UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE ENGENHARIA
CAMPUS DE ILHA SOLTEIRA

THYAGO JORGE MACHADO

RECONHECIMENTO DE VOZ DE LOCUTOR NO CONTEXTO FORENSE UTILIZANDO MÍNIMOS QUADRADOS ORDINÁRIOS

Ilha Solteira
2021

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

THYAGO JORGE MACHADO

RECONHECIMENTO DE VOZ DE LOCUTOR NO CONTEXTO FORENSE UTILIZANDO MÍNIMOS QUADRADOS ORDINÁRIOS

Tese apresentada à Faculdade de Engenharia - UNESP - Campus de Ilha Solteira, como parte dos requisitos para a obtenção do título de Doutor em Engenharia Elétrica. Área de Conhecimento: Automação.

Orientador: Jozué Viera Filho

Ilha Solteira
2021

FICHA CATALOGRÁFICA
Desenvolvido pelo Serviço Técnico de Biblioteca e Documentação

Machado, Thyago Jorge.
M149r Reconhecimento de voz de locutor no contexto forense utilizando Métodos dos Quadrados Ordinários / Thyago Jorge Machado. -- Ilha Solteira: [s.n.], 2021
69 f. : il.

Tese (doutorado) - Universidade Estadual Paulista. Faculdade de Engenharia de Ilha Solteira. Área de conhecimento: Automação, 2021

Orientador: Jozué Viera Filho
Inclui bibliografia

1. Comparação de locutor forense. 2. Fonética forense. 3. Processamento de voz. 4. Métodos dos Quadrados Ordinários (OLS). 5. Codificação Linear Preditiva (LPC).

Raiane da Silva Santos
Supervisora Técnica de Seção
Seção Técnica de Referência, Atendimento ao usuário e Documentação
Diretoria Técnica de Biblioteca e Documentação
CRB8 - 9999

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Ilha Solteira

ATESTADO DE APROVAÇÃO - DEFESA

Atestamos que **THYAGO JORGE MACHADO**, RA nº: ELE190284, RG nº 1745834-0, expedido pela SJSP/MT, defendeu, no dia 02/08/2021, a tese intitulada **Reconhecimento de voz de locutor no contexto forense utilizando Métodos dos Quadrados Ordinários**, junto ao Programa de Pós Graduação em Engenharia Elétrica, Curso de Doutorado, tendo sido 'APROVADO'.

Atestamos ainda que a obtenção do título dependerá de homologação pelo Órgão Colegiado competente.

Ilha Solteira, 02 de agosto de 2021

Márcia Regina Nagamachi Chaves
Supervisora Técnica de Seção – STPG

RESUMO

Neste trabalho tem-se como proposta o uso de uma combinação de codificação preditiva linear (LPC) e mínimos quadrados ordinários (OLS do inglês: Ordinary Least Square) como uma ferramenta de verificação de locutor para análise forense. Além disso, o método desenvolvido extrai um maior número de formantes, indispensáveis para comparações estatísticas via OLS. A estrutura proposta é robusta em certos níveis de ruído, para frases com a supressão de mudanças de palavras, e com qualidade diferente, ou mesmo diferenças significativas de tempo de áudio. A análise comparativa é avaliada pelo tempo decorrido para a obtenção dos resultados, bem como pela qualidade dos resultados. Os resultados mostram que a análise por meio de OLS gera resultados imediatos e efetivos quando comparados aos obtidos com metodologias tradicionais nos processos estudados. Depois de executar sete testes diferentes, este estudo obteve preliminarmente uma taxa de acerto de 100% considerando um conjunto de dados limitado (português do Brasil). A técnica de reconhecimento proposta estabelece confiança e similaridade sobre as quais baseiam-se os relatórios forenses, indicando verificação do locutor do discurso contestado.

Palavras-chave: Comparação de Locutor Forense; Fonética Forense; Processamento de Voz; Métodos dos Quadrados Ordinários (OLS); Codificação Linear Preditiva (LPC).

ABSTRACT

This paper proposes a combination of *Linear Predictive Coding* (LPC) and ordinary least squares (OLS) as a speaker verification tool for forensic analysis. The proposed recognition technique establishes confidence and similarity to base forensic reports, indicating and verifying the speaker of the contested discourse. Furthermore, the developed method extracts a larger number of formants, which are indispensable for statistical comparisons via OLS. The proposed framework is robust at certain levels of noise, for sentences with the suppression of word changes, and with different quality or even meaningful audio time differences. The comparative analysis is assessed for time elapsed for obtaining results, as well as results quality. Results show that the analysis, by using OLS, generates immediate, effective results when compared to those obtained with traditional methodologies on the studied process. After running seven different tests, this study preliminarily achieved a hit rate of 100% considering a limited dataset (Brazilian Portuguese).

Keywords: Forensic Speaker Comparison; Forensic Phonetics; Speech Processing; Ordinary Least Squares (OLS); *Linear Predictive Coding* (LPC).

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

ANN	Artificial Neural Networks
CNN	Redes Neurais Convolucionais
CPC	Código Penal Brasileiro
DFT	Transformada Discreta de Fourier
DNN	Deep Neural Networks
FFT	Transformada Rápida de Fourier
GMM	Gaussian Mixture Model
GUI	Interface de Usuário Convidado
HMM	Hidden Markov Model
IDVC	Inter Dataset Variability Compensation
JFA	Joint Factor Analysis
LPC	Codificação Linear Preditiva
MAPLR	Maximum a Posteriori Linear Regression
MAV	Média Móvel
MCGDZPHEC	Multi-taper Chirp Group Delay Zeros-Phase Hilbert Envelope Coefficients
MFCC	Mel-Frequency Cepstral Coefficient
MGHEC	Coeficientes do Envelope de Hilbert
MHEC	Coeficientes Médios do Envelope de Hilbert
MLLR	Maximum Likelihood Linear Regression
NIST	National Institute of Standards and Technology
NN	Neural Network
OLS	Mínimos Quadráticos Ordinários
PCA	Principal Component Analysis
PLDA	Probabilistic Linear Discriminant Analysis
PNN	Rede Neural Probabilística
SRE	Speaker Recognition Evaluation
SVM	Support Vector Machine
UBM	Universal Background Model
WCC	Within-Class Covariance Correction

SUMÁRIO

1 INTRODUÇÃO

O reconhecimento de voz é um processo de identificação do interlocutor e/ou discurso realizado. Isto é derivado de informações específicas extraídas da fala e pode ser conduzido manualmente ou automaticamente. O reconhecimento manual é subjetivo, intuitivo e sujeito a ouvir e reconhecer padrões humanos. A automatização desse processo visa trazer objetividade e rapidez ao reconhecimento. Usando formas de onda de fala, é possível projetar uma ferramenta capaz de reconhecer o que foi dito no discurso, ou para identificar o interlocutor. As aplicações são diversas e principalmente ligadas ao controle de acesso a serviços por voz, incluindo: ativação telefônica por voz; serviços bancários por rede telefônica; compras por telefone; serviços de acesso a banco de dados; informação e reserva de serviços; correio de voz; controle de segurança da informação confidencial; e acesso remoto a computadores (JUANG; FURUI, 2000).

Segundo Pegorado (2000), o reconhecimento de voz pode ser aplicado de duas maneiras: reconhecimento do discurso (identificando a fala ou fonética do sinal) e reconhecimento de locutor (para identificar o locutor). Apesar de ambos apresentarem objetivos diferentes, o discurso e o reconhecimento do locutor lidam com os mesmos dilemas. Sotaque, ruído e fonética são alguns dos fatores que tornam as ferramentas de reconhecimento de padrões aplicadas a vozes uma tarefa difícil (ABDEL-HAMID *et al.*, 2014).

Dentro da área de reconhecimento de locutor, é comum dividir o assunto em identificação e verificação (ALI *et al.*, 2014). A verificação do locutor ocorre quando o sistema deve comparar um discurso confrontado com um discurso base (referência) de tal forma que verifica se ambos foram produzidos pelo mesmo locutor. Quanto à identificação, esta compara o discurso confrontado com um banco de dados e/ou vozes de uma série de indivíduos. A maioria dos aplicativos em que a voz é usada para confirmar a identidade é classificada como verificação de locutor (FURUI, 2010). Assim, o sistema de reconhecimento deve identificar quem é o mais compatível com a voz confrontada dos indivíduos pré-selecionados (BRAID, 2003).

Deve ser observado que a identidade do falante correlaciona-se com as características fisiológicas e comportamentais do indivíduo no sistema de produção da fala. Isso pode ligar tais características aos coeficientes cepstrais e regressão da voz (CHOU, 2002).

Dessa forma, relaciona-se um aplicativo para reconhecimento de falante à fonética forense. A comparação de voz forense é baseada na comparação de uma gravação da voz de um criminoso desconhecido (a peça questionada ou traço) e uma gravação da voz de um

suspeito conhecido (a peça de comparação) (AJILI *et al.*. 2017). Para a perícia, a tomada de decisão de verificação do locutor deve apresentar um alto nível de confiança. A partir de uma pesquisa de revisão na literatura, vários trabalhos relacionados foram encontrados. Os métodos encontrados geralmente tiram proveito de formantes como uma ferramenta de verificação de locutor poderosa. Por exemplo, em Koval (2006), o autor usou trechos de áudio de 3 a 5 segundos, extraindo 5 formantes e comparando-os com os fragmentos espectrais para identificar diferenças e coincidências. Quando comparado ao mesmo trecho, foi obtido 1% de falso positivo, que aumentaram para 2% no momento em que foram comparados a diferentes trechos. Neste método não foram usadas quaisquer técnicas estatísticas no modelo, apenas técnicas visuais foram usadas com baixa taxa de erro.

Rodman *et al.* (2002) usaram formantes para identificar o locutor, aplicando uma técnica baseada em momentos espectrais atingindo uma taxa de erro de 10% a 20% na verificação do falante. Reconhecimento de locutor com uso de formantes também foi empregado em Becker *et al.* (2008). extraindo semiautomaticamente apenas F1, F2 e F3 como formantes. Os autores usaram um modelo de mistura gaussiana (GMM), que, na melhor das hipóteses, alcançou uma taxa de erro de 3% ao usar 3 formantes e suas larguras de banda correspondentes.

O trabalho de Leuzzi *et al.* (2016) também apresentou uma abordagem estatística para verificação de locutor no campo da fonética forense. Eles extraíram os formantes (F0, F1, F2, F3 e F4) usando o *software* Praat e, posteriormente, aplicaram a distância de Mahalanobis juntamente com a distribuição estatística para identificar os formantes que melhor correspondiam a assinatura de voz. Como resultado, eles alcançaram a menor taxa de erro, perto de 4%. Da mesma forma, outra abordagem usou o *software* Praat e foi baseada na taxa de probabilidade e, coincidentemente, encontraram um erro na taxa de 4% (GOLD; HUGUES, 2015).

Um sistema de verificação semiautomático de locutor baseado em uma comparação de valores de formantes, estatísticas de intervalos de tempo de chamadas telefônicas e características melódicas foi proposto em Bulgakova e Sholokhov (2016). A precisão desse sistema foi de 98,59% em um banco de dados contendo gravações de homens e 96,17% em um banco de dados contendo gravações de mulheres. Eles usaram um banco de dados com gravações com duração de 3 a 5 minutos.

As redes neurais (NN) têm desempenhado um papel importante no reconhecimento de falantes. Recentemente, algumas técnicas para verificação de locutor têm usado redes neurais profundas (DNN). Em Irum *et al.* (2019), uma revisão contendo nove técnicas diferentes

empregando DNN foi apresentada. Os resultados atingiram 0,2% e 0,88% para comparações baseadas em texto dependente e independente, respectivamente. Embora sejam bons, os resultados, não levaram em consideração o tempo de processamento, pois o método necessita de horas de aprendizagem, a depender do poder computacional. É necessário que o locutor registre a maior quantidade possível de diferentes vocabulários em uma fileira, com o objetivo de permitir que as NNs atinjam uma taxa de sucesso aumentada.

No trabalho em Devi (2019) foi apresentado um método baseado em uma combinação de NNs e lógica *fuzzy* para alcançar o reconhecimento de locutor, embora resultados positivos tenham sido obtidos (até 78%), uma desvantagem é o treinamento teve alto custo computacional. Em Chung *et al.* (2018), uma comparação aplicada de reconhecimento de voz em grande escala foi conduzida. O banco de dados continha mais de 1 milhão de declarações de mais de 6.000 palestrantes, e eles encontraram uma taxa de erro de 3,95% usando DNN (DHAKAL *et al.* 2019). Os resultados apresentados são excelentes se considerados como uma técnica de filtragem para vários locutores, no entanto, não poderia ser considerado para o campo forense, onde a precisão é fundamental.

Da mesma forma, Gujarati e Porter (2008) apresentaram uma abordagem de reconhecimento de locutor, usando uma combinação de filtro Gabor, estatísticas e redes neurais convolucionais (CNNs do inglês: Convolutional Neural Networks). Eles alcançaram uma taxa de erro de apenas 0,58% a uma velocidade de processamento de 0,3456 segundos. Apesar desses bons resultados, nenhum dos métodos apresentados aqui obteve taxa de acerto de 100%.

Pode-se observar nos trabalhos revisados que somente os testes com textos independentes e dependentes foram executados. Eles não usaram cenários como temporização de fala desigual, diferentes qualidades de som e inserção de ruído, que podem afetar a precisão dos métodos.

De acordo com Ajili *et al.* (2017), a questão da confiabilidade continua sendo um grande desafio, especialmente para sistemas de comparação de voz forense onde vários fatores de variação, como duração, ruído, conteúdo linguístico ou variabilidade dentro do falante não são levados em consideração. Além disso, os sistemas de reconhecimento de locutor apresentam uma diminuição no desempenho quando o usuário está sob condições emocionais ou de estresse (RITUERTO-GONZALEZ *et al.*, 2019).

Em Smith *et al.*. (2019), os autores indicaram que a precisão está sujeita a perturbações por ruído de fundo e diferenças no estilo de fala, que podem desempenhar um papel em casos envolvendo discriminação de voz. Nesse sentido, Krobba *et al.* (2019)

propuseram a identificar o locutor considerando a influência do ruído. Os autores usaram um método misto com base nos coeficientes gammatone multi-taper do envelope de Hilbert (MGHECs) e atraso do grupo de coeficientes *chirp multi-taper* do envelope de Hilbert de fase zero coeficientes (MCGDZPHECs). Eles alegaram uma melhoria significativa no desempenho em condições ruidosas, quando comparadas aos coeficientes médios do envelope de Hilbert convencionais (MHECs). De acordo com os resultados obtidos em (DHAKAL *et al.*. 2019), os modelos híbridos na comparação de locutor aumentam as chances de sucesso. Portanto, os autores aproveitaram a arquitetura de reconhecimento de locutor em tempo quase real que aprimorou o desempenho do reconhecimento do falante. Isso é conseguido explorando as vantagens de técnicas de extração de recursos híbridos contendo os recursos de filtros Gabor, CNNs e parâmetros estatísticas como um único conjunto de matriz.

Esta pesquisa tem como proposta o método aqui desenvolvido realizando janela de Hanning, transformada de Fourier, codificação preditiva linear (LPC) e mínimos quadrados ordinários (OLS). Portanto, este trabalho baseia-se no uso de OLS como um método de reconhecimento de locutor para análise forense, e a verificação do orador a partir das amostras coletadas no confronto, sabendo que não há banco de dados de voz oficiais, na Polícia Federal e Institutos Criminais dos Estados Brasileiros.

Por consequinte, a técnica proposta, estabelece um método diferente de reconhecimento de falantes, considerando os encontrados na literatura que podem ser aplicados ao locutor em português do Brasil. O reconhecimento de locutor no Brasil ainda apresenta a tomada de decisão baseada na análise subjetiva dos resultados usando técnicas não confiáveis (ESCH; VARY, 2009).

Em primeiro lugar, este estudo submete os discursos a formantes e extração de pitch usando codificação preditiva linear (LPC). Após a extração dessas características de fala, a técnica analisa os áudios contestados, e realiza um confronto com padrões pré-selecionados, utilizando OLS.

Como ferramenta de verificação de locutor, OLS encontrará um perfil para o áudio contestado e, posteriormente, estabelecerá similaridade entre o perfil e os padrões de fala de outros indivíduos. Dessa forma, usando a similaridade de grau dos parâmetros extraídos das falas, poderá determinar qual tem uma chance mais provável de ser reconhecido como o locutor do discurso confrontado.

A legislação brasileira possui um dispositivo que pleiteia a prescrição antes da sentença definitiva e esta é calculada com base na pena máxima (ANDREUCCI, 2018). Diz o artigo 109 do Código Penal Brasileiro (CPC) que “a prescrição, antes da sentença definitiva

[...] rege-se pelo máximo da pena privada comunicada ao crime" (BRASIL, 1940). O mesmo artigo 109 estabelece o tempo necessário a decorrer até a sentença final, para que ocorra a prescrição. Além disso, especifica que os réus menores de 21 anos, quando o crime foi cometido, ou maiores de 70, no dia da condenação, devem ter o prazo de prescrição reduzido pela metade, o que é definido, pelo CPC, como "prescrição de idade " (ENZINGER *et al.*, 2017). Isso demanda um esforço ainda maior da equipe forense, a fim de obter resultados mais rápidos.

A partir do momento em que ocorre a condenação em instância máxima da justiça brasileira (Supremo Tribunal Federal), interrompe-se a contagem regressiva da prescrição com base na pena máxima possível e a contagem regressiva com base na pena de fato aplicada ao condenado. Está tipificado no artigo 110 do CPC: "A prescrição após a sentença transitada em julgado é regulada pela pena aplicada" (BRASIL, 1940).

Como exemplo, considere um determinado processo que tenha duração aproximada de 3 a 5 anos, quando o indivíduo cometeu crime passível de pena máxima de 2 anos. Nesse caso, a prescrição ocorrerá em 4 anos. Porém, se a condenação for a pena mínima de 6 meses, a prescrição se reduz para 3 anos e o culpado ficará impune porque ocorreu a prescrição. Portanto, o tempo de conclusão do laudo pericial deve ser o menor possível, para permitir que as condenações não sejam prescritas e fazer com que os condenados cumpram suas penas de acordo com o que a justiça estabelece.

O Estado de Mato Grosso, localizado na região Centro-Oeste do Brasil, possui uma demanda reprimida de análises forenses no setor de áudio. Atualmente, o tempo médio para iniciar uma análise forense de um determinado arquivo de áudio é de cerca de um ano. Após o início dos trabalhos, são 3 a 6 meses para análise técnica e emissão de relatório final. Naturalmente, todo esse tempo se deve às técnicas utilizadas pelos especialistas, o que contribui negativamente para a eficiência da justiça brasileira. Assim, uma metodologia mais rápida e eficaz pode reduzir significativamente os tempos citados e diminuir as chances de prescrição de crimes, diminuindo a impunidade.

