UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MICHEL GOMES DE SOUZA

IDENTIFICAÇÃO E VERIFICAÇÃO DE LOCUTORES EM PORTUGUÊS E INGLÊS UTILIZANDO TRANSFER LEARNING

CAMPO MOURÃO

2022

MICHEL GOMES DE SOUZA

IDENTIFICAÇÃO E VERIFICAÇÃO DE LOCUTORES EM PORTUGUÊS E INGLÊS UTILIZANDO TRANSFER LEARNING

Identification and verification of speakers in portuguese and english using transfer learning

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Juliano Henrique Foleis

Coorientador: Prof. Dr. Diego Bertolini Gonçalves

CAMPO MOURÃO

2022

MICHEL GOMES DE SOUZA

IDENTIFICAÇÃO E VERIFICAÇÃO DE LOCUTORES EM PORTUGUÊS E INGLÊS UTILIZANDO TRANSFER LEARNING

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 13/junho/2022

Rodrigo Hübner
Prof. Dr.
Universidade Tecnológica Federal do Paraná

Rodrigo Campiolo
Prof. Dr.
Universidade Tecnológica Federal do Paraná

CAMPO MOURÃO
2022

AGRADECIMENTOS

Às instituições públicas de ensino.

RESUMO

A fala é uma das modalidades biométricas que pode ser usada para reconhecer um indivíduo. Assim, sistemas de identificação de locutores possuem aplicabilidades em problemas de autenticação, como vigilância automática e atividades forenses. Esse processo de reconhecimento é dividido em identificação e verificação do locutor. A maioria das bases de dados destinadas ao reconhecimento automático de locutores se encontram em idioma estrangeiro, como a *voxCeleb* e *Common Voice*. Por isso, optou-se pela seleção de uma base de dados com falantes brasileiros, a *Brazilian Speech Database*. Este é o primeiro trabalho a utilizar esta base aplicando métodos de identificação e verificação de locutores para avaliar as características extraídas por *transfer learning*. Posteriormente, um *subset* do *Common Voice* foi submetido aos mesmos métodos, de modo a comparar os resultados. O melhor resultado para a tarefa de identificação a base de dados em português foi de 0,70 ± 0,10 com 10 *patches* utilizando o método de *early fusion* com as características do *handcrafted*. Já para o base de dados em inglês foi de 0.68 ± 0.05 com 10 *patches* utilizando o *early fusion* de todos os extratores do método de *transfer learning*. Para o problema de verificação, a *Brazilian Speech Database* ficou com uma taxa de 0.97 ± 0.00 utilizando 10 *patches* com o MobileNet, e o *Common Voice* obteve uma taxa de 0.98 ± 0.00 com 10 *patches* para todos os descritores aplicados. Destacou-se que a complementariedade de características feita com o *early fusion* ajudou a obter resultados melhores em alguns casos. Já o uso de técnicas de extração de características utilizando o *transfer learning*, apesar de serem mais robustas e sofisticadas, apresentaram um resultado estatisticamente igual às técnicas de *handcrafted*. Um fator que pode ter influenciado os experimentos é que o *Brazilian Speech Database* é uma base de dados baseado em dependência de texto, enquanto o *Common Voice* de não dependência de texto.

Palavras-chave: transfer learning; reconhecimento do locutor; processamento de áudio; identificao de locutor; verificao de locutor.

ABSTRACT

Speech is one of the biometric modalities that can be used to recognize an individual. Thus, speaker identification systems have applicability in authentication problems, such as automatic surveillance and forensic activities. This recognition process is divided into speaker identification and verification. Most databases for automatic speaker recognition are in foreign languages, such as voxCeleb and Common Voice. Therefore, it was selected a database with Brazilian speakers, the Brazilian Speech Database. This is the first work to use this base, applying methods of identification and verification of speakers to evaluate the characteristics extracted by transfer learning from this dataset. Subsequently, a Common Voice subset was subjected to the same methods in order to compare the data. The best result for the identification task for the Brazilian database was 0.70 ± 0.10 with 10 patches using the early fusion method with the handcrafted characteristics. As for the English database, it was 0.68 ± 0.05 with 10 patches using early fusion of all extractors of the transfer learning method. For the verification problem, Brazilian Speech Database got a rate of 0.97 ± 0.00 using 10 patches with MobileNet, and Common Voice got a rate of 0.98 ± 0.00 with 10 patches for all descriptors applied. It was highlighted that the complementarity of features made with early fusion helped to obtain better results in some cases. The use of feature extraction techniques applying transfer learning, despite being more robust and sophisticated, presented a result statistically equal to the handcrafted techniques. One factor that may have influenced the experiments is that the Brazilian Speech Database is a text-dependent database, while Common Voice is a non-text-dependent database.

Keywords: transfer learning; speech recognition; audio processing; speaker identification; speaker verification.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O reconhecimento automático do locutor visa reconhecer a identidade de um indivíduo através das características intrínsecas de sua voz, como velocidade, ritmo, tom e dialeto (REYNOLDS, 2001; TAHLIRAMANI; BHATT, 2018; SHAFIK *et al.*, 2021; VACHER *et al.*, 2015; JAIN; ROSS; PANKANTI, 2006).

Há duas tarefas relacionadas ao reconhecimento automático do locutor, a saber: identificação do locutor e verificação do locutor. A identificação visa reconhecer um locutor desconhecido dentre um conjunto de diversos locutores, enquanto a verificação parte de uma suposta identidade do locutor que deve ser comprovada a partir da comparação do registro de voz deste com um conjunto de dados base (TING *et al.*, 2007). Esses processos podem ser dependentes ou independentes de texto, sendo que em um sistema dependente, o discurso de cada locutor é fixo e conhecido, já em um sistema independente é utilizada a fala espontânea sem restrições de sentenças ou frases a serem ditas (CHAUHAN; ISSHIKI; LI, 2019).

A extração de características em áudio é o processo de extrair informações relevantes de um sinal de áudio, sendo um processo importante na análise de áudio, pois permite que os sinais sejam comparados e classificados conforme as suas propriedades. Existem dois métodos principais de extração de características em áudio: o '*handcrafted*' e o '*transfer learning*'. O '*handcrafted*' é um método matemático bem definido, ao qual é passado as informações do áudio a ser extraído. Já o '*transfer learning*', também conhecido como aprendizagem por transferência, é o processo de reutilização de um modelo de aprendizagem de máquina treinado em um problema para um problema diferente. (NANNI; GHIDONI; BRAHNAM, 2017)

A maioria das bases de dados destinadas para a identificação de locutores majoritariamente se encontra em língua estrangeira, como a voxCeleb (NAGRANI *et al.*, 2019), TIMIT (GAROFOLO *et al.*, 1993), Voice Gender[1] e Common Voice[2]. Por esse motivo optou-se por escolher uma base de dados com locutores brasileiros, a *Brazilian Speech Database* (BrSd) (PAULINO *et al.*, 2018). O presente trabalho é o primeiro a usar de métodos de identificação e verificação de locutores nesta base de dados de modo a avaliar o desempenho do classificador *Support Vector Machine* (SVM) a partir de características extraídas por métodos de *transfer learning*.

1.1 Objetivos

O objetivo deste trabalho é avaliar técnicas de classificação baseadas em *transfer learning* para verificação e identificação de locutores em uma base de dados em português e comparar a mesma metodologia em uma base de dados em inglês. Os objetivos específicos são:

[1] https://www.kaggle.com/primaryobjects/voicegender
[2] https://www.kaggle.com/mozillaorg/common-voice

- Avaliar o desempenho de características *handcrafted* como *baseline* nas tarefas de identificação e verificação;
- Avaliar o desempenho de características extraídas por *transfer learning* nas tarefas de identificação e verificação;
- Avaliar o efeito da combinação de conjuntos de características no desempenho nas tarefas de identificação e verificação.

1.2 Estrutura do trabalho

No Capítulo 2, são apresentados alguns dos conceitos necessários para o entendimento do trabalho. No Capítulo 3 estão descritos alguns trabalhos de referência para a tarefa de identificação e verificação de locutores. No Capítulo 4 se encontra o método utilizado para a realização deste trabalho. Os experimentos, resultados e discussões estão dispostos no seção 4.1. Por fim, o Capítulo 6 reúne as conclusões do trabalho.

