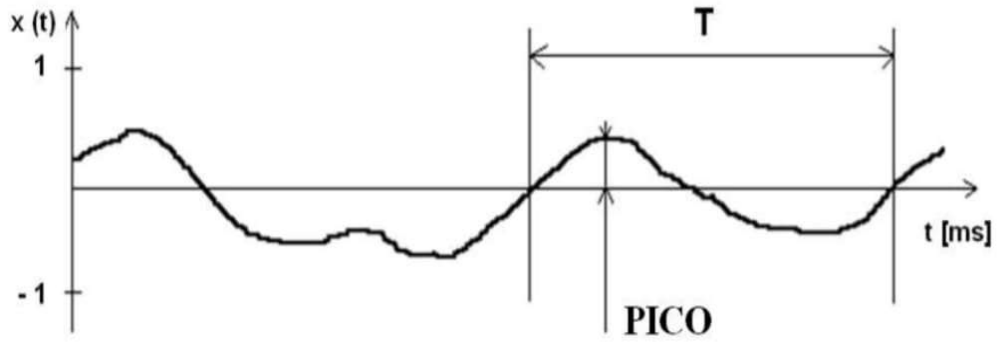
a duração de um *frame* estacionário de tamanho previamente calculado: frequências maiores devem ser aplicadas em *frames* menores; frequências menores devem ser aplicadas em *frames* maiores. Essa relação entre frequência sintonizada e tamanho da janela de observação deve obedecer a certas equações que não serão discutidas, por fugir ao foco deste trabalho.

O importante é perceber que cada frequência sintonizada é aplicada em cada *frame* do sinal, que deverá ser dividido em janelas de observação de mesmo tamanho para aquela frequência sintonizada. A próxima frequência sintonizada terá tamanho de *frame* diferente e assim sucessivamente até cobrir o espetro de interesse.

No caso da *EIH*, cada frequência também será associada a determinado limiar de audição (que não vai ser zero), pois a ideia fundamental desta técnica é imitar os limiares de audição associados a cada conjunto de células ciliadas da membrana basilar da cóclea (ver Figura 19, no Apêndice I). Em seguida, cada cruzamento crescente pelo limiar é demarcado e os seus intervalos sucessivos medidos e calculados pelo seu inverso. Este resultado deve ser alocado e posicionado em um histograma, que recebe em seguida o valor do limiar de audição da frequência sintonizada, ponderada por uma operação logarítmica.

Por outro lado, a *ZCPA* usa, para todas as frequências sintonizadas, o limiar igual a zero. Isso diminui substancialmente o custo computacional. A Figura 6 mostra graficamente como são demarcados os parâmetros usados pela *ZCPA*.

Figura 6: Demarcação da amplitude e intervalo na técnica *ZCPA*



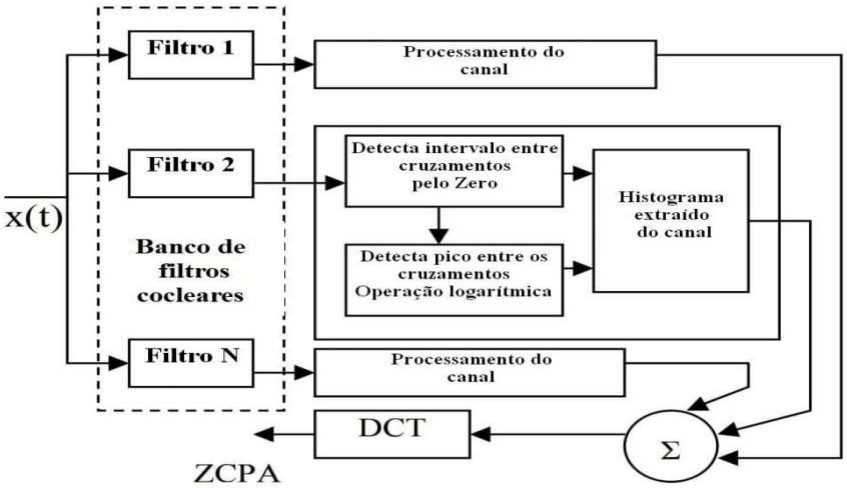
Fonte: Adaptado de KAKUR (2012)

Analisando a Figura 6, percebe-se que, na ZCPA, o intervalo calculado é entre cruzamentos positivos pelo nível zero (*Zero crossing detection*), para todas as frequências, catalogando-se no histograma (*Channel histogram*) também o valor de pico (*Peak detection/non linerity*) entre os cruzamentos, ponderados por uma operação logarítmica.

Conforme Almeida (2014), a *ZCPA* é uma simplificação da *EIH*, mas sem queda de desempenho. A diferença básica entre as duas é o limiar de audição, que, na *ZCPA*, para todas as frequências, é assumido o nível zero.

A Figura 7 mostra como ocorre a sintonização das frequências e a criação dos histogramas parciais, para cada frequência sintonizada, que serão somados, criando-se o Processamento do canal, sob o qual deverá incidir a *DCT*, axtraindo-se assim os coeficientes *ZCPA*.

Figura 7: Criação do Histograma na técnica *ZCPA*.



Fonte: Adaptado de KAKUR (2012)

Note-se na Figura 7, que , mais uma vez a DCT é usada para descorrelação e compressão da informação.

No entanto, dentre as técnicas que usam informações extraídas do espectro, a que é mais usada para separar o sinal de excitacão da resposta impulsiva do trato vocal é a análise *mel cepstral*, caracterizada pela extração dos coeficientes *MFCCs* (ALDARMAKI, 2022) (BOUROUBA, 2006).

Segundo Aldarmaki (2022), os coeficientes *MFCCs* têm a capacidade de mapear o trato vocal humano, sendo usados tanto para reconhecimento de locutor, como para reconhecimento de palavras pronunciadas, independentemente do locutor. Como o extrator de características usado neste trabalho são as *MFCCs*, maiores detalhes serão explicitados nas Seções que seguem.

4.4 PRÉ-ÊNFASE

Obtidos os ATs, faz-se necessário que estes sejam pré-processados, com o intuito de extrair dos mesmos suas características. Segundo KUMAR (2022), o espectro de frequência da voz humana é largo e dinâmico, o que facilita a sobreposição de ruídos, que contaminam o sinal. A pré-ênfase é a primeira redução de espectro, que ajuda a impedir a interferência negativa dos ruídos.

Segundo Petry (2002), a pré-ênfase é a aplicação de um filtro *FIR* de primeira ordem em toda a amostra do sinal, que deve enfatizá-lo em suas frequências mais elevadas (filtro passa-alta), eliminando uma atenuação de aproximadamente 6dB/oitava, advindas da irradiação dos lábios e do trato vocal, que traz uma distorção espectral na voz que nada contribui para a caracterização da fala. Essa pré-ênfase pode ser operacionalizada por meio da Eq. 1

$$y(n)=x(n)-\alpha\, x(n-1), \qquad \text{Eq. 1}$$

em que x(n) é o sinal amostrado, y(n) é o sinal pré-enfatizado e *"α"* é o coeficiente de pré-ênfase (entre 0,4 e 1), que geralmente é estabelecido no valor de 0,95.

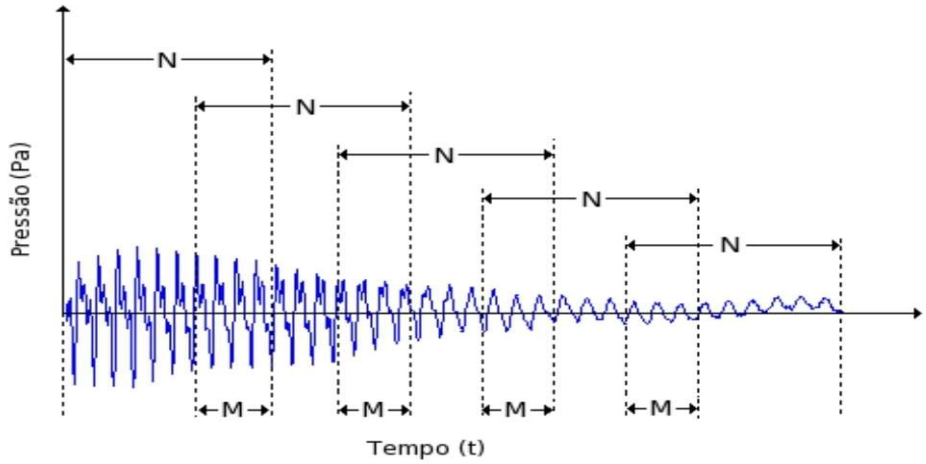
É sabido também que um sinal estacionário, ou seja, um sinal cujas formantes e frequência fundamental se mantêm em um determinado padrão ao longo do tempo, tem a característica de serem muito mais facilmente analisadas, quando comparados com sinais não-estacionários. Segundo Lathi (1989), no caso da voz humana, pode-se considerar estacionária as frequências no intervalo de 20 ms a 40 ms.

4.5 PARTIÇÃO DO SINAL EM *FRAMES* SUPERPOSTOS

Após a filtragem de pré-ênfase, parte-se para o segundo passo: particionar o sinal em pequenos pedaços (*frames*). Segundo Kumar (2022), essa partição deve ser realizada com o objetivo de selecionar partes do sinal onde o mesmo se encontra de forma estacionária.

Entretanto, esta partição deve ser realizada com a superposição parcial dos *frames,* segundo McLoughlin (2009), como esquematizado na Figura 8, a fim de se aumentar a correlação entre os períodos de tempos separados.

Figura 8: Partição do sinal em *frames* superpostos.



Fonte: Adaptado de CUADROS (2007)

Segundo McLoughlin (2009), dividir o sinal em pedaços, como na situação da Figura 8, é equivalente a multiplicar o sinal por janelas retangulares, gerando convolução no domínio da freqência, o que causa distorções intoleráveis ligadas ao fenômeno de Gibbs (RABINER, 2003) (BEZERRA, 2001) (OPPENHEIM, 2001). Só há uma forma de amenizar essas distorções: aplicar uma janela triangular para suavizar as bordas dos *frames*. É o que será visto na prósima Seção.

4.6 JANELAMENTO DOS *FRAMES*

Como visto na Seção anterior, um detalhe importante é que cada *frame* sofra um janelamento, a fim de evitar o vazamento espectral nas bordas dos sinais particionados, quando da futura análise de Fourier em cada *frame*. A janela mais indicada é a de *hamming* (OPPENHEIM, 2001) quando comparada com as demais.

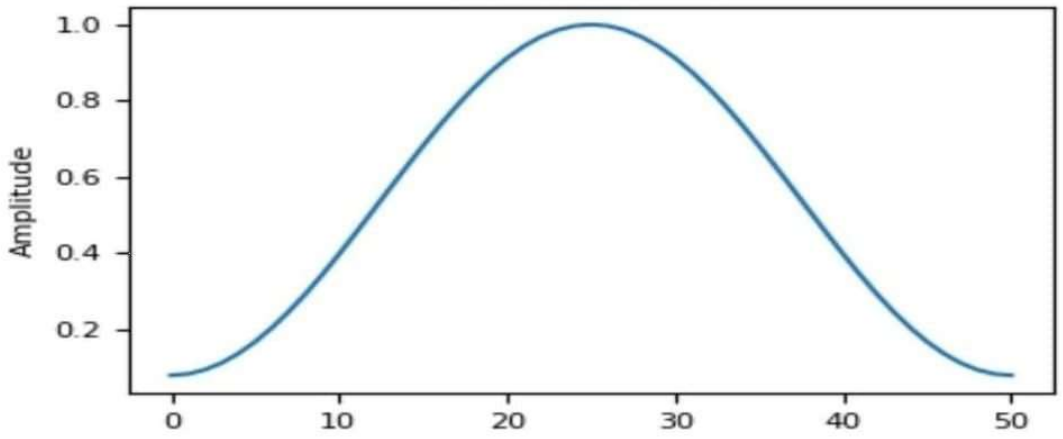
Matematicamente, a janela de *hamming* pode ser expressa conforme Eq. 2

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{n_a - 1}\right), \qquad \text{Eq. 2}$$

em que w(n) é a janela aplicada, n é o tempo e n_a é o número de amostras da janela.

A Figura 9 ilustra a janela triangular de *hamming,* cujas extremidades atenuam as bordas do *frame,* ao contrário do que ocorre com o centro da janela, cuja frequência central é enfatizada.

Figura 9: A janela de *hamming*



Fonte: Adaptado de CUADROS (2007)

Segundo Kumar (2022), a aplicação da janela minimiza os efeitos da distorção espectral, fazendo com que a borda dos frames se aproximem do nível zero. Após a aplicação da janela, todas as operações que se sequem serão aplicadas sob cada *frame,* não mais se cogitando em aplicar qualquer processamento no sinal inteiro.

4.7 ESCALAS DE FREQUÊNCIAS: HERTZ E MEL

Hertz (Hz) é uma unidade de medida de frequência usada para medir fenômenos físicos cíclicos. O valor expresso em Hz informa simplesmente quantas vezes um fenômeno periódico se repete em cada segundo. No caso da voz humana, este fenômeno periódico se propaga por diferenças de pressão irradiadas ao ar pela ação do aparelho vocal das pessoas.

O que ocorre é que o aparelho humano responsável por ouvir estes fenômenos físicos de pressão e descompressão do ar não tem uma sensibilidade linear em sintonia com a unidade de medida de frequência em Hz. Isto significa que, ao se dobrar a frequência em Hz de um som, não significa necessariamente que esta mudança será sentida pelo ouvido humano de maneira duplicada. Ela pode não ser sentida ou ser sentida de uma maneira não duplicada.

Há frequências sonoras que os seres humanos não são capazes de identificá-las em razão da própria falta de sensibilidade do aparelho auditivo, ou mesmo de um mascaramento de certas frequências em virtude da energia maior de outras frequências vizinhas.

Em outros termos, a unidade de medida de frequência em Hz só se preocupa com os fenômenos acústicos e físicos em si, desprezando a sensibilidade do ouvido humano.

Destarte, sabendo-se que a sensibilidade acústica do ouvido humano tem uma sensibilidade logarítmica, para pesquisadores que trabalham com linguística, reconhecimento de voz e doenças do aparelho auditivo ou vocal, faz-se necessário usar escalas logarítmicas mais condizentes com a sensibilidade acústica humana.

Baseado nisto, várias escalas logarítmicas foram criadas por experiências empíricas, com intuito de imitar a sensibilidade do ouvido humano.

Uma escala muito usada nos sistemas modernos de reconhecimento de locutor e de palavras na atualidade é a escala *mel*. A escala *mel* foi proposta por Stevens, Volkman e Newman em 1937. O nome *mel* advém da palavra inglesa *melody* (O'SHAUGHNESSY, 1987).

Em termos matemáticos, essas escalas psicoacústicas possuem relações com a escala tradicional em Hz. No caso da escala *mel*, segundo O'Shaughnessy (1987), é possível relacioná-la com a escala em Hz, conforme as Eq. 3 e Eq. 4

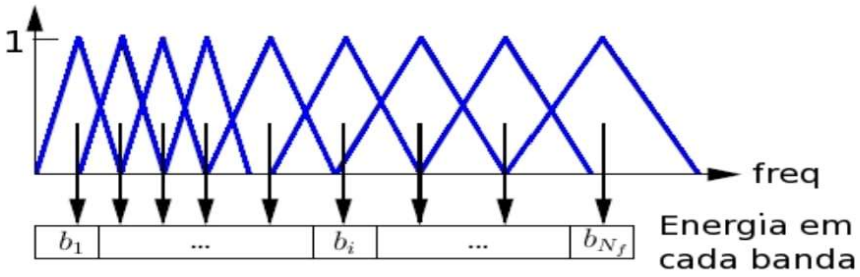
$$m = 2595 * \log10\,(1+f / 700), \qquad \text{Eq. 3}$$

$$f = 700 * (10^{m/2595} - 1), \qquad \text{Eq. 4}$$

em que *f* é a frequência em Hz e *m* é a frequência em *mels.*

Estas equações (Eq. 3 e Eq. 4) são muito importantes, pois em uma aplicação de reconhecimento de voz, faz-se necessário mapear primeiramente a faixa de frequências de interesse em Hz, transformando-a em escala *mel*, escolher um número de frequências em escala *mel* dentro do referido intervalo e depois transformar as frequências escolhidas em Hz. Segundo Aldarmaki (2022), é comum o uso de 25 filtros mapeados em escala *mel*. A Figura 10 mostra bem a aplicação do banco de filtros *mel*.