Novas metodologias foram propostas e aplicadas a diferentes grupos de pessoas. Um bom exemplo é o trabalho apresentado por Enzinger *et al.* (2017), que avalia o desempenho de um sistema acústico-fonético por meio de testes empíricos. Os resultados mostram que o desempenho obtido no auditivo-acústico-fonético-espectrográfico é inferior ao obtido nos sistemas automáticos, ou seja, a análise objetiva por meio de testes empíricos tem maior credibilidade do que a análise subjetiva por meio de gráficos.

Cinco casos de uso, já processados e condenados pelo Tribunal de Justiça do Estado de Mato Grosso no Brasil, são apresentados e analisados nesta proposta, sendo importante destacar que os relatos de casos utilizados neste trabalho são de domínio público, pois carecem de requerimento de sigilo processual.

Todos os casos envolvem verificação de voz e identificação dos envolvidos, para atestar a veracidade da gravação anexada às ações judiciais.

1.1 Objetivos

O objetivo é analisar o desempenho de uma técnica automatizada proposta para verificação forense de locutor baseada no algoritmo OLS, tanto para eficiência assertiva quanto para tempo de resposta.

Depois de realizados testes, também são feitas tentativas de analisar a robustez da técnica em relação ao texto discurso.

1.2 Constribuições

- Desenvolvimento de um novo método para verificação de locutor, que aproveita a combinação dos formantes, LPC e OLS para gerar resultados para a tomada de decisão em um contexto forense;
- Robustez do modelo desenvolvido é demonstrada pela geração de resultados positivos, mesmo com situações atípicas, como ruído, tempo de fala irregular, qualidade e independência textual;
- Para casos reais estudados e julgados pela Justiça Brasileira, o filtro de verificação de locutor possibilita que não ocorra prescrição do processo judicial.

1.3 Organização do texto

Além desta Introdução abrangendo os objetivos, justificativas, contribuições, a organização do restante do texto é a seguinte: apresenta-se no capítulo 2, o método desenvolvido é apresentado, destacando a combinação LPC e OLS; no terceiro capítulo, os resultados são apresentados, seguidos por uma comparação com outras abordagens, e, por

fim, no capítulo 4 estão as considerações finais, destacando comentários sobre a abordagem desenvolvida e trabalhos futuros.

2 RECONHECIMENTO DE LOCUTORES

O reconhecimento automático de voz é utilizado em diversas aplicações comerciais. O acesso rápido e/ou seguro a serviços ou lugares se dá por intermédio de comandos de voz. O reconhecimento do locutor ou daquilo que foi dito é papel de um sistema automático de reconhecimento de padrões, capaz de autorizar ou não um acesso do usuário, por exemplo.

Quando o reconhecimento de voz objetiva entender a fala proferida, então este é classificado como reconhecimento de discurso. Aplicações de comando de voz fazem uso dessa ferramenta para identificar um texto pré-definido, independente do locutor ou do tom de voz utilizado. Entretanto, quando o reconhecimento é aplicado para identificar o falante, a ferramenta é classificada como reconhecimento de locutor.

O reconhecimento automático de locutores em geral é mais utilizado em aplicações de segurança de acesso. Por meio de identificação precisa da identidade de um locutor, o sistema é capaz de efetuar uma autenticação e verificar quais serviços aquele usuário está autorizado a acessar.

Outro aspecto importante para o desenvolvimento de uma aplicação de reconhecimento de voz, seja ela aplicada ao discurso ou ao falante, é a forma de entrada de dados no sistema. O áudio utilizado pode apresentar as características:

- Reconhecimento dependente do texto – Quando o locutor em questão necessita dizer um texto pré-definido;
- Reconhecimento independente do texto – Quando o sistema precisa reconhecer o locutor, independente do contexto do trecho questionado;
- Tamanho do trecho – A ferramenta de reconhecimento pode receber como entrada falas grandes ou pequenas. Sua margem de acerto, tempo de processamento em geral são influenciadas de acordo com o trecho amostrado;
- Entrada sem ruído – Captura do discurso amostral realizada em ambiente controlado. Apresenta qualidade ideal que facilita o processamento e reconhecimento do padrão desejado;
- Entrada com ruído – Captura não ótima de áudio. Realizada por telefone, com som de fundo, vídeoconferência, entre outros exemplos. A ferramenta que

deseja ser robusta contra certo nível de ruído, deve ser capaz de isolar o padrão a ser reconhecido.

A combinação do objetivo de reconhecimento com as características do áudio questionado no sistema sugere possíveis aplicações onde o sistema de reconhecimento de voz pode ser aplicado. Na Seção 2.1 faz-se um levantamento do funcionamento genérico de um sistema de reconhecimento de locutores e quais as técnicas e sistemas têm sido utilizadas para esse tipo de aplicação. A área pericial apresenta um perfil específico de requisitos para o sistema de reconhecimento de voz. Na Seção 2.2 efetua-se uma análise dos sinais e dos pré-requisitos da ferramenta de reconhecimento proposta, para que esta possa ser decisiva no caráter pericial.

2.1 Funcionamento e técnicas de reconhecimento de locutores

Existem diversas propostas na literatura de sistemas e esquemas de reconhecimento de locutores. Antes de identificar tais propostas e o estado da arte, é preciso definir as etapas internas do processo de reconhecimento e eventuais técnicas que podem ser utilizadas para cada etapa. Nos artigos e trabalhos efetuam-se uma combinação do uso de cada técnica em busca de atender cada sub-aplicação do reconhecimento de locutores.

O reconhecimento de locutores pode possuir diversas etapas (AHMAD *et al.*, 2015). Entretanto, pode-se citar como etapas principais:

- Extração de Parâmetros da voz;
- Modelagem da voz;
- Análise e tomada de decisão.

A etapa de extração dos parâmetros da voz visa encontrar na fala características que definam e diferenciem a identidade de determinado falante em relação aos demais. Dentre as técnicas capazes de efetuar essa extração pode-se citar:

- Hidden Markov Model (HMM) – Técnica utilizada para mapear os sinais da fala e caracterizar os trejeitos de cada interlocutor. Para identificação de locução dependente de texto, o HMM calcula a probabilidade de que um áudio questionado seja igual a algum padrão pré-estabelecido, mesmo que o texto apresentado seja de apenas uma única palavra de contextos diferentes. Devido à simplicidade, esta técnica tem maior utilização registrada em aplicações de comandos de voz e reconhecimento de discurso (MANTRI *et al.*, 2014);

- Mel-Frequency Cepstral Coefficient (MFCC) – Para extrair um vetor de características contendo todas as informações sobre a mensagem linguística, o MFCCs imita algumas partes da produção de fala humana e da percepção da fala, além da percepção logarítmica da intensidade, do tom do sistema auditivo humano e tenta eliminar as características dependentes dos falantes, excluindo a frequência fundamental e seus harmônicos (O'SHAUGHNESSSY, 2008). Segundo Niemann (2003), para representar a natureza dinâmica da fala, os MFCCs também incluem a mudança do vetor de características ao longo do tempo como parte do vetor de características. Os MFCCs foram desenvolvidos por intermédio de estudos da percepção auditiva humana e apontou que os sinais de vozes e frequências de tons puros não seguem uma escala linear, criando uma escala chamada *mel*;
- *Linear Predictive Coding* (LPC) – Para efetuar esta técnica primeiro é realizado um pré-processamento do discurso de entrada por meio do seu prolongament, normalização, e remoção de períodos de silêncio. Ou seja, o áudio se torna mais lento e dividido em trechos de 30 ms com sobreposição de 20 ms entre os trechos, para retirar a descontinuidade do discurso. Cada trecho então é visto pela codificação LPC como parte de uma combinação linear. Cada termo ou coeficiente da combinação linear pode ser analisado para que informações como formantes ou pitch sejam extraídas (SRIVASTAVA *et al.*, 2014).
- *Fourier Based Techniques* – Tratando-se de coeficientes, as transformadas listadas aqui também possuem características de extração de parâmetros de voz (ou de qualquer forma de onda) para caracterização do sinal. O processo das transformadas ocorre de forma similar ao LPC (BEIGI, 2011). A extração dos coeficientes cepstrais, através da aplicação da Transformada de Fourier, converte o sinal de voz do domínio do tempo para domínio da frequência, no sentido de possibilitar a avaliação das características acústicas (VALIATI, 2000).
- *Wavelet Based Techniques* – De acordo com (SCHULLER, 2011), wavelets dão uma análise multi-Resolução de curto prazo de tempo, energia e frequências em um sinal de fala. Sua metodologia é passar da Transformada de Fourier para a Transformada Wavelet, após um banco de filtros que será suficiente para reconhecimento do falante. No que diz respeito à classificação, uma base Wavelet representa de uma forma única uma determinada classe de sinais que na presença de outras classes conhecidas será mais desejável.

A etapa de modelagem é apenas aplicada para sistemas de reconhecimento independente de texto, os quais se caracterizam por não dependerem de os conteúdos dos discursos padrão e questionado serem iguais. Ao invés disso, as técnicas de modelagem da voz efetuam alinhamentos sonoros de palavras ou até sílabas que são correspondentes entre os dois áudios confrontados, de forma a permitir uma posterior comparação e reconhecimento de padrão. A modelagem dos parâmetros da voz também pode auxiliar na redução de informações dos dados a serem analisados, bem como retirada de eventuais ruídos de fundo da informação. Segundo (TAFNER, 1996), o sinal digital precisa passar por um processo de refinamento que o torna apropriado para ser trabalhado.

Quanto as técnicas de modelagem, é citada a seguir:

- *I-Vector Model* – Os parâmetros de discurso previamente extraídos apresentam grande número de amostras e variáveis. Este modelo efetua a redução desses parâmetros para i-vector de matrizes T. Esse procedimento faz uso de caracterização de cada parâmetro como média gaussiana (LEI *et al.*, 2014). As entradas para o modelo i-vector podem ser tanto parâmetros de voz já extraídos quanto o próprio discurso diretamente.

Por fim, a etapa de análise estatística e de reconhecimento tem por objetivo efetuar a tomada de decisão do sistema de reconhecimento de locutor. Apesar de as aplicações de identificação e verificação de locutor obterem tomadas de decisão diferentes, as técnicas de análise costumam ser utilizadas em ambas. O conceito principal das técnicas está em comparar um ou mais padrões com o áudio questionado no reconhecimento. Algumas técnicas de análise e tomada de decisão estão listadas abaixo:

- *Gaussian Mixture Model* (GMM) – *Universal Background Model* (UBM) – O GMM-UBM efetua uma estimativa de distribuição de probabilidade associada para cada estado de HMM. O UBM é uma característica da fala utilizada em treinamento de sistemas de reconhecimento. Em geral, o áudio questionado analisado estatisticamente pelo modelo GMM é comparado às características do modelo UBM como forma de efetuar a tomada de decisão (ZHENG *et al.*, 2004).
- *Probabilistic Linear Discriminant Analysis* (PLDA) – A técnica PLDA é em geral utilizada em conjunto com o modelo i-vector para executar análise estatística e tomada de decisão de probabilidade no reconhecimento de locutores. A distribuição de probabilidade utilizada pela PLDA pode variar. Entretanto, considerando o baixo custo computacional em relação à outras distribuições, a PLDA baseada na probabilidade gaussiana tem sido mais utilizada para verificação de locutores (HANSEN and HASAN, 2015).

- *Support Vector Machine* (SVM) – As máquinas SVM pertencem a classes de modelos de aprendizagem de máquina de forma supervisionada. Utiliza análise e classificação dos padrões por meio de um sistema linear binário não-probabilístico.
- *Artificial Neural Networks* (ANN) – A utilização de redes neurais apenas começou a ser explorada recentemente na comparação de locutor. As *Deep Neural Networks* (DNN) têm sido objeto de estudo para o reconhecimento de voz como um todo.

Posterior ao levantamento das principais técnicas e do funcionamento dos sistemas de reconhecimento de locutores, pode-se listar as caracaterísticas das propostas de sistemas existentes na literatura. Os sistemas propostos são testados ou por uma base de teste própria e particular coletada pelos próprios autores, ou ainda por um sistema de teste padronizado do *National Institute of Standards and Technology* (NIST). O NIST *Speaker Recognition Evaluation* (SRE) tem o objetivo de contribuir com as pesquisas realizadas na área de reconhecimento de locutores. Uma base de dados inicial é fornecida, e os pesquisadores são desafiados a testarem seus sistemas e compararem resultados (NIST, 2016). Os desafios e as bases são renovadas bianualmente. Uma melhor descrição e uso da base de dados NIST SRE é prevista para trabalhos futuros no fechamento desta tese.

O trabalho realizado por Zheng *et al.* (2006) descrevem um sistema de identificação de locutores implementando um sistema GMM-UBM, adicionando uma adaptação de Bayes. Tal implementação utilizou sistema de teste do NIST SRE 2000. As taxas de erro caíram 31,2%.

Em Zhang *et al.* (2010) os autores descreveram um sistema de machine learning SVM e comparou sua utilização com duas técnicas: *Maximum a Posteriori Linear Regression* (MAPLR) e *Maximum Likelihood Linear Regression* (MLLR). O sistema de teste foi o NIST SRE 2008 e mostram que MAPLR tem melhor performance.

Larcher *et al.* (2013) propuseram uma ferramenta com várias técnicas de reconhecimento de locutores presentes na literatura. O ALIZE project é um *sistema open-source* que implementa técnicas como *Joint Factor Analysis* (JFA), SVM, modelo i-vector e PLDA, por exemplo.

A combinação i-vector e PLDA é investigada por diversos autores. Em Aronowitz (2014), verificou que essa combinação está sujeita a queda de performance com relação a taxa de erros quando a entrada de discurso sofre degração por ruído. A adição do sistema Inter Dataset Variability Compensation (IDVC) para efetuar possíveis compensações na entrada melhora a performance da técnica que reduz os erros em 50%. Em Garcia-Romero e Mccree, (2014) os autores resolveram o problema da queda de *performance* por meio do modelo UBM na entrada de dados e a normalização da duração do discurso entre o modelo i-vector e a

técnica PLDA. Entretanto, a performance do sistema não apresentou melhoras significativas quando aplicados ao SRE 2012. Além disso, a normalização apresentou maiores efeitos na taxa de erro do que a combinação de UBM e os i-vectors. Os autores Glembek (2014) propuseram o pré-processamento fosse realizado pela técnica Within-Class Covariance Correction (WCC) anteriormente ao modelo i-vector e assim reduções de erro chegaram a 70%.

As DNNs também têm sido foco de estudos para o reconhecimento de locutores. O trabalho de Kenny *et al.* (2014) descreveu uma proposta de uso da extração de estatística Baum-Welch, modelagem i-vector e DNN para o reconhecimento de locutores independente de texto. Testes realizados pelo NIST SRE 2012 mostram que o sistema proposto apresenta 16% de redução de erros em relação a sistemas i-vector+PLDA puros. Tal melhoria também foi percebida por Lei *et al.*. (2014) que por meio de uma sistema MFCC+i-vector+DNN atingiu redução de 30% dos erros com aplicação de testes em NIST SRE 2012. Variani *et al.*. (2014) proposeram o uso puro de DNN para investigação do reconhecimento de locutores dependente do texto. Quando comparada às abordagens i-vector+PLDA puras, a proposta apresentava reduções de até 25% na taxa de erro. Um conceito semelhante foi utilizado em Zhao *et al.* (2014). Entretanto, a abordagem sugere o uso de uma DNN apenas para retirar ruídos de fundo da entrada de dados. Posteriormente, um sistema GMM-UBM foi utilizado para reconhecer os locutores. Os resultados atestaram que a técnica possui robustez ao ruído.

Sistemas baseados em LPC apresentam custo computacional reduzido em relação aos demais. Em Srivastava (2014) utilizou-se a técnica para extração dos formantes dos discursos padrão e questionado. Posteriormente, o modelo GMM efetuou a análise dos formantes para o reconhecimento de locutores. O sistema mostrou-se menos suceptível a ruídos de fundo de até 20 dB.

A abordagem sugerida por Ahmad *et al.* (2015) apresenta complexidade em sua implementação. A extração é realizada pela técnica MFCC, modelagem pela técnica Principal Component Analysis (PCA) e análise por uma Rede Neural Probabilística (PNN). Os resultados mostram 94% de precisão.

O *software* Praat (LIESHOUT, 2003) é útil para extração de formantes utilizando LPC por Chougala e Kuntoji (2016). Os autores defendem que o uso de formantes é ideal para reconhecimento de locutores independentes de texto, principalmente quando a entrada de discurso apresenta períodos curtos.

O *software* de Praat Lieshout (2003) é também utilizado na proposta desta pesquisa para extração dos parâmetros de comparação da voz, como formantes, pitch, entre outros.

Entretanto, um sistema de análise baseado em OLS é adicionado para auxiliar a tomada de decisão por parte do perito com relação ao reconhecimento de locutores.

Assim, a análise de sinais proposta pode ser observada na seção 2.2.

2.2 Análise de sinais

A vantagem da análise de sinais é poder usar a representação no domínio da frequência de um determinado problema para simplificar a investigação matemática.

2.2.1 Transformada de Fourier e Janelamento

A transformada de Fourier é um método para representar um sinal por seus componentes de frequência (RAHMAN, 2011). Isso vem da representação da série de Fourier, que mostra a possibilidade de representar qualquer sinal pela soma de ondas simples, como senos e cossenos (BAILEY e SWARZTRAUBER, 1994). Com relação ao processamento em tempo discreto, a Transformada Discreta de Fourier (DFT) contém N-pontos de qualquer sinal x[n], definida pela Equação (1) como segue (OPPENHEIM, 2009),

$$X[k] = \sum_{n=0}^{N-1} \mathrm{x[n]}\, e^{\frac{-j2\pi kn}{N}}, \tag{1}$$

sendo k = 0, …, N-1.

Por intermédio desta operação, é possível representar os valores em x através da magnitude dos componentes deste sinal em cada valor de frequência. Esta é uma ferramenta valiosa em um contexto de processamento de dados, uma vez destacada características do sinal que não são possíveis observar no domínio de tempo. A análise das características de frequência é conhecida como análise espectral (BAILEY; SWARZTRAUBER, 1994). Para análise de áudio, a análise espectral é uma ferramenta importante, pois extrai informações sobre a origem e propriedades do arquivo de áudio, avaliando os componentes de frequência mais proeminentes do sinal. Essas frequências são conhecidas como formantes e funcionam como uma espécie de assinatura para cada apresentador, fornecendo as propriedades intrínsecas do trato vocal para cada pessoa (LEME *et al.*, 2016).

Uma função de janelamento compreende uma operação matemática e é frequentemente usada em processamento de sinais discretos. É usada para selecionar um segmento do sinal de acordo com suas propriedades. Assim, as funções janelas são

amplamente utilizadas, por exemplo, em filtros digitais, uma vez que podem restringir um sinal tanto no tempo quanto na frequência (PRABHU, 2013). Além disso, isso fornece um mecanismo de suavização de interferência nos dados, reduzindo as distorções causadas pelos efeitos de borda durante a decomposição espectral (ESSENWANGER, 1986). Consequentemente, uma série de funções de janelas vem sendo usadas, incluindo: retangular, triangular, B-sline, Parzen, Welch, Blackman e Hann (APARNA; CHITHRA, 2017).