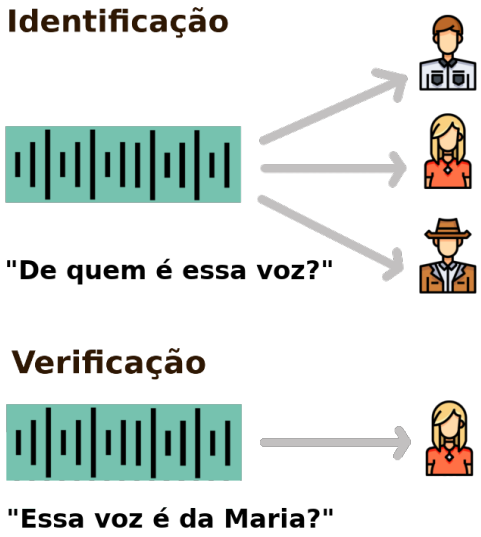
2 REFERENCIAL TEÓRICO

Neste capítulo é feito um breve estudo sobre os problemas de reconhecimento de voz, bem como uma introdução aos conceitos básicos dos descritores de áudios e seus tipos.

2.1 Verificação e identificação

O reconhecimento do locutor é o processo de verificar ou identificar um indivíduo por meio das características intrínsecas de sua voz. No processo de identificação, uma amostra da voz desconhecida é comparada com uma base de dados de locutores previamente conhecidos, visando o seu reconhecimento (REYNOLDS, 1995). O processo de verificação consiste em decidir se uma amostra de voz foi produzida por uma determinada pessoa. A Figura 1 demonstra as tarefas de identificação e verificação.

Figura 1 – Identificação e verificação de locutores.



Fonte: Autoria própria (2022).

Desde meados da década de 70 o reconhecimento de voz vem sendo um desafio para diversos pesquisadores. Velichko e Zagoruyko (1970), por exemplo, demonstraram que conseguiram reconhecer 200 palavras do idioma russo de forma aceitável utilizando ALGOL-60 e características logarítmicas do sinal acústico em 5 bandas diferentes. Ichikawa, Nakano e Nakata (1973), conseguiram resultados interessantes no reconhecimento de dígitos por meio da voz, utilizando os parâmetros de *spectrum envelope*, *cepstrum* e *partial autocorrelation coefficients* (PAC's). Na década de 80 o reconhecimento da voz passou a tentar verificar os indivíduos, Naik e Doddington (1987) utilizaram a verificação de locutor como uma alternativa para o controle de acesso, criando um protocolo simples e eficaz com resultado satisfatório, e ainda de baixo custo e fácil implementação.

2.2 Base de dados

As bases de dados de áudio são importantes para a identificação e verificação de áudio, pois fornecem um conjunto de dados com os quais as podem ser criados sistemas e métodos mais precisos. Uma base de dados de áudio fornece um conjunto de dados mais amplo e variado para ser estudado, tendo assim mais amostras para serem analisadas e um melhor entendimento do áudio. Sendo assim, iremos utilizar duas bases de dados, a *Brazilian Speech Database* (BrSd) e a *Common Voice*.

2.2.1 A Brazilian Speech Database

O BrSD é uma base de dados de falantes nativos do português, desenvolvida para ser possível fazer experimentos para identificar diversas características do indivíduo, como a idade, gênero e a emoção. A proposta de Paulino *et al.* (2018) foi coletar 400 gravações com 80 diferentes contribuidores, todos eles brasileiros e falantes de português, com idades que variam de 9 a 81 anos. Após a coleta, foi realizada uma série de experimentos de classificação utilizando métodos como o *Support Vector Machine* (SVM) e o *Multilayer Perceptron* (MLP). Partindo da análise acústica e visual para determinar as características, foram empregados descritores de textura *Local Binary Pattern* (LBP) e *Local Phase Quantization* (LPQ).

Os arquivos de dados são disponibilizados tanto em formato de áudio quanto em formato de espectrograma, sendo este último a caracterização visual do sinal de áudio. Neste caso, a transformação de áudio para imagem foi realizada pelos autores do conjunto de dados a partir do *software* livre *SoX*[1], sendo um programa de conversão de arquivos de áudio para diversos formatos, limitando a frequência dos espectrogramas a 3,4kHz e o intervalo de amplitudes de -60 dBFS a 0 dBFS (PAULINO *et al.*, 2018).

Ocorrem 3 abordagens diferentes para obter os resultados finais, sendo a classificação individual, a classificação por *early fusion*, e a classificação por *late fusion*, de modo a classificar idade, gênero e o conjunto de idade e gênero.

O melhor resultado para gênero foi de 91,2% de média com um desvio padrão de 2,62% utilizando o método de classificação SVM com *early fusion*. Já a classificação de idade foi de 88,75±5,86 % também utilizando o SVM, mas com o *late fusion* e, por fim, a classificação de idade e gênero juntos foi de 80,25±6,09 % também utilizando o SVM e o *late fusion*.

2.2.2 Common Voice

O principal propósito do *Common Voice*[2] é acelerar a pesquisa nos problemas de identificação de voz. Essa base de dados consiste em uma enorme coleção de falas transcritas

[1] http://sox.sourceforge.net/

[2] https://commonvoice.mozilla.org/

aberta ao público, sendo totalmente independente de texto, ou seja, todas as falas transcritas são únicas e sem repetição, contendo como características a transcrição do texto, idade, gênero, sotaque entre outros atributos. Além disso, o projeto também oferece uma ferramenta de gravação gratuita que pode ser usada para criar conjuntos de dados de voz.

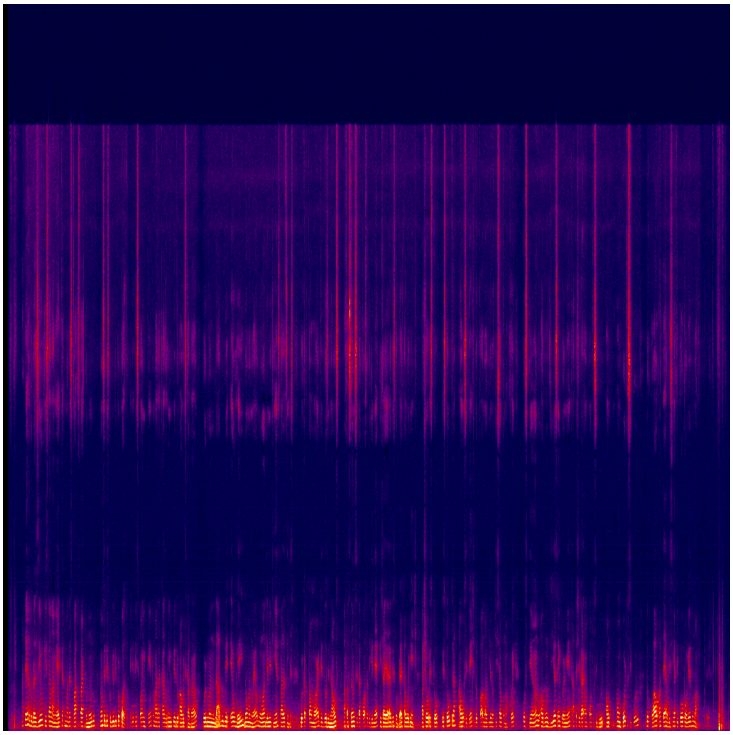
2.3 Descritores de áudio

Descritores de áudio são ferramentas analíticas que ajudam a extrair informações e características de um sinal de áudio (HERRERA; SERRA; PEETERS, 1999). Podem utilizar diversas técnicas, como análise de sinais, transformada de Fourier, espectrogramas, entre outras.

A transformada de Fourier é uma operação matemática usada na representação de sinais em um determinado período, a qual é representada por uma soma ponderada de senoides e cossenoides (BAILEY; SWARZTRAUBER, 1994). Assim, a transformada de Fourier é a representação no domínio da frequência de um sinal no domínio do tempo.

Um problema que ocorre na transformada de Fourier é a ausência das características do tempo. Por isso, se for necessário descrever variações temporais é empregado o uso da *Short-Time Fourier Transform* (STFT) (SHUKLA; TIWARI; KALA, 2010). A *Short Time Fourier Transform* (STFT) é uma técnica usada para analisar sinais de áudio no domínio do tempo e da frequência. Essa técnica usa um processo chamado análise espectral, em que um sinal é dividido em vários pedaços e cada pedaço é analisado de forma independente. Portanto, a STFT pode ser usada para gerar espectrogramas como apresentado na Figura 2.