Figura 10: Aplicação do banco de filtros *mel* em um *frame* do sinal



Fonte: Adaptado de Cuadros (2007)

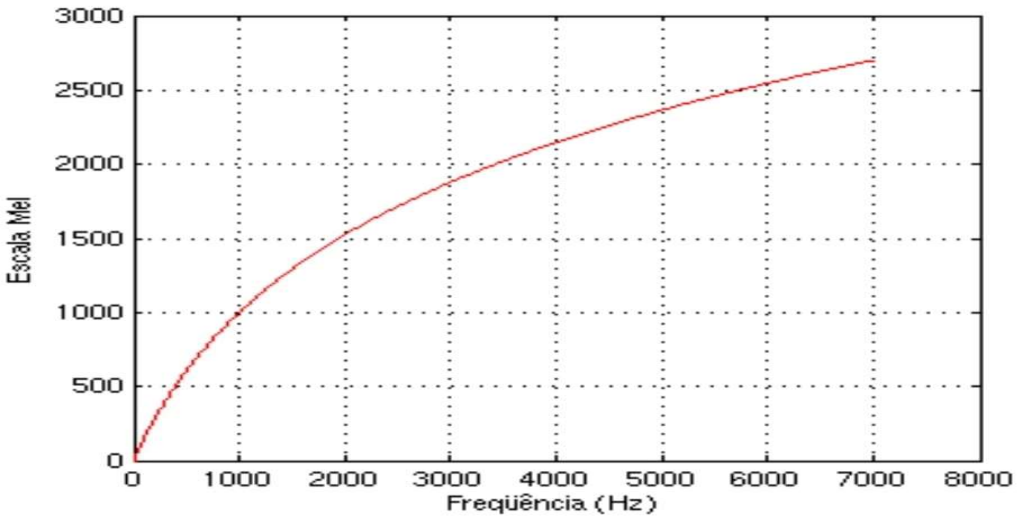
A Figura 10 tem alguns detalhes que merecem ser enfatizados. Note-se as larguras de bandas, que se iniciam praticamente iguais e vão se alargando com o aumento das frequências. Note-se que a frequência central possui a maior energia, em decorrência da aplicação da janela triangular, que atenua as raias laterais, o que dificulta o aparecimento do fenômeno do mascaramento de sinais, impedindo que se escolha frequências distintas em Hz, cuja sensibilidade do ouvido seja indiferente.

Essas frequências escolhidas dentro do intervalo de frequência máxima e mínima serão as frequências centrais dos bancos de filtros digitais passa-faixas de resposta triangular do tipo *hamming* que serão aplicadas em cada *frame* de voz.

É importante ressaltar que a aplicação desses banco de filtros *mel* devem incidir todos sobre cada *frame*, de forma serial, não podendo ser feito em cascata. Isto significa que cada filtro é aplicado sobre o *frame* original e essas filtragens parciais serão ao final somadas, criando um novo *frame* resultante, sob a qual as operações seguintes irão incidir.

A Figura 11 fornece uma ideia visual da relação entre as escalas *mel* e *Hz*.

Figura 11: Relação entre a escala Hz e escala *mel*



Fonte: Adaptado de Cuadros (2007)

Note-se que, até 1000 Hz, segundo a Figura 11, a relação entre as escalas é praticamente linear. A partir daí, a relação ganha contornos logarítmicos.

Segundo Aldarmaki (2022), a saída do banco de filtros *mel* pode ser usada diretamente na entrada de uma Rede Neural para classificação da voz, mas estes valores são ainda muito correlacionados e em grande quantidade, o que leva a necessidade de novos processamentos, como será visto adiante.

4.8 OS *MFCCS*, O *SPECTRUM* E O *CEPSTRUM*

O *spectrum* de um sinal é o mesmo sinal expresso no domínio da frequência. Isso pode ser obtido por meio da Transformada de Fourier (ALDARMAKI, 2022).

Já o *cepstrum* é exatamente a operação inversa: uma operação que atua sob um sinal no domínio da frequência, entregando como resultado um sinal no domínio do tempo. No campo de reconhecimento de voz, a operação de *cepstrum* é aplicada sob um *frame* de voz parcialmente superposto, considerado estacionário (de 20ms até 40ms), sem que este *frame* tenha se submetido a qualquer filtragem mais refinada. A seguir e em sequência, uma Transformada Discreta de Fourier (*DFT*), seguida de uma aplicação logarítmica (*LOG*) e, por fim, uma Transformada Inversa Discreta de Fourier (*IDFT*) incide também sobre o *frame* e se chega aos coeficientes *cepstrais.*

A Transformada Inversa Discreta de Fourier (*IDFT*) tem a função de retornar o sinal no domínio da frequência para o domínio do tempo, mas de maneira que seja possível identificar características relevantes ligadas ao trato vocal do indivíduo. Essas características relevantes são extraídas por meio da operação logarítmica sob o sinal no domínio da frequência, o que faz com que essas características se aproximem da resposta em frequência do trato auricular humano, que tem sensibilidade logarítmica.

De fato, algum avanço se consegue com estas operações para que se reconheça voz. Mas ainda pode ser aperfeiçoada para obter melhores resultados.

A Transformada Inversa Discreta de Fourier (*IDFT*) tem um inconveniente de retornar números complexos, o que dificulta o trabalho. Então, em sua substituição, aplica-se a Transformada Discreta do Cosseno (*DCT*), em virtude desta operação retornar dados mais comprimidos e descorrelacionados (ALDARMAKI, 2022), bem mais amigáveis de se trabalhar em termos computacionais.

Com o surgimento e estudo das escalas de frequências mais preocupadas com a sensibilidade auditiva do ser humano, como foi o caso da escala *mel*, começou-se a filtrar os *frames* de maneira mais acurada, usando frequências estratégicas, em vez de trabalhar com muitas frequências que não serviam para mapear o trato vocal das pessoas. Esse esforço foi feito com o intuito de melhorar as taxas de acerto nos processos de reconhecimento de locuções e de locutores.

Então, foi aí que uniu-se a escala *mel* com a operação de *cepstrum*, criando o *mel cepstrum*. A única diferença entre a operação de *cepstrum* para a operação de *mel cepstrum* é

que, diante de cada *frame* janelado e superposto, aplica-se um banco de filtros *mel*, antes de se aplicar o módulo da Transformada Rápida de Fourier (FFT), a operação logarítmica (*LOG*) e a Transformada Discreta do Cosseno (*DCT*).

A palavra ***ceps****trum* advém da palavra ***spec****trum*, dando a entender que seriam operações opostas. Basta que se observe as iniciais das palavras ***ceps*** e ***spec****,* que estão com as ordens das posições das letras invertidas. E é justamente isso que ocorre. Assim como a Transformada Inversa de Fourier, a Transformada Inversa do Cosseno retorna dados relevantes do sinal original (dados reais, não imaginários) para que ocorra o reconhecimento de palavras e de locutores.

Entretanto, na maioria das aplicações, usa-se a Transfaormada Discreta do Cosseno sobre o *spectrum* do sinal, ao invés da Transformada Inversa do Cosseno, o que não faz diferença em termos de reconhecimento, desde que se mantenha a escolha para todas as operações.

A sigla *MFCC*, muito usada no meio acadêmico, está relacionada com sua denominação em inglês, qual seja, *Mel Frequency Cepstral Coefficients*, que, em português, poderia ser traduzida como Coeficientes Cepstrais de Frequência Mel.

É comum nessa técnica que se despreze o primeiro coeficiente *mel cepstral*, usando-se os treze próximos coeficientes, em razão destes dados estarem fortemente relacionados ao trato vocal de cada indivíduo. O primeiro coeficiente *cepstral* estaria ligado ao canal físico de transmissão de dados. Mas há trabalhos usando diferentes quantidades de coeficientes *mel cepstrais*, dependendo da aplicação. O importante é não usar todos, pois a maior parte deles nada tem a dizer sobre a palavra falada ou ao locutor que produziu o sinal.

Esses treze primeiros coeficientes são conhecidos como estáticos, constituindo-se analogamente a uma fotografia de dados vocálicos relevantes ligados a características biométricas ao longo do sinal de voz de quem fala. Daí se justifica a razão deles serem denominados de coeficientes *mel cepstrais* estáticos.

Em linguagem de algoritmo, usando código Scilab, pode-se chegar aos coeficientes *mel cepstrais* estáticos através da seguinte operação (SKOWRONSKI, 2002) que deve incidir sobre o *frame* janelado e filtrado pelos bancos de filtros *mel:*

dct (log (abc (fft (*frame*)))),

em que

- *frame* = sinal janelado e filtrado pelos bancos de filtros *Mel;*
- fft = transformada rápida de Fourier;
- abs = módulo;

- log = logaritmo;
- dct = transformada discreta dos cossenos;

O Quadro 12 mostra a disposição dos coeficientes *mel cepstrais* estáticos em um vetor.

Quadro 12: Concatenação dos coeficientes MFCCs estáticos a nível de *frame*

Ce(1)	*Ce(2)*	*Ce(3)*	*Ce(4)*	*Ce(5)*	*Ce(6)*	*Ce(7)*	*Ce(8)*	*Ce(9)*	*Ce(10)*	*Ce(11) . . . Ce(n)*

Fonte: o autor (2022)

Sendo que *Ce(n)* é o coeficente *mel cepstral* estático e *n* é a quantidade de coeficientes adotados para compor o VCs do *frame* de voz. No caso da simulação deste trabalho, extraíu-se os 13 coeficientes iniciais, razão pela qual *n* é igual a 13.

4.8.1 Os MFCCs dinâmicos

Em adição aos coeficientes *MFCCs* estáticos, é possível extrair os coeficientes *mel cepstrais* dinâmicos de velocidade e de aceleração (ALDARMAKI, 2022). Para extrair os coeficientes de velocidade, conforme Gordillo (2018), basta que se derive a sequência de operações que levaram até aos coeficientes estáticos, usando código Scilab:

diff (dct (log (abs (fft (*frame*))))),

em que

- *frame* = sinal janelado e filtrado pelos bancos de filtros *Mel;*
- fft = transformada rápida de Fourier;
- abs = módulo;
- log = logaritmo;
- dct = transformada discreta dos cossenos;
- diff () = primeira derivada;

O Quadro 13 mostra a disposição dos coeficientes *mel cepstrais* dinâmicos de velocidade em um vetor.

Quadro 13: Concatenação dos coeficientes MFCCs de velocidade a nível de *frame*

$C_{v(1)}$	$C_{v(2)}$	$C_{v(3)}$	$C_{v(4)}$	$C_{v(5)}$	$C_{v(6)}$	$C_{v(7)}$	$C_{v(8)}$	$C_{v(9)}$	$C_{v(10)}$	$C_{v(11)}$ . . . $C_{v(n)}$

Fonte: o autor (2022)

Sendo que *Cv(n)* é o coeficente *mel cepstral* dinâmico de velocidade e *n* é a quantidade de coeficientes adotados para compor o VCs do *frame* de voz. No caso da simulação deste trabalho, extraíu-se os 13 coeficientes iniciais, razão pela qual *n* é igual a 13.

Para extrair os coeficientes de aceleração, conforme Gordillo (2018), basta que se derive a sequência de operações que levaram até aos coeficientes de velocidade, conforme código Scilab:

diff (diff (dct (log (abs (fft (*frame*)))))),

em que

- *frame* = sinal janelado e filtrado pelos bancos de filtros *Mel;*
- fft = transformada rápida de Fourier;
- abs = módulo;
- log = logaritmo;
- dct = transformada discreta dos cossenos;
- diff (diff ()) = segunda derivada;

O Quadro 14 mostra a disposição dos coeficientes *mel cepstrais* dinâmicos de aceleração em um vetor.

Quadro 14: Concatenação dos coeficientes MFCCs de aceleração a nível de *frame*

$C_{a(1)}$	$C_{a(2)}$	$C_{a(3)}$	$C_{a(4)}$	$C_{a(5)}$	$C_{a(6)}$	$C_{a(7)}$	$C_{a(8)}$	$C_{a(9)}$	$C_{a(10)}$	$C_{a(11)}$ . . . $C_{a(n)}$

Fonte: o autor (2022)

Sendo que $Ca(n)$ é o coeficente *mel cepstral* dinâmico de aceleração e *n* é a quantidade de coeficientes adotados para compor o VCs do *frame* de voz. No caso da simulação deste trabalho, extraíu-se os 13 coeficientes iniciais, razão pela qual *n* é igual a 13.

4.8.2 A janela de *liftro*

Segundo Petry (2002), é comum aplicar uma função senoidal aos coeficientes *mel cepstrais* conhecida como janela de *liftro* com o intuito de melhorar as taxas de reconhecimento em ambientes ruidosos. A finalidade desse procedimento é enfatizar os coeficientes *mel cepstrais* com maior informação espectral. Em termos matemáticos, a janela de *liftro* pode ser expressa conforme Eq. 5

$$l(n)=1+\frac{Q}{2}\sin\left(\frac{n\pi}{Q}\right), \qquad \text{Eq. 5}$$

em que

- $l(n)$ é a janela de ponderação a ser multiplicada pelo coeficiente *mel cepstral;*
- Q é uma constante, cujo valor é de 22;

- *n* é o índice do coeficiente *mel cepstral*;

Os quadros 15, 16 e 17 explicam como é feita a aplicação da janela de *liftro* aos coeficientes *mel cepstrais*.

Quadro 15: Multiplicação do fator de *liftro* pelos coeficientes MFCCs estáticos

$l_{(1)}$	$l_{(2)}$	$l_{(3)}$	$l_{(4)}$	$l_{(5)}$	$l_{(6)}$	$l_{(7)}$	$l_{(8)} \ldots l_{(n)}$
$C_{e(1)}$	$C_{e(2)}$	$C_{e(3)}$	$C_{e(4)}$	$C_{e(5)}$	$C_{e(6)}$	$C_{e(7)}$	$C_{e(8)} \ldots C_{e(n)}$
$E_{(1)}$	$E_{(2)}$	$E_{(3)}$	$E_{(4)}$	$E_{(5)}$	$E_{(6)}$	$E_{(7)}$	$E_{(8)} \ldots E_{(n)}$

Fonte: o autor (2022)

em que $E_{(n)}$ é o resultado da multiplicação dos coeficientes *mel cepstrais* estáticos $C_{e(n)}$ pelos fatores de liftro $l_{(n)}$, conforme Eq. 6

$$E_{(n)} = C_{e(n)} . l_{(n)} \quad \text{Eq. 6}$$

Quadro 16: Multiplicação do fator de *liftro* pelos coeficientes MFCCs de velocidade

$l_{(1)}$	$l_{(2)}$	$l_{(3)}$	$l_{(4)}$	$l_{(5)}$	$l_{(6)}$	$l_{(7)}$	$l_{(8)} \ldots l_{(n)}$
$C_{v(1)}$	$C_{v(2)}$	$C_{v(3)}$	$C_{v(4)}$	$C_{v(5)}$	$C_{v(6)}$	$C_{v(7)}$	$C_{v(8)} \ldots C_{v(n)}$
$V_{(1)}$	$V_{(2)}$	$V_{(3)}$	$V_{(4)}$	$V_{(5)}$	$V_{(6)}$	$V_{(7)}$	$V_{(8)} \ldots V_{(n)}$

Fonte: o autor (2022)

em que $V_{(n)}$ é o resultado da multiplicação dos coeficientes *mel cepstrais* de velocidade $C_{v(n)}$ pelos fatores de liftro $l_{(n)}$, conforme Eq. 7

$$V_{(n)} = C_{v(n)} . l_{(n)} \quad \text{Eq. 7}$$

Quadro 17: Multiplicação do fator de *liftro* pelos coeficientes MFCCs de aceleração.

$l_{(1)}$	$l_{(2)}$	$l_{(3)}$	$l_{(4)}$	$l_{(5)}$	$l_{(6)}$	$l_{(7)}$	$l_{(8)} \ldots l_{(n)}$
$C_{a(1)}$	$C_{a(2)}$	$C_{a(3)}$	$C_{a(4)}$	$C_{a(5)}$	$C_{a(6)}$	$C_{a(7)}$	$C_{a(8)} \ldots C_{a(n)}$
$A_{(1)}$	$A_{(2)}$	$A_{(3)}$	$A_{(4)}$	$A_{(5)}$	$A_{(6)}$	$A_{(7)}$	$A_{(8)} \ldots A_{(n)}$

Fonte: o autor (2022)

em que $A_{(n)}$ é o resultado da multiplicação dos coeficientes *mel cepstrais* de aceleração $C_{a(n)}$ pelos fatores de liftro $l_{(n)}$, conforme Eq. 8

$$A_{(n)} = C_{a(n)} . l_{(n)} \quad \text{Eq. 8}$$

4.8.3 A energia dos *frames*

As energias dos *frames* constituem dados relevantes para reconhecer as palavras e os locutores. Elas informam como a anergia vai mudando ao longo do sinal. Em termos matemáticos, pode-se calcular o coeficiente de energia estática dos *frames* conforme Eq. 9:

$$S=\sum_{i=1}^{n}\left(P_i\right)^2 \quad \text{com} \quad n\geq i\geq 1, \quad n\in\mathbb{Z} \quad \text{e} \quad i\in\mathbb{Z}, \qquad \text{Eq. 9}$$

em que

- *S* representa a energia do *frame;*
- *i* é a posição da amostra do *frame*;
- *n* é o tamanho do *frame*;
- *P* é a amplitude em cada posição *i;*

Da mesma forma que foi feito com os coeficientes *mel cepstrais*, é possível aplicar uma primeira derivada à expressão da Eq. 9, obtendo-se os coeficientes de energia de velocidade. Em se aplicando uma nova derivada, tem-se os coeficientes de energia de aceleração.