A função de janelamento Hann é amplamente utilizada, particularmente por causa do efeito na suavização de borda, e porque é comumente aplicada no processamento de sinais de áudio (ESCH; VARY, 2009), cuja fórmula,

$$w(n) = 0.5 * \left[1 - \cos\left(\frac{2\pi n}{N}\right)\right] = \sin^2\left(\frac{\pi n}{N}\right), \quad (2)$$

sendo, w (n) a função de janela Hanning, N é o comprimento da janela e n é o valor ao longo do intervalo de $0 \leq n \leq N$.

Se a forma de onda contiver mais de um sinal com uma pequena diferença na frequência, a resolução espectral é importante. Aqui, é melhor escolher uma janela com um estreito lóbulo principal, como a janela Hanning.

2.2.2 Codificação Preditiva Linear

LPC é uma técnica de compressão de dados amplamente utilizada em análise de áudio (SINGHAL, 1990) e é uma codificação analógica de sinal que compreende a construção de um modelo para um sinal a partir de uma função linear de seus valores anteriores. Foi introduzido pela primeira vez em 1984 como uma ferramenta para compressão de dados (BRADBURY, 2000). Usando LPC, é possível construir um envelope que representa o espectro de potência de frequências de sinal de áudio confiáveis contendo simultaneamente uma baixa taxa de bits, preservando suas características de interesse. Os coeficientes LPC são estimados da seguinte forma, conforme a Equação (3) (O'SHAUGHNESSY, 1988),

$$s_p(n) = \sum_{k}^{K} a_k * s(n-k), \quad (3)$$

Sendo, a_k são os coeficientes estimados para o modelo linear, n representa cada valor do sinal modulado, $s_p(n)$ é a amostra prevista para uma dada iteração do modelo, k é o coeficiente e K

é o número máximo de coeficientes no modelo. O modelo a ser criado toma como entrada o número de coeficientes esperados para representar o sinal de interesse, dado na Equação (4),

$$K = 4 + \left(\frac{F_s}{1000}\right), \quad 4)$$

sendo, K o número de coeficientes e F_s é a frequência de amostragem. Embora o s(n) original e $S_p(n)$ estimados estão próximos, o erro err(n) é dado na Equação (5),

$$err(n) = |s(n) - s_p(n).| \quad (5)$$

Para obter um modelo melhorado, a soma dos quadrados dos erros deve ser minimizada. Por outro lado, tomando um alto número de coeficientes, estes maximizam a precisão do envelope calculado via LPC, resultando em uma representação mais confiável do espectro de potência do sinal (CHOUGALA e KUNTOJI, 2016). Para chegar às raízes dos coeficientes, a seguinte equação pode ser usada:

$$rts = \sum_{n=1}^{N} s_p(n) * x^{N-n}, \quad 6)$$

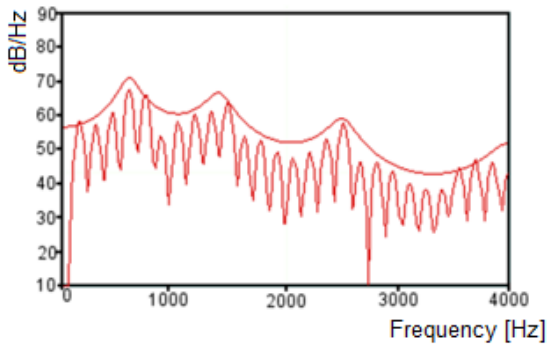
onde, s_p são os coeficientes do polinômio, rts é sua raiz, n é o coeficiente e N é o número máximo de coeficientes do polinômio.

Neste trabalho, a técnica LPC é usada para determinar as frequências dos formantes. Formantes são picos de energia de voz que definem o perfil de som de um falante (CHOUGALA; KUNTOJI, 2016). Os tons são vibrações das cordas vocais e seus modos. Também pode ser definido como a forma de onda de uma voz, amplamente divulgada como o primeiro formante, que carrega as informações mais importantes para a diferenciação do locutor. Formantes são frequências que apresentam um caráter mais proeminente quando o espectro de potência das frequências de um sinal de áudio é analisado (O'SHAUGHNESSY, 1988). Eles são caracterizados por picos que aparecem em todo o espectro. Vários estudos têm considerado a maior contribuição de até cinco dessas frequências na composição da voz, sendo a primeira de maior intensidade e denominada pitch (ou F0), e as demais (F1 a F4) de acordo (LEME *et al.*, 2016).

Com as raízes obtidas pelo modelo LPC, é possível calcular as frequências dos formantes, utilizando o arco tangente coordenado no círculo unitário (KIM *et al.*, 2006;

WALD, 1947), que refletem os picos estimados pelo modelo LPC (O'SHAUGHNESSY, 1988). Os critérios comumente usados para seleção de frequências de formantes pelo LPC consideram 90 Hz como uma frequência mínima, e a largura de banda é menor que 400 Hz (LEME *et al.*, 2016). Na Figura 1 apresenta-se o espectro obtido pela decomposição de um determinado sinal de áudio em seus componentes de frequência e o envelope por meio do LPC. Com as informações representadas no envelope LPC, é possível detectar os picos de frequência, que representam as frequências candidatas a formantes, e sua largura de banda, características necessárias para determinar os formantes.

Figura 1 - Espectro de Fourier e envelope de codificação preditiva linear (LPC) para um determinado sinal de áudio.



Fonte: Elaborado pelo próprio autor.

2.2.3 Método dos mínimos quadrados ordinários – OLS

O método proposto para comparação de áudio é baseado em OLS, levando em consideração apenas uma variável explicativa. A regressão através de OLS é uma ferramenta estatística que visa estimar a relação entre uma variável dependente e uma ou mais variáveis independentes (LEWIS-BECK e LEWIS-BECK, 2015). Neste estudo, uma comparação foi realizada entre dois segmentos de áudio (formante de referência versus formante confrontado). Portanto, a fórmula é dada considerando a equação de uma reta, vide Equação (7),

$$y = \alpha + \beta * x, \tag{7}$$

sendo, y a variável de resposta, α é o coeficiente linear, β é o coeficiente angular e x é o ponto da reta.

A qualidade do ajuste do modelo linear é calculada avaliando os quadrados residuais do ajuste. O método OLS é amplamente usado em vários contextos, como econometria, engenharia e ciência de dados. A forma como o OLS funciona é propor um modelo que se ajuste aos dados, de modo que a soma das magnitudes das distâncias entre cada ponto em relação ao modelo proposto seja a menor possível (GOLDBERGER, 1964). Isso é obtido de acordo com as equações (8) e (9),

$$y_i = \alpha + \beta * x_i + \varepsilon_i \quad (8)$$

$$\hat{\varepsilon}_i = y_i - \alpha - \beta * x_i, \quad (9)$$

sendo, $\hat{\varepsilon}_i$ o erro, e x_i e y_i são os formantes para os áudios comparados. Os estimadores de mínimos quadrados são dados pelas equações (10), (11), (12) e (13),

$$y_{médio} = \frac{(\sum_{i=0}^{n} y_i)}{n} \quad (10)$$

$$x_{médio} = \frac{(\sum_{i=0}^{n} x_i)}{n} \quad (11)$$

$$\hat{a} = y_{médio} - \hat{\beta} * x_{médio} \quad (12)$$

$$\hat{\beta} = \frac{\sum_{i=0}^{n}(x_i - x_{médio}) * (y_i - y_{médio})}{\sum_{i=0}^{n}(x_i - x_{médio})^2} = \frac{Cov(x,y)}{Var(x)} = r_{xy} * \frac{S_x}{S_y}, \quad (13)$$

sendo, r_{xy} o coeficiente de correlação de amostragem, s_x e s_y são os respectivos desvios padrão das amostras não corrigidas de x e y, $\hat{a}$ é o termo do regressor constante e $\hat{\beta}$ é o termo do regressor escalar de um modelo linear. A determinação do coeficiente R-quadrado é dada na Equação (14),

$$R^2 = r_{xy}^2. \quad (14)$$

2.2.4 Critérios de comparação estatística

Com o objetivo de comparar estatisticamente o modelo gerado, fora realizado um teste F. Isso é usado para avaliar a qualidade do ajuste de dados. Um teste F calcula as estatísticas sobre os valores das somas dos quadrados residuais no modelo testado (SEELEY e EL-BASSIOUNI, 1983). Assim, este assume um modelo arbitrário (modelo ingênuo) construído a partir dos parâmetros do modelo testado. O teste F indica se o modelo testado é capaz de se

ajustar aos dados significativamente melhor do que o modelo arbitrário. O teste F é realizado conforme a Equação (15),

$$F = \frac{\frac{RSS_1 - RSS_2}{m_2 - m_1}}{\frac{RSS_2}{n - m_2}}, \quad (15)$$

sendo que RSS representa os quadrados residuais entre o modelo arbitrário (m_1) e o modelo testado (m_2), e n é cada um dos pontos de dados comparados entre os modelos. Portanto, se o teste retorna um valor significativo, indica que o modelo de regressão linear prediz a variável de resposta melhor do que simplesmente a média dessa resposta (modelo arbitrário) (SEELEY e EL-BASSIOUNI, 1983). Os valores de significância são calculados considerando $\alpha = 90\%$ e o valor de p é representado de acordo com a Tabela 1.

Tabela 1 – Intervalos para determinação do valor p para o teste F.

Symbol	p-Value
'NS' (non-significant)	$p > 0.1$
'*'	$0.05 < p \leq 0.1$
''**	$0.01 < p \leq 0.05$
'*'**	$p \leq 0.01$

Fonte: Elaborado pelo autor.

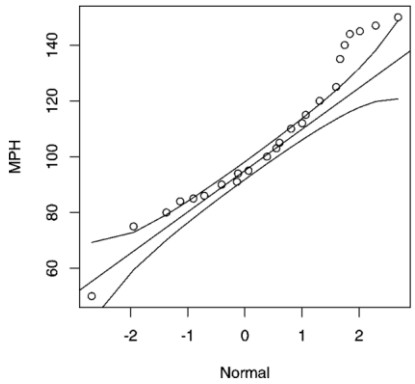
2.2.5 Q-Q Plot

Um gráfico Q-Q é uma ferramenta usada para comparar duas distribuições de probabilidade (LOY *et al.*. 2016). Isso é mostrado por meio de uma representação gráfica da dispersão entre os quantis dessas duas distribuições, tentando avaliar como uma distribuição se ajusta a outra (MARDEN, 2004). Assim, se ambas as distribuições de probabilidade são idênticas, esta dispersão é dada em linha reta, onde seus quantis também devem ser idênticos (LOY *et al.*, 2016). No entanto, qualquer desvio de uma linha reta expressa características que indicam a diferença entre essas duas distribuições.

O uso de um gráfico Q – Q ajuda a comparar os ajustes de um grupo de dados à distribuição normal, representando a distribuição gráfica dos quantis da distribuição em questão e os quantis de uma distribuição teórica normal (MARDEN, 2004). Além disso, é útil comparar quaisquer das distribuições dadas e avaliar a extensão de qualquer semelhança ou diferença, de acordo com o espalhamento em relação à linha reta. É também um método muito útil para detecção de *outliers* no grupo de dados (LOY *et al.*, 2016). Na Figura 2 mostra-se a comparação da dispersão para um dado em relação a uma distribuição normal. A

linha reta se destaca, com as linhas curvas representando o desvio padrão da distribuição em ambos os lados.

Figura 2 - Gráfico Q-Q para dispersão de dados em relação a uma distribuição normal.



Fonte: Elaborado pelo próprio autor.

2.3 Reconhecimento de Locutores na área pericial forense brasileira

O reconhecimento forense de locutores tornou-se uma ferramenta importante nas esferas judiciais e administrativas, pois o uso exclusivo da voz humana como instrumento é cada vez maior, nos mais diversos métodos de comunicação existentes.

Para contribuir com as soluções dos problemas, existem palavras-chaves que são diferentes para a identificação de falantes forenses em cada idioma, estas determinadas por fonólogos, linguistas, cientistas e engenheiros, no processo científico de comparação de locutores.

Existem diversos tipos de procedimentos atualmente utilizados em vários sistemas jurídicos para comparação de locutor. Porém, predomina a aplicação de método subjetivo em vez do objetivo, permitindo a cada profissional na área contra-argumentar suas metodologias e proferir suas próprias conclusões, inviabilizando o senso comum, carecendo de materiais técnicos computacionais capazes de atingir resultados quantitativos sólidos, os quais serão debatidos neste trabalho.

Argumenta-se que, embora as ferramentas baseadas em computador para análises acústicas e novos dados estatísticos sobre algumas características específicas dos falantes possam auxiliar o especialista em tirar conclusões, um dispositivo de identificação de voz totalmente automático certamente não está à vista, devido ao grande número de variáveis e distorções do sinal de fala que são típicas da situação forense.

Em um caso forense, não existe um texto pré-selecionado ou pré-organizado que possa ser usado por ferramentas de ponta para fazer comparações palavra a palavra. Em muitos

casos, não existe um único aspecto da amostra de fala, incluindo a duração geral do áudio, sobre a qual o especialista pode exercer uma influência. As pessoas suspeitas se recusam a produzir qualquer amostra de discurso, o que, é claro, pode fazer na maioria dos sistemas legais, não produzir prova contra si mesmo.

Desta forma, para captar o padrão de voz do indivíduo que não está disposto a contribuir com os peritos, deve-se ter uma metodologia que possa vir a ser usada com ruídos externos, fala pausada, com intuito de executar a perícia do arquivo digital, ponto este também trabalhado trabalho para que exista embasamento técnico e científico corroborando em resultados periciais conclusivos.

A aplicação de reconhecimento de voz na área pericial é utilizada de forma subjetiva para identificar locutores. Como forma de provar que um discurso pré-amostrado é realmente de um indivíduo, a captura não controlada de um segundo discurso muitas vezes se faz necessária. Diz-se então, que o processo de reconhecimento de locutor na área pericial apresenta um perfil específico de entrada de dados:

- Independente do texto – O reconhecimento de locutor em questão deve ser realizado independentemente das diferenças entre o discurso pré-amostrado e o questionado;
- Pode apresentar ruído – A ferramenta deve ser robusta a eventuais ruídos presentes em ambos os discursos, pré-amostrado e/ou questionado;
- Tamanho curto de trecho – Como forma de melhorar o nível de certeza de identificação a amostragem do áudio questionado deverá apresentar trechos curtos.

A área pericial forense faz uso do reconhecimento manual de locutores como forma de julgar se um áudio questionado pode ser comparado com a voz de um suspeito. A técnica manual ainda é utilizada por não haver ferramentas automáticas que tragam confiabilidade alta para efetuar tal tarefa, sem o veredito e experiência de um perito. Entretanto, a área forense pode se apoiar em ferramentas técnicas de análise como auxílio ao laudo pericial (BRAID, 2003). Atualmente, um sinal de áudio a ser analisado pela perícia em geral apresenta-se em gravações curtas, de locais ruidosos e/ou geradas através de gravadores e telefones. A análise ocorre em duas etapas:

- Análise perceptual humana – Por meio da audição humana, o perito pode efetuar suas interpretações aos estímulos quanto à idade, gênero, saúde e até identidade de um falante interlocutor de determinado discurso. Entretanto, as considerações baseadas nesse tipo de análise são pautadas na experiência do perito;

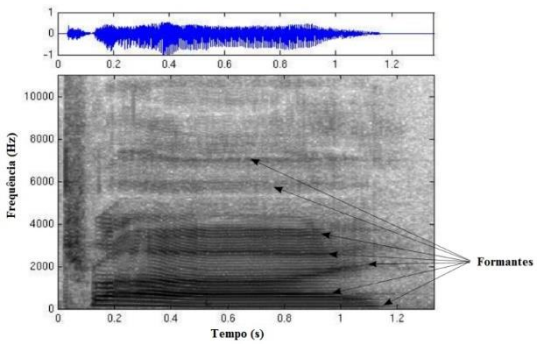
- Análise acústica – Por meio de ferramentas de *software*, pode-se extrair e até analisar de forma estatística e dando recursos técnicos, e as vezes objetivos, de se comparar as vozes do áudio questionado com o suspeito.

A análise acústica em geral se utiliza de espectrogramas para extrair elementos e parâmetros das falas. Entre os elementos analisados neste trabalho pode-se citar:

- Formantes – Os formantes de uma fala podem ser extraídos das regiões ressonantes do discurso, presentes em um espectrograma. Quando o trato vocal muda seu contorno, como na Figura 3, a região de ressonância também muda (BEIGI, 2011);
- *Pitch* – O *pitch* é uma relação de formato e tensão da frequência fundamental de vibração das cordas vocais durante a fala. Essa relação sobre influência do fluxo de saída de ar juntamente com a configuração do trato vocal em enunciados (BEIGI, 2011);
- Intensidade – Intensidade de vibração das cordas vocais durante o *pitch*. Existe uma relação direta entre os parâmetros *pitch* e intensidade (*loudness*) (BEIGI, 2011).

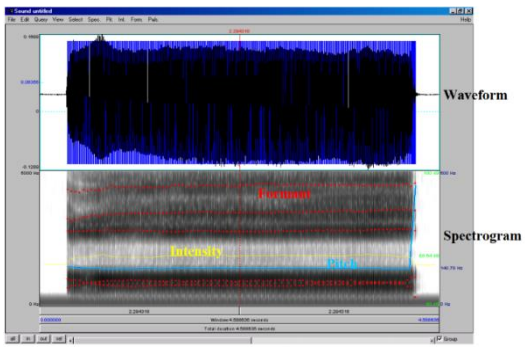
Na área forense brasileira, esses parâmetros são extraídos da fala usando *software*s semelhantes ao Praat (LIESHOUT, 2003), conforme ilustrado na Figura 4.

Figura 3 – Representação de Formante.



Fonte: Beigi (2011).

Figura 4 – Formantes, Intensidade e Pitch identificados pelo *software* Praat.



Fonte: Adaptado de Lieshout (2003).

Na seção 2.4 apresenta-se duas análises onde a metodologia de reconhecimento de locutores é aplicada na área forense brasileira.

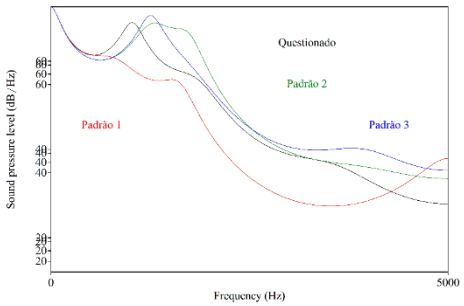
2.4 Análise de casos da metodologia de Reconhecimento de Locutores

Esta seção descreve o método de reconhecimento de locutores utilizado atualmente pelos Peritos Oficiais Criminais da Polícia Federal Brasileira para complementar o laudo pericial com recursos técnicos acústicos sobre o discurso questionado.