Figura 2 – Espectrograma de um áudio da BrSD.



Fonte: Autoria própria (2022).

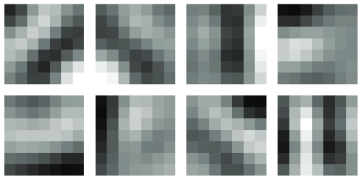
A partir do espectrograma da voz de um dado locutor é realizado o processo de extração de modo a obter os vetores de características. Esses vetores também são chamados descritores de áudio, pois representam as informações mais relevantes do sinal de áudio em um espaço de menor dimensionalidade. Os descritores de áudio utilizados neste trabalho são: *handcrafted* e os baseados em *transfer learning*.

2.3.1 Descritores *handcrafted*

Os descritores *handcrafted* são utilizados para os mais diversos problemas que envolvem áudio, quase sempre utilizando como base as imagens de espectrogramas (COSTA *et al.*, 2011). Neste trabalho, são utilizados os descritores de textura *Binarized Statistical Image Features* (BSIF), *Local Binary Pattern* (LBP) e o *Local Phase Quantization (LPQ)*.

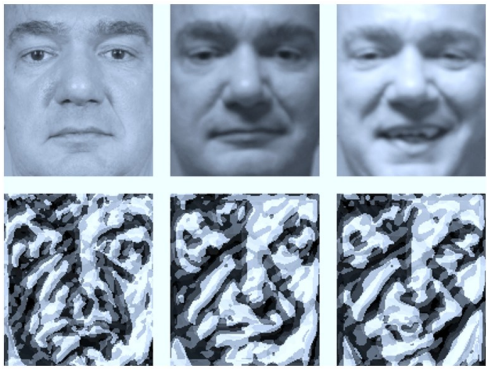
O *Binarized Statistical Image Features* (BSIF) é um descritor que analisa o código binário de cada pixel baseado em *Independent Component Analysis* (ICA) e usa filtros de diversos tamanhos, usados para extrair informações sobre formas, texturas e outras características de uma imagem. O BSIF é conhecido por apresentar características da textura de forma robusta quando comparado com outros descritores (KANNALA; RAHTU, 2012). A Figura 3 ilustra os diferentes tipos de filtros e a Figura 4 demonstra a aplicação do BSIF em imagens.

Figura 3 – Filtros 9x9 do BSIF.



Fonte: Kannala e Rahtu (2012, p. 1364).

Figura 4 – Imagens com a aplicação do BSIF.

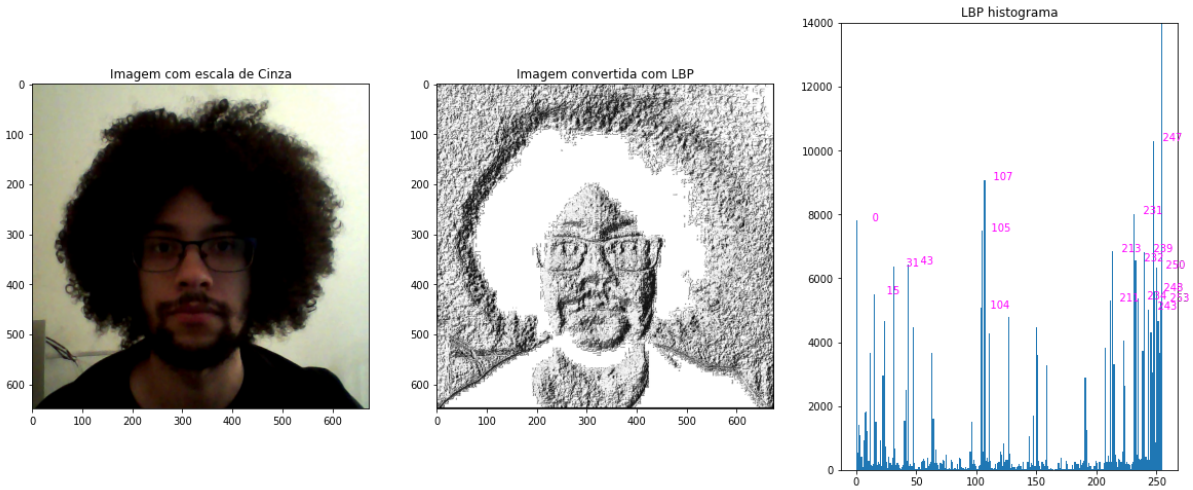


Fonte: Fonte: Kannala e Rahtu (2012, p. 1365).

O *Local Binary Pattern* (LBP) é um extrator de textura simples, e eficiente. Gera um histograma de padrões uniformes utilizando imagens em escala de cinza, baseado na vizinhança de cada píxel. LBP analisa uma imagem em um *grid* de pixels e compara os pixels vizinhos

para criar um código binário. Esse código binário é então usado para representar a imagem em um formato que pode ser analisado por um classificador. Sendo seus principais parâmetros o N representando o número de vizinhos considerados ao calcular o valor do LBP e R o raio do círculo usado para determinar os vizinhos. Isto permite capturar as características de textura da imagem de forma eficiente (PAULINO *et al.*, 2018), conforme exemplificado na Figura 5.

Figura 5 – Imagens com a aplicação do LBP.



Fonte: Autoria própria (2022).

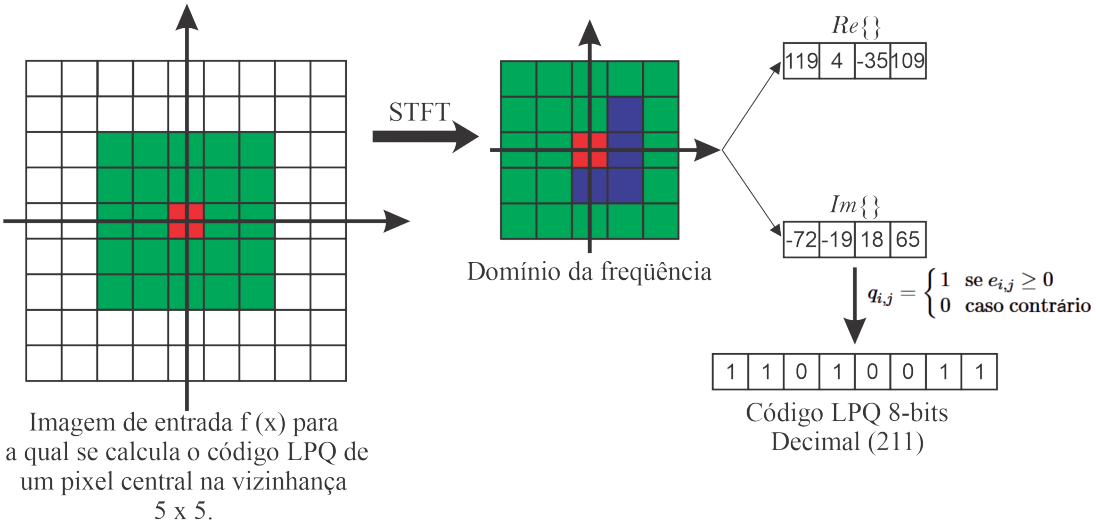
Já o *Local Phase Quantization* (LPQ) é um descritor de textura projetado para lidar com imagens borradas. Ele apresenta um comportamento similar ao LBP, divergindo apenas na metodologia empregada para o cálculo da parametrização da vizinhança. O LBP utiliza operações simples, enquanto o LPQ opera sobre a *Discrete Fourier Transform* (DFT) 2D (GONZALEZ-SOLER *et al.*, 2020). Seu principal parâmetro é a variável *winSize* caracterizando o tamanho da janela de análise utilizada pelo algoritmo. A Figura 6 demonstra os passos necessários para a obtenção do LPQ.

2.3.2 Descritores baseados em *Transfer Learning*

Transfer learning é uma técnica que transpõe as características previamente aprendidas em um contexto específico e as aplica na resolução de um problema em um contexto diferente (ZHUANG *et al.*, 2021). Essa técnica vem sendo utilizada em muitos problemas de classificação com resultados satisfatórios (LU *et al.*, 2021). A Figura 7 mostra um exemplo em que as características aprendidas para diferenciar entre vários animais, são usadas para classificar chihuahuas que não estavam presentes no problema de classificação original. Neste trabalho são avaliados dois conjuntos de características extraídas por esse método, o *MobileNet* e o *VGG16*.