4.8.4 Construindo o vetor de características do sinal

Os 13 coeficientes *MFCCs* estáticos $E_{(n)}$, *os* 13 coeficientes *de* velocidade $V_{(n)}$ e os 13 coeficientes de aceleração $A_{(n)}$, bem como a energia do *frame (S)* devem ser concatenados nesta ordem, formando o VCs do *frame*, conforme Quadro 18.

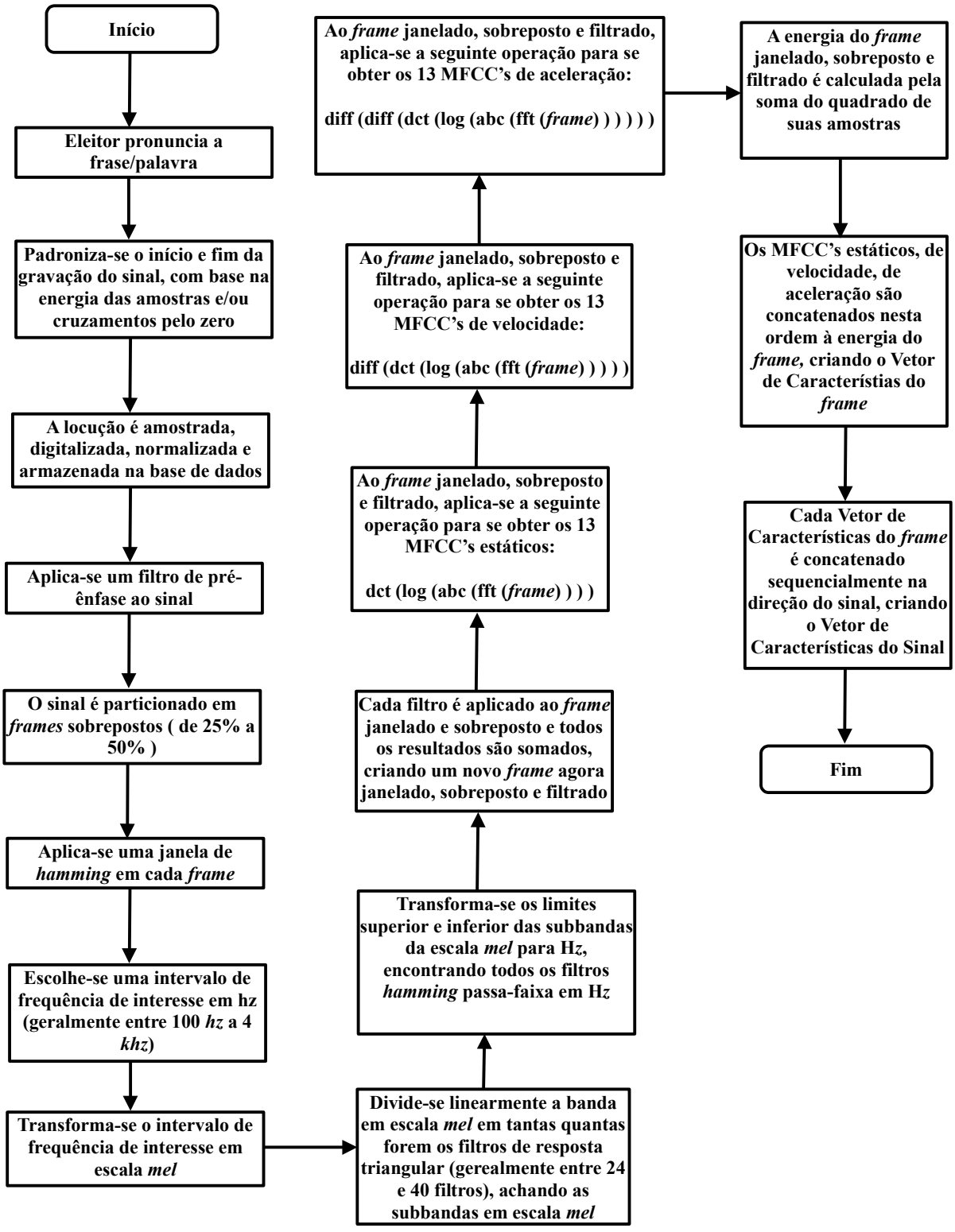
Quadro 18: VCs do *Frame* de sinal (40 características)

Coefs. Estáticos													Coefs. de Velocidade													Coefs. de Aceleração													Energia
E	E	E	E	E	E	E	E	E	E	E	E	E	V	V	V	V	V	V	V	V	V	V	V	V	V	A	A	A	A	A	A	A	A	A	A	A	A	S
De 1 a 13													De 1 a 13													De 1 a 13													De 1 a 1

Fonte: o autor (2022)

Note-se que outros atributos, além dos estabelecidos no Quadro 18, podem ser adicionados. Não existe um modelo de quantidade de atributos fixos. O importante é manter a ordem deles ao longo do *frame.* A Figura 12 mostra o fluxograma geral até se chegar ao VCs do sinal.

Figura 12: Fluxograma de extração dos coeficientes *MFCCs*



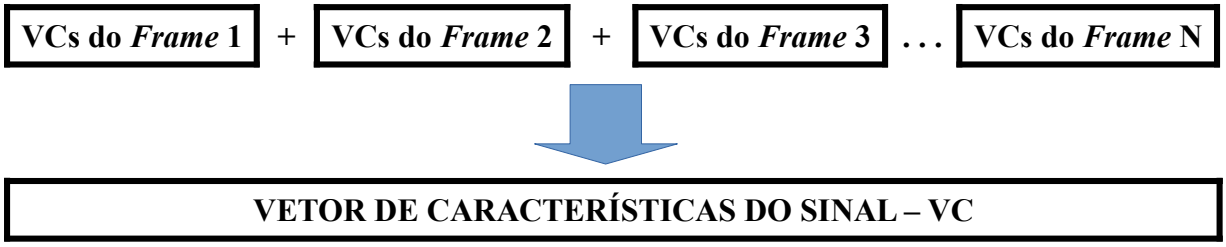
OBS: fft = transformada rápida de fourier; **abs** = módulo; **log** = logarítmo; **dct** = transformada discreta do cosseno; **diff** = derivada.

Fonte: o autor (2022)

Conforme Figura 12, para se criar o VCs do sinal inteiro, os VCs dos *frames* devem ser concatenados de forma sequencial e gravados em uma base de dados para posterior consulta e

processamento visando reconhecer palavras e locutores. A Figura 13 ilustra como deve ser a concatenação dos VCs dos *Frames* para formar o VC do Sinal.

Figura 13: Concatenação dos VCs dos *frames* para criar o VC do sinal



Fonte: o autor (2022)

Em Figura 13, o sinal de "+" significa concatenação, não soma algébrica.

4.8.5 Cálculo da correlação, da distância euclidiana e dos limiares

Extraídos os vetores de características dos ATs ou dos APs, vislumbra-se duas opções tecnológicas amplamente usadas na atualidade para classificação e reconhecimento de voz: as Redes Neurais (*ANN*) e os Modelos Ocultos de Markov (*HMMs*). Entretanto, como o objetivo do trabalho é estabelecer uma solução no âmbito da urna eletrônica (que comporta no máximo 400 eleitores) e levando em consideração o fato de que a urna eletrônica não trabalha em rede, tendo como gargalo a limitação de recursos, optou-se pela elaboração de um um modelo mais customizado, deixando que sistemas mais complexos e robostos (como os *HMMs* ou as Redes Neurais) possam utilizar e processar os VCs, atuando na gigantesca base de dados da Justiça Eleitoral em busca de fraudes e duplicidades de inscrições, antes de serem inseminados nas urnas eletrônicas, por meio da detecção de semelhanças entre vozes.

Destarte, optou-se por gravar a pronúnica de 20 ATs para cada locução, extrair os seus respectivos VCs, e, por meio de correlação estatística, estabelecer parâmetros de verossimilhança e LCs para reconhecimento de palavras e locutores por meio da máxima verossimilhança (*template matching,* que significa correspondência de modelo, em português).

A correlação linear consiste numa operação matemática cuja saída é o coeficiente de correlação de *Pearson,* que se encontra no intervalo de -1 até 1. Quanto mais próximo de 1 a

correlação, maior a verossimilhança da sequência de coeficientes *mel cepstrais* e de energias, o que indica a autenticidade de uma determinada palavra ou certo locutor. A descrição matemática para se chegar ao coeficiente de correlação pode ser vista na Eq. 10:

$$\rho = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}, \qquad \text{Eq. 10}$$

em que

- ρ é o coeficiente de correlação *de Pearson;*
- n é o número de coeficientes de um vetor de características;
- x_i é o iésimo coeficiente de um vetor x*;*
- y_i é o iésimo coeficiente de um vetor y*;*
- $\overline{x}$ é a média aritmética dos coeficientes do vetor de características x*;*
- $\overline{y}$ é a média aritmética dos coeficientes do vetor de características y*;*

Outra operação matemática que pode ser usada é a distância euclidiana ou distorção vetorial, que entrega um valor que mede a distância entre dois vetores de mesmo tamanho. Entretanto, a lógica dos resultados, quando se compara com a correlação estatística, é invertida: quanto menor a distância entre dois vetores, maior será sua verossimilhança. A Eq. 11 mostra como se calcula a distância euclidiana entre dois vetores x e y:

$$d(x, y) = \sum_{i=1}^{n} \sqrt{(x_i - y_i)^2} \qquad \text{Eq. 11}$$

em que

- x é um vetor de tamanho n;
- y é outro vetor de tamanho n;
- x_i é o valor do vetor x na posição i;
- y_i é o valor do vetor y na posição i;
- d é a distorção (ou distância) vetorial entre o vetor x e o vetor y;
- i é a posição do vetor;

Se a distância euclidiana entre dois vetores x e y de mesmo tamanho for igual a zero, isto significa que $x = y$. Entretanto, se a correlação entre dois vetores x *e* y de mesmo tamanho for diferente de zero, isto significa que estes vetores são distintos ($x \neq y$).

O Quadro 19 procura simplificar a compreensão dos parâmetros de comparação para escolha de reconhecimento de locutor ou de palavra.

Quadro 19: Diferenças entre a Correlação e a Distância Euclidiana

Correlação	Distância Euclidiana
Quanto maior o seu valor (de -1 até +1), maior o grau de verossimilhança entre dois vetores	**Quanto maior o seu valor (de 0 a qualquer valor positivo), menor a verossimilhança entre dois vetores**
Quanto menor o seu valor (de -1 até +1), menor o grau de verossimilhança entre dois vetores	**Quanto menor o seu valor (de 0 a qualquer valor positivo), maior a verossimilhança entre dois vetores**

Fonte: o autor (2022)

Caso se adote a distorção vetorial no lugar da correlação estatística como parâmetro de comparação, a lógica do resultado é invertida em relação à correlação. Ou seja, ao se chegar a um limiar X com a distorção vetorial, significa que, para que uma palavra seja reconhecida, o valor da distorção de prova só pode ter como resultado um valor que não ultrapasse o valor X. Caso ocorra a ultrapassagem, não haveria o reconhecimento, devendo o sistema solicitar ao locutor nova pronúncia da locução.

Neste experimento a máxima correlação estatística indicará quem está falando ou o que está se falando, dentro do universo de palavras previamente estabelecido. Entretanto, a máxima correlação por si só seria insuficiente, caso se queira um sistema mais funcional, pois não se pode permitir que o usuário diga qualquer coisa ao microfone e o sistema identifique apenas algo mais parecido com o que fora dito. O ideal é que se o usuário pronuncia uma palavra não prevista e não cadastrada no sistema, este solicite ao usuário para que se repita a locução, tendo em vista que não se teria alcançado um LC para reconhecimento previamente calculado.

É óbvio que não se pode tolerar que o sistema solicite de forma indefinida ao locutor a repetição da pronúncia da palavra ou frase que o habilita ao acesso do sistema, pois, neste caso, pode estar havendo uma tentativa de fraude de um impostor tentando se passar pelo locutor legítimo. É preciso então estabelecer o número máximo de tentativas (por exemplo 3 tentativas) para que o locutor possa ser habilitado.

É importante enfatizar que a solução proposta emgloba dois tipos de identificação de voz: uma de locutor; e outra de palavras ditas por este locutor. Ou seja, primeiro se reconhece biometricamente o locutor e em seguida se reconhece palavras deste locutor para dar comandos à Urna Eletrônica.

Para autenticação de locutor, usuou-se como parâmetro uma mesma palavra pronunciada por todos os locutores (no caso desta simulaçção, a frase PAC ou a locução CONFIRMA). Neste caso, as autocorrelações são feitas entre palavras pronunciadas pelo

mesmo locutor e as correlações cruzadas são feitas entre palavras pronunciadas por locutores diferentes.

Para o reconhecimento de palavras (no caso, as palavras que invocam os comandos da urna eletrônica), todas as correlações são feitas entre palavras pronuncidas pelo mesmo locutor, de sorte que as autocorrelações são feitas entre as mesmas palavras e as correlações cruzadas entre palavras distintas.

Se o método de extração das características da voz (no caso deste trabalho, as *MFCCs*) estiverem corretos, a maior parte das autocorrelações devem estar acima de um determinado limiar e as correlações cruzadas devem estar abaixo desse limiar. Para que o sistema seja robusto, a média das autocorrelações deve ser não apenas mais alta que as correlações cruzadas, mas deve guardar a maior distância possível da média das correlações cruzadas.

Uma forma simples de se calcular o LC é considerá-lo como o menor valor, dentre as possíveis autocorrleações. Ocorre que este recurso torna o sistema muito rígido, podendo aumentar a taxa de erros. Só se justificaria sua implementação no caso em que haja apenas um ou poucos usuários optantes pelo reconhecimento de voz como fator de autenticação.

Uma forma mais técnica e usual de se calcular o LC é por meio do cálculo da média e desvio padrão entre as correlações cruzadas e autocorrelações, conforme Eq. 12:

$$LC = \frac{\left(S_a \times \overline{X}_c + S_c \times \overline{X}_a\right)}{\left(S_a + S_c\right)}, \qquad \text{Eq. 12}$$

em que

- LC é o limiar de correlação (valor mínimo para ser reconhecido);
- S_a é o desvio padrão entre as possíveis autocorrelações entre os VCs;
- S_c é o desvio padrão entre as possíveis correlações cruzadas entre os VCs;
- $\overline{X}_a$ é a média entre as possíveis autocorrelações entre os VCs;
- $\overline{X}_c$ é a média entre as possíveis correlações cruzadas entre os Vcs;

4.9 CONCLUSÃO DO CAPÍTULO

Neste capítulo, é feito um resumo sobre as principais técnicas usadas para extração de dados biométricos da voz humana, segundo o atual estado da arte.

Em seguida, adentra-se na parte técnica propriamente dita, explicando as várias etapas, em ordem de execução, para se chegar aos VCs, usando a extração dos *MFCCs*, técnica usada neste trabalho para caracterização das palavras e frases treinadas.

CAPÍTULO 5

5. TESTES E RESULTADOS

A simulação feita neste trabalho usou 8000 Hz de taxa de amostragem, 16 bits de resolução e os áudios foram gravados em formato *wav* em todos os treinamentos, bem como os vetores de características se submeteram aos mesmos parâmetros.

Para reconhecimento de locutor, foram feitas duas simulações. Uma com a frase PAC. E outra com a palavra CONFIRMA. Estas simulações foram feitas com o intuito de explorar as diversas possibilidades que existem para se implementar este sistema.

Para todos os AT, tanto para reconhecimento de locutor, com para reconhecimento de palavras, foram captados 20 áudios. No caso de identificação de locutor, fez-se necessária a captação de áudios de voluntários, que se disponibilizaram a enviar amostras das locuções, enviando-as pelo aplicativo *whatsapp* em formato *.ogg*. Esses arquivos foram baixados e convertidos em formato *.wav*, com parâmetros semelhantes aos áudios de treinamento do autor deste trabalho.

O locutor *"jj"*, autor deste trabalho, pronunciou a frase PAC 20 vezes. Os voluntários foram 20 do sexo masculino e 20 do sexo feminino, cada um pronunciando a frase PAC uma única vez. As comparações para fins de reconhecimento de locutor foram feitas partindo do princípio de que o locutor reivindicante é o locutor *"jj"*.

Tudo o que foi dito em relação à frase PAC nos dois últimos parágrafos, pode ser estendido para a palavra CONFIRMA, de sorte que a única diferença é o tamanho do sinal. Chega-se ao total de **120 áudios (40 áudios do locutor *"jj"* e 80 áudios dos demais locutores)** e seus respectivos VCs para cálculos de correlações, limiares e probabilidades de erros para reconhecimento de locutor.

Para identificação de palavras do mesmo locutor, a fim de simular o reconhecimento de palavras isoladas, o autor deste trabalho gravou 20 vezes cada uma das palavras descritas em Tabela 8 (240 áudios), Tabela 10 (240 áudios) e Tabela 11 (240 áudios), totalizando **720 áudios** e seus respectivos VCs para cálculos de correlações, limiares e probabilidades de erros para reconhecimento de palavras.

