
UNIVERSIDADE FEDERAL DE OURO PRETO
DEPARTAMENTO DE CIÊNCIA DE COMPUTAÇÃO
REDES NEURAIS E APRENDIZAGEM EM PROFUNDIDADE - PCC117

Projeto de Pesquisa

COMPARAÇÃO ENTRE RANDOM FOREST E REDES
NEURAIS CONVOLUCIONAIS PARA
IDENTIFICAÇÃO DE FALANTES EM CONDIÇÕES NÃO
CONTROLADAS:
UM ESTUDO COM VoxCELEB1

Aluno:

Pâmela Murta

Professores:

Prof. Eduardo José da Silva Luz

Prof. Vander Luis de Souza Freitas

1 Introdução

A análise forense de conversas digitais tem ganhado relevância crescente em contextos judiciais e investigativos, conforme apontam Morrison (2009). Mensagens de voz provenientes de aplicativos como WhatsApp, Telegram e outras plataformas de comunicação tornaram-se fontes importantes de informação em investigações criminais, demandando identificação precisa dos falantes envolvidos. Nesse cenário, a perícia de áudio enfrenta desafios significativos relacionados à escala, à objetividade e à eficiência dos métodos tradicionais de análise.

Atualmente, a identificação de falantes em gravações de áudio é realizada predominantemente de forma manual por peritos especializados. Esse processo apresenta limitações importantes que comprometem sua aplicabilidade prática. Primeiro, trata-se de procedimento altamente demorado: a análise de conversas extensas pode consumir horas de trabalho especializado. Segundo, o método está sujeito a vieses e inconsistências inerentes à percepção humana, o que afeta a reproduibilidade e a confiabilidade dos resultados. Terceiro, a abordagem manual não é escalável para grandes volumes de dados, cenário cada vez mais comum em investigações que envolvem múltiplos dispositivos e longas janelas temporais. Estudos demonstram que a análise manual de apenas 100 mensagens de áudio pode demandar mais de 8 horas de trabalho pericial Gold & French (2011), evidenciando a necessidade de automação.

Técnicas tradicionais de reconhecimento de falantes, como Gaussian Mixture Models Universal Background Model (GMM-UBM) Reynolds et al. (2000) e iVectors Dehak et al. (2011), foram amplamente aplicadas com sucesso em ambientes controlados. No entanto, essas abordagens demonstram desempenho degradado quando expostas às condições reais que caracterizam comunicações digitais. Áudio comprimido com diferentes codecs, qualidade variável decorrente de condições de gravação não padronizadas, curta duração dos segmentos (tipicamente entre 5 e 30 segundos) e presença de ruído ambiental não controlado constituem desafios que comprometem a eficácia desses métodos clássicos.

Mais recentemente, técnicas de Deep Learning têm demonstrado resultados superiores em tarefas de identificação de falantes. Redes Neurais Convolucionais (CNNs) Abdel-Hamid et al. (2014) e arquiteturas recorrentes como LSTMs Graves & Jaitly (2014) mostraram capacidade notável de aprender representações robustas diretamente de características acústicas, superando métodos tra-

dicionais em diversos benchmarks internacionais. Trabalhos como os de [Snyder et al. \(2018\)](#) e [Variani et al. \(2014\)](#) consolidaram o Deep Learning como estado-da-arte em reconhecimento de falantes, abrindo caminho para novas aplicações práticas.

1.1 Lacuna Identificada na Literatura

Apesar dos avanços recentes, este trabalho identifica uma lacuna importante na literatura: faltam estudos que comparem sistematicamente aprendizado de máquina clássico versus Deep Learning especificamente no regime de poucos falantes, cenário típico de contextos forenses. A maioria dos trabalhos que empregam Deep Learning — como x-vectors e ECAPA-TDNN — requer milhares de falantes para treinamento efetivo, um cenário irrealista para perícias onde tipicamente se analisa conversas entre 3 e 10 suspeitos.

O dataset VoxCeleb1 [Nagrani et al. \(2017\)](#), amplamente utilizado em pesquisas de reconhecimento de falantes, contém áudio de entrevistas capturadas em condições não controladas, incluindo qualidade variável, ruído ambiental e reverberação. Embora não seja constituído especificamente por gravações de aplicativos de comunicação, suas características reproduzem fielmente os desafios encontrados em comunicações digitais reais. Compreender como diferentes arquiteturas de aprendizado de máquina lidam com essas condições torna-se crucial para o desenvolvimento de aplicações forenses práticas e confiáveis.

1.2 Objetivos e Hipóteses de Pesquisa

O objetivo principal deste trabalho consiste em comparar sistematicamente duas abordagens de aprendizado de máquina para identificação de falantes no regime de poucos falantes. As abordagens selecionadas são: (1) Random Forest utilizando features agregadas, representando a baseline clássica de engenharia manual de características; e (2) CNN 1D processando sequências temporais de features, representando a abordagem de Deep Learning com aprendizado de representações.

Para conduzir essa comparação, será utilizado um subset do VoxCeleb1 dataset contendo entre 5 e 10 falantes, simulando cenários forenses típicos onde o número de suspeitos é naturalmente limitado. Essa configuração experimental permite investigar questões fundamentais que orientam a pesquisa. Primeiro: qual a vantagem real de aprender representações automaticamente (CNN)

versus empregar engenharia manual de features (RF) quando se dispõe de poucos falantes? Segundo: features temporais processadas como sequências são superiores a estatísticas agregadas para discriminação de identidade vocal? Terceiro: como cada abordagem lida com a variabilidade de qualidade de áudio característica de cenários reais?

Com base nessas questões, foram formuladas três hipóteses de pesquisa que nortearão a análise experimental:

H1 (Superioridade das CNNs): CNNs superarão Random Forest em pelo menos 10 a 15 pontos percentuais de acurácia no regime de poucos falantes, demonstrando que o aprendizado automático de representações é efetivo mesmo na ausência de milhares de classes para treinamento.