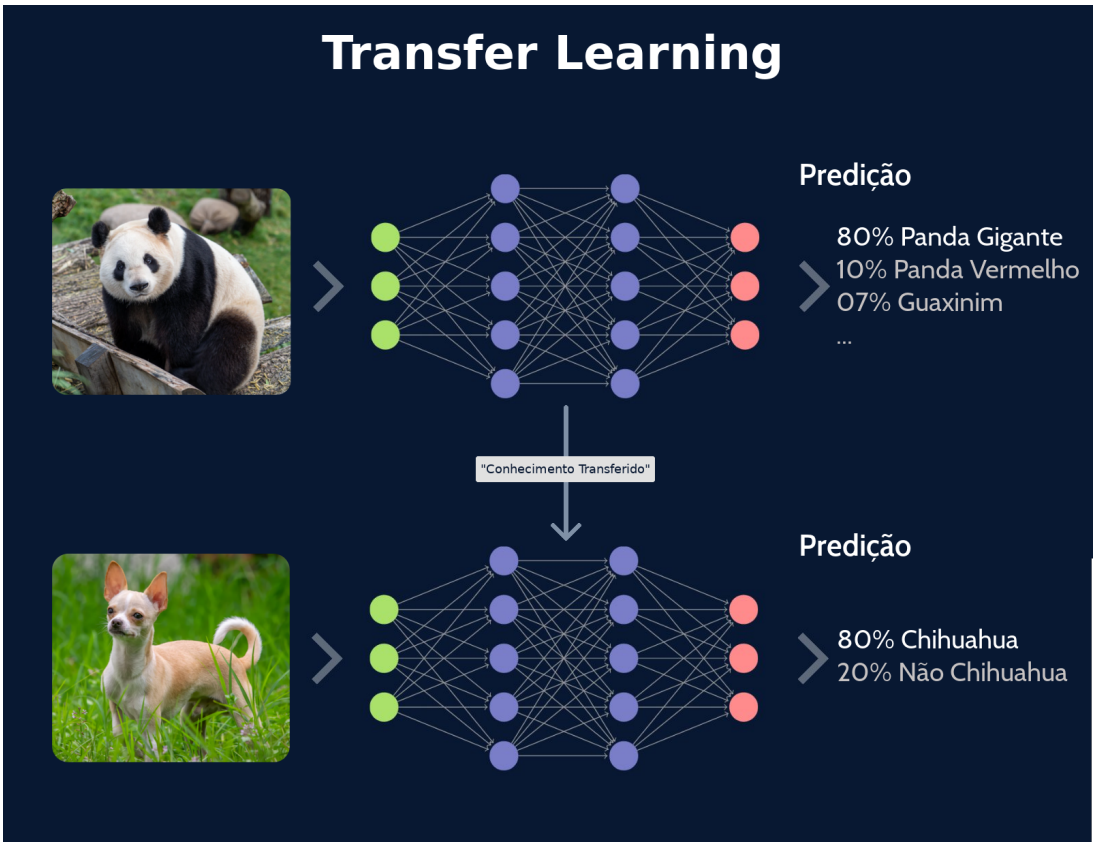
O modelo *MobileNet* surgiu da necessidade de se ter modelos de rede neural convolucional pequenos e de baixa latência que podem ser facilmente incorporado em qualquer aplicativo de visão computacional de dispositivos móveis. O seu diferencial é que além de focar

Figura 6 – Método do LPQ



Fonte: Belahcene *et al.* (2016, p. 2).

Figura 7 – Exemplo de transferência de conhecimento entre redes neurais.

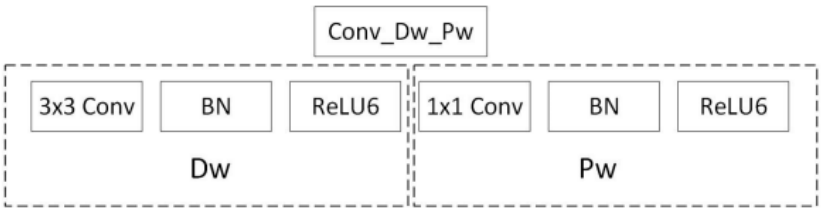


Fonte: Autoria própria (2022).

no tamanho ela considera a latência, diminuindo os seus parâmetros sem sacrificar a acurácia. As restrições de recursos (latência, tamanho) podem ser escolhidas para cada caso específico. Essa rede é usada para detecção de objetos, reconhecimento de faces e lugares (HOWARD *et al.*, 2017).

A Figura 8 mostra a estrutura convolutiva básica do MobileNet. Sendo a variável Conv_-Dw_Pw a profundidade e a separabilidade da estrutura convolucional. Sua estrutura é composta por camadas *depth-wise* (Dw) e camadas *point-wise* (Pw). Dw são camadas profundas convolucionais usando *kernels* de 3×3, enquanto a Pw são camadas convolucionais utilizando *kernels* de 1×1. Cada resultado de convolução é tratado pelo algoritmo de normalização em lote e com função de ativação *Rectified Linear Unit* (ReLU) (LI *et al.*, 2018).

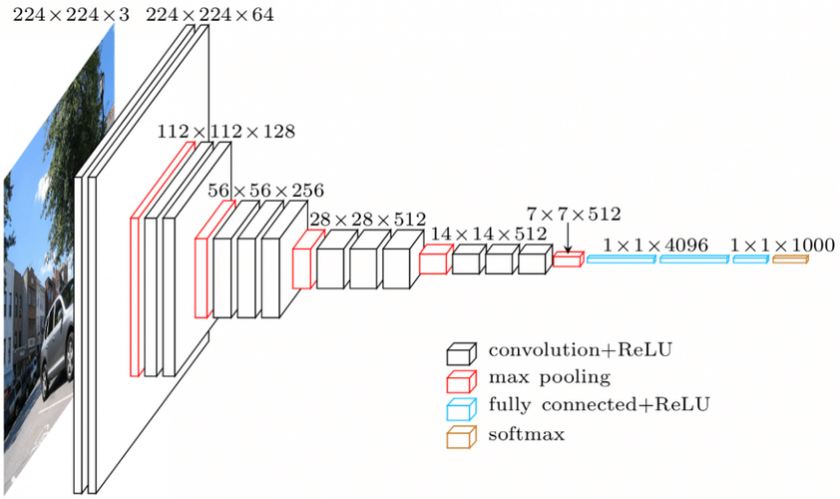
Figura 8 – Estrutura convolucional básica do MobileNet.



Fonte: Li *et al.* (2018, p. 6).

A VGG-16 é uma rede neural desenvolvida por Simonyan e Zisserman (2015) para a competição ILSVRC 2014 [3]. Esta rede neural contém 16 camadas convolucionais e 3 camadas totalmente conectadas que contêm a maioria dos parâmetros da rede. A classificação é realizada pela função de ativação *soft-max*, usada para emitir as probabilidades de saída (LOUKADAKIS; CANO; O'BOYLE, 2018). A Figura 9 demonstra a arquitetura da rede neural.

Figura 9 – Arquitetura da rede neural VGG-16



Fonte: Loukadakis, Cano e O'Boyle (2018, p. 4).

2.4 *Suport Vector Machine*

O *Support Vector Machine* é uma técnica de aprendizado de máquina baseada na teoria de aprendizado estatístico. Foi projetado originalmente para separar linearmente por um

[3] https://image-net.org/challenges/LSVRC/

hiperplano de duas classes. O SVM visa enfatizar a diferença entre dados pertencentes a diferentes classes, calculando o chamado “limite de decisão”, também conhecido como vetores de suporte (CORTES; VAPNIK, 1995). Entretanto, em diversas situações reais os dados não tendem a ser linearmente separáveis, mesmo assim, o SVM consegue mapear o conjunto de dados de entrada para um espaço dimensional maior, em que os dados podem ser separados linearmente.

3 TRABALHOS RELACIONADOS

Neste capítulo apresenta trabalhos que seguem a mesma perspectiva que o tema proposto.