Para extração dos coeficientes *MFCCs*, adotou-se 24 filtros de resposta triangular, que atuaram em uma faixa de frequência de interesse compreendida entre 100 Hz até 4000 Hz. O tamanho dos *frames* foram padronizados em 20 ms, superpostos em 10 ms (50%), tanto para palavras de 2 a 4 sílabas, como para a frase PAC. O coeficiente de pré ênfase foi estabelecido no valor de 0.95.

Como a taxa de amostragem de todos os áudios foi de 8000 Hz, conclui-se que cada *frame* possui 160 amostras do sinal. Devido à necessidade de superposição dos *frames* em 50%, estas 160 amostras inicais são incrementadas em um *looping* de metade do seu valor (80 amostras, ou 10 ms) para delimitação da posição dos *frames* ao longo do sinal.

5.1 USANDO A FRASE PAC

Todas as frases PAC foram gravadas em um tempo fixo de 1,725125 segundos, equivalentes a 13600 amostras, ocupando um espaço de 26,9 kB por locução.

As 13600 amostras do sinal foram suficientes para pronunciar naturalmente a frase PAC, como se esta fosse a resposta a uma pergunta sobre qual o nome do locutor reivindicante.

Na Tabela 4, demonstra-se um resumo que quantifica e informa os parâmetros adotados na simulação deste trabalho, em relação ao tamanho, quantidade de amostras e ocupação de espaço que seria necessária para armazenar cada AT e seu respectivo VC na base da Justiça Eleitoral.

Tabela 4: Parâmetros de captação das locuções da frase PAC

Áudio			**VC extraído do áudio**	
Tempo do áudio	**Nº de amostras**	**Memória por áudio**	**Nº de dados biométricos**	**Memória por VC**
1,7 s	**13600**	**26,9 kB**	**6800**	**13,3 kb**

Fonte: o autor (2022)

Quantificar a memória necessária para extrair o VCs dos áudios, como é exposto na Tabela 4, é de suma importância, pois são esses vetores que deverão ser gravados na urna eletrônica.

Como mencionado anteriormente, geralmente se usa como dados relevantes os 13 primeiros *MFCCs*, conhecidos como estáticos, os 13 primeiros *MFCCs* dinâmicos de

velocidade, os 13 primeiros *MFCCs* dinâmicos de aceleração e, no caso desta simulação, a anergia total do *frame* também foi levada em conta para construir o VC. Há trabalhos com menos coeficientes, dependendo da aplicação.

Portanto, de cada *frame* de 160 amostras, foram extraídos 40 coeficientes (13+13+13+1) *MFCCs*, incluindo o coficiente de energia. Basta que se multiplique 170 (número de *frames* superpostos) por 40 para saber o tamanho do VCs da frase PAC: 6800 coeficientes.

A Tabela 5 mostra um panorama dos resultados das possíveis autocorrelações (correlações entre os VCs extraídos do mesmo locutor, no caso do locutor *"jj"*) e das possíveis correlações cruzadas (correlações entre os VCs extraídos do locutor *"jj"* e os 40 outros locutores), todos pronunciando a frase PAC.

A variável LC, presente na Tabela 5 é calculada conforme Eq. 12 da Seção 4.8.5.

Tabela 5: Correlação entre VCs para identificação do locutor *jj* (frase PAC)

20 VCs/texto	380 autocorrelações			800 correlações cruzadas			LC	FN	FP	PE%	PA%
	AC	$\overline{X}_a$	S_a	CC	$\overline{X}_c$	S_c					
PAC JJ PAC M PAC H	380	0,61	0,04	800	0,37	0.06	0,51	4	25	4,18%	95,82
TOTAL	380	-	-	800	-	-	-	4	25	4,18%	95,82

Fonte: o autor (2022)

A variável FN da Tabela 5 representa a quantidade de falsos negativos, ou seja, o número de autocorrelações que não atingiram o limiar de correlação (LC), e a variável FP representa a quantidade de falsos positivos, ou seja, o número de correlações cruzadas que ultrapassaram o limiar de correlação (LC). A variável AC representa a quantidade de autocorrelações e a variável CC representa a quantidade de correlações cruzadas.

As variáveis PE e PA da Tabela 5 são as probabilidades de erro e de acerto, respectivamente, e podem ser obtidas conforme Eq. 13 e Eq. 14 respectivamente

$$PE = \frac{FN}{AC} + \frac{FP}{CC} \qquad \text{Eq. 13}$$

$$PE = PA - 1 \qquad \text{Eq. 14}$$

Note-se que, dentre as possíveis autocorrelações (400) da Tabela 5, 20 têm como resultado o valor 1, pois decorrem de correlação entre VCs extraídos dos mesmos arquivos de áudios, razão pela qual usou-se as autocorrelações cujo resultado foi diferente de 1, ou seja, 380 autocorrelações. Também foram calculadas as médias ($\overline{X}_a$ e $\overline{X}_c$) e os desvios padrões (S_a

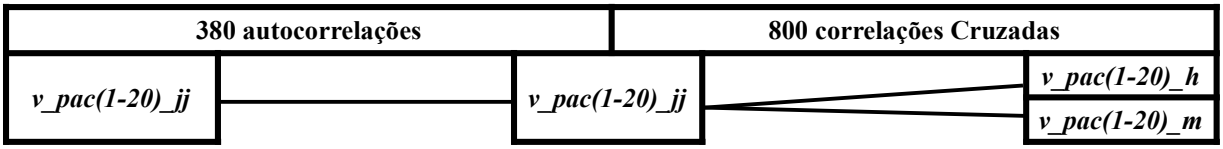
e S_c) das autocorrelações e das correlações cruzadas respectivamente para fins de cálculo do LC.

No caso das correlações cruzadas da Tabela 5, não existem resultados com valor igual a 1, pois são entre locutores diferentes. Ao todo são 800 correlações cruzadas: 400 correlações entre os VCs do locutor *"jj"* e os VCs das locutoras do sexo feminino; 400 correlaçoes entre os VCs do locutor *"jj"* e os VCs dos locutores do sexo masculino). Note-se que a média das correlações cruzadas ($\overline{X}_c = 0{,}37$) é bem inferior à média das autocorrelações ($\overline{X}_a = 0{,}61$**),** o que se constitui em um ótimo resultado e um forte indício de que os atributos*,* usando extração de *MFCCs,* foram extraídas de maneira correta, apesar da probabilidade de erro de 4,18%.

A lógica da Tabela 5 explicada nos cinco últimos parágrafos é importante para entender as demais tabelas deste capítulo, pois seguem a mesma lógica.

A Figura 14 procura simplificar o processo para melhor compreensão, de maneira que cada reta ligando o conjunto de VCs representa 380 correlações (se forem autocorrelações) e 400 correlações (se forem correlações cruzadas).

Figura 14: Correlações para identificação de locutor usando a frase PAC



Fonte: o autor (2022)

Note-se na Figura 14 a padronização do nome dos arquivos *wav.,* gravados na base de dados, onde *"v"* significa vetor de características, *"pac"* significa a frase PEDRO ÁLVARES CABRAL, os números significam a ordem temporal em que as locuções foram pronunciadas. As letras *"m"* e *"h"* significam o sexo dos diferentes locutores identificados por números de 1 até 20.

5.2 USANDO A PALAVRA CONFIRMA

Todas as locuções da palavra CONFIRMA foram gravadas em um tempo fixo de 0,825125 segundos, equivalentes a 6400 amostras, ocupando um espaço de 12,9 kB por locução.

As 6400 amostras do sinal devem ser suficientes para pronunciar naturalmente a palavra CONFIRMA, como se ao locutor tivessem dirigido a pergunta: seu nome é Pedro Álvares Cabral? E o locutor respondesse: CONFIRMA.

Essas 6400 amostras foram divididas em *frames* de 160 amostras, que equivalem a 20 ms, em taxa de amostragem de 8000 Hz, com superposição dos *frames* em 50% (80 amostras).

A Tabela 6 tem a mesma lógica da Tabela 4, sendo aplicada à palavra CONFIRMA.

Tabela 6: Parâmetros de captação das locuções das palavras de 2 a 4 sílabas

Áudio			VC extraído do áudio	
Tempo do áudio	**Nº de amostras**	**Memória por áudio**	**Nº de dados biométricos**	**Memória por VC**
0,82s	**6400**	**12,9 kB**	**3200**	**6,29 kB**

Fonte: o autor (2022)

Note-se, na Tabela 6, a memória de 6,29 kB utilizada para armazenamento do VC, arquivo relativamente pequeno, ideal para ser inseminado na urna eletrônica usando o método proposto neste trabalho.

A Tabela 7 tem a mesma lógica da Tabela 5, mas usando a palavra CONFIRMA.

Tabela 7: Correlação entre VCs para identificação do locutor *jj* (CONFIRMA)

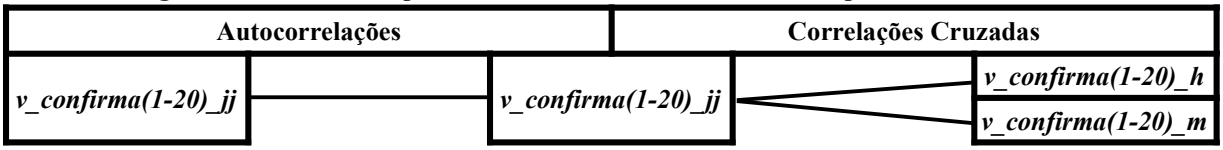
20 VCs/texto	380 autocorrelações			800 correlações cruzadas			LC	FN	FP	PE%	PA%
	AC	$\overline{X}_a$	S_a	CC	$\overline{X}_c$	S_c					
CONFIRMA JJ CONFIRMA M CONFIRMA H	**380**	**0,71**	**0,041**	**800**	**0,49**	**0,068**	**0,63**	**14**	**12**	**5,18%**	**94,82**
TOTAL	**380**	**-**	**-**	**800**	**-**	**-**	**-**	**14**	**12**	**5,18%**	**94,82**

Fonte: o autor (2022)

Note-se na Tabela 7 que o LC para a palavra CONFIRMA é maior que o LC da frase PAC, na Tabela 5, o que torna o reconhecimento mais difícil usando a palavra CONFIRMA.

A Figura 15 tem a mesma lógica da Figura 13, sendo que nela foram usadas a palavra CONFIRMA, no lugar da frase PAC.

Figura 15: Correlações para identificação de locutor usando a palavra CONFIRMA



Fonte: o autor (2022)

A escolha entre uma palavra (no caso, a palavra CONFIRMA) ou uma frase contendo três palavras (no caso, a frase PAC) é de suma importância, pois esta escolha se reflete no

tempo gasto para o reconhecimento do eleitor no dia da eleição, bem como no custo computacional.

5.3 USANDO AS PALAVRAS PARA COMANDAR A URNA

Identificado o eleitor por meio da frase PAC ou da palavra CONFIRMA, o eleitor estará habilitado a votar, dando comandos de voz à urna eletrônica mediante sinais sonoros. A princípio, é preciso afirmar que todos os parâmetros usados para a palavra CONFIRMA, na Seção 5.2, foram extendidos a todas as palavras para comandar a urna eletrônica.

Ao se usar as palavras que seriam logicamente relacionadas aos comandos da urna eletrônica, como as palavras da Tabela 8, não se garante o sigilo do voto, em razão da coincidência de significado dos comandos da urna eletronica com o significado das palavras pronunciadas.

Entretanto, realizou-se a extração dos VCs das palavras da Tabela 8 para se averiguar a influência negativa na questão da probabilidade de erro, causada pela proximidade fonética entre as palavras escolhidas, principalmente entre as palavras DOIS e OITO, entre as palavras BRANCO e CINCO, entre as palavras TRÊS e SEIS e entre as palavras SETE e ZERO.

Tabela 8: Correlações entre VCs de palavras escolhidas sem critério fonético

20 VCs/texto	4.560 autocorrelações			52.800 correlações cruzadas			LC	FN	FP	PE%	PA%
	AC	$\overline{X}_a$	S_a	CC	$\overline{X}_c$	S_c					
ZERO	380	0,61	0,06	4400	0,39	0,10	0,52	38	508	21,55	78,45
UM	380	0,72	0,03	4400	0,41	0,08	0,63	2	31	1,24	98,76
DOIS	380	0,54	0,05	4400	0,34	0,09	0,46	46	436	22,02	77,98
TRÊS	380	0,61	0,04	4400	0,30	0,14	0,53	28	318	14,60	85,40
QUATRO	380	0,62	0,06	4400	0,33	0,11	0,52	24	59	7,66	92,34
CINCO	380	0,60	0,05	4400	0,36	0,10	0,52	26	218	11,80	88,20
SEIS	380	0,60	0,03	4400	0,32	0,13	0,54	14	292	10,33	89,67
SETE	380	0,64	0,05	4400	0,40	0,07	0,54	16	42	5,17	94,83
OITO	380	0,65	0,08	4400	0,39	0,07	0,52	36	181	13,59	86,41
NOVE	380	0,65	0,04	4400	0,37	0,08	0,55	10	13	2,93	97,07
BRANCO	380	0,55	0,11	4400	0,29	0,12	0,42	46	565	25,00	75,00
CORRIGE	380	0,72	0,03	4400	0,34	0,08	0,61	0	0	0	100,00
TOTAL	4560	-	-	52800	-	-	-	286	2663	11,32	88,68

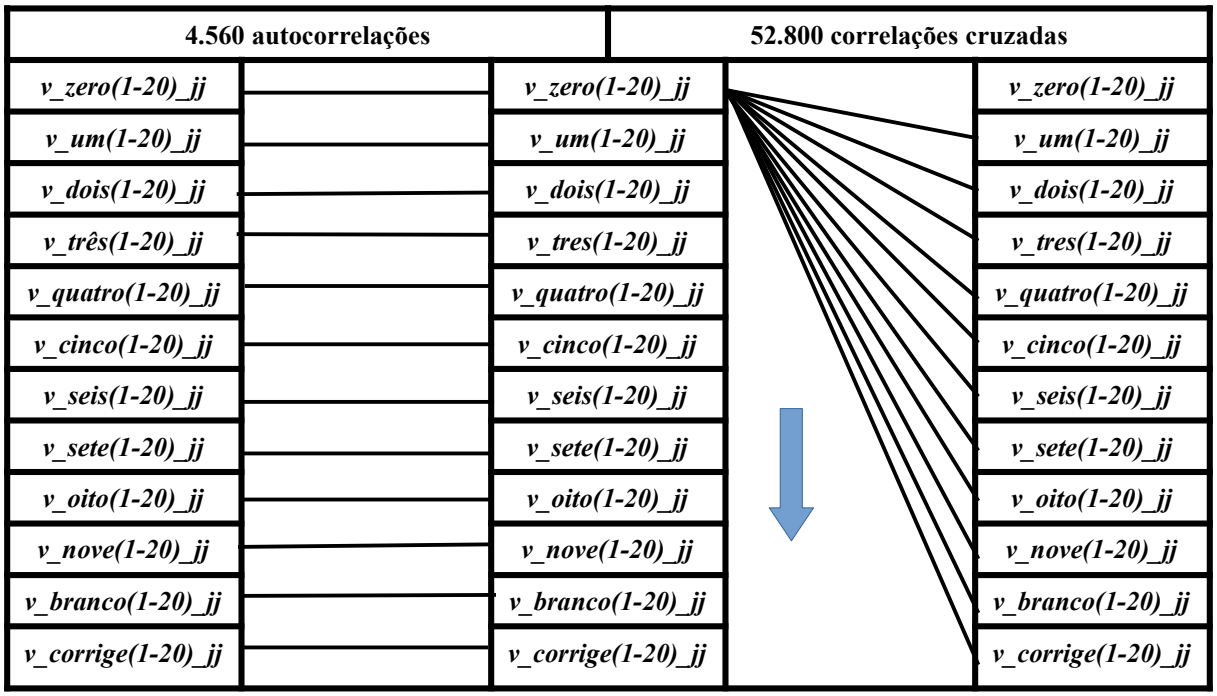
Fonte: o autor (2022)

Para se verificar estatísticamente a influência negativa da proximidade fonética entre palavras escolhidas para um sistema de comandos vocálicos, basta atentar para a quantidade

de FP da Tabela 8, onde se constata que as palavras foneticamente parecidas tiveram as maiores quantidades de falsos positivos. Veja-se por exemplo que a palavra CORRIGE, por ser foneticamente diferente de todas as demais, não teve nenhum falso positivo nem nenhum falso negativo.

A Figura 16 tenta simplificar a forma como foram feitas as possíveis correlações, por meio da mesma lógica da Figura 14, mas usando palavras diferentes, pronunciadas pelo mesmo locutor.

Figura 16: Correlações para identificação de palavras



Fonte: o autor (2022)

Na Figura 16, note-se que, do lado esquerdo, estão posicionadas as autocorrelações (entre mesmas palavras pronunciadas pelo locutor "*jj"*) e do lado direito estão posicionadas as correlações cruzadas (entre palavras distintas pronunciadas pelo locutor "*jj"*). Conforme visto na explicação do sentido da Figura 13, cada reta representa 380 correlações, no caso das autocorrelações, e 400 correlações, no caso das correlações cruzadas.