H2 (Vantagem das Sequências Temporais): Features processadas como sequências temporais completas (CNN) serão mais discriminativas que estatísticas agregadas (RF), uma vez que capturam a dinâmica prosódica da fala — ritmo, entonação, padrões de pausas — que são perdidos na agregação estatística.

H3 (Robustez a Variações de Qualidade): CNNs demonstrarão maior robustez a variações de qualidade presentes no VoxCeleb1 (ruído, reverberação, compressão variável), devido à sua capacidade de aprender features invariantes a essas perturbações através do processo de treinamento.

1.3 Contribuições Esperadas

Este trabalho apresenta cinco contribuições principais para a área de identificação automática de falantes:

Primeiro, fornece a primeira análise detalhada comparando Random Forest versus CNNs especificamente no regime de poucos falantes (5 a 10 indivíduos), regime típico de cenários forenses mas subexplorado na literatura de Deep Learning, que tradicionalmente foca em milhares de classes.

Segundo, realiza uma investigação rigorosa de como features hand-crafted (MFCCs agregados por estatísticas descritivas) se comparam com representações aprendidas automaticamente (espectrogramas processados por CNN) em condições de dados limitados, questão central para aplicações práticas.

Terceiro, utiliza um dataset público e amplamente aceito (VoxCeleb1), garantindo que os re-

sultados sejam verificáveis e reproduzíveis por outros pesquisadores. As condições acústicas não controladas do dataset aproximam-se das condições reais encontradas em comunicações digitais, aumentando a validade externa dos achados.

Quarto, disponibilizará um pipeline completo e open-source, implementado de forma modular e bem documentada, abrangendo todas as etapas: pré-processamento de áudio, extração de features, treinamento de modelos, avaliação quantitativa e análise de resultados. Esse material facilitará a replicação e extensão do trabalho por outros grupos de pesquisa.

Quinto, os insights metodológicos e os resultados empíricos serão diretamente aplicáveis ao desenvolvimento de ferramentas automatizadas para análise pericial de áudio. A metodologia proposta é generalizável para diversas fontes de áudio, incluindo comunicações digitais capturadas de aplicativos como WhatsApp, Telegram e chamadas VoIP, ampliando seu impacto prático.

1.4 Organização do Artigo

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta uma revisão da literatura sobre identificação de falantes, cobrindo tanto métodos clássicos quanto abordagens recentes de Deep Learning para processamento de áudio. A Seção 3 descreve detalhadamente a metodologia proposta, incluindo o dataset selecionado, os procedimentos de pré-processamento, a extração de features acústicas e as arquiteturas dos modelos a serem comparados. Por fim, a Seção 5 apresenta as considerações finais e aponta direções futuras para a continuidade desta linha de pesquisa.

2 Trabalhos Relacionados

Esta seção revisa os principais trabalhos relacionados à identificação de falantes, organizados por abordagem metodológica. Primeiro, são apresentados os métodos clássicos que dominaram a área por décadas. Em seguida, discutem-se as contribuições do Deep Learning para processamento de áudio. Por fim, contextualiza-se o uso de datasets públicos e identificam-se as lacunas que este trabalho pretende preencher.

2.1 Métodos Clássicos de Identificação de Falantes

GMM-UBM e i-vectors. Reynolds et al. (2000) propuseram o modelo GMM-UBM (Gaussian Mixture Model - Universal Background Model), que se consolidou como referência em reconhecimento de falantes por mais de uma década. O método modela a distribuição de features acústicas utilizando misturas de Gaussianas. Um modelo universal, treinado com dados de múltiplos falantes, serve como baseline estatística. Modelos específicos de cada falante são então adaptados a partir desse modelo universal, capturando características individuais da voz.

Posteriormente, Dehak et al. (2011) introduziram os i-vectors, revolucionando a área ao representar falantes como vetores de baixa dimensionalidade em um espaço de variabilidade total. Essa abordagem dominou competições internacionais como o NIST Speaker Recognition Evaluation até o advento do Deep Learning. Apesar de seu sucesso em ambientes controlados, os i-vectors apresentam duas limitações significativas. Primeiro, requerem grande quantidade de dados de treinamento para estimar adequadamente o espaço de variabilidade total. Segundo, seu desempenho degrada substancialmente em condições de áudio degradado, caracterizadas por baixa relação sinal-ruído, reverberação ou compressão agressiva.

Random Forest para Classificação de Áudio. Embora menos comum que GMM-UBM em reconhecimento de falantes, Random Forests têm sido aplicados com sucesso em diversas tarefas de classificação de áudio, conforme documentado por Breiman (2001). A principal vantagem dessa abordagem reside em sua robustez natural a overfitting, decorrente do ensemble de árvores de decisão treinadas em subconjuntos aleatórios dos dados. Adicionalmente, Random Forests oferecem interpretabilidade através da análise de importância de features, permitindo identificar quais características acústicas são mais discriminativas. Neste trabalho, Random Forest é utilizado como baseline representativo de métodos não-neurais, fornecendo um ponto de comparação claro com as abordagens de Deep Learning.

2.2 Deep Learning para Identificação de Falantes

Redes Neurais Profundas (DNNs). O trabalho pioneiro de Variani et al. (2014) demonstrou que DNNs podem aprender embeddings discriminativos diretamente de features acústicas, superando i-vectors em datasets desafiadores. A proposta de d-vectors mostrou que representações

aprendidas por redes profundas capturam informações de identidade do falante de forma mais robusta que features hand-crafted. Essa descoberta motivou uma série de trabalhos subsequentes explorando arquiteturas neurais para a tarefa.