3.1 Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms

Os sinais de fala são usados como a principal fonte de entrada da interação humano-computador, podendo assim desenvolver diversas aplicações com base nesse quesito, como fala automática, reconhecimento de emoções, de gênero e de idade. Classificar falantes por idade e gênero é uma tarefa difícil no processamento de fala devido às deficiências dos métodos atuais de extração de características e modelos de classificação.

Para este problema, a proposta de Tursunov *et al.* (2021) foi criar uma rede neural convolucional (CNN) com um módulo de multi atenção (MAM) para extrair as características espaciais e temporais dos dados de entrada de forma eficaz.

O sistema de classificação de idade e gênero proposto foi testado usando o Common Voice e um conjunto de dados de reconhecimento de fala da língua coreana desenvolvido pelos pesquisadores.

O modelo sugerido alcançou 96%, 73% e 76% de acurácia para o reconhecimento de gênero, idade e idade-gênero, respectivamente, usando o Common Voice conjunto de dados. Já os resultados do conjunto de dados de reconhecimento de fala coreano foram de 97%, 97% e 90% para gênero, idade e reconhecimento idade-gênero, respectivamente.

3.2 Handcrafted vs Non-Handcrafted Features for Computer Vision Classification

Nanni, Ghidoni e Brahnam (2017) apresentam uma pesquisa exploratória, que tenta identificar a eficácia de diferentes métodos a partir de vários conjuntos de extratores de características separados e combinados, para quantificar os resultados finais em várias tarefas de classificação de imagens.

Fazendo o uso de visão computacional, foram selecionados vários *datasets* para testar os métodos em um classificador SVM, utilizando os vetores de características extraídos com o LPQ, variações do LBP, bem como características obtidas a partir de métodos de *transfer learning*.

Como o tamanho dos vetores de característica é demasiado grande, isso torna o trabalho de processamento mais dispendioso. O uso de redutores de dimensionalidade como o *Principal Component Analysis* (PCA) e o *Discrete Cosine Transform* (DCT) foi empregado.

Os autores demonstram que o modelo de *transfer learning* supera as abordagens baseadas em características *handcrafted*. Já a fusão de características *handcrafted* demonstrou ser promissora, alcançando resultados satisfatórios.

3.3 Speaker Recognition using fusion of features with Feedforward Artificial Neural Network and Support Vector Machine

Chauhan, Isshiki e Li (2020) apresentaram um estudo comparativo de desempenho de modelos que utilizam diversas combinações de características diferentes, entre identificação e verificação de locutores.

Para a classificação, são empregados o SVM e *Artificial Neural Network* (ANN). Já as características extraídas foram *Mel Frequency Cepstral Coefficient* (MFCC), *Linear Predictive Coding* (LPC) e *Perceptual Linear Prediction* (PLP) do dataset ELSDSR, o qual é exclusivamente de falantes da língua inglesa.

Constatou-se que características combinadas para identificação de locutor teve um aumento de desempenho de 3% a 5% do que as características isoladas. Os autores justificam que isso ocorre porque o uso do conjunto de características terá mais informações importantes devido à sua complementariedade. Um ponto importante é que a classificação com o Modelo de Rede Neural demonstrou um melhor desempenho do que o SVM para esse conjunto de dados.

Já o presente trabalho emprega sobre a base de dados BrSd os métodos *handcrafted* e métodos de *transfer learning*, combinando características e comparando os resultados com a base de dados *Common Voice* de língua inglesa. Desta forma, foi possível avaliar o impacto de diferentes características assim como se existem grandes variações de desempenho usando diferentes línguas.

4 MÉTODO

O propósito deste trabalho é avaliar métodos de identificação e verificação de locutores em uma base de dados em língua portuguesa, a *Brazilian Speech Database* (PAULINO *et al.*, 2018) e na base de dados de lingua inglesa a *Common Voice*. Para isto, são empregados diferentes descritores de características baseados em abordagens *handcrafted* e *transfer learning*, além de técnicas de fusão de características.

O desempenho do BrSd foram comparados aos resultados obtidos de um *subconjunto* do Common Voice submetido aos mesmos métodos. Outro método empregado nesse *subconjunto* foi a replicação dos áudios, repetindo parte do espectrograma para ficar parecido com a duração dos audios do BrSd, pois a maioria dos áudios da base de dados original tem em média 10 segundos, já o BrSd, 15 segundos de duração. Os detalhes do *subconjunto* podem ser observadas na Tabela 1.

Tabela 1 – Subset do *Common Voice* baseado no BrSd.

Idade	Masculino	Feminino	Total
teens	10	10	20
twenties	12	12	24
thirties	4	4	8
fourties	4	3	7
fifties	2	1	3
sixties	5	7	12
seventies	3	2	5
eighties	0	1	1

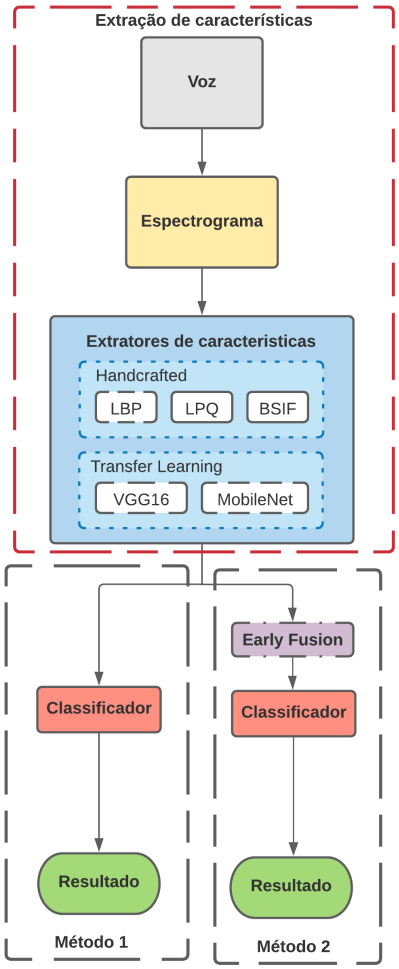
Fonte: Autoria própria (2022).

A Figura 10 apresenta o método proposto para este trabalho. A primeira etapa é feita a geração dos espectrogramas dos áudios, a seguir é realizada a extração de características de todos os áudios da base de dados, logo após são realizados os seguintes passos:

1. O áudio é transformado em uma imagem de espectrograma;
2. O espectrograma é dividido em 1, 3, 5 e 10 pedaços verticais, chamados *patches*;
3. Os conjuntos de características (LBP, LPQ, BSIF, VGG-16 e MobileNet) são extraídos separadamente de cada *patch* gerado no passo anterior sendo armazenados em vetores de características.

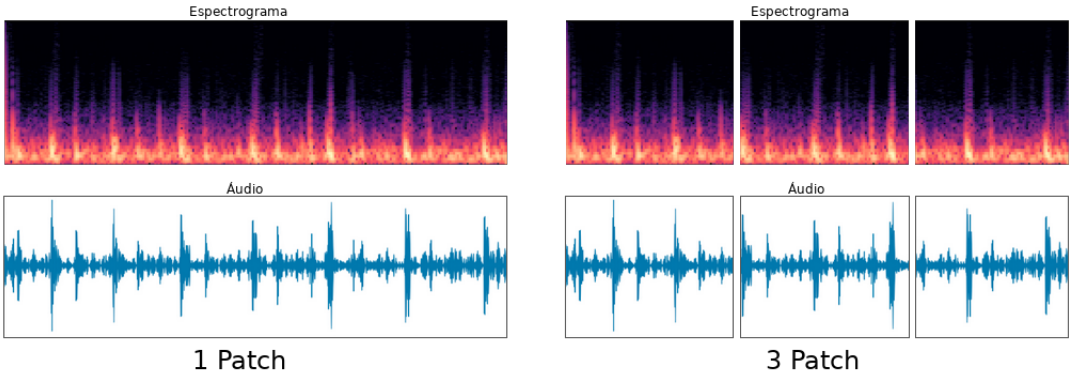
O áudio é um fenômeno que varia no tempo e a representação com o espectrograma transpõe essa característica. Queremos utilizar esse fenômeno no nosso classificador, por isso a extração de pedaços pequenos desse áudio, o chamado *patch*, podem conter características únicas nesses espaços de tempo menor. Uma representação do conceito de *patch* pode ser observada na Figura 11.