Como do lado das autocorrelações existem 12 retas, basta multiplicar por 380, para que se encontre o total de correlações, obtendo-se um total de 4.560 autocorrelações. Como do lado das correlações cruzadas exitem 11 retas, basta multiplicar por 400, para se obter 4.400 correlações por VC, que multiplicadas por 12, chega-se ao resultado total de 52.800 correlações cruzadas.

A Tabela 9 mostra o resultado de todas as possíveis correlações entre os VCs extraídos das palavras TRÊS e SEIS, catalogando a probabilidade de erro, caso um usuário pronuncie cada uma delas.

Tabela 9: Correlações entre VCs das palavras TRÊS e SEIS

20 VCs/texto	760 autocorrelações			800 correlações cruzadas			LC	FN	FP	PE%	PA%
	AC	$\overline{X}_a$	S_a	CC	$\overline{X}_c$	S_c					
TRÊS	380	0,61	0,047	400	0,57	0,049	0,59	126	141	68,6	31,4
SEIS	380	0,60	0,038	400	0,57	0,049	0,59	132	141	70,1	29,9
TOTAL	760	-	-	800	-	-	-	258	282	69,36	30,64

Fonte: o autor (2022)

Segundo os resultados da Tabela 9, a probabilidade de erro, ao se pronunciar a palavra TRÊS e a palavra SEIS é de 69,36%, o que torna o sistema inviável em termos práticos. Isso mostra que escolher palavras foneticamente parecidas para comandar máquinas pode ser uma estratégia inviável.

Os resultado das Tabelas 8 e 9 só vêm a corroborar a assertiva de Ynoguti (1999), que afirma que quanto maior o número de palavras parecidas em um sistema de reconhecimento de voz, mais difícil é o seu reconhecimento. A solução para se resolver o impasse proposto neste trabalho, como já anteriormente debatido, é substituir as palavras por outras foneticamente distintas, como as que estão na Tabela 10.

Tabela 10: Correlações entre VCs de palavras escolhidas com critério fonético

20 VCs/texto	4.560 auto correlações			52.800 correlações cruzadas			LC	FN	FP	PE%	PA%
	AC	$\overline{X}_a$	S_a	CC	$\overline{X}_c$	S_c					
FEIJÃO	380	0,71	0,06	4400	0,34	0,09	0,56	10	59	3,98	96,02
ABELHA	380	0,68	0,04	4400	0,34	0,08	0,55	7	16	2,21	97,79
URSO	380	0,69	0,03	4400	0,31	0,09	0,58	0	5	0,12	99,88
BOLA	380	0,59	0,05	4400	0,26	0,09	0,48	2	46	1,58	98,42
PÁSSARO	380	0,65	0,04	4400	0,29	0,06	0,51	2	0	0,53	99,47
MAÇÃ	380	0,73	0,03	4400	0,35	0,11	0,64	8	2	2,16	97,84
ARCO-ÍRIS	380	0,68	0,05	4400	0,29	0,10	0,56	10	0	2,64	97,36
JACARÉ	380	0,75	0,04	4400	0,31	0,12	0,64	2	2	0,58	99,42
TIGRE	380	0,57	0,1	4400	0,20	0,09	0,37	36	85	11,41	88,59
NAVIO	380	0,65	0,06	4400	0,38	0,07	0,53	18	23	5,26	94,74
SERRA	380	0,69	0,03	4400	0,38	0,11	0,61	10	36	3,45	96,55
CORRIGE	380	0,72	0,03	4400	0,30	0,08	0,60	0	0	0,00	100,00
TOTAL	4560	-	-	52800	-	-	-	105	274	2,82	97,18

Fonte: o autor (2022)

Segundo a Tabela 10, a quantidade total de falsos positivos (FP) é apenas 274, quantitativo bem inferior à quantidade total de falsos positivos (FP) da Tabela 8, que teve 2.663 falsos positivos. Isso se reflete na taxa de probabilidade de erro total, que caiu de 11,38% (Tabela 8) para 2,82% (Tabela 10).

Após a escolha do número do candidato, ou da opção de anular o voto ou votar em branco, o eleitor precisa confirmar essa escolha ou reiniciar todo o processo. Neste momento, é recomendável disponibilizar uma nova tabela de correspondência, com apenas duas opções de comando: *confirmar* ou *reiniciar*. Isso é plenamente justificável em razão de segurança e aumento da taxa de acerto do sistema, pois, segundo Ynoguti (1999), quanto menor o número de palavras possíveis de serem reconhecidas, maior será a taxa de acerto do sistema. Em outras palavras, uma das formas de reduzir a taxa de erro de um sistema de reconhecimento de voz por palavras isoladas é justamente reduzir o número de padrões de voz a ser reconhecido.

A Tabela 11 mostra o desemprenho do sistema de reconhecimento de duas palavras foneticamente distintas, uma usada para confirmação (CONFIRMA) e outra usada para correção (PÁSSARO). A palavra usada para invocação do comando de *corrigir* não poderia ser a palavra CORRIGE, em razão de sua semelhança fonética com a palavra CONFIRMA.

Tabela 11: Correlações entre VCs das palavras usadas para confirmação final

20 VCs/texto	760 autocorrelações			800 correlações cruzadas			LC	FN	FP	PE%	PA%
	AC	$\overline{X}_a$	S_a	CC	$\overline{X}_c$	S_c					
CONFIRMA	380	0,65	0,04	400	0,27	0,03	0,46	0	0	0,00	100
PÁSSARO	380	0,71	0,041	400	0,27	0,03	0,46	0	0	0,00	100
TOTAL	760	-	-	800	-	-	-	0	0	0,00	100

Fonte: o autor (2022)

Pelos resultados da Tabela 11, adotando-se as palavras CONFIRMA e PÁSSARO, a probabilidade de erro de reconhecimento em um ambiente controlado é de 0%, justamente em razão da redução do número de padrões a serem reconhecidos, aliada a uma escolha de palavras estrategicamente escolhidas em razão da diferenciação acústica entre as mesmas.

5.4 USANDO APENAS OS COEFICIENTES MFCCs ESTÁTICOS

Na Tabela 10, foram usados VCs montados pela concatenação de coeficientes MFCCs estáticos, de velocidade, de aceleração e de energia estática. A Tabela 12 mostra os mesmos resultados, usando VCs elaborados pela extração apenas dos coeficientes MFCCs estáticos.

Note-se que a redução dos atributos diminui necessariamente o custo computacional da aplicação, o que se constitui em um fator interessante, no caso da urna eletrônica, em razão da natural escassez de recursos por ela requerida.

O VC extraído, para cada palavra pronunciada, usando apenas MFCCs estáticos, passa de 3200 atributos para 1040 atributos e a memória ocupada pelo VC passa de 6,29 kB para 2,07kB.

Tabela 12: Usando apenas os coeficientes MFCCs estáticos

20 VCs/texto	4560 auto correlações			52800 correlações cruzadas			LC	FN	FP	PE%	PA%
	AC	$\overline{X}_a$	S_a	CC	$\overline{X}_c$	S_c					
FEIJÃO	380	0,74	0,05	4400	0,42	0,10	0,63	51	16	14,1	85,9
ABELHA	380	0,72	0,04	4400	0,44	0,08	0,61	38	22	10,5	89,5
URSO	380	0,71	0,03	4400	0,40	0,08	0,62	52	2	13,7	86,3
BOLA	380	0,61	0,04	4400	0,33	0,10	0,52	164	16	43,5	56,5
PÁSSARO	380	0,67	0,04	4400	0,35	0,06	0,54	1	2	0,3	99,7
MAÇÃ	380	0,74	0,03	4400	0,44	0,11	0,66	88	6	23,3	76,7
ARCO-ÍRIS	380	0,68	0,05	4400	0,36	0,09	0,57	5	8	1,5	98,5
JACARÉ	380	0,74	0,04	4400	0,41	0,11	0,64	80	10	21,3	78,7
TIGRE	380	0,60	0,08	4400	0,31	0,08	0,46	173	36	46,3	53,7
NAVIO	380	0,67	0,05	4400	0,45	0,08	0,58	212	32	56,5	43,5
SERRA	380	0,73	0,03	4400	0,48	0,10	0,66	118	20	31,5	68,5
CORRIGE	380	0,72	0,03	4400	0,34	0,08	0,61	0	2	0,05	99,95
TOTAL	4560	-	-	52800	-	-	-	982	172	21,9	78,1

Fonte: o autor (2022)

Entretanto, conforme Tabela 12, apesar da melhora no quesito custo cumputacional, há uma considerável queda de probabilidade de acerto do sistema, indo de 97,18% para 78,1%.

5.5 REDUZINDO O NÚMERO DE PALAVRAS

Uma forma não computacional de reduzir a taxa de erro de um sistema de reconhecimento de voz para execução de treze comandos distintos, como é o caso da urna eletrônica, é reduzir o máximo possível o número de palavras do dicionário treinado na base do sitema. Note-se que, no caso da urna eletrônica, seriam preciso, numa primeira análise, treze palavras para invocação dos seus comandos.

Entretanto, a escolha dos comandos pela pronúncia das palavras podem ser realizadas em etapas, associando cada palavra a mais de um comando. No caso da urna eletrônica, ao se reduzir de treze palavras para seis palavras, a associação entre palavras e comandos deve ser de uma palavra para dois comandos, em duas etapas. Em outras palavras, é possível escolher o número do candidato em uma eleição usando voz com apenas seis palavras previamente treinadas.

Como o mais conveniente é escolher seis palavras foneticamente distintas, pode-se ecolhê-las tendo como base os resultados de menor erro obtidos na Tabela 12, como é sugerido no Quadro 20.

Quadro 20: 1ª etapa

PÁSSARO	ABELHA	JACARÉ	URSO	ARCO-ÍRIS	MAÇÃ
0 ou *6*	*1* ou *7*	*2* ou *8*	*3* ou *9*	*4* ou *branco*	*5* ou *corrige*

Fonte: o autor (2022)

Segundo o Quadro 20, um eleitor que queira votar em branco teria que pronunciar a palavra ARCO-ÍRIS. Essa foi a primeira etapa. Em seguida uma nova correspondência deve ser impressa, como a do Quadro 21.

Quadro 21: 2ª etapa

PÁSSARO	URSO
4	*branco*

Fonte: o autor (2022)

Como o eleitor quer votar em branco, conforme Quadro 21, deve pronunciar a palavra URSO. Note-se que, para garantia do sigilo, apenas as palavras da 1ª etapa precisam ser permutadas, mantendo-se os pares de comandos nas mesmas posições.

Para demonstrar o desempenho do sistema usando as seis palavras do Quadro 20, foi elaborado a Tabela 13, que foi construída tendo como base os VCs contendo apenas os coeficientes MFCCs estáticos

Tabela 13: Usando apenas MFCCs estáticos com 6 palavras

20 VCs/texto	2280 auto correlações			12000 correlações cruzadas			LC	FN	FP	PE%	PA%
	AC	$\overline{X}_a$	S_a	CC	$\overline{X}_c$	S_c					
ABELHA	380	0,72	0,04	2000	0,40	0,08	0,60	0	12	0,6	99,4
URSO	380	0,71	0,03	2000	0,38	0,06	0,59	6	0	1,6	98,4
ARCO-ÍRIS	380	0,68	0,05	2000	0,39	0,06	0,55	3	4	0,9	99,0
PÁSSARO	380	0,67	0,04	2000	0,35	0,07	0,55	1	2	0,4	99,6
MAÇÃ	380	0,74	0,039	2000	0,45	0,08	0,64	8	4	2,3	97,7
JACARÉ	380	0,74	0,04	2000	0,40	0,09	0,63	20	6	5,6	94,4
TOTAL	2280	-	-	12000	-	-	-	38	28	1,9	98,1

Fonte: o autor (2022)

A Tabela 13 (que usou seis palavras extraindo apenas coeficientes MFCCs estáticos) mostra a probabilidade de acerto do modelo, no caso 98,1%, superior aos 78,1% da Tabela 12 (que usou doze palavras extraindo apenas coeficientes MFCCs estáticos) e superior aos 97,18% da Tabela 10 (que usou doze palavras extraindo coeficientes MFCCs estáticos, de velocidade e de aceleração).

Até agora a melhor opção, em termos de desempenho e custo computacional, foi o uso de seis palavras com extração apenas dos coeficientes MFCCs estáticos.

Para saber se o uso das seis palavras com extração dos coeficientes MFCCs estáticos, de velocidade e de aceleração seria mais viável que extrair apenas os coeficientes MFCCs estáticos, elaborou-se a Tabela 14

Tabela 14: Usando MFCCs estáticos, de velocidade e de aceleração com 6 palavras

20 VCs/texto	2280 auto correlações			12000 correlações cruzadas			LC	FN	FP	PE%	PA%
	AC	$\bar{X}_a$	S_a	CC	$\bar{X}_c$	S_c					
ABELHA	380	0,68	0,04	2000	0,30	0,07	0,53	0	0	0	100
URSO	380	0,69	0,03	2000	0,29	0,06	0,54	0	0	0	100
ARCO-ÍRIS	380	0,68	0,05	2000	0,33	0,06	0,53	0	4	0,2	99,8
PÁSSARO	380	0,65	0,04	2000	0,28	0,07	051	0	2	0,1	99,9
MAÇÃ	380	0,73	0,03	2000	0,36	0,08	0,62	0	0	0	100
JACARÉ	380	0,75	0,04	2000	0,32	0,1	0,63	0	0	0	100
TOTAL	2280	-	-	12000	-	-	-	0	6	0,05	99,95

Fonte: o autor (2022)

Comparando-se os resultados entre as Tabelas 13 e 14, nota-se que a probabilidade de acerto usando seis palavras com uso de extração de coeficeintes MFCCs estáticos, dinâmicos e de velocidade (Tabela 14) é ligeiramente superior quando se usa as mesmas seis palavras, extraindo-se apenas os coeficientes MFCCs estáticos (Tabela 13). Passou-se de 98,1% para 99,5 % de probabilidade de acerto, um aumento de 1,4%, considerado pequeno, quando se leva em conta o alto custo computacional.

Os dados levam a crer que, com uso de seis palavras foneticamente distintas, a extração das MFCCs estáticas é suficiente para se ter um ótimo desempenho, em ambientes controlados.

5.6 UMA TRAVA DE SEGURANÇA PARA CONFIRMAR O VOTO

O comando final de confirmar é o momento mais importante e crítico do sistema. No momento de escolhê-lo usando voz, uma série de fatos inesperados podem acontecer, ainda que estatisticamente improváveis, diante da implementação de um tratamento acústico explicitado no Apêndice II. Por exemplo, alguém no local de votação pode ter sido preso em flagrante, causando tumulto na seção eleitoral e, com isso, o estado emocional do eleitor foi

severamente alterado e ele trocou a palavra que invoca o comando *reiniciar* pela palavra que invoca o comando de *confirmar*. Caso isso ocorra, não tem como reverter o erro.

Daí, sugere-se que, a função final de confirmar o voto só seja executada, depois do eleitor pronunciar a palavra correspondente por três vezes seguidas.

Caso uma delas seja diferente ou tenha sido reconhecida de forma errada, o sistema deve reiniciar o processo de confirmação exigindo a pronúncia da palavra CONFIRMA por mais três vezes seguidas.

E se o eleitor demorar muito repetindo as locuções sem ser reconhecido ou suas locuções não estiverem atingindo os LC em razão de rouquidão ou inflamação na garganta? É o que será visto na próxima seção.

5.7 UMA SOLUÇÃO DE CONTINUIDADE

Uma série de acontecimentos podem fazer com que um eleitor tenha dificuldades em não ter sua voz reconhecida no dia da eleição. Exemplos: o eleitor pode ter contraído uma doença em suas cordas vocais (um câncer ou um cisto), modificando radicalmente as componentes expectrais de sua voz; o ambiente da seção eleitoral pode ter se tornado muito barulhento; os ruídos internos do *hardware* da urna eletrônica se tornaram demasiadamente altos a ponto de inviabilizar o reconhecimento de voz, causando muitos falsos positivos e falsos negativos.

Essas dificuldades não podem ser obstáculos para que o eleitor não possa votar, pois o direito ao voto é considerado um direito fundamental do ser humano.

Então, neste caso, o presente trabalho de reconhecimento de voz também poderia ajudar este eleitor a votar sem tocar na urna eletrônica e usando sua voz, sem extração dos coeficientes *MFCCs* ou outras técnicas de reconhecimento de voz.

Usando a mesma lógica das permutações das palavras, é possível fazer com que o Presidente de Mesa ou alguém de confiança indicada pelo próprio eleitor, escute a pronúncia das palavras pronunciadas por este e invoque as funções da urna eletrônica acionando o teclado do terminal do mesário, sem saber quais comandos foram acionados. A solução proposta de uso do teclado ratifica o pensamento de Ynoguti (1999), que se manifesta afirmando que é improvável que o reconhecimento de fala possa substituir completamente os tradicionais teclados.

Mesmo assim, é possível, sem que haja qualquer processamento digital de voz, fazer com que alguém de confiança do eleitor vote em nome dele, garantindo-se o sigilo total do voto e com segurança.