X-vectors e TDNNs. [Snyder et al. \(2018\)](#) propuseram x-vectors, uma arquitetura baseada em Time-Delay Neural Networks (TDNNs) que rapidamente se tornou estado-da-arte em reconhecimento de falantes. X-vectors processam segmentos de áudio de comprimento variável através de múltiplas camadas temporais, extraíndo embeddings de dimensão fixa via statistical pooling. Essa abordagem demonstrou excelente capacidade de generalização, inclusive em condições acústicas adversas.

Estendendo o trabalho de x-vectors, [Desplanques et al. \(2020\)](#) desenvolveram o modelo ECAPA-TDNN, incorporando mecanismos de atenção que permitem à rede focar em regiões temporais mais discriminativas do sinal. ECAPA-TDNN obteve resultados competitivos no benchmark VoxCeleb, consolidando-se como uma das arquiteturas mais poderosas disponíveis. No entanto, tanto x-vectors quanto ECAPA-TDNN requerem datasets extensos — milhares de falantes — para treinamento efetivo, o que limita sua aplicabilidade em cenários forenses caracterizados por poucos falantes e dados limitados.

2.3 CNNs para Processamento de Áudio

Convoluçãoes para Reconhecimento de Fala. [Abdel-Hamid et al. \(2014\)](#) foram pioneiros na aplicação de CNNs para reconhecimento de fala, demonstrando que operações convolucionais capturam eficientemente padrões espectrais e temporais presentes em espectrogramas. CNNs 1D, aplicadas diretamente a sequências de features como MFCCs, apresentam a vantagem de menor complexidade computacional comparado a CNNs 2D sobre espectrogramas completos, tornando-as atraentes para aplicações com restrições de recursos.

[Palaz et al. \(2015\)](#) exploraram CNNs 1D para processamento direto do sinal de áudio bruto (end-to-end learning), eliminando completamente a necessidade de engenharia manual de features. Embora promissora do ponto de vista teórico, essa abordagem apresenta dois desafios práticos. Primeiro, requer quantidades muito grandes de dados para convergir adequadamente. Segundo, demanda considerável poder computacional devido ao processamento de sinais de alta taxa de

amostragem, o que pode inviabilizar sua aplicação em cenários com recursos limitados.

2.4 RNNs e LSTMs para Modelagem Temporal

LSTMs para Sequências de Áudio. Graves & Jaitly (2014) demonstraram que Recurrent Neural Networks (RNNs), particularmente LSTMs, são eficazes para modelar dependências temporais de longo prazo em sinais de fala. LSTMs bidirecionais (BiLSTMs) processam informação em ambas as direções temporais, capturando contexto passado e futuro simultaneamente. Essa característica é especialmente relevante para tarefas de processamento de linguagem natural e reconhecimento de fala, onde o contexto bidirecional melhora significativamente a performance.

A arquitetura LSTM original, introduzida por Hochreiter & Schmidhuber (1997), foi projetada especificamente para lidar com o problema de gradientes desvanecentes que afeta RNNs tradicionais. Através de gates (portas) que controlam o fluxo de informação, LSTMs conseguem capturar dependências que se estendem por centenas de timesteps. Essa capacidade é particularmente relevante para características prosódicas da fala, como padrões de entonação e ritmo, que se manifestam em escalas temporais longas.

2.5 Arquiteturas Híbridas CNN-LSTM

Modelos híbridos que combinam CNNs para extração de features locais com LSTMs para modelagem temporal têm obtido resultados promissores em diversas tarefas de processamento de áudio. Sainath et al. Sainath et al. (2015) demonstraram que CNNs seguidas de LSTMs superam arquiteturas isoladas em reconhecimento de fala, especialmente em ambientes ruidosos.

A intuição por trás dessa arquitetura híbrida é complementar. CNNs extraem features invariantes localmente, capturando padrões espectrais característicos independentemente de sua posição temporal exata. LSTMs, por sua vez, capturam a evolução temporal dessas features, modelando prosódia e dinâmica vocal que se manifestam em escalas de tempo mais longas. Para identificação de falantes, essa combinação é particularmente relevante: tanto características espectrais instantâneas (timbre, formantes) quanto padrões temporais (ritmo de fala, entonação) são discriminativos e complementares.

2.6 Datasets Públicos para Speaker Recognition

VoxCeleb. Nagrani et al. (2017) introduziram o VoxCeleb1, um dataset de larga escala contendo mais de 100.000 segmentos de áudio de 1.251 celebridades, extraídos de vídeos do YouTube. VoxCeleb caracteriza-se por condições não controladas que refletem cenários reais: variações significativas de qualidade, ruído ambiental diverso, reverberação, distância variável do microfone e compressão com diferentes codecs. Essas características fazem do VoxCeleb um benchmark desafiador, ideal para avaliar a robustez de sistemas de reconhecimento de falantes.

Diferentemente de datasets controlados como TIMIT, onde as gravações seguem protocolos estritos em ambientes acústicos padronizados, VoxCeleb captura os desafios encontrados em cenários reais de uso. Áudio comprimido, ambientes acústicos diversos — desde estúdios de televisão até espaços ao ar livre — e variações de equipamento de captura tornam o dataset um proxy apropriado para comunicações digitais, embora não utilize especificamente codecs de VoIP como Opus.

VoxCeleb tem sido utilizado como benchmark em centenas de publicações científicas sobre reconhecimento de falantes, permitindo comparações diretas entre diferentes abordagens metodológicas. Sua disponibilidade pública gratuita garante reproduzibilidade dos experimentos — requisito crítico para pesquisa científica rigorosa — e facilita a validação independente de resultados reportados.

2.7 Lacunas Identificadas e Posicionamento deste Trabalho

Embora exista literatura extensa sobre Deep Learning para reconhecimento de falantes, a revisão realizada identificou quatro lacunas importantes que motivam este trabalho.

Primeiro, poucos estudos compararam sistematicamente Random Forest (representante de aprendizado de máquina clássico) versus CNNs no contexto específico de poucos falantes. A maior parte da literatura foca em comparações entre diferentes arquiteturas de Deep Learning (CNN vs LSTM vs Transformers), negligenciando a comparação fundamental com métodos não-neurais em regimes de dados limitados.