Figura 10 – Fluxograma da metodologia.



Fonte: Autoria própria (2021).

Figura 11 – Representação dos *patches* de um espectrograma.



Fonte: Autoria própria (2021).

Os parâmetros utilizados nos extratores de características empregados neste trabalho podem ser observados na Tabela 2.

Tabela 2 – Dimensões dos vetores de características e principais parâmetros.

Descriptor	Parâmetros	Dimensão
LBP	N = 8; R = 2	59
LPQ	winSize = 7	256
BSIF	filtro = ICAtextureFilters-11X11-8bit	256
VGG16	Simonyan e Zisserman (2015)	512
MobileNet	Howard *et al.* (2017)	1280

Fonte: Fonte: Autoria própria (2021).

Para avaliar o efeito da combinação destes diferentes conjuntos de características em ambas tarefas (identificação e verificação), optamos por avaliar dois métodos diferentes, mostrados na Figura 10. No primeiro método, utilizado como base de comparação para os demais, cada conjunto de características é avaliado separadamente, no segundo método, conhecido como *early-fusion*, diferentes combinações de vetores de características são concatenados em um único vetor.

O classificador escolhido foi o SVM, por ser um poderoso algoritmo amplamente usado em diversas tarefas de classificação. Conhecido por obter bons resultados em vários problemas relacionados a reconhecimento de padrões (LU; ZHANG; LI, 2003).

Para estimar a capacidade de generalização do método proposto, foi avaliado um protocolo de validação cruzada em 5 vias. A métrica de avaliação final dos resultados foi a média harmônica entre a precisão e o *recall*, conhecida como F1-Score (PEDREGOSA *et al.*, 2011). O teste T de *Student* foi usado para avaliar se as diferenças nos resultados obtidos são estatisticamente significativas.

4.1 Experimentos

Este experimento utilizou a base de dados do BrSD para a identificação e verificação de locutores de modo a obter o desempenho nessas tarefas. Para obter os dados do desempenho, foram feitas a divisão sem sobreposição dos espectrogramas em *patches* para em seguida ser realizada a extração dos vetores de características dos mesmos.

O próximo passo foi a normalização com o *z-score* dos vetores de características. Para chegar nos melhores hiper parâmetros de cada modelo, foi realizado uma busca exaustiva com os valores de parâmetros especificados. Os valores dos parâmetros avaliados podem ser observados na Tabela 3. Para estimar a capacidade de generalização dos modelos foi realizada a validação cruzada em 5 vias.

Nesta etapa foram avaliadas 80 classes (locutores), onde cada locutor tem 5 exemplos de frases *(folds)*, sendo cada uma dessas frases transformada em espectrograma e divida em 4 *patches* diferentes, (1, 3, 5 e 10 patches).

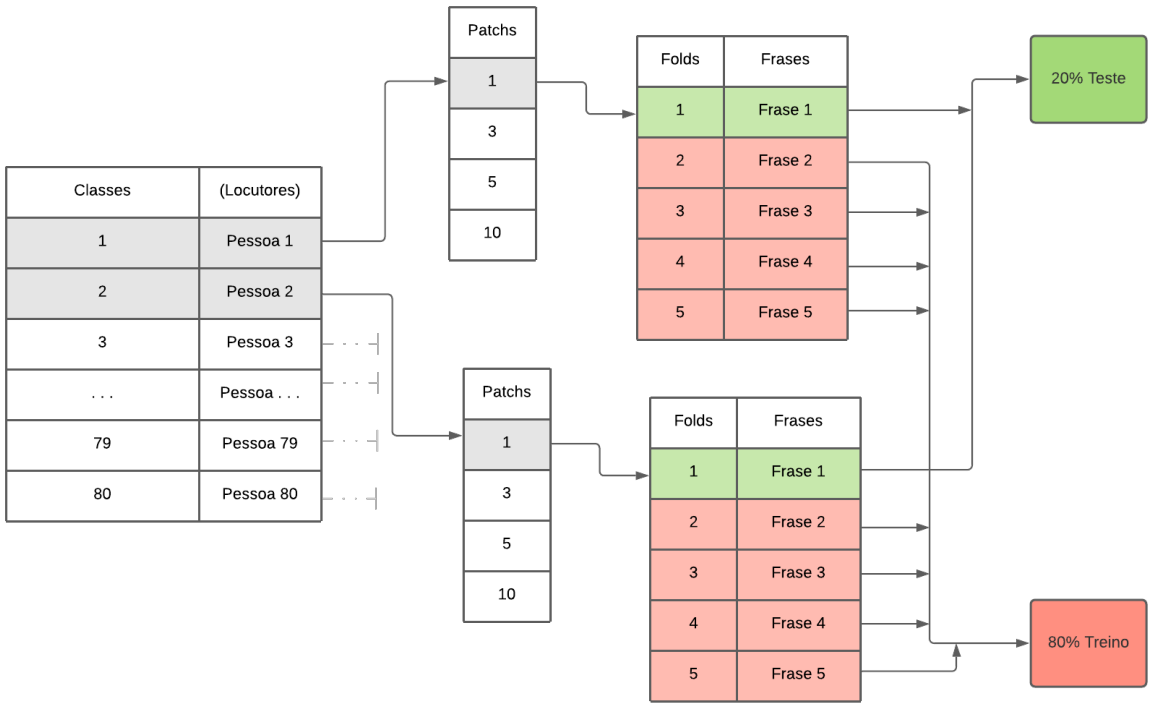
Tabela 3 – Valores avaliados para os parâmetros C e Gamma do SVM.

Parametros	Valores			
C	1	10	100	1000
Gamma	auto	2e-1	2e-2	**2e-3**

Fonte: Autoria própria (2022).

Para cada locutor foi escolhido 20% dos áudios para compor o teste, os 80% restantes foi usado para o treino do classificador. A Figura 12 detalha a divisão entre os dados de testes e os dados de treino.

Figura 12 – Fluxograma simplificado da escolha do treino e teste para a tarefa de identificação.



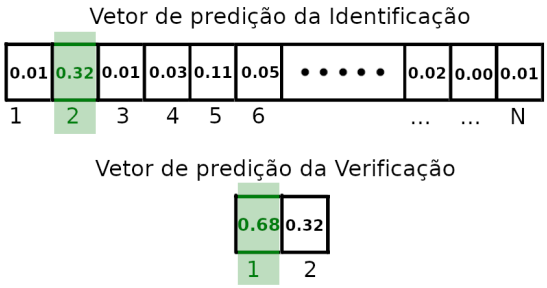
Fonte: Autoria própria (2021).

A Figura 13 demonstra os vetores de predição para os problemas de identificação e de verificação respectivamente. Para o problema de identificação, é criado um vetor de predição com o tamanho igual a quantidade de pessoas da amostra escolhida, no caso 80. Cada posição (ou índice) do vetor representa um dos indivíduos, incluindo aquele a ser identificado. Partindo de um áudio do qual se deseja identificar o locutor, a probabilidade de cada indivíduo ser o locutor em questão é preenchida pelo classificador em sua posição correspondente no vetor. Assim, o indivíduo associado a posição com a maior probabilidade, é identificado como o locutor do áudio.

Já no problema de verificação o modelo de classificador é personalizado para cada indivíduo. Desta forma, o vetor de predição é binário, sendo que um dos índices corresponde a classe de interesse (que se deseja verificar) e o outro índice a uma classe impostora. Quando a probabilidade da classe de interesse é superior ao da classe impostora, o indivíduo avaliado

é verificado como o locutor. Neste trabalho, a classe impostora pode conter 5 ou 79 indivíduos aleatórios da base de dados.

Figura 13 – Vetores de predições dos problemas de identificação e verificação.



Fonte: Autoria própria (2022).

5 RESULTADOS E DISCUSSÕES

Neste capitulo são apresentados resultados e discussões, realizados na tarefa de identificação e verificação de locutores.

5.1 Identificação de locutores

As Tabelas 4 e 5 apresenta os resultados por descritor de textura para todas as quantidades de *patches* avaliadas no problema de Identificação nas bases de dados BrSd e *Common Voice* respectivamente. Sendo essas tabelas correspondendo aos resultados obtidos com o Método 1, que utiliza as características individualmente de cada descritor de textura para classificar. O método é representado na Figura 10.