Esta técnica utiliza o *software* de áudio forense Praat, que desenha as curvas dos formantes dos objetos questionados e os padrões dos falantes suspeitos, fazendo com que o especialista visualize as curvas mais próximas do material verificado para poder emitir uma afirmação.

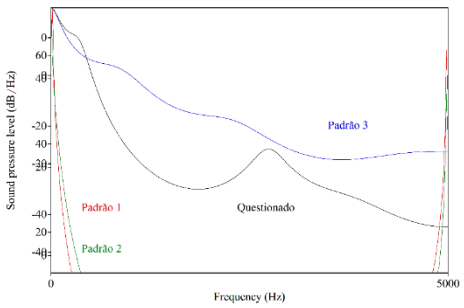
Observando a Figura 5, o padrão 3 apresenta curva semelhante ao formante do áudio questionado. Então, neste caso, a conclusão trivial do perito seria que existe grande probabilidade de o falante do áudio questionado ser o suspeito interlocutor do padrão 3. Entretanto, tal análise apresenta muita subjetividade e está sujeita a experiência do perito encarregado. Isso se deve porque nem sempre as curvas dos formantes do áudio questionado e seu provável autor são parecidas, conforme ilustrado na Figura 6. De acordo com esta segunda forma de onda, já não é trivial apontar uma conclusão pericial. Afinal nenhum padrão dos suspeitos corresponde diretamente ao áudio questionado.

Figura 5 – Comparação trivial de curvas de formantes.



Fonte: Elaborado pelo próprio autor.

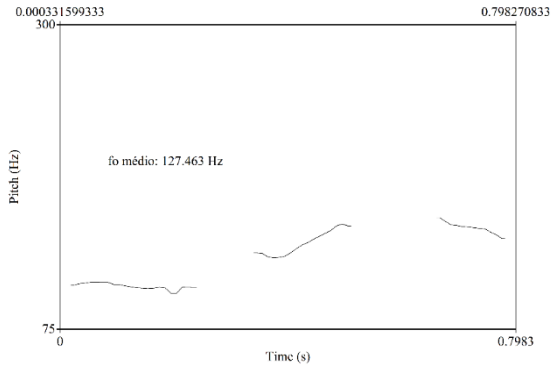
Figura 6 – Comparação não trivial de curvas de formantes.



Fonte: Elaborado pelo próprio autor.

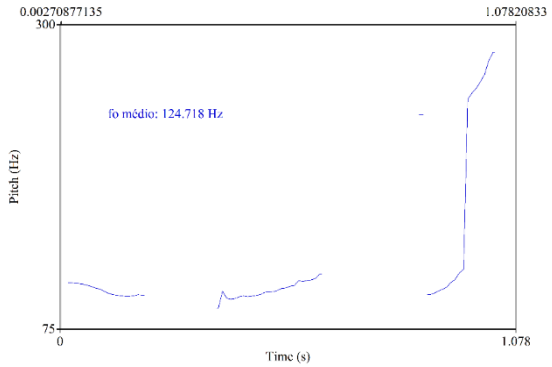
O programa Praat, também possibilita a análise pelo pitch, frequência primordial ou F0. Nas Figuras 7 e 8 que os valores do pitch médio são equivalentes, porém suas curvas indicam desproporcionalidade, tendo no padrão 3 pico próximo a 300 Hz.

Figura 7 – Pitch (F0) médio questionado.



Fonte: Elaborado pelo próprio autor.

Figura 8 – Pitch (F0) médio do padrão 3.



Fonte: Elaborado pelo próprio autor.

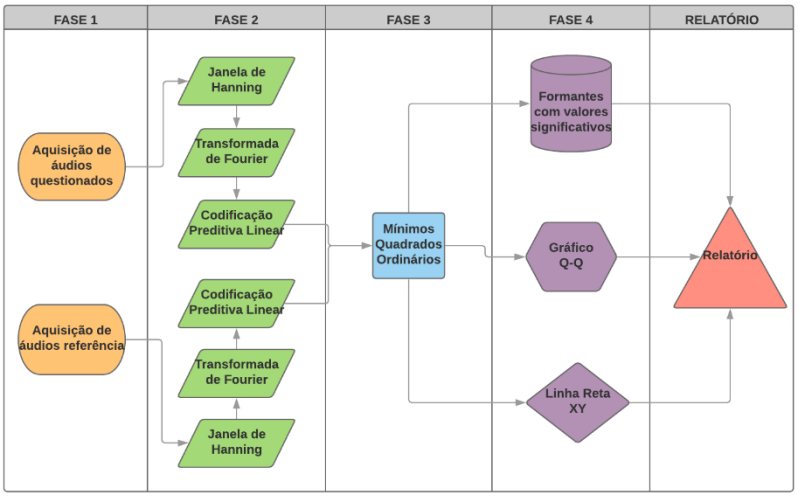
A metodologia aqui descrita é aplicada para os casos de reconhecimento de locutores na área forense brasileira. Na sua totalidade baseada em conclusões subjetivas, e por interpretação dos gráficos, a análise de perfil se mostra suscetível à perda de robustez para caraterísticas da fala como regionalismos, sotaques, cacoetes, tiques entre outros. Estas variáveis, mesmo que presentes no parecer técnico, dificultam que dois peritos efetuem a mesma análise.

Logo, como forma de melhorar e padronizar a análise acústica dos áudios da área forense, neste trabalho apresenta-se uma proposta que adiciona recursos de análise e que de forma objetiva auxiliem o perito a ter laudos mais robustos contra falhas.

3 METODOLOGIA PROPOSTA

Para melhor entender a metodologia proposta, realizou-se um fluxograma apresentado na Figura 9.

Figura 9 – Estrutura desenvolvida para comparação forense de locutor.



Fonte: Elaborado pelo autor.

Com o método proposto, espera-se produzir o resultado com o objetivo de confirmar se o áudio confrontado corresponde à referência. Essa abordagem pode ser resumida da seguinte maneira. Primeiro, a mesma frase produzida no áudio contestado é gravada por uma suposta referência (Fase 1). Em seguida, ele executa o janelamento seguido pela transformada de Fourier. Posteriormente, o algoritmo LPC extrai os formantes (Fase 2). Na próxima etapa, utiliza-se o modelo OLS para comparar estatisticamente se cada formante apresenta algum grau de significância (Fase 3). Na Fase 4, o gráfico Q-Q e as linhas retas XY são plotados para cada formante no confronto entre os dois áudios. Na Fase 5, a metodologia apoia a produção de um relatório final, que pode ser analisado por um perito forense.

Cada etapa da metodologia está descrita nas seções posteriores.

3.1 Fase 1: Aquisição de áudios contestados e referência

Os arquivos de áudio foram adquiridos e capturados por meio de aparelho celular em ambiente controlado, com o telefone posicionado próximo ao locutor. Os arquivos foram amostrados a 48 kHz no modo de áudio estéreo. O indivíduo executou as frases propostas utilizadas para realizar a comparação com o falante. As frases gravadas foram as seguintes (em português brasileiro):

1 “O rato roeu a roupa do rei de Roma”;

2 “O macaco mordeu a macacada no monte Maia”;

3 “O macaco mordeu o sapato”.

É importante ressaltar que este estudo empregou frases curtas para comparação de falantes, pois isso reduziria significativamente a taxa de erro. Os desempenhos dos sistemas de verificação automática de locutor degradam-se, devido à redução na quantidade de voz usada para inscrição e verificação. Combinar vários sistemas (com base em diferentes recursos e classificadores) pode reduzir consideravelmente a taxa de erro de verificação do locutor com enunciados curtos (PODDAR *et al.*, 2019).

O processamento de áudio foi realizado usando um aplicativo desenvolvido em Matlab 2018b. Primeiro, os dados foram carregados por uma interface gráfica a usuário (GUI) do próprio Matlab, onde o usuário poderia selecionar o arquivo de áudio desejado. O *software* é capaz de suportar qualquer formato de arquivo de áudio, como m4a, ogg, wav, mp3 e mp4. No entanto, o formato padrão m4a foi o escolhido. Depois de carregar o arquivo de áudio, o usuário pode optar por reproduzir a faixa selecionada para verificar o arquivo.

3.2 Fase 2: Extração de Formantes

O arquivo de áudio é então segmentado em partes menores de duração ajustável, de acordo com a seguinte equação:

$$t_{jan} = \left(\frac{0.45}{pitch_{floor}}\right) * 1000, \tag{16}$$

sendo, t_{jan} é o tamanho da janela em segundos, e $pitch_{floor}$ é a menor frequência esperada para aquele arquivo de áudio analisado, selecionada pelo usuário.

Os segmentos possuem uma sobreposição (janela Hanning) para permitir uma maior continuidade entre os dados, e seu valor pode ser selecionado pelo usuário. É definido com uma sobreposição de 90% como padrão. Por meio da Equação (16), o tempo de janelamento é mantido abaixo dos intervalos de 25 ms, que é o valor mínimo necessário recomendada para a extração das características espectrais referentes aos sinais de áudio (DRESCH, 2015).

Após a segmentação de dados, a decomposição do espectro de potência das frequências para cada segmento foi realizada. Para tanto, foi aplicada uma janela de Hanning para suavizar os efeitos da borda que podem distorcer o sinal. Em seguida, a transformada rápida de Fourier (FFT) foi calculada para os segmentos. Como resultado, um novo conjunto de dados foi obtido abrangendo as potências do espectro no intervalo de 0–24 kHz.

O próximo passo foi estimar o número de coeficientes para cada janela de dados, visando extrair o *pitch* e as frequências dos formantes do sinal via LPC (O'Shaughnessy, 1988). Em seguida, as frequências foram obtidas de acordo com os critérios de frequência mínima (conforme o estágio de janelamento do sinal) e com a largura de banda máxima aceitável (Leme *et al.*. 2016). Foram consideradas as frequências que correspondiam a ambos os critérios, sendo a primeira pitch (F0) e as subsequentes F1 a F4. Na literatura foi demonstrado que estas são passíveis de englobar a informação espectral referente a cada vogal (LEME *et al.*, 2016). Contudo, o *software* desenvolvido é executado na extração do maior número possível de formantes do áudio (aproximadamente 16–20 formantes). É importante mencionar que o número de formantes depende das propriedades espectrais das vozes de cada indivíduo. A extração automática das propriedades acústicas ocorreu por meio da técnica LPC, que reduziu os dados coletados por regressão linear. Em seguida, os dados obtidos foram armazenados para comparação entre os falantes. A comparação foi realizada em pares por meio da análise dos dados dos formantes através do algoritmo OLS.

3.3 Fase 3: OLS e comparação estatística

Os dados foram agrupados de acordo com cada formante específico e comparados via OLS por meio de regressão linear (Goldberger, 1964). O modelo linear implementado retorna as características relacionadas à comparação de cada um dos formantes de áudio comparados. Ele faz a avaliação resultante comparando os valores de p que resultam do teste F realizado pelo modelo. Para julgar os resultados obtidos pelo modelo implementado, foi testado o valor de significância retornado pela análise de pitch, que foi o primeiro dos critérios. Em seguida, a análise foi concluída garantindo formantes que ajudaram na comparação entre os palestrantes. O tom compreende a frequência mais relevante para a comparação. Esperava-se que o aumento do número de formantes auxiliares retornasse valores significativos pelo modelo, produzindo uma quantidade maior de evidências que as vozes comparadas são oriundas do mesmo falante.

Um gráfico Q-Q foi usado para corroborar que os resíduos dos formantes comparados tiveram uma distribuição semelhante. Isso mostrou que compreendiam faixas de frequências semelhantes, independentemente do resultado que apresentassem no modelo. Finalmente, a linha XY foi traçada confrontando cada formante do áudio contestado com aquele de referência. Isso demonstrou como cada ponto dos dados estava relacionado à sua respectiva distância da linha XY do modelo OLS.

Por meio da análise do processamento do *software* via gráfico e valores absolutos dos parâmetros, o perito pode encontrar indícios objetivos e técnicos que apontam para qual dos suspeitos apresenta características da fala mais próximas do áudio questionado. No Capítulo 4 são efetuados testes que avaliam o funcionamento e robustez da técnica proposta.

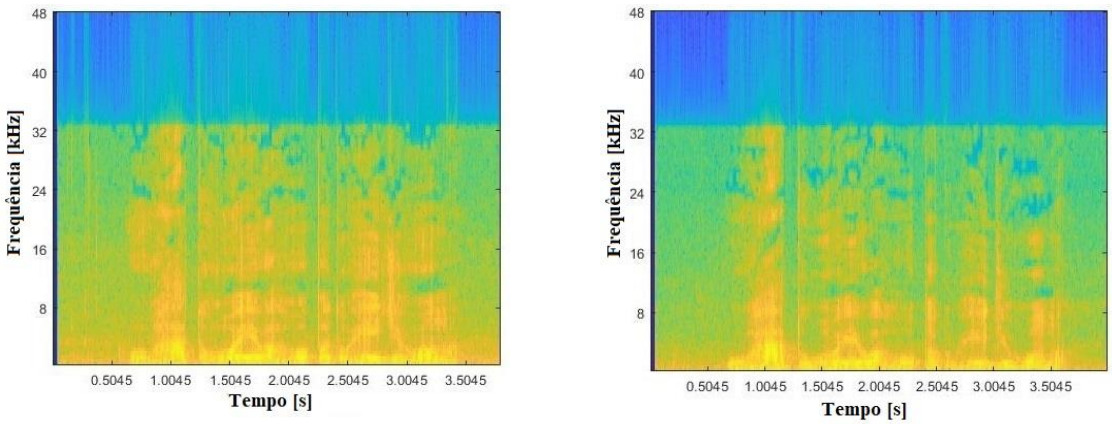
4 TESTES E RESULTADOS

Uma série de testes aplicados e seus resultados são apresentados neste capítulo.

4.1 Resultados experimentais

Para verificar se um áudio confrontado corresponde a uma referência, a primeira etapa consiste na análise do padrão de espectro. Os áudios foram adquiridos conforme apresentado na Fase 1. Na Figura 10 (a) mostra-se o espectro de frequências gerado para um determinado suspeito, enquanto na Figura 10(b) mostra-se o espectro para a referência (ambos relacionados à sentença # 1). Os resultados foram obtidos aplicando a janela Hanning (90% de sobreposição) e a FFT (Fase 2). Conforme indicado pela Figura 10, a análise de áudios utilizando apenas o espectro de frequências, além de ser complexa, torna-se muito subjetiva e tende a introduzir erro, acabando por exigir mais informações de dados para a tomada de decisão.

Figura 10 - Espectro de frequência de áudio para: (a) confrontado; (b) referência.



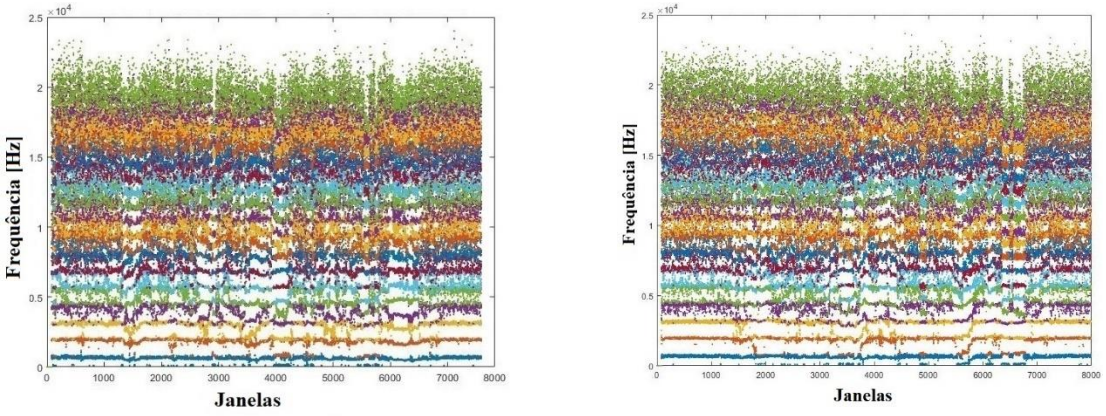
(**a**) (**b**)

Fonte: Elaborado pelo autor.

É importante ressaltar que outros fatores importantes para a abordagem proposta são como os formantes são posicionados e com a quantidade de composição. Assim, o algoritmo LPC foi conduzido. Na Figura 11 mostra-se o comportamento de cada

formante com suas respectivas frequências em relação com a frase #1, em 11 (a) apresenta-se os áudios contestados e em 11 (b) os áudios de referência.

Figura 11 - Formantes (N = 19) extraídos: (a) confrontado; (b) referência.

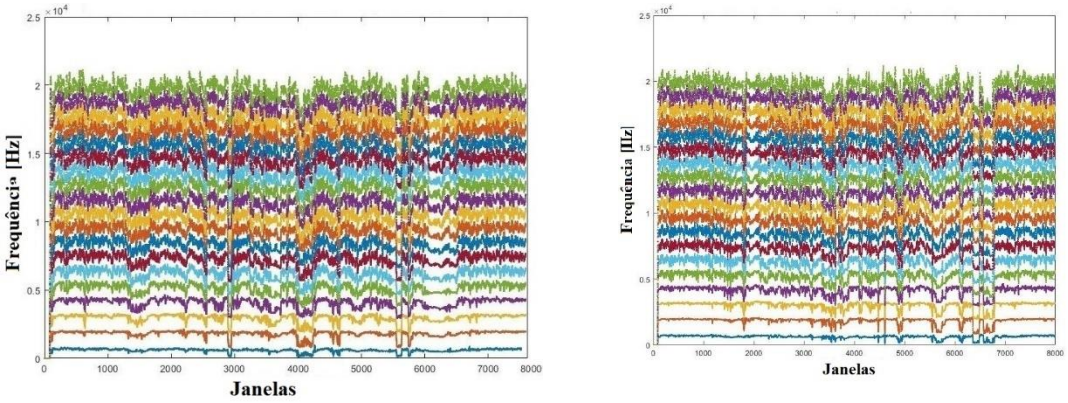


(**a**) (**b**)

Fonte: Elaborado pelo autor.

Note que ainda não é possível estabelecer um parâmetro para um resultado confiável baseado apenas em gráficos. Para aprimorar a visualização, foi realizado um filtro de média móvel (MAV). Este filtro é usado para clarear os pontos, tornando-os mais contínuos. Na Figura 12 (a) mostra-se os formantes extraídos do áudio contestado, enquanto na Figura 12 (b) apresenta-se aqueles extraídos do áudio de referência, após a aplicação MAV (ambos relativos à Sentença nº 1). É notável que certificar a probabilidade entre os espectros dos formantes, mesmo após a aplicação do filtro MAV, se apresenta como uma tarefa difícil. Dessa forma, foi necessário quantificar essas diferenças sutis estatisticamente, visando obter uma análise muito mais precisa, conforme exigido no contexto forense.