Segundo, trabalhos influentes como x-vectors e ECAPA-TDNN requerem milhares de falantes para treinamento adequado. Cenários forenses típicos, no entanto, envolvem entre 3 e 10 suspeitos — um regime de dados drasticamente diferente onde não está claro se Deep Learning mantém suas

vantagens.

Terceiro, falta análise detalhada comparando engenharia manual de features (MFCCs agregados por estatísticas descritivas) versus aprendizado de representações end-to-end (espectrogramas processados por CNN). Essa comparação é fundamental para entender quando a complexidade adicional do Deep Learning se justifica na prática.

Quarto, embora VoxCeleb1 seja amplamente utilizado, poucos estudos analisam explicitamente como diferentes métodos — clássicos versus Deep Learning — lidam com a variabilidade de qualidade inerente ao dataset. Essa análise é crucial para aplicações forenses, onde a qualidade do áudio raramente é controlada.

Este trabalho preenche essas lacunas ao fornecer uma comparação sistemática e rigorosa entre Random Forest e CNN 1D em um subset do VoxCeleb1, focando explicitamente no regime de poucos falantes relevante para aplicações forenses. A utilização do VoxCeleb1 garante tanto reproduzibilidade quanto realismo: suas condições acústicas variáveis, embora não sejam WhatsApp ou Opus especificamente, capturam desafios similares aos encontrados em comunicações digitais reais não controladas.

3 Metodologia

Esta seção descreve detalhadamente a metodologia experimental proposta. Primeiro, apresenta-se o dataset selecionado e sua justificativa. Em seguida, descrevem-se os procedimentos de pré-processamento de áudio. Posteriormente, detalha-se a extração de features acústicas. Por fim, especificam-se as arquiteturas dos modelos a serem comparados, a configuração de treinamento e as métricas de avaliação.

3.1 Dataset e Seleção de Falantes

Fonte de Dados. O dataset selecionado para este estudo é o VoxCeleb1 [Nagrani et al. \(2017\)](#), um dataset público amplamente utilizado pela comunidade científica de reconhecimento de falantes. VoxCeleb1 contém mais de 100.000 vídeos do YouTube de 1.251 celebridades, com áudio extraído de entrevistas, aparições em programas de televisão e eventos públicos.

Para simular cenários forenses realistas — onde o número de suspeitos é naturalmente limitado — será selecionado um subset contendo entre 5 e 10 falantes. A seleção priorizará falantes com maior número de amostras disponíveis no dataset, garantindo quantidade suficiente para treinamento, validação e teste. Essa abordagem é consistente com aplicações práticas de perícia criminal, onde tipicamente se analisa conversas entre poucos indivíduos identificados previamente pela investigação.

Características do Dataset:

- **Fonte:** VoxCeleb1 development set
- **Formato Original:** Áudio extraído de vídeos (m4a/aac)
- **Taxa de Amostragem:** Normalizada para 16 kHz
- **Duração dos Segmentos:** Variável (5 a 30 segundos), ajustada via padding ou truncamento
- **Condições Acústicas:** Não controladas — ambientes de entrevistas, estúdios de televisão, eventos públicos (ruído de fundo variável, qualidade não uniforme)
- **Falantes Selecionados:** 5 a 10 falantes com ≥ 100 amostras cada
- **Total de Amostras:** 500 a 1000 amostras (dependendo do número final de falantes)

Justificativa para Escolha do VoxCeleb1. A seleção deste dataset fundamenta-se em quatro fatores principais. Primeiro, por ser público e amplamente disponível, VoxCeleb1 permite que outros pesquisadores repliquem os experimentos e validem independentemente os resultados obtidos. Segundo, possibilita comparação direta com baselines reportados na literatura científica, contextualizando a performance dos métodos propostos. Terceiro, suas condições acústicas variáveis — qualidade não uniforme, presença de ruído, reverberação, compressão com diferentes codecs — reproduzem fielmente os desafios encontrados em comunicações digitais reais. Quarto, por utilizar dados de domínio público (celebridades em vídeos do YouTube), evita questões éticas relacionadas à privacidade de conversas particulares que surgiriam ao utilizar gravações reais de aplicativos de mensagens.

Analogia com Comunicações de WhatsApp. Embora VoxCeleb1 não utilize especificamente o codec Opus empregado pelo WhatsApp, as condições de áudio presentes no dataset — qualidade

variável, compressão, ruído ambiental — são análogas aos desafios enfrentados em comunicações digitais reais. Trabalhos futuros poderão aplicar compressão Opus artificialmente ao VoxCeleb1 para estudar especificamente o impacto desse codec, mas essa análise está fora do escopo inicial deste estudo.

Divisão dos Dados. O subset selecionado será dividido de forma estratificada em três conjuntos:

- **Treinamento:** 70% das amostras de cada falante
- **Validação:** 15% das amostras de cada falante
- **Teste:** 15% das amostras de cada falante

A divisão estratificada garante representação proporcional de todos os falantes em cada conjunto, essencial para avaliação imparcial dos modelos. Essa abordagem previne viés onde um falante específico dominaria um split, o que comprometeria a capacidade de generalização avaliada no conjunto de teste.

3.2 Pré-processamento de Áudio

Antes da extração de features, todos os áudios passarão por uma pipeline padronizada de pré-processamento, garantindo consistência e qualidade dos dados de entrada.

Conversão de Formato. Os áudios originais serão convertidos para formato WAV com taxa de amostragem de 16 kHz e canal mono. A conversão será realizada utilizando a biblioteca FFmpeg, preservando ao máximo a qualidade original e garantindo compatibilidade com as ferramentas de processamento subsequentes.