Para o eleitor, em termos práticos, não se muda muita coisa, além do que já foi visto até aqui. Simplesmente ele irá pronunciar as palavras que deseja, conforme tabela de correspondência privada a ele disponibilizada na tela da urna eletrônica após cada ação de comando. A diferença então estará na atuação de uma segunda pessoa (podendo ser um dos mesários, se assim requerer o eleitor), que ouvirá a locução pronunciada pelo eleitor e acionará a combinação de teclas, conforme tabela de correspondência pública impressa no caderno de votação.

Note-se que, para garantia do sigilo do voto, deve existir novamente a ação indissociável de duas tabelas de correspondência. Uma fixa, pública, impressa no caderno de votação, gravada internamente na urna eletrônica e conhecida por todos, inclusive pelo Presidente de Mesa. E outra tabela de correspondência privada, impressa na tela da urna eletrônica, que só o eleitor tem acesso a ela, que permuta após cada ação de comando executada.

Aqui vale esclarecer que há a necessidade de uma pequena modificação no número de locuções. Isso se deve ao fato de que o terminal do mesário não possui a tecla que invoca o comando *branco.* Destarte, há a necessidade de mais uma combinação de teclas que deve ser acionada pelo Presidente de Mesa, para haver encaixe posicional lógico entre as locuções das correspondências privadas e os 13 comandos básicos da urna eletrônica. No caso, adicionou-se a palavra TATU.

Outro detalhe neste procedimento é que, como não há processamento digital de áudio que reconheça palavras (pois o reconhecimento neste caso é feito pelo Presidente de Mesa ou por outra pessoa indicada pelo eleitor), não há necessidade de criar dois tipos de correspondência (as de escolha e as de confirmação/reinício), como foi feito na seção anterior, de maneira que todos os comandos devem estar disponibilizados numa única correspondência.

É claro que, para acionamento deste procedimento, deve haver solicitação livre e consciente do eleitor, de forma que o Presidente de Mesa informa pelo teclado do terminal do mesário a invocação do mecanismo, da mesma forma que se habilita atualmente outras funções, como a suspensão do voto.

Feito isto, é preciso estabelecer combinações de teclas no terminal do mesário que deverão acionar comandos da urna eletrônica relacionadas com as palavras pronunciadas pelo eleitor.

O Quadro 22 estabelece relações posicionais que não permutam entre teclas do terminal do mesário e palavras, gravadas internamente na urna eletrônica e impressas no caderno de votação para consulta do mesário.

Quadro 22: Correspondência pública

LOCUÇÕES	FEIJÃO	ABELHA	URSO	BOLA	PÁSSARO	ARCOÍRIS	MAÇÃ	JACARÉ	TIGRE	NAVIO	SERRA	ONÇA	TATU
COMBINAÇÃO DE TECLAS	12	21	34	43	56	65	78	89	90	18	72	63	54
FUNÇÃO	Selecionar a correspondente *string* na correspondência pública e descobrir a sua posição na correspondência privada. Em seguida, acionar o respectivo comando.												

Fonte: o autor (2022)

Note-se que a combinação de teclas, no Quadro 22, nada está relacionada com os números dos candidatos escolhidos pelo eleitor, pois este só deverá pronunciar as locuções discriminadas na 1ª linha.

A título de exemplificação, suponha-se que o eleitor chegou em frente à urna eletrônica e solicitou ao Presidente de Mesa o presente mecanismo de votação mediante pronúncia de palavras sem reconhecimento digital de voz. O Presidente de Mesa invoca o procedimento e a 1ª correspondência privada é impressa no terminal do eleitor, conforme Quadro 23.

Quadro 23: 1ª correspondência privada

BOLA	PÁSSARO	ARCO-ÍRIS	MAÇÃ	JACARÉ	TIGRE	NAVIO	SERRA	ONÇA	URSO	FEIJÃO	ABELHA	TATU
0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*branco*	*corrige*	*confirma*

Fonte: o autor (2022)

Caso o eleitor esteja votando para o cargo de prefeito e deseja escolher o número 25, ele tem que pronunciar a palavra ARCO-ÍRIS, de acordo com Quadro 23. Note-se que o Presidente de Mesa não tem acesso a esta correspondência privada, mas à correspondência pública impressa no caderno de votação, conforme Quadro 22.

Após ouvir a pronúncia da palavra ARCO-ÍRIS, o Presidente de Mesa deverá consultar a correspondência pública (Quadro 22) impressa no manual do mesário e pressionar a combinação de teclas correspondentes à locução ARCO-ÍRIS, no caso, a tecla 6, seguida da tecla 5, formando o número 65 e, em seguida, deverá pressionar a tecla confirma do terminal

do mesário. A escolha do número 65 seguida da tecla confirma executará outro comando para verificar em que posição se encontra a *string* ARCO-ÍRIS na correspondência privada, acionando o respectivo comando, no caso em tela, a escolha do número 2.

Note-se que a escolha das teclas no terminal do mesário pelo Presidente de Mesa nada tem a ver com a escolha do eleitor, sendo impossível ao Presidente de Mesa saber qual o comando o eleitor escolheu.

Destarte, por meio do mecanismo acima, o eleitor escolhe o número 2, primeiro número do seu candidato, sem que ninguém saiba ou tenha a certeza disso, nem mesmo o Presidente de Mesa ou alguém de confiança do eleitor.

Em seguida, nova correspondência privada é gerada, como no Quadro 24, com uma nova permutação das palavras, para escolha do segundo número, no caso o número 5.

Quadro 24: 2ª correspondência privada

ONÇA	FEIJÃO	ABELHA	URSO	JACARÉ	MAÇÃ	NAVIO	TATU	BOLA	TIGRE	ARCO-ÍRIS	PÁSSARO	SERRA
0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*branco*	*corrige*	*confirma*

Fonte: o autor (2022)

No caso, o eleitor teria que pronunciar a palavra MAÇÃ, conforme 2ª correspondência privada (Quadro 24).

Após ouvir a pronúncia da palavra MAÇÃ, o Presidente de Mesa deverá consultar a correspondência pública (Quadro 22) impressa no manual do mesário e pressionar a combinação de teclas correspondentes à locução MAÇÃ, no caso, a tecla 7, seguida da tecla 8, formando o número 78 e, em seguida, deverá pressionar a tecla confirma.

A escolha do número 78, seguida da tecla confirma, deverá executar outro comando para verificar em que posição se encontra a *string* MAÇÃ, na correspondência privada, acionando o respectivo comando, no caso em tela, a escolha do número 5.

Em seguida, percebendo que os dígitos do número do candidato ao cargo de Prefeito já foram completados, a urna eletrônica fará a seguinte pergunta:

“Você quer votar no candidato cujo número é 25?”

Nova correspondência então é gerada e disponibilizada ao eleitor, conforme Quadro 23.

Quadro 25: 3ª Correspondência privada

ARCO-ÍRIS	SERRA	ABELHA	URSO	BOLA	MAÇÃ	NAVIO	TATU	JACARÉ	TIGRE	ONÇA	PÁSSARO	FEIJÃO
0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*branco*	*corrige*	*confirma*

Fonte: o autor (2022)

Diante da pergunta e possíveis respostas disponibilizadas pela urna eletrônica, se o eleitor perceber que a tela indicou corretamente o número do seu candidato, ele deverá pronunciar ao microfone a palavra que invoca o comando *confirma*, conforme Quadro 25.

No caso, o eleitor teria que pronunciar a palavra FEIJÃO, conforme 3ª correspondência privada (Quadro 25).

Após ouvir a pronúncia da palavra FEIJÃO, o Presidente de Mesa deverá consultar a correspondência pública (Quadro 22) impressa no manual do mesário e pressionar a combinação de teclas correspondentes à locução FEIJÃO, no caso, a tecla 1, seguida da tecla 2, formando o número 12 e, em seguida, deverá pressionar a tecla confirma.

A escolha do número 12, seguida da tecla confirma, deverá executar outro comando para verificar em que posição se encontra a *string* FEIJÃO, na correspondência privada, acionando o respectivo comando, no caso em tela, a escolha da opção de confirmar o número escolhido.

Estes procedimentos são instantâneos e praticamente em tempo real em virtude da velocidade computacional. A escolha dos dois números como no exemplo anterior não deve durar mais que 15 segundos.

Note-se que este procedimento pode inclusive ser uma opção a ser disponibilizada a qualquer eleitor, independentemente dele ter ou não ter fornecido seus áudios de treinamento. O eleitor que não forneceu seus áudios de treinamento poderia optar por este procedimento em razão de algum receio sanitário, em razão de possíveis contaminações do leitor biométrico ligadas a doenças endêmicas ou pandêmicas transmitidas por contato. A adoção deste procedimento poderia inclusive reduzir gastos, em razão da desnecessidade de tratamento acústico ou uso de microfones dinâmicos.

5.8 CONCLUSÃO DO CAPÍTULO

Neste capítulo, são feitas simulações e testes, catalogando os resultados em tabelas. Baseado nestes resultados, abre-se um leque de possibilidades de possíveis implementações, procurando eliminar ou diminuir possíveis problemas por ocasião da implementação da presente ideia.

É sugerida uma trava de segurança, com um intuito de possibilitar que o comando final de confirmar o voto do eleitor fosse reconfirmado, pois, como se sabe, esta decisão é irreversível, não podendo o eleitor requerer mudança de voto após a urna eletrônica reconhecer a locução CONFIRMA para finalizar a sua escolha.

Propõe-se uma solução de continuidade, caso ocorra algum problema no reconhecimento de voz, ligado a ruídos internos do *hardware,* doenças nas cordas vocais ou mesmo em razão de solicitação livre e espontânea do eleitor em realizar o procedimento.

Os resultados das correlações entre os vetores de características extraídos dos áudios de treinamento mostram que, em ambiente controlado, ao se escolher 12 palavras foneticamente distintas entre si para comandar a urna eletrônica, tem-se um ganho na taxa de acerto de 88,68% para 97,18% quando se extrai os coeficientes MFCCs estáticos e dinâmicos.

Ao se escolher apenas 6 palavras foneticamente distintas, extraindo-se somente os coeficientes MFCCs estáticos, há um ganho na taxa de acerto de 78,1% para 98,1%, o que demonstra a eficácia da estratégia. Ao se acrescentar nesta última estratégia a extração dos coeficientes MFCCs dinâmicos de velocidade e de aceleração, obtem-se um ganho na taxa de acerto de 98,1% para apenas 99,95%, não se justificando o aumento do custo computacional.

Os testes demonstram que a melhor estratégia foi escolher 6 palavras foneticamente distintas com extração unicamente dos coeficientes MFCCs estáticos.

CAPÍTULO 6

6. CONCLUSÕES E TRABALHOS FUTUROS

Neste capítulo, apresentam-se as conclusões derradeiras, enfatizando as principais contribuições desta dissertação, frente aos resultados alcançados e à literatura consultada, propondo várias ações de pesquisas para avançar nos objetivos ligados a este trabalho.

6.1 CONCLUSÕES E CONTRIBUIÇÕES DESTA DISSERTAÇÃO

Tendo em vista os resultados das correlações, aponta-se a escolha da técnica de extração dos coeficientes *MFCCs* como uma técnica adequada, em ambientes controlados, para extração de dados biométricos a fim de caracterizar a voz de eleitores, com a finalidade de que os mesmos possam ser autenticados usando voz, contribuindo para o melhoramento da Justiça Eleitoral, no quesito acessibilidade (para eleitores deficientes, em razão da ausência dos braços).

A estratégia de substituir as locuções que seriam normalmente consideradas para manipular a urna eletrônica ou qualquer máquina (as locuções, UM, DOIS, TRÊS…), por outras acusticamente distintas entre si, como sugerido na Seção 5.3, mostrou-se bem sucedida neste trabalho, frente aos resultados das correlações encontradas. Esta estratégia parece ser uma novidade nunca catalogada antes na literatura especializada, razão pela qual deve-se considerá-la com uma contribuição.

Visando a aumentar ainda mais a taxa de acerto, constatou-se que, ao se reduzir o número de padrões de palavras a serem reconhecidas, houve um ganho na taxa de acerto, aliada a uma economia de custo computacional, possibilitado pela não necessidade de extração dos coeficientes *MFCCs* dinâmicos.

Outra contribuição está relacionada à estratégia de permutar as palavras que invocam os comandos da urna eletrônica após cada ação de comando, o que se constitui em uma forte barreira de segurança e sigilo, de forma que ela se consubstancia em um duplo fator de

autenticação simultâneo, na qual o usuário pode ser testado, em um só momento, sobre a veracidade de sua senha numérica e sua biometria vocal.

6.2 TRABALHOS FUTUROS

- Pesquisar sobre a existência de técnicas e suas performances que possam detectar se um áudio gravado foi produzido diretamente por um ser humano ou se ele foi reproduzido por alguma máquina. Caso não exista essa técnica, pensar em criar uma com o objetivo de evitar que uma pessoa se passe por outra usando cópia gravada de sua voz em sistemas de reconhecimento biométrico usando voz;

- Criar um sistema de presença biométrica usando voz em encontros, reuniões ou cursos virtuais (por meio da pronúncia da locução PRESENTE, por exemplo, ou por meio de código de autenticação numérica informada por voz, enviado por mensagem eletrônica ao celular), com o intuito de evitar que pessoas estejam assistindo aulas ou participando de reuniões virtuais sem estarem de corpo presente, apenas com a foto do perfil ou por meio de outra pessoa que não deveria assumir o lugar daquela; o código numérico seria informado por meio de locuções permutadas, da mesma forma que proposto na Seção 3.2.

- Criar um sistema de acesso e autenticação de usuário, usando senhas e correspondências sonoras criadas pelo usuário dos mais diversos tipos (notas musicais diferentes, com instrumentos diferentes, tonalidades diferentes, sons de pássaros diferentes, letras do alfabeto, etc) produzidos pelo celular. Exemplo: Dó – 0, Ré – 1, Mi – 2, Fá – 3, Sol – 4, Lá – 5, Si – 6, Dó# - 7, Ré# - 8, Sol# - 9; Lá # - corrige, Si2 – confirma;

- Fazer testes com usuários que imitam vozes, com o intuito de saber se a técnica de extração dos coeficientes *mel cepstrais* consegue identificar a tentativa de fraude;

- Fazer testes com taxas de amostragens maior do que 8000 Hz, *frames* maiores, banco de filtros *mel* em maior número, para saber se essas mudanças afetariam ou não as taxas de erro quando se usa a técnica de extração dos coeficientes *MFCCs*;

- Saber se o envelhecimento natural afeta o reconhecimento de voz e até quanto tempo as amostras de voz poderiam servir como parâmetro biométrico robusto. De quanto em quanto tempo a pessoa teria que revisar seus dados biométricos vocais?;

- Investigar até que ponto doenças respiratórias como a gripe afetariam o reconhecimento biométrico de voz ou se não haveria perda de desmpenho no reconhecimento em virtude da doença;

- Criar um sistema que atue em caixas eletrônicos para fazer operações bancárias usando voz, sem precisar tocar em teclas e nem ser reconhecido pelas imagens das impressões digitais, mas pela voz do usuário. O usuário apenas teria que informar, mediante código de locuções vocais, qual seria o seu CPF, por meio da estratégia de permutação descrita neste trabalho;

- Criar um sistema que atue em caixas eletrônicos para fazer operações bancárias usando áudios gravados no celular do usuário, sem precisar tocar em teclas do caixa eletrônico e nem ser reconhecido pelas impressões digitais, mas usando função de correlação, que tem certa imunidade a ruídos, sem usar os coeficientes *MFCCs*, o que dispensaria, em tese, qualquer tratamento acústico;

REFERÊNCIAS

ALDARMAKI, Hanan et al. Unsupervised automatic speech recognition: A review. **Speech Communication,** 2022.

ALMEIDA, C. R. **Extratores de características acústicas inspirados no sistema periférico auditivo**, p. 13-17, Dissertação de Mestrado, Universidade Federal de Sergipe, São Cristóvam, SE, 2014.

BEZERRA, M.R. **Reconhecimento Automático de Locutor para Fins Forenses, Utilizando Técnicas de Redes Neurais**, Dissertação de Mestrado, IME, Rio de Janeiro, 2001.

BOERSMA, P.; WEENINK, **D. Praat: doing phonetics by computer.** Disponível em: <https://www.fon.hum.uva.nl/praat/>. Acesso em: 05 de nov. de 2022.

BOUROUBA, E.-Hocine; BEDDA, Mouldi; DJEMILI, Rafik. Isolated words recognition system based on hybrid approach DTW/GHMM. **Informatica**, v. 30, n. 3, 2006.

CAMPBELL, Joseph P. Speaker recognition: A tutorial. **Proceedings of the IEEE**, v. 85, n. 9, p. 1437-1462, 1997.

REBILLARD, G. **Fonctionnement de la cochlée.** 2021. Disponível em: <http://www.cochlea.eu/cochlee/fonctionnement> Acesso em: 05 de nov. de 2022.

CUADROS, D.R.C. **Reconhecimento de voz e de locutor em ambientes ruidosos: comparação das técnicas MFCC e ZCPA**, p.12, p.16, p.17, p.98. Dissertação de Mestrado, Universidade Federal Fluminense, Niteroi-RJ, 2007.