Figura 12 - Suavizando os formantes usando um filtro MAV para: (a) áudio confrontado; (b) áudio de referência.



(**a**) (**b**)

Fonte: Elaborado pelo autor.

Para obter um método robusto de verificação do locutor, neste trabalho é proposto a utilização do modelo OLS, para comparação dos formantes de referência foram comparados ao áudio confrontado. O pequeno valor de p resultou em uma significância mais alta. Nesta abordagem, a maior significância foi representada por ´ *** `. É importante destacar que o pitch é o formante principal, e se não apresentar ´ *** ´, a comparação é considerada negativa, exigindo que os formantes restantes apresentem somente um não significativo (NS) e os demais pelo menos um ´ * ´, exceto uma pequena sequência de NS nos últimos formantes. Dessa forma, na Tabela 2 apresentam-se os resultados obtidos por meio de OLS, comparando o áudio confrontado com a referência (ambos para a frase #1). Após a análise dos resultados, fica evidente que os 19 formantes apresentaram altos níveis de significância. Quanto menor o valor p, mais forte é a correlação entre os pares (formante F3).

Tabela 2 – Níveis de significância obtidos pelo modelo OLS considerando os áudios analisados.

Formante	Valor de p	Significância
Pitch (F0)	4,45191E-17	***
F1	3,0643E-08	***
F2	5,21042E-15	***
F3	3,28572E-22	***
F4	8,03937E-11	***
F5	1,63395E-11	***
F6	9,19947E-14	***
F7	8,38252E-14	***
F8	2,15671E-10	***
F9	1,75295E-08	***
F10	7,46944E-09	***
F11	6,53178E-08	***

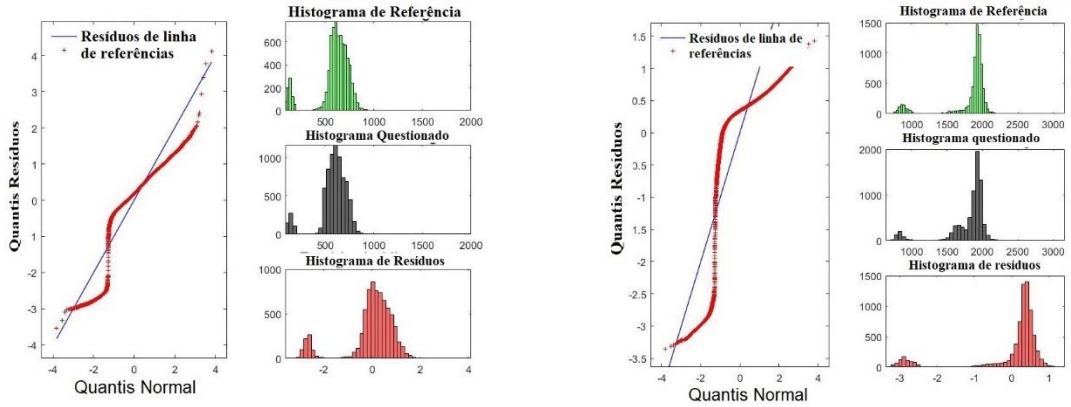
F12	7,18641E-07	***
F13	2,64407E-07	***
F14	2,29227E-06	***
F15	1,23279E-05	***
F16	1,24595E-06	***
F17	1,37544E-05	***
F18	0,000179185	***

Fonte: Elaborado pelo autor.

4.2 Validando o modelo

Para comparar distribuições normais, o gráfico Q – Q pode ser utilizado. O objetivo foi verificar se o modelo apresenta distribuição semelhante, se o histograma da referência era próximo ao do áudio confrontado. Se um resultado afirmativo for alcançado, isso fornece um resultado positivo para a comparação de locutores, garantindo que a distribuição normal seja confrontada com o mesmo segmento de áudio. Os histogramas residuais, por sua vez, são inversamente proporcionais à qualidade do formante. Isso significa que correspondências menores de histogramas levam a um valor p melhor, aumentando a confiabilidade do modelo. Para fins de brevidade, apenas alguns formantes são mostrados. Nas Figuras 13 (a) e (b) mostram os gráficos Q – Q para os formantes F0 (pitch) e F1, respectivamente. Doravante, todos os números apresentados nesta seção estão relacionados à sentença #1. Ao examinar a Figura 13 (a), fica evidente que existem quatro pontos de intersecção dos resíduos com a reta: um de alta, dois de média e um de baixa intensidade. Por outro lado, na Figura 13 (b) apresenta-se dois pontos de intersecção, um de alta e um de baixa intensidade. Para concluir, F0 apresenta maior grau de significância em comparação com F1, conforme apresentado anteriormente na Tabela 2 (valor de p).

Figura 13 - Gráfico Q-Q dos formantes comparando o áudio confrontado com os áudios de referência e confrontado: (a) formante F0 e (b) formante F1.

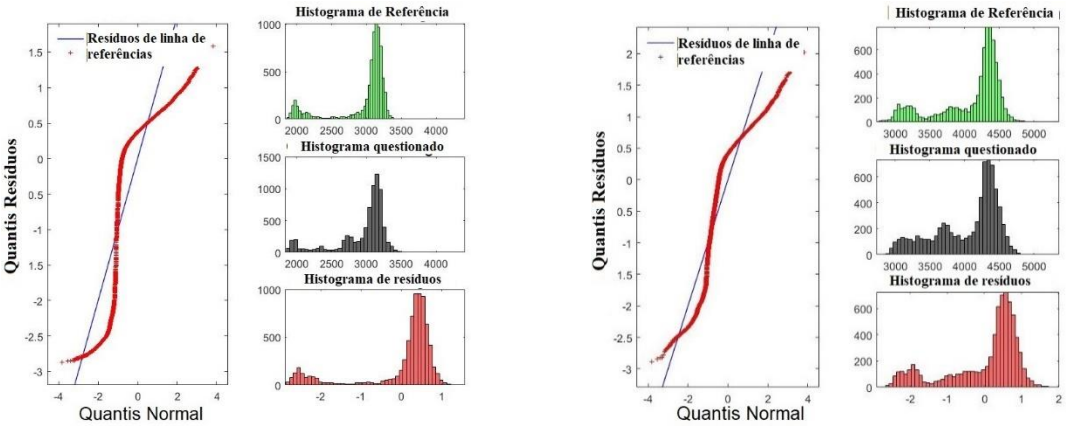


(**a**) (**b**)

Fonte: Elaborado pelo autor.

Nas Figuras 14 (a) e (b) mostra-se os gráficos Q-Q para os formantes F2 e F3, respectivamente. Ao examinar a Figura 14 (a), fica evidente que existem três pontos onde os resíduos se cruzam com a linha reta: um é muito alto e dois de alta intensidade. Por outro lado, a Figura 14 (b) apresenta três interseções altas. Esse resultado é compatível com os da Tabela 2, onde F3 apresentou o maior valor para p.

Figura 14 - Gráfico Q-Q dos formantes comparando o áudio confrontado com os áudios de referência: (a) formante F2 e (b) formante F3.

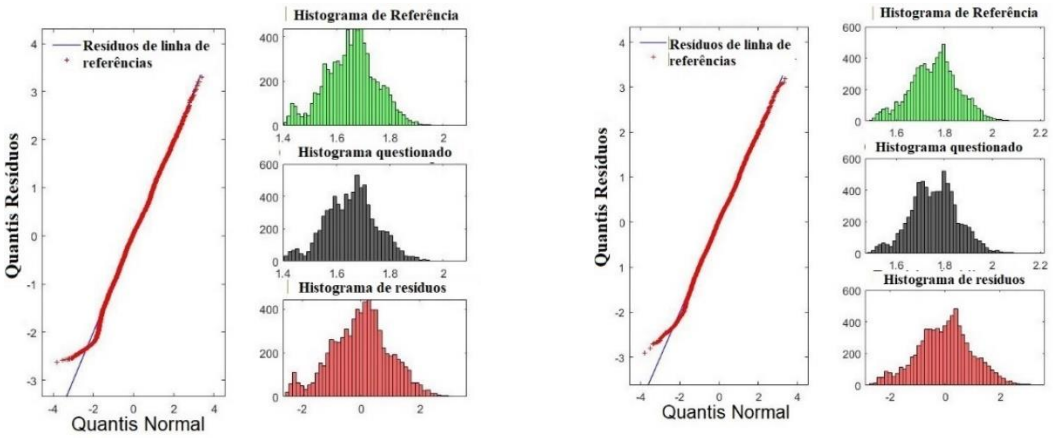


(**a**) (**b**)

Fonte: Elaborado pelo autor.

Na Figuras 15 (a) e (b) apresenta-se os gráficos Q-Q para os formantes F15 e F16, respectivamente. Ao examinar as Figuras 15 (a) e (b), fica evidente que a linha dos resíduos coincide com a linha de referência, mostrando que as significâncias são mais fracas do que as apresentadas para F3, por exemplo. Esses resultados são compatíveis com os apresentados na Tabela 2, onde F15 e F16 apresentam valores de p menores.

Figura 15 - Gráfico Q-Q dos formantes comparando o áudio confrontado com os áudios de referência: (a) formante *F15* e (b) formante *F16*.

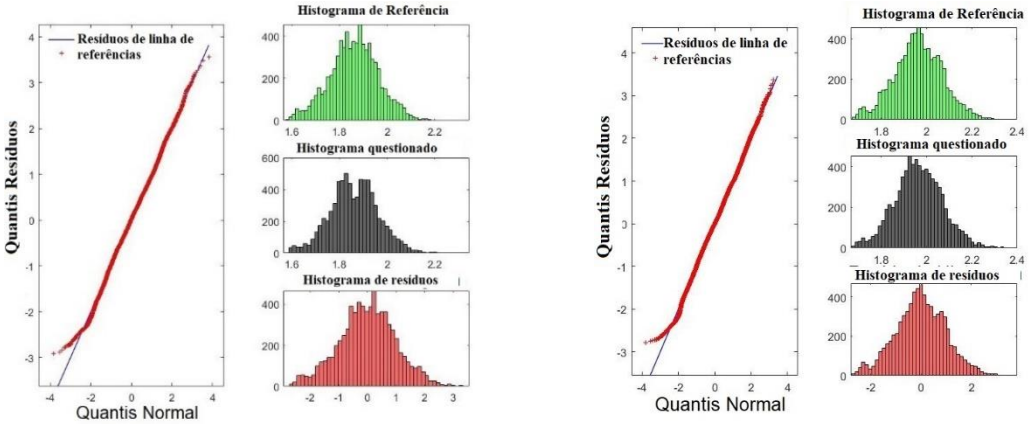


(**a**) (**b**)

Fonte: Elaborado pelo autor.

Nas Figuras 16 (a) e (b) mostram-se os gráficos Q-Q para os formantes F17 e F18, respectivamente. Da mesma forma, os resultados indicam que a linha dos resíduos coincide com a linha de referência, mostrando que as significâncias são mais fracas do que para F2 e F3, por exemplo. Esses resultados são compatíveis com os apresentados na Tabela 2, onde F17 e F18 apresentaram os menores valores de p. Espera-se que os últimos quatro formantes tenham resultados menos satisfatórios em comparação com os primeiros formantes. No entanto, eles podem ser importantes para verificar os falantes com tons de voz semelhantes.

Figura 16 - Gráfico Q-Q dos formantes comparando o áudio confrontado com os áudios de referência: (a) formante F17 e (b) formante F18.



(**a**) (**b**)

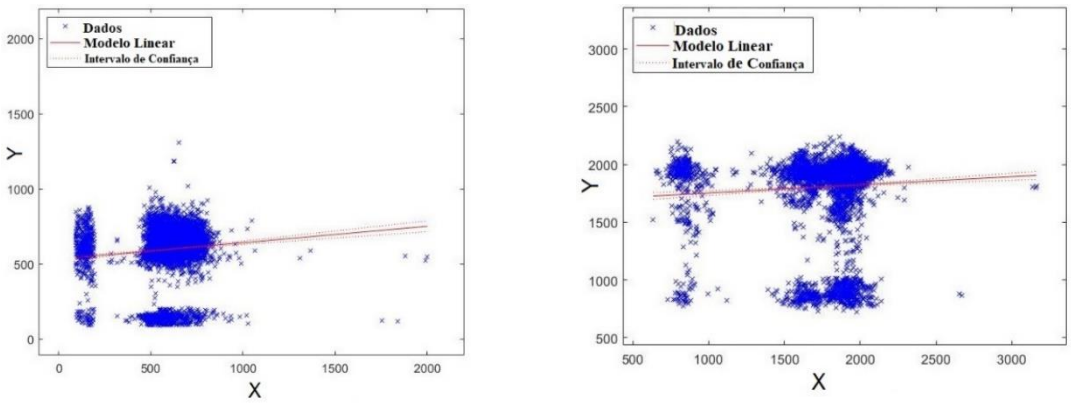
Fonte: Elaborado pelo autor.

É importante ressaltar que embora alguns formantes no gráfico Q–Q sejam quase coincidentes com a linha de referência, isso não desqualifica os resultados do modelo. O

gráfico Q – Q é uma comparação de como os dados se ajustam em relação à norma. Assim, divergências em relação à linha central indicam que há diferença no regime de frequência dos formantes comparados, mas não sugerem problemas com a qualidade do modelo. Além disso, a semelhança no perfil dos resíduos indica que os formantes apresentam padrões semelhantes. Porém, a comparação por meio de um modelo linear pode (ou não) retornar um valor significativo, independentemente. Assim, no gráfico Q – Q mostra-se que a faixa de frequência comparada entre os formantes é semelhante, o que reforça o método extraído dos formantes usando o LPC.

Aqui, o modelo obtido por meio de um gráfico XY também foi avaliado. Esperava-se que os dados estivessem próximos da linha reta, o que indicaria um bom ajuste para o modelo. Como consequência, fica implícito que o nível de significância do formante é mais forte. Nas Figuras 17 (a) e (b) mostram-se os gráficos para os formantes F0 e F1, respectivamente. Conforme observado em tais figuras, os dados estão aglomerados próximos à reta, mostrando que os modelos para os formantes F0 e F1 têm forte significância.

Figura 17 - Linha reta XY para: (a) formante F0 e (b) formante F1.

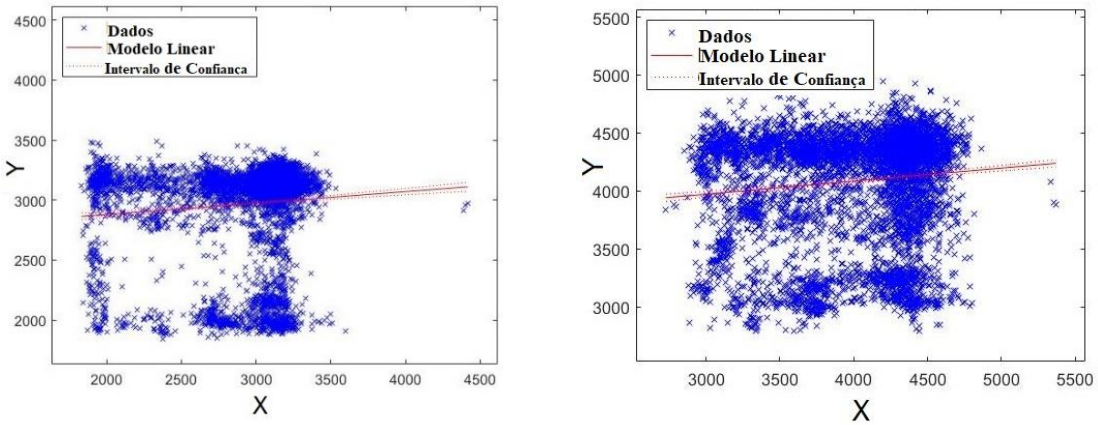


(**a**) (**b**)

Fonte: Elaborado pelo autor.

Nas Figuras 18 (a) e b mostram-se os gráficos para os formantes F2 e F3, respectivamente. Conforme observado, mais uma vez, os dados estão aglomerados próximos à reta, mostrando que os modelos para os formantes F2 e F3 possuem forte significância.

Figura 18 - A reta XY para: (a) formante F2 e (b) formante F3.

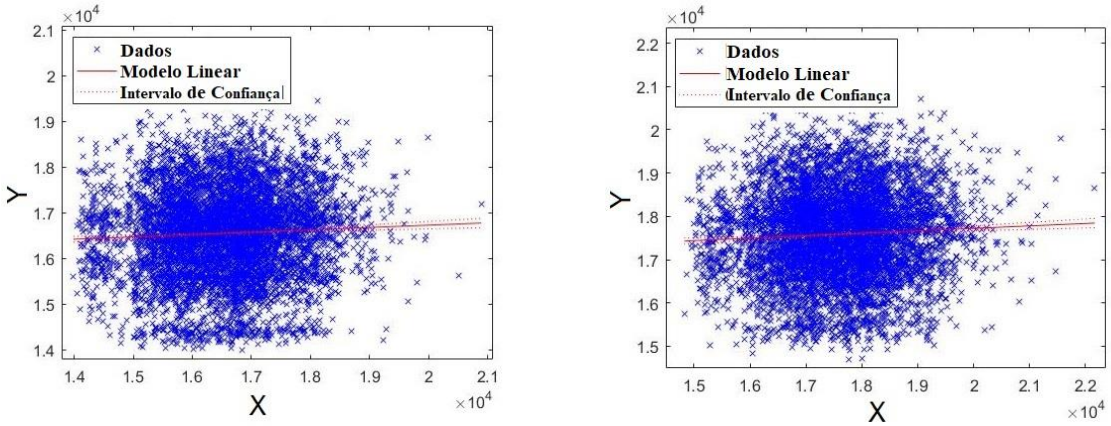


(**a**) (**b**)

Fonte: Elaborado pelo autor.

Os últimos quatro formantes também foram avaliados. Primeiro, nas Figuras 19 (a) e (b) mostram os gráficos para os formantes F15 e F16, respectivamente. Diferentemente dos resultados já apresentados, os dados estão espalhados pela reta, mostrando que os modelos para os formantes F15 e F16 têm menor significância em comparação com F0, F1, F2 e F3 (Figuras 11 e 12). Resultados semelhantes também foram mostrados para valores p e gráfico Q-Q. Uma atenção especial deve ser dada às escalas de eixos das Figuras 19 e 20, pois são multiplicadas por 10^4, ao contrário das Figuras 17 e 18.

Figura 19. A reta XY para: (a) formante F15 e (b) formante F16.