Voice Activity Detection (VAD). Será aplicada detecção automática de atividade vocal para remover segmentos de silêncio, focando os modelos apenas em regiões com presença efetiva de fala. A função `librosa.effects.split()` será utilizada com threshold de energia de 20 dB, valor empiricamente adequado para separar fala de silêncio em condições variadas de ruído.

Normalização de Amplitude. Cada áudio será normalizado individualmente para amplitude máxima de 1.0, garantindo consistência no range dinâmico entre diferentes gravações. Essa nor-

malização compensa variações de volume de gravação que não são características do falante, mas sim do equipamento ou distância do microfone.

3.3 Extração de Features Acústicas

Três tipos complementares de features acústicas serão extraídos de cada segmento de áudio, capturando diferentes aspectos da voz humana relevantes para identificação de falantes.

1. MFCCs (Mel-Frequency Cepstral Coefficients). MFCCs são amplamente utilizados em reconhecimento de fala e falantes por capturarem características espetrais relevantes para a percepção auditiva humana. A escala mel aproxima a resposta não-linear do sistema auditivo humano às frequências. Os seguintes parâmetros serão utilizados:

- **n_mfcc:** 40 coeficientes
- **n_fft:** 2048 (janela de FFT)
- **hop_length:** 512 (deslocamento entre janelas)
- **window:** Hann (janelamento)

Adicionalmente, serão calculadas as derivadas primeira (Δ -MFCCs) e segunda ($\Delta\Delta$ -MFCCs) dos MFCCs, capturando a dinâmica temporal de como as características espetrais evoluem. Essas derivadas são importantes porque a fala não é estática: a transição entre fonemas carrega informação discriminativa complementar aos valores instantâneos.

2. Características de Pitch (F0). O pitch fundamental (F0) é altamente discriminativo entre falantes, correlacionando-se diretamente com características fisiológicas das cordas vocais. As seguintes estatísticas de pitch serão extraídas de cada segmento:

- Média de F0
- Desvio padrão de F0
- Valores mínimo e máximo de F0

O algoritmo pYIN [Mauch & Dixon \(2014\)](#), disponível via Librosa, será utilizado para estimação robusta de pitch, configurado para range de 80 a 400 Hz, adequado para abranger vozes masculinas e femininas.

3. Features Espectrais Complementares. Para capturar aspectos adicionais do espectro de frequências, três features espetrais serão extraídas:

- **Spectral Centroid:** Centro de massa do espectro, relacionado ao brilho percebido da voz
- **Spectral Rolloff:** Frequência abaixo da qual está concentrada 85% da energia espectral
- **Zero Crossing Rate:** Taxa de cruzamentos por zero, relacionada à presença de componentes de alta frequência

Representação Final das Features. Para os modelos de Deep Learning (CNN), as features serão mantidas como sequências temporais completas com shape (T, F) , onde T representa o número de frames temporais e F o número de features. Para garantir consistência de dimensionalidade entre amostras de diferentes durações, será aplicado padding (preenchimento com zeros) ou truncagem para comprimento fixo $T_{max} = 100$ frames.

Para o modelo baseline Random Forest, as sequências temporais serão agregadas calculando estatísticas descritivas — média, desvio padrão, mínimo, máximo — de cada feature ao longo do tempo, resultando em vetores de dimensão fixa. Essa agregação é necessária porque Random Forest opera sobre vetores de features fixas, não sobre sequências.

3.4 Arquiteturas de Modelos

Serão comparadas duas abordagens fundamentalmente diferentes de aprendizado de máquina para identificação de falantes, representando paradigmas distintos de processamento de informação.

3.4.1 Baseline: Random Forest com Features Agregadas

Random Forest representa a abordagem clássica de engenharia manual de features combinada com ensemble de árvores de decisão. Nessa abordagem, as features temporais são agregadas usando estatísticas descritivas, transformando sequências variáveis em vetores de dimensão fixa.

Processo de Engenharia de Features. Para cada feature temporal extraída (MFCCs, pitch, features espetrais), serão calculadas quatro estatísticas descritivas:

- Média (μ) — valor central da distribuição
- Desvio padrão (σ) — variabilidade temporal
- Mínimo — valor extremo inferior
- Máximo — valor extremo superior

Considerando 40 MFCCs, 4 estatísticas de pitch e 3 features espetrais, a dimensionalidade final do vetor de features será: $(40 + 4 + 3) \times 4 = 188$ features agregadas por amostra.

Configuração do Modelo Random Forest:

- **n_estimators:** 150 árvores de decisão
- **max_depth:** 20 (profundidade máxima de cada árvore)
- **min_samples_split:** 2
- **criterion:** Gini impurity
- **Input:** Vetor de 188 features agregadas

Justificativa Metodológica. Random Forests foram selecionados como baseline por três razões principais. Primeiro, são naturalmente robustos a overfitting devido ao ensemble de árvores treinadas em subconjuntos aleatórios dos dados. Segundo, requerem pouca engenharia de hiperparâmetros comparado a métodos neurais, funcionando bem com configurações padrão. Terceiro, fornecem interpretabilidade através da análise de importância de features, permitindo identificar quais características acústicas são mais discriminativas. Essa abordagem foi amplamente utilizada em problemas de classificação de áudio antes da popularização do Deep Learning.

3.4.2 Deep Learning: CNN 1D com Features Sequenciais

CNNs 1D processam diretamente sequências temporais de features, aprendendo representações discriminativas automaticamente através de múltiplas camadas convolucionais hierárquicas. Dife-

rentemente de Random Forest, CNNs operam sobre sequências completas, preservando a dinâmica temporal.

Arquitetura Proposta:

A arquitetura será composta por três blocos convolucionais seguidos de um classificador denso. Cada bloco extrai features progressivamente mais abstratas.