Observa-se na Tabela 4 que os resultados sofrem uma variação do F1-Score dependendo dos descritores e número de *patches*. Dentre esses valores destaca-se como o menor valor do *F1-Score* obtido para o descritor VGG-16 e 1 *patch* por áudio (0,35 ± 0,09) e o maior valor para a método de extração BSIF com 10 *patches* por áudio (0,66 ± 0,06). Já os descritores baseados em *transfer learning* (VGG-16, MobileNet), apesar de serem computacionalmente mais complexos comparados aos métodos *handcrafted* (LBP, LPQ, BSIF), não apresentaram resultados estatisticamente superiores.

Tabela 4 – *F1-scores* por descritor de áudio para todas as quantidades de *patches* avaliadas no BrSd para a identificação de locutores.

Descritor de Textura	Patchs			
	1	3	5	10
LBP	0,42 ± 0,06	0,53 ± 0,03	0,57 ± 0,04	**0,57 ± 0,03**
LPQ	0,37 ± 0,02	0,56 ± 0,07	**0,62 ± 0,09**	0,61 ± 0,07
BSIF	0,45 ± 0,06	0,62 ± 0,06	0,65 ± 0,07	**0,66 ± 0,06**
VGG-16	0,35 ± 0,09	0,52 ± 0,08	0,58 ± 0,09	**0,59 ± 0,09**
MobileNet	0,43 ± 0,18	0,56 ± 0,12	0,58 ± 0,09	**0,58 ± 0,06**

Fonte: Autoria própria (2022).

Tabela 5 – F1-scores por descritor de áudio para todas as quantidades de *patches* avaliadas no *Common Voice* para a identificação de locutores.

Descritor de Textura	Patchs			
	1	3	5	10
LBP	0.09 ± 0.08	0,23 ± 0,07	**0,24 ± 0,06**	0,22 ± 0,07
LPQ	0,16 ± 0,07	0,22 ± 0,06	**0,25 ± 0,08**	0,24 ± 0,06
BSIF	0,13 ± 0,08	0,21 ± 0,07	**0,24 ± 0,07**	0,19 ± 0,07
VGG-16	0,30 ± 0,05	0,52 ± 0,03	**0,58 ± 0,03**	0,56 ± 0,04
MobileNet	0,39 ± 0,04	0,52 ± 0,07	0,52 ± 0,03	**0,58 ± 0,05**

Fonte: Autoria própria (2022).

As Tabelas 6 e 7 descrevem os resultados de três combinações de descritores de áudio para todas as quantidades de *patches* por áudio avaliado no problema de Identificação para o BrSd e o *Common Voice*. Esta Tabela corresponde aos resultados obtidos com o Método 2 representado na Figura 10.

Na Tabela 6 o melhor resultado obtido foi com o método de *early fusion* utilizando os descritores LBP, LPQ e BSIF, com o F1-Score de 0,70 ± 0,10, para 5 *patchs* por áudio. Este resultado sugere que as predições de cada descritor são complementares, conforme a teoria de combinação de descritores. Neste cenário, nenhum dos descritores de áudio isolados se destacou em relação aos outros. Entretanto, quando combinados utilizando o método de *early fusion*, observou-se uma melhora na acurácia.

Tabela 6 – *F1-scores* usando o método de *Early Fusion* para a identificação de locutores com o BrSd. (1) Descritores LBP, LPQ e BSIF, (2) Descritores MobileNet e VGG-16, (3) Todos os descritores.

Descritor de Textura	Patchs			
	1	3	5	10
Early Fusion (1)	0,45 ± 0,08	0,63 ± 0,08	0,67 ± 0,07	**0,70 ± 0,10**
Early Fusion (2)	0,42 ± 0,22	0,60 ± 0,15	0,63 ± 0,13	**0,63 ± 0,09**
Early Fusion (3)	0,47 ± 0,08	0,61 ± 0,10	0,64 ± 0,11	**0,68 ± 0,08**

Fonte: Autoria própria (2022).

Tabela 7 – F1-scores usando o método de Early Fusion para a Identificação de locutores com o *Common Voice*. (1) Descritores LBP, LPQ e BSIF, (2) Descritores MobileNet e VGG-16, (3) Todos os descritores.

Descritor de Textura	Patchs			
	1	3	5	10
Early Fusion (1)	0,21 ± 0,08	0,25 ± 0,07	**0,28 ± 0,09**	0,24 ± 0,06
Early Fusion (2)	0,36 ± 0,09	0,58 ± 0,08	0,60 ± 0,03	**0,68 ± 0,05**
Early Fusion (3)	0,19 ± 0,15	0,44 ± 0,13	0,46 ± 0,14	**0,51 ± 0,15**

Fonte: Autoria própria (2022).

Analisando as Tabelas 4 e 6, nota-se que, também conforme o número de *patches* por áudio aumenta, o F1-score médio de todos os métodos aumentam. Entretanto, os ganhos parecem não aumentar mais além de 5 *patches* por áudio. Por outro lado, quanto mais *patches* por áudio são usados para treinar os modelos, maior é o custo computacional. Por isto, os resultados com 5 *patches* por áudio apresentam a melhor relação entre custo computacional e o F1-Score.

Porém, na Tabela 5 e 7, os valores do *Common Voice* ficaram bem abaixo do que os retratados no BrSd, sendo o menor valor da Tabela 5 o F1-Score de 0,09 ± 0,08, para 1 *patchs* do descritor de textura LBP e o maior valor o F1-Score de 0,58 ± 0,03, para o VGG-16 com 5 *patchs*. Apesar do baixo F1-Score para a tarefa de identificação no *Common Voice* a Tabela 7

demonstra que a complementariedade de características pode ser benéfico para o problema, sendo 0,68 ± 0,05 o maior valor para o F1-Score com a junção das características dos descritores de *transfer learning*.

5.2 Verificação de locutores

O resultado de verificação de locutores para o BrSd podem ser observados nas Tabelas 8 e 10. Estes resultados foram obtidos com o método de classes impostoras, comparando o locutor de interesse com dois modelos de classificadores, um com 5 indivíduos impostores e outro com 79. O mesmo se aplica para as Tabelas 9 e 11 que representam os F1-Score para o *Common Voice*.

Percebe-se que na Tabela 8 os maiores F1-Scores estão localizados nos *patchs* 5 e 10. Na Tabela 10 o mesmo não ocorre. A Tabela 11 apresenta os melhores F1-Scores com o *patch* 10, e já na Tabela 10 isso já não ocorre. Contudo, estatisticamente não há diferenças entre o número de *patchs* entre as Tabelas 8 e 11. A existência de dependência de texto que existe na base de dados do BrSd e a não dependência de texto no *Common Voice* pode ser um dos fatores que contribuíram para esses resultados entre as bases de dados.

Tabela 8 – F1-scores por descritor de áudio para todas as quantidades de *patches* avaliadas e quantidade de classes no BrSd para a verificação de locutores.

Descritor de Textura	Quantidade de Classes	Patchs			
		1	3	5	10
LBP	5	0,85 ± 0,03	0,91 ± 0,04	**0,95 ± 0,02**	0,94 ± 0,02
	79	0.94 ± 0,01	0,95 ± 0,01	0,96 ± 0,00	**0,96 ± 0,00**
LPQ	5	0,93 ± 0,02	0,96 ± 0,03	**0,96 ± 0,01**	0,95 ± 0,02
	79	0,94 ± 0,01	0,94 ± 0,01	0,95 ± 0,01	**0,96 ± 0,00**
BSIF	5	0,95 ± 0,02	0,97 ± 0,01	**0,97 ± 0,01**	0,95 ± 0,01
	79	0,95 ± 0,01	0,95 ± 0,01	0,95 ± 0,01	**0,95 ± 0,01**
VGG-16	5	0,84 ± 0,03	0,91 ± 0,03	**0,92 ± 0,01**	0,91 ± 0,03
	79	0,93 ± 0,01	0,95 ± 0,01	0,96 ± 0,01	**0,96 ± 0,00**
MobileNet	5	0,89 ± 0,04	0,94 ± 0,03	**0,96 ± 0,02**	0,93 ± 0,01
	79	0,95 ± 0,01	0,96 ± 0,01	0,95 ± 0,00	**0,97 ± 0,00**

Fonte: Autoria própria (2022).