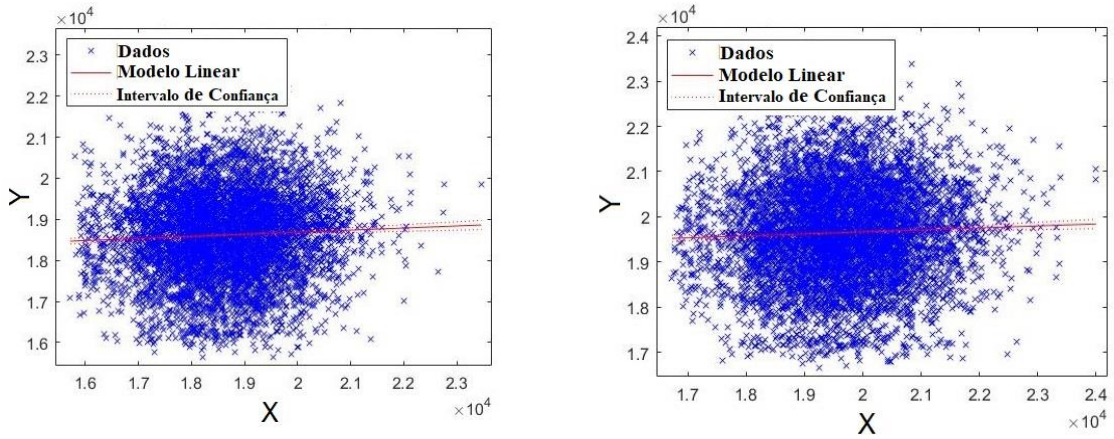
(**a**) (**b**)

Fonte: Elaborado pelo autor.

Em segundo lugar, nas Figuras 20 (a) e (b) mostram os gráficos para os formantes F17 e F18, respectivamente, e conforme observado nessas figuras, os dados estão espalhados pela reta, mostrando que os modelos para os formantes F17 e F18 apresentam menor significância

em comparação com F0, F1, F2 e F3 (Figuras 17 e 18). Resultados semelhantes já foram mostrados para valores p e gráficos Q-Q.

Figura 20 - Linha reta XY para: (a) formante F17 e (b) formante F18.



(**a**) (**b**)

Fonte: Elaborado pelo autor.

4.3 Resultados práticos

Para validar o método proposto, foram realizados testes práticos, considerando um banco de dados composto por 26 falantes (13 homens e 13 mulheres), onde todos os 26 como suspeitos foram classificados ordinalmente (ver Tabela 3). Outro áudio foi gravado a partir de um destes e tomado como referência. A coluna "Tempo" refere-se à duração do áudio gravado. Um cenário interessante foi realizado, em que os suspeitos falaram a frase nº 2 (como áudio de referência), e o áudio contestado da frase nº 3 foi gravado. Na Tabela 3 mostra-se os resultados dos valores de p comparando todos os suspeitos e formantes.

Na Tabela 3, é evidente que o suspeito 3 atingiu um valor de significância mais alto de `***` e apenas um formante, NS. Esses resultados estão em um nível de confiança aceitável para certificar positivamente que o suspeito nº 3 é o falante responsável, de acordo com o áudio contestado. Ao analisar a Tabela 3, pode-se considerar que o suspeito nº 11 pode ser acusado de falar. É importante destacar que o valor de significância retornado pela análise de pitch (F0) deve ser levado em consideração no primeiro critério de análise. Como o valor de p é NS para F0, essa hipótese deve ser rejeitada. Esse teste também levou em consideração uma frase com supressão e troca de palavras, bem como o pico de frequências sendo diferente para cada personagem narrado. Esta análise permite a validação do modelo desenvolvido, levando à verificação de qual locutor é o responsável, mesmo sob alguma interferência no áudio gravado.

Tabela 3 – Os resultados da comparação entre os suspeitos (Sentença # 3) e a referência (Sentença # 2).

Suspeito	**Tempo(s)**	**Pitch**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**	**F16**	**F17**	**F18**
1	4,52644	NS	***	***	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS		
2	5,715986	*	***	**	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	NS
3	**4,5**	*******	*******	**NS**	*****	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	
4	4,457347	***	***	***	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
5	2,942653	***	***	**	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
6	4,244014	***	***	***	***	***	***	***	***	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
7	4,265351	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
8	3,732018	***	NS	*	NS	NS	NS	NS	**	**	**	**	**	***	***	***	**	*	*	**
9	3,092018	***	***	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
10	4,154921	NS	NS	NS	*	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	NS			
11	4,03068	NS	*	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
12	5,118685	NS	***	NS	*	***	***	***	***	***	**	***	**	**	**	**	*	**	*	NS
13	3,412018	NS	NS	***	***	***	***	***	**	**	*	*	*	NS	NS	NS	NS	NS	NS	NS
14	3,966667	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
15	3,433333	***	*	NS	NS	NS	NS	**	**	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS
16	3,54	**	***	***	***	***	***	***	***	**	**	**	***	**	*	NS	NS	NS	NS	NS
17	4,99068	*	NS	*	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	
18	4,18	*	***	***	***	***	**	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
19	4,5	***	***	***	***	**	***	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS

Fonte: Elaborado pelo autor.

Na Tabela 4 mostra-se a comparação para outro cenário, em que a frase nº 1 é gravada para os áudios contestados e de referência. A comparação confrontou vários suspeitos, incluindo alguns dos usados anteriormente na Tabela 3. Mais uma vez, o suspeito nº 3 alcançou significância máxima em todos os formantes (um resultado positivo), validando a hipótese de que o áudio confrontado lhe pertence.

Tabela 4 – Os resultados do confronto entre os suspeitos (Sentença # 1) e referência (Sentença # 1).

Suspeito	Tempo(s)	Pitch	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
3	**3,736961**	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******	*******
4	2,686667	***	***	**	NS	*	NS	NS	NS	*	**	**	**	***	***	***	***	***	**	**
6	3,795986	***	***	***	***	***	***	***	***	***	**	**	***	***	***	**	*	NS	NS	NS
10	4,64254	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	*	*	**	**		
12	4,265351	***	***	NS	NS	NS	NS	NS	*	NS	**	*	NS	NS	**	NS	*	*	**	*
16	3,582653	***	***	NS	NS	NS	**	**	***	***	***	**	**	**	*	*	NS	*	NS	NS
17	3,86	***	*	***	***	***	*	**	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS

18	3,54	NS	*	***	NS	NS	***	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
19	3,22	NS	NS	***	***	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
20	5,097347	***	***	**	NS	**	NS	NS	NS	NS	NS	NS	NS	*	*	NS	NS	*	NS	NS
21	2,857347	***	***	***	NS	NS	NS	*	NS	*	**	**	**	NS	NS	NS	NS	NS	NS	NS
22	2,644014	***	***	***	***	**	**	NS	NS	NS	NS	NS	NS	NS	NS	*	**	*	**	*
23	4,03068	**	***	*	NS	NS	NS	NS	**	**	**	*	*	**	*	NS	*	NS	NS	NS
24	3,412018	***	***	NS	**	**	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
25	3,113333	NS	NS	NS	NS	NS	NS	NS	NS	NS	**	***	**	**	**	**	*	NS	NS	NS
26	2,835986	***	*	**	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS

Fonte: Elaborado pelo autor.

Na Tabela 5 apresenta-se um novo cenário, no qual o objetivo era investigar a correlação do áudio de referência da frase 1 com ele próprio. Era esperado que todos os formantes fossem identificados positivamente, demonstrando então que os áudios eram idênticos. Conforme mostrado na Tabela 5, os valores de p são 0 para todos os formantes. Isso demonstra claramente que os dois áudios comparados são iguais, sinalizando que o modelo proposto também pode detectar com precisão neste cenário.

Tabela 5 - Resultados da correlação entre o áudio de referência, para a frase 1, consigo mesmo.

	Tempo(s)	**Pitch**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**	**F16**	**F17**	**F18**
Valor - p		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	3,736961	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***

Fonte: Elaborado pelo autor.

Posteriormente, verifica-se o efeito da qualidade dos áudios no método proposto, com os resultados apresentados na Tabela 6. Primeiramente, é importante destacar que 128 kbps é a qualidade padrão. Aqui, este estudo comparou os áudios contestados em qualidades baixas (64 kbps), médias (128 kbps) e altas (256 kbps) com a referência na qualidade padrão (128 kbps). Os testes foram conduzidos para o suspeito # 3. Ao examinar a Tabela 6, fica evidente que os áudios amostrados nas mesmas frequências tendem a obter melhores resultados, tendo 19 formantes extraídos. Por outro lado, apenas 16 formantes foram obtidos ao considerar uma qualidade superior para o áudio contestado (256 kbps). Apesar deste ponto, os resultados positivos verificam o locutor. Pelo contrário, o áudio de baixa qualidade (64 kbps) para o áudio contestado diminuiu o número de formantes para 17. Embora alguns valores de p não significativos tenham sido obtidos nos últimos formantes, o alto número de representações, “***”, garantiu que o falante identificado fosse positivo.

Tabela 6 - Os resultados do confronto considerando as diferentes qualidades dos áudios (Sentença # 1 para o suspeito # 3).

Qualidade	Tempo(s)	Pitch	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
256kbps	3,9	***	***	***	***	***	***	***	***	***	***	***	***	**	**	***	**			
128kbps	3,817347	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
64kbps	3,752185	***	***	***	***	***	***	***	***	***	**	**	**	*	*	NS	NS	NS		

Fonte: Elaborado pelo autor.

Outro teste fundamental consiste em investigar se diferentes tempos de áudio influenciam nos resultados do modelo desenvolvido. Consequentemente, o tempo do áudio de referência foi considerado muito lento, lento e muito rápido. Os resultados são apresentados na Tabela 7. O teste também foi realizado para o suspeito nº 3, considerando o tempo de duração do áudio contestado (3,817347 s) para a frase nº 1.

Tabela 7 - Os resultados da referência, em diferentes tempos, com o áudio contestado com 3,817347s (Sentença nº 1 para o suspeito nº 3).

Velocidade	Tempo(s)	Pitch	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
Muito Lento	12,414671	***	***	***	***	***	***	***	***	***	***	***	***	***	***	**	**	**	**	
Lento	5,417347	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	
Muito Rápido	2,708005	***	NS	**	***	***	***	***	***	***	***	***	***	***	***	***	***	**	**	

Fonte: Elaborado pelo autor.

Na Tabela 7, é evidente que os valores de p reduziram significativamente. Em primeiro lugar, deve-se notar que o formante F18 não foi reconhecido. Em segundo lugar, se o tempo de áudio for reproduzido em câmera lenta, o nível de significância diminui para os últimos formantes. Porém, o resultado ainda é positivo para verificação do locutor. Por outro lado, quando o áudio possui tempo curto, ou é acelerado, o nível de significância dos formantes oscila, onde até F1 obtém NS. No entanto, os resultados ainda confirmam positivamente o locutor.

Por fim, foram examinados os diferentes níveis de ruído do áudio contestado, seguido do confronto estatístico. O teste também foi realizado para o suspeito nº 3, considerando ruído marrom, rosa e branco, para a frase nº 1. Isso representa 50%, 5% e 1% do espectro. Esses resultados são apresentados na Tabela 8. Em primeiro lugar, examinando a Tabela 8, é notável que após a inserção do ruído, o número máximo de formantes reduziu-se para 10. Apesar disso, os resultados puderam verificar positivamente a identidade do falante, visto que todos

os outros valores de p obtiveram "***". No geral, todos os sete testes experimentais atingiram 100% de precisão, demonstrando que o método é adequado para verificar o falante em um contexto forense.

Tabela 8 - Resultado do confronto de áudios sofridos com inserção de diferentes níveis de ruído (Sentença nº 1 para suspeito nº 3).

Barulho	**Tempo(s)**	**Pitch**	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**	**F13**	**F14**	**F15**	**F16**	**F17**	**F18**
Brown50	3,817347	***	***	***	***	***	***	***	***	***	***									
Pink5	3,817347	***	***	***	***	***	***	***	***	***	***									
White1	3,817347	***	***	***	***	***	***	***	***	***	***									

Fonte: Elaborado pelo autor.

Para avaliar cenários reais e validar melhor o método, outros áudios contendo várias condições (áudios de referência) foram testados. Esses registros foram obtidos após mais de 30 dias do áudio contestado, todos para a frase nº 1. Na Tabela 9 mostram-se os resultados obtidos. Primeiro, foi realizada uma interceptação telefônica. Assim, a ligação foi realizada na rua considerando um cenário ruidoso (carros, vento e vozes das pessoas). Devido à limitação do canal telefônico, apenas 5 formantes foram possíveis de serem obtidos. Após a aplicação do método proposto, obteve-se boa significância em todos os formantes (resultado positivo), validando a hipótese de que o áudio confrontado pertence àquele suspeito. Em outro dia, uma mensagem do WhatsApp foi gravada (em casa). Nesse cenário, as crianças brincavam e os ruídos de fundo da TV estavam presentes. Mais uma vez, resulta em uma correspondência positiva. Não obstante, uma nova gravação de áudio foi realizada em um restaurante durante uma festa de aniversário (gravador Sony Icd-PX470). Da análise, foram obtidos 18 formantes.

Conforme mostrado na Tabela 9, o método proposto também pode detectar com alta precisão neste cenário. Posteriormente, foi realizado outro registro em uma sala com o microfone do computador. Também havia ruído de fundo causado pelo ar condicionado, mouse e teclado. Conforme anteriormente obtidos, os resultados garantem que o locutor seja identificado positivamente. É importante ressaltar que o número de formantes é intrinsecamente limitado ao tipo de gravador de áudio utilizado. Além disso, os resultados demonstraram que o método é adequado para verificar o locutor em diversos cenários de ruído e para diferentes dispositivos de gravação.

Tabela 9 - Resultados do confronto entre os suspeitos (Sentença nº 1) e, referência (Sentença nº 1) registrados após 30 dias, considerando ruídos e diversos aparelhos gravadores.

Condições	Tempo(s)	Pitch	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
Celular_Rua	3,646	***	***	***	NS	***														
WhatsAppCasa	3,284	***	***	***	*	**	***	**	***	***	***	***	***	**	***	**	**	**		
Sony_Restaurante	3,903	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
PC_Escritório	4,139	***	***	***	***	***	***	**	***	***	***	***	**	**	**	**	**	**	**	

Fonte: Elaborado pelo autor.

Por fim, o método proposto foi avaliado preliminarmente utilizando o conjunto de dados LibriSpeech (PANAYOTOV *et al.*, 2015). Este conjunto de dados é de domínio público e contém 1983 falantes, todos em inglês. Eles conduziram todas as gravações durante meses, com sotaques diferentes, misturando feminino e masculino, etc. Em primeiro lugar, o arquivo com o nome 9023-296467 foi considerado como áudio questionado. Dessa forma, 20 áudios diferentes foram considerados como referências (Tabela 10). Com base na metodologia proposta, este estudo observou que os áudios nomeados como 9023-296468 e 7177-258977 são os melhores candidatos a palestrante. O locutor 9023-296468 obteve um valor de significância superior de `***` e apenas um formante NS, sendo o falante correspondente para o teste (verificação positiva). É importante destacar que os áudios para 9023-296468 e 9023-296467 foram gravados pela mesma caixa de som em um horário diferente.

Para obter mais detalhes sobre a formação do conjunto de dados, os leitores podem explorar a seguinte referência (PANAYOTOV *et al.*, 2015). No entanto, os leitores devem prestar atenção ao melhor valor p (F0) que foi obtido para o falante nº 9 (10^{-7}), que pode ser considerado um valor baixo em uma condição sem ruído. Comparando-o com a Tabela 2, em que 10^{-22} foi obtido para o melhor valor p, o inglês pode ser considerado uma língua não adaptativa para o método proposto, uma vez que se inserir ruído, todos os valores p aumentam para um valor não significativo em vários formantes. O motivo pode ser explicado desde a história da língua portuguesa que surgiu no século XIII, passando por uma evolução latina. A língua portuguesa brasileira possui um vocabulário complexo, possuindo um sistema fonético simétrico e equilibrado com as notas finais mais claras do que no português europeu (TEYSSIER, 1982). Por outro lado, a história da língua inglesa começa no ano 499 D.C. Nesse sentido, é muito mais antigo que a língua portuguesa e, desde então, teve influências dos anglo-saxões e posteriormente dos franceses. Nesse contexto, o sistema de som vocálico mudou substancialmente a partir da correlação entre ortografia e pronúncia, distanciando-se

de outras línguas da Europa Ocidental (SHUTZ, 2013). Portanto, os resultados aqui alcançados são legítimos para o português brasileiro e reforçam fortemente a potencialidade deste trabalho, uma vez que o conjunto de dados de várias vozes está em inglês, o que poderia fornecer distanciamento do português brasileiro. Da mesma forma, diversos métodos encontrados na literatura, para identificação de falantes, podem não funcionar com precisão para o português brasileiro.

Tabela 10 - Resultados obtidos após a aplicação do método proposto ao conjunto de dados LibriSpeech.

Teste #	Nome do Áudio	Tempo(s)	Valor – p para F0	Pitch	F1	F2	F3	F4
1	9023-296468	11,670	3.90716×10^{-6}	***	NS	***	***	**
2	7859-102518	12,520	0.502099189	NS	**	NS	NS	NS
3	7720-105167	10,160	0.014028753	**	NS	NS	***	**
4	8447-284436	9,980	0.00285753	***	***	**	NS	NS
5	14-212	13,570	0.225905533	NS	NS	NS	NS	NS
6	3906-184005	8,620	0.596373964	NS	NS	*	NS	NS
7	4044-9010	9,040	0.563113699	NS	**	**	*	**
8	6300-39660	11,070	0.049729008	**	NS	NS	NS	NS
9	7177-258977	12,730	1.64337×10^{-7}	***	**	**	NS	NS
10	8262-279161	10,980	0.004817798	***	NS	NS	NS	NS
11	9000-282380	11,250	0.246992694	NS	***	***	**	***
12	152-87733	12,950	0.347284969	NS	NS	NS	*	*
13	218-131205	12,730	0.482158941	NS	NS	NS	NS	NS
14	398-123602	11,920	0.13380562	NS	NS	NS	NS	NS
15	444-138076	11,170	0.66022574	NS	NS	NS	NS	
16	511-131226	11,040	0.033288607	**	***	*	NS	NS
17	639-124526	12,450	0.255417428	NS	NS	NS	NS	NS
18	766-127193	10,980	0.891719784	NS	***	NS	NS	NS
19	850-131003	11,140	0.001656815	***	NS	**	NS	NS
20	949-134657	9,460	0.080615174	*	*	NS	NS	NS

Fonte: Elaborado pelo autor.

4.4 Casos de Uso

Caso 1: Violência doméstica definida pela lei Maria da Penha (lei brasileira de proteção à mulher) onde, recorrentemente, a esposa foi vítima de calúnia e violência psicológica e não aguentava mais viver neste ambiente, exigindo medidas de proteção contra o marido como

também divórcio. Por meio do celular, a mulher conseguiu registrar dois episódios em que o agressor a teria insultado psicologicamente. Sob juramento, o ex-marido nega que a voz contida na unidade de armazenamento seja sua, exigindo, assim, uma verificação de áudio.