- **Camada de Entrada:** (T_{max}, F) — sequência temporal de features

- $T_{max} = 100$ frames temporais (padded/truncated)
 - $F = 40$ (apenas MFCCs) ou $F = 47$ (MFCCs + pitch + spectral completo)

- **Bloco Convolucional 1:**

- Conv1D(64 filtros, kernel=3, padding='same')
 - BatchNormalization()
 - Activation: ReLU
 - MaxPooling1D(pool_size=2)
 - Dropout(0.3)

- **Bloco Convolucional 2:**

- Conv1D(128 filtros, kernel=3, padding='same')
 - BatchNormalization()
 - Activation: ReLU
 - MaxPooling1D(pool_size=2)
 - Dropout(0.3)

- **Bloco Convolucional 3:**

- Conv1D(256 filtros, kernel=3, padding='same')
 - BatchNormalization()
 - Activation: ReLU

- GlobalAveragePooling1D()
- **Classificador Denso:**
 - Dense(128 neurônios, activation='relu')
 - Dropout(0.5)
 - Dense(num_speakers, activation='softmax')

Total de Parâmetros Treináveis: Aproximadamente 180.000 parâmetros.

Fundamentação Teórica da Arquitetura. Cada componente da arquitetura CNN foi selecionado por razões específicas fundamentadas na literatura de Deep Learning:

As **Convoluçãoes 1D** extraem padrões locais invariantes ao longo da dimensão temporal, capturando características espectrais discriminativas independentemente de sua posição exata na sequência. Isso permite que a rede reconheça padrões como formantes ou transições fonéticas mesmo quando ocorrem em momentos diferentes da fala.

BatchNormalization estabiliza o processo de treinamento ao normalizar as ativações entre camadas, acelera a convergência permitindo taxas de aprendizado maiores, e age como regularizador implícito reduzindo overfitting.

MaxPooling serve três propósitos: reduz a dimensionalidade temporal, tornando o processamento mais eficiente; introduz invariância a pequenas translações temporais; e força a rede a aprender representações mais robustas.

GlobalAveragePooling substitui o tradicional Flatten + Dense, oferecendo duas vantagens: permite processar inputs de comprimento variável (flexibilidade) e reduz drasticamente overfitting ao eliminar parâmetros treináveis nessa camada.

Dropout constitui regularização explícita, desativando aleatoriamente neurônios durante o treinamento para prevenir overfitting. Isso é crucial dado que o dataset é relativamente pequeno comparado aos padrões de Deep Learning.

Múltiplas Camadas criam uma hierarquia de features: camadas iniciais detectam padrões simples (bordas espectrais, transições abruptas), enquanto camadas profundas combinam esses padrões em representações de alto nível (características vocais complexas específicas de cada falante).

Comparação Conceitual entre Abordagens:

Tabela 1: Random Forest vs CNN 1D – Diferenças Fundamentais

Aspecto	Random Forest	CNN 1D
Features	Agregadas (hand-crafted)	Sequenciais (aprendidas)
Informação Temporal	Estatísticas (μ, σ)	Dinâmica completa
Representação	Fixa (188 dims)	Hierárquica (64→128→256)
Interpretabilidade	Alta (importância)	Baixa (black-box)
Generalização	Depende de features	Aprende features
Dados Necessários	Menos (~20/falante)	Mais (~50/falante)

3.5 Configuração de Treinamento

Todos os modelos de Deep Learning serão treinados seguindo a mesma configuração padronizada, garantindo comparabilidade dos resultados:

- **Otimizador:** Adam com learning rate inicial de 0.001
- **Função de Perda:** Categorical Cross-Entropy
- **Batch Size:** 32 amostras
- **Épocas:** Máximo de 100, com Early Stopping (paciente de 15 épocas sem melhoria)
- **Callbacks Utilizados:**
 - ReduceLROnPlateau (reduz learning rate em fator 0.5 após 5 épocas sem melhoria)
 - ModelCheckpoint (salva automaticamente o melhor modelo por validation loss)
- **Regularização:** Dropout nas camadas especificadas, L2 weight decay = 0.0001

Infraestrutura Computacional. O treinamento será realizado utilizando Google Colab com GPU NVIDIA (Tesla T4 ou similar), que fornece recursos computacionais suficientes para as arquiteturas propostas.

3.6 Métricas de Avaliação

Os modelos serão avaliados utilizando métricas padrão para classificação multiclasse, fornecendo visão abrangente da performance:

- **Accuracy (Acurácia):** Proporção de predições corretas sobre o total de amostras
- **Precision, Recall, F1-Score:** Calculados tanto na forma macro (média simples entre classes) quanto weighted (ponderada pelo número de amostras)
- **Confusion Matrix:** Visualização de confusões entre pares de falantes
- **Per-Speaker Accuracy:** Acurácia calculada individualmente para cada falante

Análise de Significância Estatística. Para determinar se diferenças observadas entre modelos são estatisticamente significativas, serão aplicados testes estatísticos apropriados:

- **Teste t pareado:** Se as distribuições de performance forem aproximadamente normais
- **Teste de Wilcoxon:** Para distribuições não-normais (teste não-paramétrico)
- **Intervalos de Confiança:** 95% para todas as métricas reportadas

3.7 Experimentos Planejados

Seis experimentos principais foram planejados para investigar sistematicamente diferentes aspectos da comparação entre Random Forest e CNN.

Experimento 1: Comparação Baseline vs Deep Learning. Treinar e avaliar Random Forest e CNN 1D no mesmo subset do VoxCeleb1, comparando performance com análise estatística rigorosa (teste t pareado ou Wilcoxon, dependendo da normalidade das distribuições).

Experimento 2: Ablation Study – Contribuição de Features. Investigar a contribuição individual de cada tipo de feature para ambos os modelos:

- Apenas MFCCs (40 coeficientes)
- MFCCs + Δ + $\Delta\Delta$ (120 coeficientes)
- MFCCs + Pitch (44 features)
- MFCCs + Pitch + Spectral (47 features – conjunto completo)

Essa análise permite identificar quais features são mais discriminativas e se CNNs conseguem aproveitar melhor features complementares que Random Forest.