DAN, Z. Speaker Recognition Based on LS-SVM. **The 3rd International Conference on Innovative Computing Information and Control**, p. 25–28, 2008.

FARRELL, K. R.; MAMMONE, R.; ASSALEH, K. Speaker Recognition Using Neural Networks and Conventional Classifiers. **IEEE Trans. Speech, and Audio Processing**, v. 2, n. 1, p.194–205, 1994.

FIORILLO, Luca et al. COVID-19 surface persistence: a recent data summary and its importance for medical and dental settings. **International journal of environmental research and public health**, v. 17, n. 9, p. 3132, 2020.

GORDILLO, Christian Dayan Arcos. **Realce e Reconhecimento de Voz Contínua em Ambientes Adversos**. 2018. Tese de Doutorado. PUC-Rio.

KACUR, Juraj; VARGA, Mario; ROZINAJ, Gregor. ZCPA features for speech recognition. In: **2012 IX International Symposium on Telecommunications (BIHTEL).** IEEE, 2012. p. 1-4.

KUMAR, Tapesh; MAHRISHI, Mehul; MEENA, Gaurav. A comprehensive review of recent automatic speech summarization and keyword identification techniques. **Artificial Intelligence in Industrial Applications**, p. 111-126, 2022.

LATHI, B. P. **Digital and Analog Communication Systems.** 1989.

LIN, Shoufeng. **Logarithmic Frequency Scaling and Consistent Frequency Coverage for the Selection of Auditory Filterbank Center Frequencies**. arXiv preprint arXiv:1801.00075, 2017.

MANNELL, R. **FFT and LPC spectrum settings.** 2020. Disponível em:<https://www.mq.edu.au/about/about-the-university/our-faculties/medicine-and-health-sciences/departments-and-centres/department-of-linguistics/our-research/phonetics-and-phonology/speech/acoustics/speech-acoustics/fft-and-lpc-spectrum-settings> Acesso em: 05 de nov. de 2022.

MCLOUGHLIN, Ian. **Applied speech and audio processing: with Matlab examples**. Cambridge University Press, 2009.

OLIVEIRA, Marcos Paulo Barros. **Verificação Automática de locutor, Dependente do Texto, Utilizando Sistemas Híbridos MLP/HMM.** Dissertação de Mestrado – Instituto Militar de Engenharia / IME – 2001.

OPPENHEIM, Alan V.; BUCK, John R.; SCHAFER, Ronald W. **Discrete-time signal processing. Vol. 2**. Upper Saddle River, NJ: Prentice Hall, 2001.

O'SHAUGHNESSY, Douglas. Speech Communication, Human and Machine Addison Wesley. **Reading MA,** p. 40, 1987.

PARANAGUÁ, Evandro David Silva. **Reconhecimento de locutores utilizando modelos de markov escondidos contínuos**. Mestrado em ciências em engenharia, Instituto Militar de Engenharia, Rio de Janeiro, p. 33, 1997.

PETRY, A. **Reconhecimento Automático de Locutor Utilizando Medidas Invariantes Dinâmicas Não-Lineares. Tese (Doutorado)** — Programa de Pós-Graduação em Ciência da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil, 2002.

PEACOCKE, Richard D.; GRAF, Daryl H. **An introduction to speech and speaker recognition**. In: Readings in Human–Computer Interaction. Morgan Kaufmann, 1995. p. 546-553.

PUJOL, Rémy. **Oreille.** 2016. Disponível em: < http://www.cochlea.eu/oreille-generalites> Acesso em: 05 de nov. de 2022.

RABINER, Lawrence; JUANG, Biing-Hwang. **Fundamentals of speech recognition**. Prentice-Hall, Inc., 1993.

REYNOLDS, D. A.; ROSE, R. C. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models**. IEEE Trans. Speech Audio Process,** v. 3, p. 72–83, 1995.

SCILAB. Disponível em: <https://www.scilab.org/>. Acesso em: 05 de nov. de 2022.

SHAO, Yang; WANG, Deliang. Robust speaker recognition using binary time-frequency masks. In: **2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings**. IEEE, 2006. p. I-I.

SILVA, Marco Aurélio Botelho. **Uma contribuição para a caracterização do sinal de voz envelhecida**, p.5. Dissertação de Mestrado, Universidade Federal Fluminense, Niteroi-RJ, 2010.

SKOWRONSKI, Mark D.; HARRIS, John G. Aumento da largura de banda do filtro mfcc para reconhecimento de fonemas com ruído robusto. In: **2002 IEEE International Conference on Acoustics, Speech, and Signal Processing** . IEEE, 2002. pág. I-801-I-804.

TISBY, Naftali Z. On the application of mixture AR hidden Markov models to text independent speaker recognition. **IEEE Transactions on Signal Processing,** v. 39, n. 3, p. 563-570, 1991.

TRIBUNAL REGIONAL ELEITORAL DO RJ. **Urna eletrônica brasileira**. 2020. Disponível em: < https://www.tre-rj.jus.br/imagens/fotos/2020-01-09-urna-portal-eleicoes-tse/> Acesso em: 05 de nov. de 2022.

TRIBUNAL SUPERIOR ELEITORAL. **Biometria: identificação do eleitor pelas digitais garante mais segurança às eleições.** 2017. Disponível em:<https://www.tse.jus.br/imprensa/noticias-tse/2017/Marco/biometria-identificacao-do-eleitor-pelas-digitais-garante-mais-seguranca-as-eleicoes/>. Acesso em: 05 de nov. de 2022.

TRIBUNAL SUPERIOR ELEITORAL. **Estatística do Eleitorado Ceará,** Outubro de 2018. Disponível em: <https://sig.tse.jus.br/ords/dwapr/seai/r/sig-eleitor-eleitorado-mensal/home?p0_mes=10&session=102814702250128>. Acesso em: 05 de nov. de 2022.

TRIBUNAL SUPERIOR ELEITORAL. **Biometria.** 2021. Disponível em:<https://www.tse.jus.br/eleitor/biometria/biometria>. Acesso em: 05 de nov. de 2022.

TRIBUNAL SUPERIOR ELEITORAL. **Estatística do Eleitorado Nacional,** Novembro de 2022. Disponível em: <https://sig.tse.jus.br/ords/dwapr/seai/r/sig-eleitor-eleitorado-mensal/home?session=232031254512737>. Acesso em: 19 de dez. de 2022.

WANG, N. Robust Speaker Recognition using Both Vocal Source and Vocal Tract Features Estimated from Noisy Input Utterances. **IEEE International Symposium on Signal Processing and Information Technology, 2007.**

YNOGUTI, Carlos Alberto. **Reconhecimento de fala contínua usando modelos ocultos de Markov.** Universidade Estadual de Campinas, Campinas, São Paulo, 1999.

APÊNDICE I
A COMUNICAÇÃO SONORA HUMANA

O ser humano produz sinais sonoros através da movimentação das pregas vocais, que, por meio de compressões e descompressões das partículas do ar, transmitem-nos de forma irradiada, de sorte que, se estes sinais chegarem ao aparelho auditivo (ouvido) de outro ser humano com potência suficiente, possam ser interpretados.

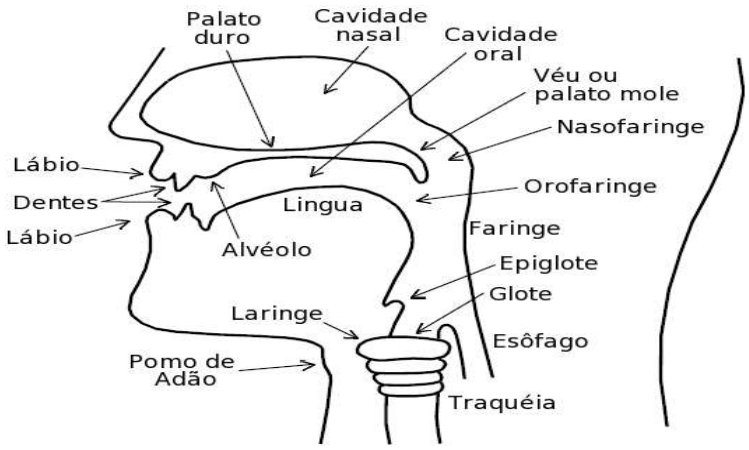
A quantidade de compressões e descompressões que cada prega vocal é capaz de promover por segundo é a frequência sonora medida em Hz. O valor desta frequência, bem como sua potência, é função de características biológicas e genéticas de cada ser humano.

A voz humana pode ser considerada complexa, no que diz respeito às suas características acústicas. Isto quer dizer por exemplo que uma determinada palavra pronunciada em um dado instante, apesar de ter um mesmo sentido sendo pronunciada em um momento posterior, para quaisquer pessoas que a escutem e a pronunciem, pode ser bem diferente quando se parte para uma análise de frequências que compõem o sinal da fala. Em outras palavras, o trato vocal humano não é um sistema invariante no tempo.

Esse fenômeno é decorrência de vários fatores: estado emocional do locutor no momento da fala; distância de seu trato vocal até o microfone onde se capta o áudio; presença de diversas espécies de ruídos que contaminam o sinal, advindos tanto do ambiente, quanto dos equipamentos eletrônicos.

A Figura 17 mostra uma simplificação do conjunto de órgãos que estão direta ou indiretamente relacionados com a produção da fala humana.

Figura 17: Aparelho vocal humano



Fonte: Adaptado de SILVA (2010)

A voz é produzida pela vibração das cordas vocais situadas na Glote e essa energia advém da passagem de ar que vem dos pulmões. A intensidade e a velocidade da passagem de ar é conscientemente controlada pela pessoa, através do diafragma, o principal músculo do corpo responsável pela respiração humana.

A corda vocal é o primeiro e o principal órgão responsável pela característica sonora de cada pessoa. No caso de indivíduos do sexo masculino, as cordas vocais são mais grossas, possuindo mais músculos, o que leva essa categoria de indivíduos a gerarem sons com frequências situadas numa faixa mais grave (80 Hz até 150 Hz), quando comparadas com o sexo oposto. Já indivíduos do sexo feminino, ou menores de 13 anos de idade, possuem cordas vocais mais finas, com menos músculos, o que leva a essa categoria de indivíduos a produzirem sons em faixas de frequências mais agudas (150 Hz até 250 Hz), quando comparadas aos indivíduos masculinos adultos.

A partir daí, o som produzido vai passar pelo primeiro filtro natural construído pela própria natureza humana e influenciado por características genéticas de cada ser humano, que é a faringe. Esta atenuará ou aumentará o nível sonoro de determinadas frequências, de acordo com o formato, tamanho e demais características biológicas e genéticas de cada pessoa.

Após a faringe, o ar, dependendo do sinal que o indivíduo queira transmitir, nasal ou oral, poderá tomar dois caminhos: ou vai totalmente para a orofaringe (sons orais); ou o som vai dividido entre orofaringe e nasofaringe (sons nasais). Essas estruturas são intermediárias e têm a função de moldar novas frequências, servindo de órgãos de passagem para a cavidade oral e nasal, respectivamente.

Em qualquer das situações acima, se a língua ficar parada, deixando o fluxo do ar passar pela boca, o som produzido será uma das vogais orais (Á, É, Í, Ó, Ú, Â, Ê, Ô). Se acontecer a mesma coisa, mas parte do ar subir para a cavidade nasal, será produzido uma das vogais nasais (AN, EN, IN, ON, UN). Se houver uma ação da língua com intuito de colocar obstáculos à passagem livre do ar, será produzido uma das diversas consoantes.

Pode-se dizer, de uma forma geral, que vogais são sinais sonoros que passam livremente pelo trato vocal e nasal e que as consoantes são sinais sonoros que têm certa dificuldade de passar pelo trato vocal, em virtude de obstáculos criados pela língua, em conjunto com os lábios e dentes.

Daí, pode-se chegar à conclusão de que, em qualquer sinal sonoro produzido por um ser humano, a maior quantidade de energia se encontra nas vogais, sejam elas orais ou nasais. Trechos do sinal sonoro que possuem consoantes não mudas (tipo P e K por exemplo),

possuem os menores níveis de energia, quando comparadas com as partes do sinal em que há produção de vogais.

No caso das vogais, sejam elas nasais ou orais, o que as diferencia uma das outras é justamente o tamanho da abertura que permite a passagem ao ar para que elas sejam transmitidas. A vogal A por exemplo é a vogal onde o ar passa mais livremente pelo trato vocal. Já a vogal U é a vogal onde o ar passa pelo trato vocal de forma menos livre, em razão da articulação dos lábios, que diminuem a passagem do ar.

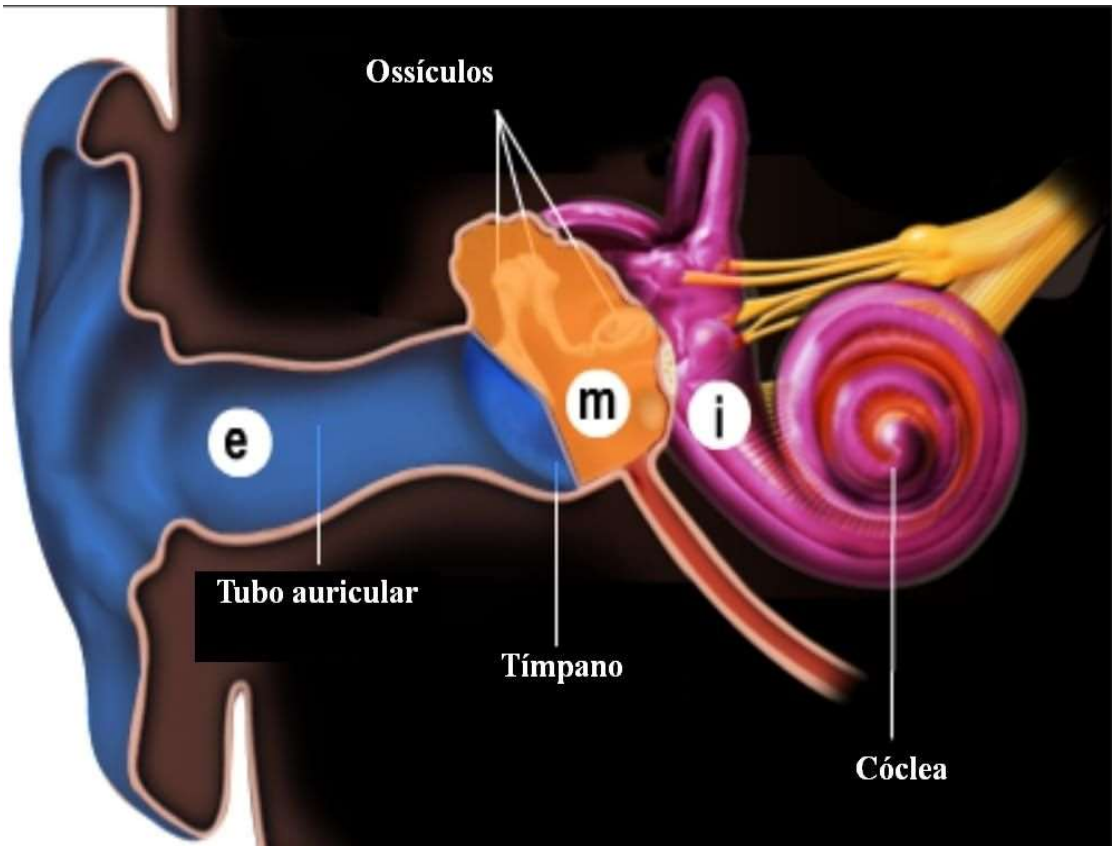
No caso das consoantes, o que as diferenciam umas das demais é a forma como os obstáculos criados pela ação da língua, dos lábios e dos dentes atuam, ou de forma isolada, ou em conjunto. Por exemplo: a consoante P é formada pelo obstáculo criado pela passagem do ar criada pelos lábios, que se fecham, impedindo a passagem do ar e depois soltando. Daí, pode-se saber por que razão uma pessoa que sofreu um grave ferimento nos lábios, em razão de um acidente de carro, tem sua voz consideravelmente modificada.

O OUVIDO HUMANO

O ouvido humano pode ser dividido em três partes: ouvido externo; ouvido médio e ouvido interno. O ouvido externo tem a função de captar as ondas sonoras, conduzindo-as pelo canal auditivo e as enviando ao tímpano, órgão sensitivo que tem a mesma função de um diafragma, ou seja, de transmitir pulsos mecânicos do ar através de vibrações proporcionais às descompressões e compressões do sinal sonoro. O ouvido médio se estende do tímpano até a tuba auditiva e é composto por 3 ossículos denominados de martelo, bigorna e estribo. Estes ossículos recebem as vibrações da membrana timpânica, mas amplificando-os em até 22 vezes, o que explica a capacidade do ser humano de conseguir escutar sons baixos. O ouvido interno é formado pelo labirinto (formado por 3 canais) e pela cóclea. Esta é formada por mais de 15000 células ciliadas (como fios de cabelo), que vibram de acordo com a movimentação do líquido coclear, que se movimenta conforme as vibrações dos ossículos. A função destas células é transformar estes movimentos em sinais elétricos proporcionais aos movimentos do líquido coclear e enviá-los ao cérebro por meio de canais nervosos para serem interpretados.