Caso 2: Negociação comercial firmada por meio de ligação telefônica, confirmando dívida de K milhões de reais (moeda brasileira) que deveria ser paga em 5 partes iguais (5 x K / 5). Nos autos da ação encontra-se registrado o pagamento de uma única parte, que teria sido a primeira, e o credor requer, na justiça, as outras 4 partes restantes. O suposto devedor, por sua vez, alegou por escrito que não reconhece a dívida exigida e nega ter feito qualquer acordo ou ter confessado qualquer dívida, seja por escrito ou por telefone. Como o arquivo de áudio está em seu poder, o credor incluiu a gravação da ação como prova do acordo e o juiz pediu perícia para comprovar a veracidade dos fatos.

Caso 3: Assinatura de revista mensal de moda por meio de pagamento corporativo com cartão de crédito realizado via ligação telefônica, com dados confirmados por voz, no valor de X reais. A empresa não reconhece a transação reivindicada pelo cliente, mas considerando que possui apenas 21 funcionários, todos do sexo masculino, decidiu verificar as gravações no período de assinatura. Com todos os arquivos de áudio, a empresa solicitou uma análise forense para verificar se um de seus funcionários fez a transação.

Caso 4: Um policial abordou um motociclista que dirigia em alta velocidade e afirmou ter recebido uma oferta de suborno para liberá-lo imediatamente sem fazer o teste do bafômetro. O policial gravou a conversa durante a abordagem e, assim que pediu os documentos do homem, informou ao motociclista que estava sendo gravado. O motociclista negou a autoria do áudio e foi solicitada uma análise forense.

Caso 5: Em uma solicitação de compra via WhatsApp de uma empresa privada, uma troca de mensagens de voz entre uma funcionária da empresa e uma cliente desencadeou um processo por assédio sexual da funcionária contra a cliente. Ao reclamar sobre a situação comprometedora de vexame e constrangimento, o empregado exigiu na Justiça uma indenização por danos morais. O cliente negou veementemente ter enviado esse tipo de conteúdo a funcionária e uma especificação de prova foi exigida.

A - Resultados para o Caso 1

Na Tabela 11 apresentam-se os resultados da aplicação da metodologia proposta para o Caso 1. Ao analisar os resultados, verificou-se que o formante F18 não foi obtido, possivelmente devido ao sufocamento do celular durante a discussão. No entanto, a suspeita

foi considerada positiva pelo fato de obter "***" para vários formantes. O pequeno valor de p resultou em uma significância mais alta. Nesta abordagem, a maior significância foi representada por "***".

Tabela 11 – Os resultados obtidos após a aplicação do método proposto ao Caso 1.

Suspeito	Tempo(s)	Pitch	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
A1	18,3569	***	***	***	**	***	**	***	***	***	**	**	***	***	***	***	**	***	***	

Fonte: Elaborado pelo autor.

B- Resultados para o Caso 2

A Tabela 12 mostra o resultado para o Caso 2. Com base nos resultados obtidos, nota-se que, apesar de a conversa ser longa, as características marcantes da voz comunicante foram validadas com áudio padrão, confirmando que a voz em questão pertence ao suspeito (Caso 2). As conversas telefônicas tendem a ter uma qualidade inferior devido às limitações técnicas do canal (O'SHAUGHNESSY, 1988). Por isso, os resultados baseados na metodologia proposta não apresentaram significância uniforme para todos os formantes.

Tabela 12 - Resultados obtidos após a aplicação do método proposto ao Caso 2.

Suspeito	Tempo(s)	Pitch	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
B1	31,0560	***	**	***	***	NS	**	**	*	**	***	**	**	**	***	***	***	**	*	**

Fonte: Elaborado pelo autor.

C- Resultados para o Caso 3

Na Tabela 13 apresentam-se os resultados obtidos para o Caso 3. É importante mencionar que, ao contrário dos outros casos, que tiveram apenas uma suspeita, aqui foram coletados arquivos de áudio dos 21 funcionários do sexo masculino da empresa. Todos os suspeitos negaram totalmente sobre a autoria dos áudios. Assim, todos foram considerados suspeitos e foi realizada a coleta de áudios para as 21 vozes comunicantes.

Com base nos resultados obtidos, percebe-se que os suspeitos C17 e C20 apresentam maior nível de significância ao considerar todos os formantes. Portanto, eles podem ser classificados como autores dos áudios. Porém, analisando minuciosamente os formantes, percebe-se que o suspeito C20 apresenta uma maior quantidade de p-Valores "***" se comparado ao C17. Além disso, C17 apresentou um "NS". Assim, seria possível descartar C17 e afirmar que o áudio pertence a C20 suspeito. Ao verificar que o *software* entrega resultados objetivos, é possível classificar como caso falso positivo o áudio pertencente a C17, pois, apesar de a significância ser relativamente baixa, não poderia apenas utilizando os

filtros executar a rejeição automática. Considerando 21 testes realizados e um caso de falso positivo, o resultado é uma acurácia de 96% para o método proposto, o que é considerado suficiente para análises forenses (MORRISON, 2011). A suspeita da possível falha metodológica é que os formantes confrontados têm semelhança, mesmo baixa, e a técnica mostra que, apesar da semelhança ser pequena, foi observada, mas não o suficiente para validar que C17 é o suposto autor após a leitura das informações.

Tabela 13 - Resultados obtidos após a aplicação do método proposto ao Caso 3.

Suspeito	Tempo(s)	Pitch	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
C1	3,2182	NS	*	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	
C2	6,7436	*	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS				
C3	3,0043	**	NS	NS	*	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	
C4	3,3206	**	NS	***	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
C5	3,2415	***	**	**	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS			
C6	3,2573	**	***	***	NS	NS	*	***	*	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
C7	4,7832	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	
C8	3,2551	***	NS	*	NS	NS	NS	NS	*	**	**	NS	*	***	*	NS	NS	NS	NS	*
C9	3,4461	***	**	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS					
C10	3,4249	NS	NS	NS	*	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	NS			
C11	3,4270	NS	*	*	NS	***	***	NS	NS	NS	NS	NS	NS	NS	***	***	NS	**	NS	***
C12	3,4268	NS	***	NS	*	**	NS	NS	NS	NS	**	**	*	**	NS	*	NS	*	*	NS
C13	2,4598	NS	NS	*	***	NS	NS	NS	NS	**	*	*	*	NS	NS	NS	NS	NS	NS	NS
C14	3,0185	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
C15	3,3577	***	*	NS	NS	NS	NS	**	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS	NS
C16	3,3351	*	**	NS	***	NS	NS	**	***	**	**	NS	***	**	*	NS	NS	NS	NS	NS
C17	3,3365	**	*	*	NS	**	*	*	*	*	*	**	**	*	*	*	*	*	*	*
C18	4,8113	*	***	NS	***	NS	**	*	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
C19	3,3365	NS	***	***	NS	**	***	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
C20	3,4141	***	***	**	***	***	***	***	**	***	***	***	***	**	***	***	***	***	***	**
C21	3,4102	NS	NS	*	*	NS	NS	NS	NS	NS	NS	*	NS	NS	NS	NS	NS	NS	NS	NS

Fonte: Elaborado pelo autor.

D- Resultados para o Caso 4

Na Tabela 14 apresentam-se os resultados obtidos para o Caso 4. O fato de o gravador estar muito próximo da voz comunicante suspeita, sem sufocamento, gerou ótimos resultados com variáveis excelentes em todos os formantes verificados, o elevado número de "***" garantiu que o orador fosse identificado positivamente.

Tabela 14 - Resultados obtidos após a aplicação do método proposto ao Caso 4.

Suspeito	Tempo(s)	Pitch	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
D1	64,3721	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***

Fonte: Elaborado pelo autor.

E- Resultados do Caso 5

Na Tabela 15 apresentam-se os resultados obtidos com a aplicação da metodologia proposta para o Caso 5. É importante mencionar que o WhatsApp possui um codificador, o que reduz a qualidade dos arquivos de áudio, assim como de imagens e vídeos, com o objetivo de facilitar a troca de informações entre as extremidades (KARPISEK *et al.*, 2015) Os resultados da compressão resultaram significativamente na não obtenção dos formantes F15 a F18. Porém, os demais formantes foram suficientes para atestar a voz positiva do suspeito, conforme observado na Tabela 15.

Tabela 15 - Resultados obtidos após a aplicação do método proposto ao Caso 5.

Suspeito	Tempo(s)	Pitch	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18
E1	5,1623	***	***	**	***	**	***	**	**	**	NS	***	*	**	**	**				

Fonte: Elaborado pelo autor.

A comparação de locutor foi fundamental para a solução de controvérsias, pois foi a prova chave para validar os fatos apresentados por uma das partes. É importante mencionar que em todas as ações desses processos, os diferentes juízes proferiram as sentenças com base no trabalho forense aplicado às análises de áudio, as quais foram preponderantes para a elucidação da verdade. A seguir, é apresentado um breve resumo das conclusões, retiradas das sentenças judiciais, de cada caso aqui analisado.

Caso 1: Após trabalho de reconhecimento da locução, o ex-marido foi condenado por ter praticado violência psicológica e foi punido com pena em que tem medida de proteção para manter-se afastado por pelo menos 500 m de qualquer familiar da ex-mulher. A visita da criança será feita com a presença de assistente social, além de ter sido condenado por calúnia. No entanto, isso foi convertido em pagamento de cesta básica para assistência a instituições filantrópicas e o divórcio ainda está em curso.

Caso 2: Depois de julgado, o proprietário do crédito venceu e o imóvel está programado para ir a leilão para pagar a dívida do devedor atualizado. Isso foi possível devido ao fato do reconhecimento de que a gravação de voz ser uma confissão de dívida. O crime de furto foi

prescrito em razão do tempo excessivamente prolongado para compor a prova pericial e para julgar, não houve prisão, apenas o pagamento da dívida.

Caso 3: O funcionário responsável pela ligação foi demitido da empresa e teve que pagar uma indenização de 3 salários mínimos à revista por falso testemunho. Ele não foi preso, porque o crime prescreveu. Nesse caso específico, apenas para a realização do trabalho forense, desde a chegada do pedido até a conclusão da ação, demorou aproximadamente um ano e meio. Tempo suficiente para o criminoso ficar impune.

Caso 4: O motociclista foi preso e teve sua carteira de motorista retida, no entanto, ele pagou uma fiança e foi libertado. A ação sobre o crime de corrupção foi prescrita devido à negligência no seguimento das vias judiciais. A perícia forense no áudio contribuiu em parte significativa pelo excessivo atraso.

Caso 5: O flerte invasivo foi considerado uma contravenção penal de importunação ofensiva ao pudor, definida pelo artigo 61 da Lei de Contravenções Penais (Lei nº 3688/41). Nesse caso, a pena imposta foi de multa para a ação penal e o mesmo valor para a ação cível por dano moral.

Dessa forma, é importante realizar a perícia técnica em áudio, porém o tempo prescricional da pena precisa ser minuciosamente observado, evitando a sensação de impunidade.

Na Tabela 16 informa-se a correlação entre penalidades quanto às suas prescrições seja de maneira regular ou por idade e a Tabela 17 apresenta a comparação dos tempos para os dois métodos (usual e proposto) destacando os tempos de trabalho forense e de prescrição. É importante mencionar que o método usual aplicado é ouvir, inúmeras vezes, com o ouvido do expert forense, treinado para identificar qualquer ponto da fala que possa ser característico no áudio padrão, para posterior localização e confronto com o áudio disputado, exigindo mais tempo.

Tabela 16 - Correlação entre penalidades quanto às suas prescrições.

Penalidade	Prescrição Regular	Prescrição Idade
0 - 12 meses	3 anos	18 meses
>1 anos < 2 anos	4 anos	2 anos
>2 anos < 4 anos	8 anos	4 anos
>4 anos < 8 anos	12 anos	6 anos
>8 anos < 12 anos	16 anos	8 anos
>12 anos	20 anos	10 anos

Fonte: Elaborado pelo autor.

Tabela 17 - Analogia do tempo total de trabalho e prescrição com os métodos atuais aplicados e testados.

Caso #	Tempo Completo de Trabalho		Prescrição	
	Método Usual	Método Testado	Método Usual	Método Testado
1	3 meses	1 dia	Não	Não
2	4 meses	1 dia	Sim	Não
3	6 meses	3 dias	Sim	Não
4	3 meses	1 dia	Sim	Não
5	3 meses	1 dia	Não	Não

Fonte: Elaborado pelo autor.

É importante mencionar que, para iniciar um trabalho de comparação de locução, independente da metodologia a ser aplicada, o tempo de coleta do material a ser confrontado com o áudio disputado dura cerca de uma hora para cada comunicante. Isso deve-se ao fato de ser necessário gravar várias vezes a locução. Porém, se for considerado toda a metodologia de comparação de locução, ao aplicar a metodologia aqui proposta, e a diferença de tempo é expressiva, pode ser reduzida em até 99% em relação aos métodos usuais. Dessa forma, a fila de espera, que hoje é de até um ano para o próximo expert forense começar a trabalhar, poderia ser drasticamente reduzida. Isso é possível se for considerado que o perito estará ocupado, em média, 1 dia para cada caso com poucas comparações, ou seja, mais tempo para dedicar-se a um novo trabalho, com menos tempo de processamento e filtro de narração.

Considerando, por exemplo, que se o Caso 3 tivesse sido analisado por meio da comparação de locutor proposta neste relatório, a condenação do caso poderia não ter sido prescrita. Mesmo o Caso 3 estava a uma semana de seu prazo final, com a metodologia aqui proposta, o criminoso teria sua pena a cumprir, inibindo essa lacuna de impunidade que existe no Código Penal Brasileiro.

É importante destacar também que a aplicação de técnica forense mais rápida pode levar o suspeito a uma condenação mais precoce, o que é tão importante quanto a preocupação com a prescrição.

Na Tabela 18 mostra-se a comparação da eficiência no resultado assertivo para o Caso 3. Analisando os resultados apresentados na Tabela 18, nota-se que apenas um dos comunicantes para o Caso 3 apresentou falso positivo. Em suma, a eficiência observada foi de 96% para o método testado, ou seja, dos 25 testes, apenas 1 deles apresentou falso positivo. É relevante citar que este resultado pode ser considerado robusto o suficiente para aplicações na área forense (MORRISON, 2011).

Tabela 18: Comparação da eficiência no resultado assertivo para o Caso 3: considerando as metodologias usuais e testadas.

Comunicação de Voz	Método	
	Usual	Testado
Caso #1 – A1	Correto	Correto
Caso #2 – B1	Correto	Correto
Caso #3 – C1	Correto	Correto
Caso #3 – C2	Correto	Correto
Caso #3 – C3	Correto	Correto
Caso #3 – C4	Correto	Correto
Caso #3 – C5	Correto	Correto
Caso #3 – C6	Correto	Correto
Caso #3 – C7	Correto	Correto
Caso #3 – C8	Correto	Correto
Caso #3 – C9	Correto	Correto
Caso #3 – C10	Correto	Correto
Caso #3 – C11	Correto	Correto
Caso #3 – C12	Correto	Correto
Caso #3 – C13	Correto	Correto
Caso #3 – C14	Correto	Correto
Caso #3 – C15	Correto	Correto
Caso #3 – C16	Correto	Correto
Caso #3 – C17	Correto	Errado
Caso #3 – C18	Correto	Correto
Caso #3 – C19	Correto	Correto
Caso #3 – C20	Correto	Correto
Caso #3 – C21	Correto	Correto
Caso #4 – D1	Correto	Correto
Caso #5 – E1	Correto	Correto

Fonte: Elaborado pelo autor.

Analisando especificamente o Caso 3, C17 (falso positivo), o profissional forense poderia facilmente distinguir por meio de audição que o autor do crime era o comunicante C20 em vez de C17. Isso porque, apesar de ambos os áudios serem masculinos, o áudio questionado era mais agudo em comparação com o áudio de referência. Assim, apesar da metodologia não ter identificado claramente o autor do áudio, o especialista forense poderia facilmente identificar o verdadeiro comunicante.

É importante mencionar também que o método testado consiste em um verdadeiro filtro para que o responsável pela comparação tenha um escopo menor de áudios para trabalhar, agilizando assim o resultado. Por exemplo, para o Caso 3, que continha 21 comparações de narrações a serem realizadas, em apenas 3 dias utilizando a metodologia aqui apresentada, este valor foi reduzido para 2 indivíduos possíveis (Suspeitos # 17 e # 20).

É importante observar que a redução do tempo de análise pode significar economia, pois uma ação judicial acarreta custos com juízes, promotores e forenses e, quanto mais longo o curso normal pelos canais judiciais, até a sentença final, maior serão as despesas processuais. Assim, um método capaz de reduzir uma tarefa de meses em dias pode afetar significativamente a redução dos custos processuais.

4.5 Comparação com outras soluções de última geração

Praat é um pacote de *software* de código aberto usado para a análise de parâmetros acústicos, como formantes. Este *software* é usado de forma abrangente em aplicações linguísticas e forenses em todo o mundo. Ele pode realizar a comparação de falantes usando uma comparação gráfica de espectros e valores absolutos. Sua principal desvantagem é a baixa precisão, dificultando o uso em nível forense. Apesar disso, a metodologia proposta para comparação de falantes desenvolvida, baseada na extração de formantes via LPC seguida da aplicação do algoritmo OLS, é inédita na literatura forense.

Para avaliar o método proposto, na Tabela 19 apresenta-se uma comparação entre o método proposto e o Praat em termos de tempo de execução, número de formantes, linhas e formatos suportados. É importante ressaltar que Praat extrai apenas formantes, enquanto o método proposto faz isso e avalia estatisticamente o modelo para fornecer um resultado preciso para especialistas forenses. Examinando a Tabela 19, pode-se verificar que o tempo de execução para extrair os formantes é uma tarefa demorada para o método proposto em comparação com o *software* Praat. Isso pode ser explicado se for analisado o número de formantes extraídos pelo método proposto, que é quase quatro vezes maior. Da mesma forma, o número de linhas é aproximadamente sete vezes maior. Por outro lado, o método proposto é mais robusto, levando a um resultado muito mais confiável. Mais uma vez, é importante destacar que esta comparação de tempo tem apenas o propósito de gerar os formantes e parte dos gráficos na mesma medida que o *software* Praat pode fornecer.

Tabela 19 - Comparação da extração dos formantes para o método proposto e o *software* Praat.

Métodos	**Tempo de Execução (s)**	**Número de Formantes**	**Linhas**	**Formatos Suportados**
Método Presente	27,557 s	19	7,590	TODAS EXTENSÕES DE ÁUDIO
PRAAT	0,1 s	5	1,193	AIFC, AIFF, FLAC, NEXT/SUN, NIST, MP3 e WAV

Fonte: Elaborado pelo autor.