Experimento 3: Análise de Robustez a Variabilidade. Investigar performance em subsets do VoxCeleb1 estratificados por qualidade de áudio:

- Alta qualidade: Segmentos com SNR > 20 dB
- Média qualidade: Segmentos com SNR entre 10 e 20 dB
- Baixa qualidade: Segmentos com SNR < 10 dB

A hipótese é que CNNs serão mais robustas devido ao aprendizado automático de features invariantes a ruído.

Experimento 4: Variação de Hiperparâmetros da CNN. Explorar a arquitetura ótima através de grid search limitado:

- Profundidade: 2, 3 ou 4 blocos convolucionais
- Kernel size: 3, 5 ou 7
- Número de filtros: [32, 64, 128] vs [64, 128, 256] vs [128, 256, 512]
- Dropout rate: 0.2, 0.3 ou 0.5

Experimento 5: Análise de Curvas de Aprendizado. Treinar modelos com diferentes quantidades de dados (20, 40, 60, 80, 100 amostras por falante) para entender data efficiency:

- Quantos dados Random Forest precisa para saturar (deixar de melhorar)?
- Quantos dados CNN precisa para superar Random Forest?
- Qual a taxa de melhoria com mais dados para cada abordagem?

Experimento 6 (Opcional): Comparação com Baselines Publicados. Se o tempo permitir, comparar a CNN proposta com resultados reportados em papers que utilizaram VoxCeleb1, fornecendo contexto adicional sobre performance relativa aos métodos estado-da-arte.

Tabela 2: Cronograma de execução do projeto (8 semanas)

Semana	Atividade	Duração
1	Setup do ambiente e análise exploratória	1 semana
2	Extração completa de features	1 semana
3	Baseline Random Forest	1 semana
4-5	CNN 1D (implementação + tuning)	2 semanas
6	Experimentos comparativos	1 semana
7	Análise de resultados	1 semana
8	Escrita final e preparação de apresentação	1 semana

3.8 Cronograma de Execução

A simplificação para dois modelos principais permite dedicar duas semanas completas ao desenvolvimento e refinamento da CNN, garantindo exploração adequada de hiperparâmetros e análise profunda da performance. Esse tempo adicional é essencial para obter resultados rigorosos e confiáveis.

4 Resultados

Este trabalho espera fornecer evidências empíricas sobre a eficácia relativa de Deep Learning (CNNs) comparado a métodos clássicos (Random Forest) para identificação de falantes no regime de poucos falantes característico de aplicações forenses. Com base nas hipóteses formuladas e na revisão da literatura, os seguintes resultados são antecipados.

H1: Superioridade das CNNs. Espera-se que CNNs superem Random Forest em pelo menos 10 a 15 pontos percentuais de acurácia. Essa melhoria virá de três capacidades fundamentais das CNNs: aprendizado automático de features hierárquicas sem necessidade de engenharia manual, processamento de sequências temporais completas preservando dinâmica prosódica, e captura conjunta de padrões espectrais e temporais através de convoluções.

Performance Esperada:

H2: Vantagem das Sequências Temporais. Features processadas como sequências temporais completas (CNN) deverão demonstrar maior poder discriminativo que estatísticas agregadas (Random Forest). Essa vantagem será especialmente pronunciada para falantes com características prosódicas distintas — ritmo de fala, padrões de entonação, distribuição de pausas — que são

Tabela 3: Performance esperada dos modelos no conjunto de teste

Modelo	Accuracy	F1 (macro)	Parâmetros	Tempo/época
Random Forest	~75%	~0.73	–	0.05s
CNN 1D	~88%	~0.86	180K	2.5s

completamente perdidas quando se agregam MFCCs em estatísticas descritivas.

H3: Robustez a Variações de Qualidade. CNNs deverão demonstrar maior robustez quando testadas em diferentes níveis de qualidade de áudio, com degradação de performance menor que Random Forest à medida que a relação sinal-ruído diminui. Essa robustez virá da capacidade das CNNs de aprender automaticamente features invariantes a perturbações através do processo de treinamento em dados diversos.

Análises Adicionais Esperadas. Além das comparações quantitativas principais, este trabalho produzirá análises complementares que fornecerão insights sobre o comportamento dos modelos:

A **Matriz de Confusão** permitirá identificar pares de falantes frequentemente confundidos pelos modelos, possibilitando investigação das características acústicas que tornam esses falantes similares. Espera-se que CNNs apresentem confusões mais localizadas (entre falantes realmente similares), enquanto Random Forest pode apresentar confusões mais distribuídas.

Análise de **Importância de Features** será conduzida para Random Forest, identificando quais features agregadas são mais discriminativas. Para CNN, técnicas de visualização como Grad-CAM serão exploradas para entender quais regiões temporais do sinal a rede considera mais relevantes para a decisão.

Curvas de Aprendizado deverão revelar padrões distintos: CNN requerendo mais dados para convergir mas atingindo performance superior com datasets maiores, enquanto Random Forest converge rapidamente mas satura mais cedo, não se beneficiando tanto de dados adicionais.

Finalmente, análise de **Trade-offs Práticos** documentará que Random Forest é significativamente mais rápido para treinar e inferir, mais interpretável através de importância de features, e requer menos dados para resultados razoáveis. CNN, por outro lado, é mais precisa mas computacionalmente cara e opera como black-box menos interpretável. Esses trade-offs são cruciais para decisões de implementação em sistemas forenses reais.