A Figura 18 mostra o aparelho auditivo humano. Nele, pode-se constatar a presença do ouvido externo (letra “e”), ouvido médio (letra “m”) e ouvido interno (letra “i”). É no ouvido interno que se encontra a cóclea, prinicipal órgão responsável pela sensibilidade acústica. Dentro da cóclea existe a membrana basilar, repleta das células ciliadas anteriormente discutidas.

Figura 18: Funcionamento do ouvido humano



Fonte: Adaptado de PUJOL (2016)

A cóclea funciona como um conjunto de filtros passa-faixas, que identifica amplitudes e frequências dos sons codificados no movimento do líquido coclear. Esses filtros passa-faixas da membrana basilar são conhecidos como bandas críticas.

Em termos computacionais e quando se intenta reconhecer voz, o estudo destas bandas críticas está relacionada com a sensibilidade auditiva do ser humano (psicoacústica) e se constitui em fator de suma importância, pois, muitas frequências que compõem um sinal de voz humana não precisam ser processadas, haja vista que nada informam sobre quem está falando ou o que se está falando. Só serviriam para encher o canal de informações inúteis, aumentando o processamento e as taxas de erros de reconhecimento. É extremamente necessário, em processamento de voz, enxugar o sinal com base nestas bandas críticas, com o intuito de facilitar o processamento digital necessário para extração de parâmetros relevantes que indiquem o que ou quem está falando. Daí a necessidade de se ter uma noção mais precisa sobre as referidas bandas críticas.

AS BANDAS CRÍTICAS

Bandas críticas são um conjunto de frequências que o ouvido humano é incapaz de distingui-las umas das outras. Isto significa que, em um sinal sonoro produzido por um ser humano, há várias faixas de frequências que são imprestáveis em processamento de sinais, tanto para compreensão dos áudios, como para reconhecimento de locutor e de locuções.

A largura de banda destas frequências em Hz é informada pela sigla ERB (em inglês, *equivalent retangular bandwith*, que, traduzindo para o português significa Largura de Banda Retangular Equivalente. Obs: Não confundir com o ERB (estação rádio base) da telefonia celular.

Em termos fisiológicos, cada banda crítica corresponde a 1,3 mm da cóclea. Em termos matemáticos (LIN, 2017), pode-se calculá-la, por meio da Eq.15.

$$ERB = 21{,}4 \times \left[\left(4{,}37 \times \frac{f_c}{1000} \right) + 1 \right], \qquad \text{Eq. 15}$$

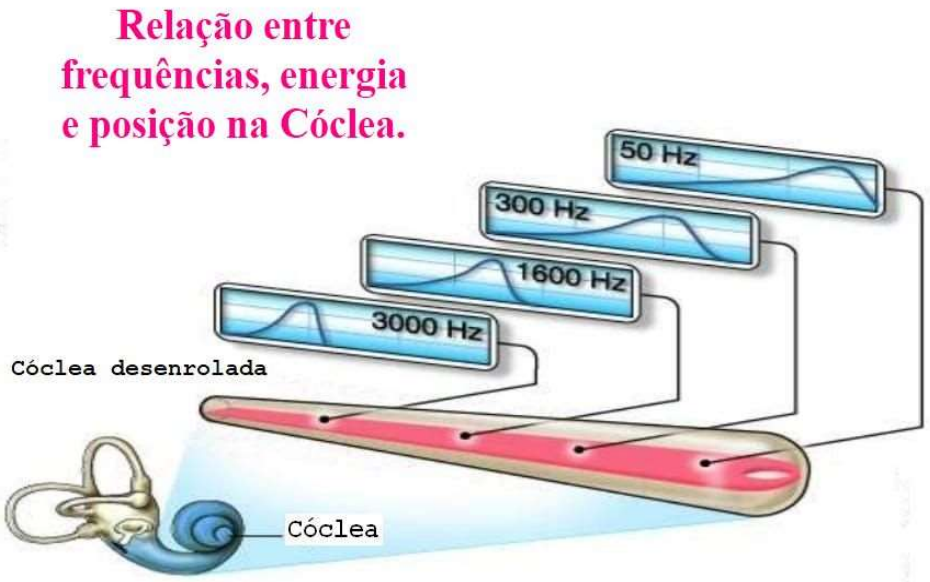
em que *fc* é frequência central e ERB é a banda crítica.

Por exemplo, se se quer saber em que banda crítica a frequência central de 50 Hz estaria, basta substituir a frequência central na Eq.15 e encontrar o valor da ERB. Fazendo f_c = 50 Hz, obtem-se algo próximo a 26 Hz. Isso significa que as rais laterais são de 13 Hz (a metade de 26 Hz). Portanto, basta diminuir a frequência central de 50 Hz do valor de 13 Hz e somar a frequência central de 50 Hz com 13 Hz, que se encontrará, respectivamente, os limites inferior e superior da faixa procurada, ou seja 37 Hz – 63 Hz.

O resultado quer dizer que qualquer sinal sonoro com frequências entre 37 Hz – 63 Hz será interpretado pelo ouvido humano como se fosse de 50 Hz. Para que o ouvido humano reconheça sons diferentes de 50 hz, a faixa de frequência do sinal deve se situar fora do intervalo de 37 Hz – 63 Hz.

A Figura 19 tenta representar o funcionamento da membrana basilar da cóclea desenrolada.

Figura 19: Funcionamento da Cóclea humana



Fonte: Adaptado de REBILLARD (2021)

Segundo a Figura 19, frequências centrais estão sendo associadas a determinadas posições da membrana ciliar, além do fato de estarem associadas a determinados níveis de energia como exigência para serem percebidas, de maneira que frequências mais altas requerem menos energia e frequências mais baixas requerem maior nível de energia para serem percebidas.

É importante enfatizar que não existe um número fixo de bandas críticas, pois ao se escolher aleatoriamente intervalos de frequências superpostas, sempre serão encontradas novas frequências centrais e consequentemente novos valores para ERB (bandas críticas). Estas bandas são instantaneamente selecionadas pelo ouvido humano, em cada instante de tempo que ele percebe sons compostos por várias frequências.

De forma geral, abaixo de 500 Hz, as bandas críticas são em torno de 100 Hz. Acima de 500 Hz, a banda crítica é 20% maior que a anterior.

APÊNDICE II
CAPTAÇÃO DOS ÁUDIOS E TRATAMENTO ACÚSTICO

A captação dos áudios, tanto na fase de treinamento, como na fase de prova, constitui-se em momento crucial para que o sistema de reconhecimento de locutor e de palavras funcione com sucesso. Na fase de treinamento, esta captação deve ocorrer antes da eleição, por ocasião do alistamento ou revisão eleitoral, momento em que o eleitor deverá fornecer ao Cartório Eleitoral as 20 amostras de treinamento de cada uma das 13 palavras usadas para invocar os comandos da UE. Na fase de prova, a captação ocorre no dia da eleição, quando o eleitor pronunciará as locuções para serem gravadas e processadas, possibilitando ao eleitor ser identificado por biometria de voz e, em seguida, votar usando RPI.

Uma série de recomendações e ações devem ser tomadas para que esses áudios sejam captados da melhor forma possível.

Captar as locuções de treinamento, transformando-as em sinais digitais para compor a base de dados da Justiça Eleitoral, é a primeira fase para se reconhecer biometricamente o eleitor e poder fazer com que ele possa votar usando voz, além de poder descobrir multiplicidade de inscrições (ou fraudes) na base de dados.

Isso implica necessariamente adotar uma taxa de amostragem, ou seja, é preciso definir quantas vezes por segundo o sinal originalmente analógico deve ser amostrado para ser transformado em um sinal digital.

Como os ATs comporão a base de dados da Justiça Eleitoral, recomenda-se que eles sejam amostrados a uma taxa alta, como a taxa de 44100 Hz, a fim de captar o maior nível de informação dos áudios, diferenciando locutores e palavras.

Essa alta taxa de amostragem se justifica também em razão de que a forma de se extrair os VCs podem, ao longo do tempo, serem modificadas, melhoradas ou substituídas, além do fato de que certos parâmetros podem também ser modificados no momento de se extrair os VCs. Exemplo de parâmetros que podem ser modificados a cada eleição: tamanho dos *frames (* nesta simulação, adotou-se *frames* de 20 ms, mas podem variar entre 20 ms e 40 ms) e tamanho da superposição dos *frames* (em nossa simulação, adotou-se *frames* superpostos em 50%, mas pode variar para menos).

Portanto, deve-se encarar os ATs como fontes de informações densas e brutas (por isso a taxa alta de 44100 Hz), guardadas no banco de dados da Justiça Eleitoral. Estes áudios podem ser reamostrados em 8000 Hz (ou taxa menor que 44100 Hz) visando à redução de consumo de memória na urna eletrônica, se esse for o caso.

Por outro lado, quando se fala em varrer os dados da base de dados da Justiça Eleitoral em busca de semelhanças entre VCs extraídos de ATs para encontrar multiplicidades de inscrições (possíveis fraudes), o ideal não é ter VCs extraídos de ATs com taxas de amostragem baixas, mas altas, pois lida-se no caso com um gigantesco banco de dados e quanto maior as informações sobre o trato vocal dos eleitores, melhor vai ser a acurácia do sistema.

Toda esta logística se justifica em razão de que a qualidade de um áudio digital advindo de um sinal analógico captado com taxa de amostragem de 44100 Hz e reamostrado em 8000 Hz é bem mais rico em informaçõe precisas do que um áudio captado originalmente em 8000 Hz.

Com relação aos APs no dia da eleição, não resta dúvida de que eles devem ser captados na mesma taxa de amostragem dos áudios de onde foram extraídos os VCs inseminados na urna eletrônica.

Portanto, os ATs captados e gravados em 44100 Hz de amostragem devem gerar dois tipos de VCs: um para ser inseminado na urna eletrônica (extraídos de áudios reamostrados a 8000 Hz); e outro extraído diretamente dos áudios amostrados a 44100 Hz, para atuarem no banco de dados da Justiça Eleitoral, detectando multiplicidades de inscrições.

Após capturar os ATs do eleitor, o servidor do Cartório Eleitoral deve fazer pelo menos 3 pequenas votações simuladas com o eleitor para everiguar se houve, por parte dele, uma compreensão total do funcionamento do sistema, bem como se ele foi capaz de reproduzir os áudios gravadas de forma que o sistema esteja reconhecendo seus comandos e escolhendo os números mediante código sonoro permutado de forma satisfatória.

Caso negativo, a melhor opção é orientar o eleitor que ele não deve votar por meio daquele sistema.

Indispensável também realizar as devidas correlações entre os VCs dos eleitores pertencentes à seção eleitoral do eleitor, checando se o eleitor estaria bem identificado em sua seção eleitoral, evitando problemas no dia da eleição.

Com relação ao tempo de gravação, na pior das hipóteses, estima-se um tempo de 20 minutos para gravação dos 240 áudios e mais 10 minutos para que o servidor da Justiça Eleitoral explique ao eleitor a lógica de funcionamento do sistema e faça uma pequena simulação de votação para testar se seria viável para ele usar o método biométrico da voz no dia da eleição.

Pessoas com problemas vocais, como, por exemplo, portadores de câncer na garganta ou que sofram de algum cisto nas cordas vocais, precisariam aderir a outro método usando voz

(sugere-se a adoção do método proposto na Seção 5.7), mas sem a gravação dos treinamentos e sem o reconhecimento de sua voz no dia da eleição.

É importante salientar que o servidor do cartório deve orientar o eleitor sobre determinados parâmetros e comportamentos que devem ser seguidos no momento de pronunciar as locuções. Por exemplo, em relação à distância da boca ao microfone, recomenda-se uma distância padronizada para todos os eleitores. Outra recomendação é não movimentar a cabeça no momento da fala para evitar diferenças grandes entre as amostras de treinamento no que concerne ao nível de energia ao longo do sinal, pois se a boca estiver muito longe do microfone, a energia tende a diminuir naquele trecho do sinal.

Também reforça-se a necessidade de enfatizar as vogais das locuções, onde se concentra as maiores informações e energias ao longo do sinal. Há pessoas que falam muito baixo e demasiadamente rápido, omitindo a pronúncia das sílabas. Isso deve ser evitado, tanto no momento de fornecer os ATs, quanto no momento de fornecer os APs, no dia da eleição.

Deve-se alertar ao eleitor de que ele deve ir à seção eleitoral no dia da eleição com a intenção de ser reconhecido, ou seja, ele deve se esforçar para tentar repetir o que ele pronunciou e gravou no Cartório Eleitoral, por ocasião de sua revisão biométrica ou alistamento. Se o eleitor for com a intenção de não ser reconhecido, modificando intencionalmente sua voz, o algoritmo gravado na urna não o reconhecerá.

A regra deve ser que o eleitor seja reconhecido biometricamente e possa votar usando sua voz sem erro atribuído ao mecanismo da urna. E em reconhecimento de voz, existe um gargalo que precisa ser combatido: o ruído.

O dia da eleição não costuma ser um dia barulhento a ponto de prejudicar um sistema de reconhecimento de voz. Até por que é proibido por lei fazer propaganda eleitoral no dia da eleição, constituindo-se em um crime de boca de urna para todo aquele que desobedeça a essa regra.

Entretanto, não se pode impedir também que exista um ar-condicionado ou um ventilador funcionando na seção eleitoral no dia do pleito. Inviável também impedir que mesários, fiscais de partido e eleitores da fila de espera para votar possam conversar de maneira obviamente comedida. Aparelhos eletrônicos devem ser desligados e entregues aos mesários, enquanto o eleitor vota. Isso é lei. Portanto, ruídos de celulares também não deverão ser um problema.

Outro detalhe é que não há a necessidade de transformar todas as seções eleitorais, adapatando-as à reconhecimento de voz, pois, como já foi visto, a biometria das impressões digitais funciona para a maioria dos eleitores. No máximo, bastaria que uma seção eleitoral,

em cada local de votação, possuísse uma seção eleitoral habilitada para reconhecimento de voz, com dois microfones e um amplificador atuando dentro da cabine de votação. Esta deveria ser um pouco aumentada em termos de comprimento e largura, para abranger tais dispositivos e também revestida internamente por espuma acústica ou material semelhante.

A parte de cima da cabine de votação e a parede que fica próxima ao terminal do eleitor devem também ser revestidas com material que evite reverberação e reflexão, com intuito de que apenas a voz do eleitor seja captada.

O tamanho e as especificações desses utensílios devem ser especificados e atestados por estudos técnicos aprovados por um laudo de um profissional especializado em acústica e captação de áudio.

Estes dispostivos de tratamento acústico devem ser instalados pelos mesários e por eles retirados de maneira fácil e rápida ao término da eleição, não devendo, de forma alguma, serem incorporados às paredes das salas, pois esses locais não são de propriedade da Justiça Eleitoral, mas requisitados.

A entrada e a saída do eleitor à cabine de votação adaptada ao reconhecimento de voz devem continuar totalmente abertas, inclusive para cadeirantes e acompanhantes, se autorizados.

O projeto acústico deve se preocupar em impedir que ruídos se misturem com a voz do eleitor, não em impedir que o som produzido pelo eleitor se irradie pela seção eleitoral, pois quanto mais alto o eleitor falar, ainda que por meio de amplificação sonora em cascata, melhor será o seu reconhecimento biométrico por voz.

Em caso de problemas de ruídos internos no equipamento de *hardware*, não havendo urna eletrônica para substituir, a opção ventilada na Seção 5.7 (Uma Solução de Continuidade) pode ser adotada se solicitada pelo eleitor. Isto requer que a voz do eleitor seja ouvida pelo Presidente de Mesa, através de um *headfone* (fone de ouvido), caso o eleitor possua uma voz muito baixa.

Os microfones, tanto o que capta a locução direta do eleitor, como o que capta o som amplificado, devem ser do tipo dinâmico (em inglês, *shotgun*), conhecido popularmente como microfone de palco. Eles têm a propriedade técnica de captar a voz em uma única direção: a direção da voz de quem está de frente para ele.

O microfone dinâmico poderia ser colocado em cima da urna eletrônica, com o respectivo pedestal e mecanismo de adequação à altura do eleitor. O ajuste da altura poderia ser feito pelo secretário da seção eleitoral ou por um dos dois mesários. A função deste primeiro microfone seria de repassar a voz do locutor para um amplificador. Este

amplificador, por sua vez, teria a função de repassar a voz amplificada para um segundo microfone, que se encarregaria de captar a voz amplificada e enviá-la à CPU da urna eletrônica. Este mecanismo pode ser uma forma indireta de redução do ruído, caso exista.

Para eleitores cadeirantes que prefiram usar a voz para votar, sugere-se então que, em cada local de votação, exista uma seção eleitoral especializada em captação e reconhecimento de voz, sendo disponibilizada mais um tipo de microfone conhecido popularmente como *Headset*. Isso ajudaria o cadeirante a não se esforçar demasiadamente para se aproximar do microfone, que provavelmente estaria hasteado por meio de um pedestal ou embutido na própria urna eletrônica.