No geral, o tempo total para inserir dados, selecionar as opções desejadas, extrair os formantes, realizar o OLS, comparar os áudios e gerar todos os gráficos pode levar até 112 s, o que é considerado rápido, pois realiza inúmeras tarefas matemáticas e plota aproximadamente 44 gráficos. Enquanto o *software* Praat suporta sete formatos de áudio, o método desenvolvido suporta todos os arquivos de áudio, fornecendo uma vantagem significativa, mesmo quando o áudio contestado é gravado por uma unidade onde precisa ser exportado para um formato de áudio diferente. Neste caso, seria necessário converter o áudio para algum formato compatível com Praat. Isso pode resultar em perda de dados e/ou picos de suavização dos formantes, o que seria estritamente inadmissível em um contexto forense.

5 CONCLUSÃO

Neste trabalho foi apresentado um método baseado em LPC-OLS aplicado à verificação da identidade do locutor em um contexto forense. Até o momento, não há evidências de qualquer outro método que enfoque as mesmas questões apresentadas nesta tese. A principal contribuição desta abordagem consiste em um método confiável e eficiente para extrair formantes, seguido de teste estatístico do modelo (usando OLS) que fornece um resultado preciso para o especialista forense.

Como resultado, o método proposto é capaz de verificar o locutor com sucesso e com maior taxa de acerto ao usar LPC-OLS. É importante destacar que, após a execução de sete testes diferentes, o método rendeu 100% de acerto (que incluiu supressão e troca de palavras, uma frase sem ajustes, diferentes qualidades de áudio, velocidades de fala distintas e níveis de ruído desiguais).

No entanto, é importante destacar que a principal contribuição é para o português brasileiro em que esta abordagem se mostrou bem-sucedida após ter sido validada preliminarmente, com uma base de dados limitada a cinco casos de uso. Além disso, os resultados reforçam fortemente a potencialidade deste trabalho, uma vez que o conjunto de dados de várias vozes está em inglês, o que poderia fornecer distanciamento do português brasileiro. Da mesma forma, diversos métodos encontrados na literatura para verificação de falantes podem não funcionar com precisão para o português brasileiro.

Comparando o método com o *software* Praat, pode-se observar claramente que a abordagem proposta apresenta um maior número de formantes, o que reduz as chances de falsos positivos na comparação de locutor. Além disso, o grande número de linhas utilizadas para extrair os formantes produz uma matriz significativa de robustez incremental para o modelo proposto. Ao contrário do *software* Praat, o método desenvolvido suporta todos os formatos de áudio que podem prevenir a perda de dados ou suavizar os picos do formante durante uma determinada conversão de formato.

Em resumo, os resultados deste estudo mostraram a eficiência, precisão e robustez do método proposto na detecção da identidade do locutor, considerando os estudos encontrados na literatura forense.

Apesar das vantagens, melhorias no método proposto ainda precisam ser investigadas. Assim, trabalhos futuros podem ser concluídos para avaliar a sensibilidade do método desenvolvido para extrair mais formantes, por exemplo, considerar outros níveis de ruído e diferentes qualidades de áudio. Além disso, muito mais pesquisas devem ser conduzidas sobre

a avaliação da comparação de formantes por meio de OLS, considerando o aumento do tamanho do conjunto de dados e a otimização do consumo de tempo. Destaca-se a formação de um banco de dados para o falante do Português Brasileiro contendo áudios gravados após diferentes períodos de tempo, sob várias condições acústicas, para cada grupo étnico, sexo e faixa etária, ruído e vários dispositivos de gravação.

REFERÊNCIAS

ABDEL-HAMID, O.; MOHAMED, A.; JIANG, H.; DENG, L.; PENN, G.; YU, D. Convolutional Neural Networks for Speech Recognition. **IEEE/ACM Trans. Audio Speech Lang. Process**., Piscataway, v. 22, 1533–1545, 2014.

AJILI, M.; BONASTRE, J.F.; KHEDER, W.B.; ROSSATO, S.; KAHN, J. Homogeneity Measure Impact on Target and Non-target Trials in Forensic Voice Comparison. In INTERSPEECH, 2017, Stockholm. **Proceedings** […] [*S. l.: s. n.*], 2017. p. 2844–2848.

ALI, H.; AHMAD, N.; ZHOU, X.; IQBAL, K.; ALI, S.M. DWT features performance analysis for automatic speech recognition of Urdu. **SpringerPlus,** Heidelberg, v. 3, p. 1–10, 2014.

ANDREUCCI, R. A. **Manual de direito penal**. São Paulo: Saraiva, 2018.

APARNA, R.; CHITHRA, P.L. Role of windowing techniques in speech signal processing for enhanced signal cryptography. In: FUJITA, H. et al. (ed). **Advanced Engineering Research and Applications**. London: Springer, 2017. v. 5, cap. 28, p. 446–458.

BAILEY, D.H.; SWARZTRAUBER, P.N. A fast method for the numerical evaluation of continuous Fourier and Laplace transforms. **SIAM J. Sci. Comput**., Philadelphia, v. 15, p. 1105–1110, 1994.

BECKER, T.; JESSEN, M.; GRIGORAS, C. Forensic speaker verification using formant features and Gaussian mixture models. In: ANNUAL CONFERENCE OF THE INTERNATIONAL SPEECH COMMUNICATION ASSOCIATION, INTERSPEECH, 2008, Brisbane. **Proceedings** [...] [*S. l.: s. n.*], 2008. p. 1505–1508.

BEHLAU, M. S. **Uma análise das vogais do português brasileiro falado em São Paulo:** perceptual, espectrográfica de formantes e computadorizada de frequência fundamental. 1984. 123 f. Dissertação (Mestrado) - Escola Paulista de Medicina, Universidade Federal de São Paulo (UNIFESP), São Paulo, 1984.

BOSCHI, J. A. P. **Das penas e seus critérios de aplicação**. Porto Alegre: Livraria do Advogado, 2018.

BRADBURY, J. ***Linear Predictive Coding***. New York: McGraw-Hill, 2000.

BRAID, A.C.M. **Fonética forense.** 2nd ed. Campinas: Millennium, 2003.

BRASIL. **Decreto-Lei 2.848, de 07 de dezembro de 1940**. Código Penal. Diário Oficial da União, Rio de Janeiro, 31 dez. 1940.

BULGAKOVA, E.V.; SHOLOKHOV, A.V. Semi-automatic speaker verification system. **Sci. Tech. J. Inf. Technol. Mech**. Heidelberg, v. 16, p. 284–289, 2016. DOI:10.17586/2226-1494-2016-16-2-284-289.

CHOU, W.; RECCHIONE, M.C.; ZHOU, Q. Automatic Speech/Speaker Recognition over Digital Wireless Channels. U.S. Patent 6,336,090, 1 January 2002.

CHOUGALA, M.; KUNTOJI, S. Novel text independent speaker recognition using LPC based formants. *In*: INTERNATIONAL CONFERENCE ON ELECTRICAL, ELECTRONICS, AND OPTIMIZATION TECHNIQUES (ICEEOT), Chennai. **Proceedings** […] [S. l.: s. n.], 2016. p. 510–513.

CHUNG, J.S.; NAGRANI, A.; ZISSERMAN, A. VoxCeleb2: Deep Speaker Recognition. *In*: INTERSPEECH 2018, Hyderabad. **Proceedings** […] [*S. l.; s. n.*], 2018. p. 1086–1090, DOI: 10.21437/Interspeech.2018-1929.

DEVI, J.S. Language and Text Independent Speaker Recognition System using Artificial Neural Networks and *Fuzzy* Logic. **Int. J. Recent Technol. Eng**., Damkheda, v. 7, p. 327–330, 2019.

DHAKAL, P.; DAMACHARLA, P.; JAVAID, A.Y.; DEVABHAKTUNI, V. A Near Real-Time Automatic Speaker Recognition Architecture for Voice-Based User Interface. **Mach. Learn. Knowl. Extr.,** Basel, v. 1, p. 504-520, 2019.

DRESCH, A.A.G. **Método para reconhecimento de vogais e extração de parâmetros acústicos para analises forenses**. 2015. Master's Thesis, Universidade Tecnológica Federal do Paraná, Apucarana, 2015.

ENZINGER, Ewald, and Geoffrey Stewart Morrison. "Empirical test of the performance of an acoustic-phonetic approach to forensic voice comparison under conditions similar to those of a real case." **Forensic Science International,** Shannon, v. 277, p. 30-40, 2017.

ESCH, T.; VARY, P. Efficient musical noise suppression for speech enhancement system. *In*: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2009, Taipei. **Proceedings** […] [S. l.: s. n.], 2009. p. 4409–4412.

ESSENWANGER, O. M. **Elements of statistical analysis**. General Climatology, London: Elsevier, 1986. 424 p. (World Survey of Climatology, 1B).

FURUI, S. Speaker Recognition in Smart Environments. *In*: **Human-Centric Interfaces for Ambient Intelligence.** Cambridge: Academic, 2010. Cap. 7, p. 163–184.

GOLD, E.; HUGHES, V. Front-end approaches to the issue of correlations in forensic speaker comparison. *In*: INTERNATIONAL CONGRESS OF PHONETIC SCIENCES (ICPhS), 18 th., 2015, Glasgow. **Proceedings** […] [*S. l.: s. n.*], 2015.

GOLDBERGER, A.S. **Econometric theory**: goldberger econometric theory. New York: John Wiley & Sons, 1964.

GOMES, G. M. P. Reconhecimento de locutor, com palavras isoladas, pelo método do alinhamento temporal dinâmico. *In*: SBT, 11., 1993, Natal. **Anais** […][*S. l.: s. n.*], 1993. p. 520-527.

GOODMAN-DELAHUNTY, Jane, and Mandeep K. DHAMI. A forensic examination of court reports. **Australian Psychologist,** Abingdon, v. 48, n. 1, p. 32-40, 2013.

GUJARATI, D.N.; Porter, D.C. **Econometria básica**. 5th ed. Bookman: Porto Alegre, 2008.

GUJARATI, D. **Econometria básica**. 3 ed. São Paulo: Pearson Education do Brasil, 2000.

NIEMANN, H. **Klassifikation von mustern**. 2nd ed. Berlin, New York, Tokyo: Springer, 2003.

HEIJ, C.; DE BOER, P.; FRANSES, P. H.; KLOEK, T.; VAN DIJK, H. K. **Econometric Methods with Applications in Business and Economics**. Oxford: [s. n.], 2004.

IOVAN, C. *et al.*. Moving and calling: mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. **Geographic information science at the heart of Europe,** [*s. l.*], p. 247-265, 2013.

IRUM, A.; SALMAN, A. Speaker verification using deep neural networks: a review. **Int. J. Mach. Learn. Comput**., Singapore, v. 9, p. 20–25, 2019. DOI:10.18178/ijmlc.2019.9.1.760.

MARKEL, J.D.; GRAY, A.H. Jr. **Linear prediction of speech. springer** [*S. l.*]: Science & Business Media, 2013. 290 p.

JUANG, B.H.; FURUI, S. Automatic recognition and understanding of spoken language—A first step toward natural human-machine communication. **Proc. IEEE**, Pistacaway, v. 88, 1142–1165, 2000.

KARPISEK, F.; BAGGILI, I.; BREITINGER. F. WhatsApp network forensics: Decrypting and understanding the WhatsApp call signaling messages. **Digital Investigation** v. 15, p. 110-118, 2015.

KIM, C.; SEO, K.D.; SUNG, W. A robust formant extraction algorithm combining spectral peak picking and root polishing. **Eurasip J. Appl. Signal Process**. [*s. l.*], p. 1-16, 2006. DOI: 10.1155/ASP/2006/67960.

KOVAL, S. Formants matching as a robust method for forensic speaker identification. In: INTERNATIONAL CONFERENCE ON SPEECH AND COMPUTER (SPECOM), 2006, St. Petersburg. **Proceedings** […] [*S. l.: s. n.*], 2006.

KROBBA, A.; DEBYECHE, M.; SELOUANI, S.A. Multitaper chirp group delay Hilbert envelope coefficients for robust speaker verification. **Multimed. Tools Appl**., New York, v. 78, p. 19525, 2019. DOI:10.1007/s11042-019-7154-y.

LADEFOGED, P. **A course in phonetics**. 3. ed.. [*S. l.*]: Harcourt Brace Jovanovich Inc, 1993.

LEE, D. D. Learning the parts of objection by non-negative matrix factorization. **Nature,** London, v. 401, p. 788-791, 1999.

LEME, A.L.M.; MARCELINO, M.A.; PRADO, P.P.L.D. Margens de tolerância e valores de referência para os formantes de vogais orais para uso em terapias de voz para surdos em computador comercial. **CoDAS**, [*s. l.*], v. 28, p. 610–617, 2016,

LEUZZI, F.; TESSITORE, G.; DELFINO, S.; FUSCO, C.; GNEO, M.; ZAMBONINI, G.; FERILLI, S. A statistical approach to speaker identification in forensic phonetics field. *In:* THE INTERNATIONAL WORKSHOP ON NEW FRONTIERS IN MINING COMPLEX PATTERNS, 2016, Riva del Garda. **Proceedings** […] [*S. l.: s. n.*], 2016. DOI: 10.1007/978-3-319-61461-8_5.

LEWIS-BECK, C.; LEWIS-BECK, M. **Applied regression**: an introduction. Thousand Oaks: Sage, 2015. v. 22.

LONG, C. J; DATTA, S. Wavelet based feature extraction for phoneme recognition. *In*: INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE, 4., 1996, [*s. l.*]. **Proceedings** […] [*S. l.: s. n.*], 2016. p. 264-267, v. 1.

LOY, A.; FOLLETT, L.; HOFMANN, H. Variations of Q–Q Plots: the power of our eyes! **Am. Stat**., Alexandria, v. 70, p. 202–214, 2016

MACHADO, Thyago J *et al.*. Forensic speaker verification using ordinary least squares. **Sensors**, Basel, v. 19, n. 20, p. 4385, 2019. DOI:10.3390/s19204385

MAJEWSKI, W.; HOLLIEN, H. Euclidean distance between long-te1m speech spectra as a criterion for speaker identification. **Speech Communication Seminar**, Stockolm, p. l-3, 1974.

MARDEN, J.I. Positions and QQ plots. **Stat. Sci**., Shaker Heights, v. 19, p. 606–614, 2004.

MARKEL, J. D.; DAVIS, S. B. Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. **IEEE Trans**., [*s. l.*], ASSP-27, v. l, p. 74-82, 1979.

MORRISON, G. S. Measuring the validity and reliability of forensic likelihood-ratio systems. **Science & Justice**, London, v. 51, n. 3, p. 91-98, 2011.

GHITZA, O. Auditory models and human performance in tasks related to speech coding and speech recognition. **IEEE Transactions on Speech and Audio Process-ing**, 2., n. 1, p. 115-131, 1994.

OPPENHEIM, A.V.; BUCK, J. R.; SCHAFER, R.W. **Discrete-time signal processing**. 3rd ed. Upper Saddle River: Prentice Hall, 2009.

O'SHAUGHNESSY, D. *Linear Predictive Coding*. **IEEE Potentials,** Pistacaway, v. 7, p. 29–32, 1988.

O'SHAUGHNESSSY, D. Invited paper: Automatic speech recognition: History, methods and challenges. **Pattern Recognition**, Amsterdam, v. 41, n. 10, p. 2965–2979, 2008.

PANAYOTOV, V.; CHEN, G.; POVEY, D.; KHUDANPUR, S.; VASSIL, P. LIBRISPEECH: An ASR corpus based on public domain audio books. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2015, Brisbane. **Proceedings** […] [*S. l.: s. n.*], 2015. p. 5206–5210.

PEGORARO, T.F. **Algoritmos robustos de reconhecimento de voz aplicados a verificação de locutor.** 2000. 101 f. Dissertação (Mestrado), Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, 2000. Available online: http://www.repositorio.unicamp.br/handle/REPOSIP/259689 (accessed on 26 July 2018).

PODDAR, A.; SAHIDULLAH, M.; SAHA, G. Quality measures for speaker verification with short utterances. **Digit. Signal Process,** Waltham, v. 88, p. 66–79, 2019. DOI: 10.1016/j.dsp.2019.01.023.

PRABHU, K.M.M. **Window functions and their applications in signal processing**. Boca Raton: CRC, 2013.

RAHMAN, M. **Applications of fourier transforms to generalized functions**. Southampton: WIT, 2011.

RITUERTO-GONZÁLEZ, E.; MÍNGUEZ-SÁNCHEZ, A.; GALLARDO-ANTOLÍN, A.; PELÁEZ-MORENO, C. Data augmentation for speaker identification under stress conditions to combat gender-based violence. **Appl. Sci**., Bucharest, v. 9, p. 2298, 2019.

RODMAN, R.; MCALLISTER, D.; BITZER, D.; CEPEDA, L.; ABBITT, P. Forensic speaker identification based on spectral moments. **Forensic Linguist**., Sheffield, v. 9, p. 22–43, 2002. DOI:10.1558/sll.2002.9.1.22.

SCHULLER, B. Voice and speech analysis in search of states and traits. *In*: SALAH, A. A.; GEVERS, T. (ed.) **Computer analysis of human behaviour, advances in pattern recognition**. Berlin: Springer, 2011. Cap. 9, p. 227-253.

SCHÜTZ, R. História da Língua Inglesa. English Made in Brazil. [*S. l.: s. n.*], 2013. Available online:: https://www.sk.com.br/sk-historia-da-lingua-inglesa.html. Accessed on: 29 November 2018.

SEELY, J.F.; EL-BASSIOUNI, Y. Applying Wald's variance component test. **Ann. Stat.** Shaker Heights, v. 11, p. 197–201, 1983.

SINGHAL, S. High quality audio coding using multipulse LPC. *In*: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 1990, Albuquerque. **Proceedings** […] [S. l.: s. n.], 1990. p. 1101–1104.

SMITH, H.M.J.; BAGULEY, T.S.; ROBSON, J.; DUNN, A.K.; STACEY, P.C. Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance. **Appl. Cogn. Psychol**., Oxford, v. 33, p. 272–287, 2019. DOI:10.1002/acp.3478.

SNELL, R.C.; MILINAZZO, F. Formant location from LPC analysis data**. IEEE Trans. Speech Audio Process**. Piscataway, v. 1, p. 129–134, 1993.

STIGLER, S. M. **The History of statistics**: the measurement of uncertainty before 1900. [*S. l.*]: Harvard University Press, 1986.

TAFNER, M. A. **Reconhecimento de palavras faladas isoladas utilizando redes neurais artificiais.** 1996. 102 f. Dissertação (Mestrado em Engenharia de Produção), Universidade Federal de Santa Catarina, Florianópolis, 1996.

TEYSSIER, P. 1982. História da língua portuguesa: Lisboa: Sá da Costa Editora.
The Journal of the Acoustical Society of America, Melville, v. 55, n. 6, p. 1304-1312, 1974.

W. H. LAWTON (1971). Self-modeling curve resolution. **Technometrics**, Philadelphia, v.13, n.3, p. 617-622, 1971.

WALD, A. A note on regression analysis. **Ann. Math. Stat**., Shaker Heights, v. 18, p. 586–589, 1947.