Contribuição Científica Ampla. Além dos resultados quantitativos diretos, este trabalho con-

tribuirá com análise detalhada metodológica e insights práticos. Primeiro, fornecerá diretrizes baseadas em evidências sobre quando usar ML clássico versus Deep Learning para identificação de falantes em contextos forenses com poucos falantes. Segundo, disponibilizará código open-source reproduzível implementando ambas as abordagens de forma modular e bem documentada. Terceiro, oferecerá insights sobre o que CNNs aprendem automaticamente comparado a features hand-crafted, através de análises de interpretabilidade. Esses elementos aumentarão o impacto do trabalho além dos resultados numéricos específicos.

5 Considerações Finais

Este trabalho propõe uma comparação sistemática e rigorosa entre aprendizado de máquina clássico (Random Forest com features agregadas) e Deep Learning (CNNs 1D processando sequências temporais) para identificação automática de falantes. A pesquisa utiliza o VoxCeleb1 dataset — público, amplamente validado pela comunidade científica — garantindo reproduibilidade dos experimentos e permitindo comparação direta com baselines reportados na literatura.

A escolha por focar em um subset contendo entre 5 e 10 falantes do VoxCeleb1 foi motivada pela necessidade de simular cenários forenses realistas. Diferentemente de trabalhos que treinam modelos com milhares de falantes, este estudo concentra-se no regime de dados limitados característico de aplicações periciais, onde tipicamente se analisa conversas entre poucos suspeitos identificados pela investigação. Essa abordagem é mais relevante para aplicações práticas de perícia do que os cenários de larga escala frequentemente explorados na literatura de reconhecimento de falantes.

A metodologia proposta estabelece uma comparação entre duas filosofias fundamentalmente diferentes de processamento de informação. De um lado, Random Forest representa a tradição de engenharia manual de features combinada com agregações temporais, transformando sequências variáveis em vetores de características fixas. De outro, CNNs exemplificam o paradigma moderno de aprendizado de representações hierárquicas diretamente de sequências temporais completas. Essa comparação permitirá responder questões práticas relevantes: em que circunstâncias a complexidade adicional do Deep Learning se justifica comparada a abordagens clássicas mais simples e interpretáveis?

A simplificação metodológica para dois modelos principais — ao invés de múltiplas arquiteturas neurais competindo simultaneamente — permite análise mais profunda e rigorosa de cada abordagem. Haverá tempo adequado para exploração cuidadosa de hiperparâmetros, ablation studies detalhados investigando contribuição de diferentes features, e interpretação minuciosa dos resultados. Essa escolha prioriza qualidade e profundidade sobre quantidade, alinhando-se com o prazo do projeto e garantindo contribuições científicas sólidas.

As condições acústicas variáveis presentes no VoxCeleb1 — qualidade não uniforme, presença de ruído ambiental, reverberação, compressão com diferentes codecs — capturam fielmente os desafios encontrados em comunicações digitais reais. Embora o dataset não utilize especificamente o codec Opus empregado pelo WhatsApp, suas características são análogas e suficientemente desafiadoras para validar os métodos propostos. Trabalhos futuros poderão aplicar compressão Opus artificialmente ao VoxCeleb1 para estudar especificamente o impacto desse codec, expandindo os achados iniciais.

Como direções futuras, pretende-se explorar cinco extensões principais deste trabalho. Primeiro, após validar o baseline CNN proposto, investigar arquiteturas mais complexas incluindo LSTMs e modelos híbridos CNN-LSTM que capturam dependências temporais de longo prazo. Segundo, explorar transfer learning utilizando modelos pré-treinados no VoxCeleb1 completo, avaliando se conhecimento adquirido com milhares de falantes transfere efetivamente para o regime de poucos falantes. Terceiro, coletar e analisar áudios reais de WhatsApp com codec Opus para validação adicional em condições totalmente realistas. Quarto, aplicar técnicas de explicabilidade (XAI) para aumentar a interpretabilidade das CNNs, crucial para aceitação em contextos forenses onde decisões precisam ser justificadas. Quinto, comparar diretamente com métodos estado-da-arte como x-vectors e ECAPA-TDNN no mesmo subset, contextualizando a performance dos métodos propostos.

Este projeto estabelece a base sólida para uma linha de pesquisa contínua em aplicações de Deep Learning para análise forense de áudio. A ênfase em praticidade — através do foco em poucos falantes —, reproduzibilidade — via uso de dataset público e código open-source — e aplicabilidade real — simulando condições não controladas de comunicações digitais — diferencia este trabalho e maximiza seu impacto potencial tanto científico quanto prático.

Referências

- Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10), 1533–1545. doi: 10.1109/TASLP.2014.2339736
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. doi: 10.1023/A:1010933404324
- Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788–798. doi: 10.1109/TASL.2010.2064307
- Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. interspeech* (pp. 3830–3834). doi: 10.21437/Interspeech.2020-2650
- Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech Language and the Law*, 18(2), 293–307. doi: 10.1558/ijssl.v18i2.293
- Graves, A., & Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *Proc. international conference on machine learning (icml)* (Vol. 32, pp. 1764–1772).
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735
- Mauch, M., & Dixon, S. (2014). pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *Proc. ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 659–663). doi: 10.1109/ICASSP.2014.6853678
- Morrison, G. S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, 49(4), 298–308. ((Citado como 2018 no texto, mas publicado em 2009)) doi: 10.1016/j.scijus.2009.09.002

- Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. In *Proc. interspeech* (pp. 2616–2620). doi: 10.21437/Interspeech.2017-950
- Palaz, D., Magimai-Doss, M., & Collobert, R. (2015). Analysis of CNN-based speech recognition system using raw speech as input. In *Proc. interspeech* (pp. 11–15). doi: 10.21437/Interspeech.2015-3
- Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3), 19–41. doi: 10.1006/dspr.1999.0361
- Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. In *Proc. ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4580–4584). doi: 10.1109/ICASSP.2015.7178838
- Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *Proc. ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5329–5333). doi: 10.1109/ICASSP.2018.8461375
- Variani, E., Lei, X., McDermott, E., Moreno, I. L., & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *Proc. ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4052–4056). doi: 10.1109/ICASSP.2014.6854363