Portanto, foi identificado que para algumas combinações de características utilizando o *early fusion* ajudou a obter melhores resultados em ambas as bases de dados.

Nota-se que, em geral, conforme o número de *patches* por áudio aumenta, o F1-score médio de todos os métodos aumentam. Entretanto, os ganhos parecem não aumentar mais além de 5 *patches* por áudio.

Tabela 9 – F1-scores por descritor de áudio para todas as quantidades de *patches* avaliadas e quantidade de classes no *Common Voice* para a verificação de locutores.

Descritor de Textura	Quantidade de Classes	Patchs			
		1	3	5	10
LBP	5	**0,83 ± 0,03**	0,80 ± 0,02	0,80 ± 0,02	0,79 ± 0,01
	79	0,96 ± 0,00	0,97 ± 0,00	0,98 ± 0,00	**0,98 ± 0,00**
LPQ	5	**0,84 ± 0,03**	0,82 ± 0,02	0,82 ± 0,02	0,80 ± 0,02
	79	0,96 ± 0,01	0,97 ± 0,00	0,97 ± 0,00	**0,98 ± 0,00**
BSIF	5	**0,84 ± 0,02**	0,82 ± 0,01	0,82 ± 0,02	0,81 ± 0,01
	79	0,96 ± 0,01	0,97 ± 0,00	0,97 ± 0,00	**0,98 ± 0,00**
VGG-16	5	0,81 ± 0,01	0,87 ± 0,01	**0,90 ± 0,01**	0,87 ± 0,01
	79	0,94 ± 0,01	0,97 ± 0,00	0,97 ± 0,00	**0,98 ± 0,00**
MobileNet	5	0,81 ± 0,02	**0,89 ± 0,01**	0,88 ± 0,02	0,88 ± 0,01
	79	0,94 ± 0,01	0,97 ± 0,00	0,98 ± 0,00	**0,98 ± 0,00**

Fonte: Autoria própria (2022).

Tabela 10 – F1-scores usando o método de *Early Fusion* para a verificação de locutores com o BrSd. (1) Descritores LBP, LPQ e BSIF, (2) Descritores MobileNet e VGG-16, (3) Todos os descritores.

Descritor de Textura	Quantidade de Classes	Patchs			
		1	3	5	10
Early Fusion (1)	5	0,95 ± 0,02	0,97 ± 0,01	**0,97 ± 0,01**	0,96 ± 0,01
	79	**0.95 ± 0,00**	0,94 ± 0,00	0,94 ± 0,01	0,95 ± 0,01
Early Fusion (2)	5	0,86 ± 0,03	0,92 ± 0,04	0,94 ± 0,02	**0,94 ± 0,02**
	79	0,93 ± 0,03	0,95 ± 0,01	0,96 ± 0,00	**0,97 ± 0,01**
Early Fusion (3)	5	0,87 ± 0,01	0,93 ± 0,03	**0,96 ± 0,02**	0,96 ± 0,02
	79	0,96 ± 0,02	0,96 ± 0,00	**0,96 ± 0,00**	0,96 ± 0,01

Fonte: Autoria própria (2022).

Por outro lado, quanto mais *patches* por áudio são usados para treinar os modelos, maior é o custo computacional. Por isto, os resultados com 5 *patches* por áudio apresentam a melhor relação entre custo computacional e o F1-Score.

O melhor resultado para a tarefa de identificação para a base de dados em português foi de 0,70 ± 0,10 com 10 *patches* utilizando o método de *early fusion* com as características do *handcrafted*. Já para a base de dados em inglês foi de 0.68 ± 0.05 com 10 *patches* utilizando o *early fusion* de todos os extratores do método de *transfer learning*.

Para o problema de verificação, a *Brazilian Speech Database* ficou com uma taxa de 0,97 ± 0,00 utilizando 10 *patches* com o MobileNet, e o *Common Voice* obteve uma taxa de

Tabela 11 – F1-scores usando o método de *Early Fusion* para a verificação de locutores com o *Common Voice*. (1) Descritores LBP, LPQ e BSIF, (2) Descritores MobileNet e VGG-16, (3) Todos os descritores.

Descritor de Textura	Quantidade de Classes	Patchs			
		1	3	5	10
Early Fusion (1)	5	**0,84 ± 0,02**	0,83 ± 0,03	0,83 ± 0,03	0,83 ± 0,02
	79	0,96 ± 0,01	0,97 ± 0,00	0,97 ± 0,01	**0,97 ± 0,00**
Early Fusion (2)	5	0,79 ± 0,02	0,88 ± 0,01	0,89 ± 0,01	**0,89 ± 0,00**
	79	0,93 ± 0,01	0,97 ± 0,01	0,97 ± 0,01	**0,98 ± 0,00**
Early Fusion (3)	5	0,79 ± 0,01	0,85 ± 0,05	0,86 ± 0,04	**0,87 ± 0,03**
	79	0,97 ± 0,00	0,96 ± 0,00	0,97 ± 0,00	**0,98 ± 0,00**

Fonte: Autoria própria (2022).

0,98 ± 0,00 com 10 *patches* para todos os descritores aplicados. Destacou-se que a complementariedade de características feita com o *early fusion* ajudou a obter resultados melhores em alguns casos.

6 CONCLUSÃO

Neste trabalho foram avaliadas diversas estratégias para as tarefas de identificação e verificação de locutores em uma base de áudios em português e inglês utilizando métodos de extração de características mais simples e conhecidos, os *handcrafted*, e métodos mais elaborados, conhecidos como *transfer learning*.

Ao comparar a métrica F1-Score da tarefa de identificação e de verificação, observa-se a superioridade dos índices obtidos pela verificação. Logo a metodologia empregada resultou em uma melhor assertividade na verificação de locutores.

Foi notável como a complementariedade de características feita com o *Early Fusion* ajudou a obter resultados melhores em algumas combinações de descritores de características. Já os descritores baseados em *transfer learning* (VGG-16, MobileNet), apesar de serem computacionalmente mais complexos comparados aos métodos handcrafted (LBP, LPQ, BSIF), e de se esperar resultados melhores como Nanni, Ghidoni e Brahnam (2017) demonstra em seu trabalho, não apresentaram resultados estatisticamente superiores neste cenário em específico.

Um fator a considerar é que o *Brazilian Speech Database* é uma base de dados baseada em dependência de texto e o *Common Voice* de não dependência de texto. Sendo essa caraterística do texto um ponto a ser pesquisado em trabalhos futuros, assim como, testes com múltiplas combinações de classificadores, chamadas *bagging*, aliadas a redução de dimensionalidade.

REFERÊNCIAS

BAILEY, D. H.; SWARZTRAUBER, P. N. A Fast Method for the Numerical Evaluation of Continuous Fourier and Laplace Transforms. **SIAM Journal on Scientific Computing**, v. 15, n. 5, p. 1105–1110, set. 1994. ISSN 1064-8275, 1095-7197. Disponível em: http://epubs.siam.org/doi/10.1137/0915067.

CHAUHAN, N.; ISSHIKI, T.; LI, D. Speaker recognition using lpc, mfcc, zcr features with ann and svm classifier for large input database. *In*: **2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)**. Singapore: IEEE, 2019.

CHAUHAN, N.; ISSHIKI, T.; LI, D. Speaker recognition using fusion of features with feedforward artificial neural network and support vector machine. *In*: **2020 International Conference on Intelligent Engineering and Management (ICIEM)**. London, UK: IEEE, 2020. p. 170–176.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, Springer, v. 20, n. 3, p. 273–297, 1995.

COSTA, Y. M. G. *et al.* Music genre recognition using spectrograms. *In*: **2011 18th International Conference on Systems, Signals and Image Processing**. Sarajevo, Bosnia and Herzegovina: IEEE, 2011. p. 1–4.

GAROFOLO, J. S. *et al.* Philadelphia: Linguistic data consortium,. **Computer Science and Language**, Web Download, 1993.

GONZALEZ-SOLER, L. J. *et al.* **Texture-based Presentation Attack Detection for Automatic Speaker Verification**. 2020.

HERRERA, P.; SERRA, X.; PEETERS, G. Audio Descriptors and Descriptors Schemes in the Context of MPEG-7. *In*: **ICMC: International Computer Music Conference**. Bejing, China: [*s.n.*], 1999. p. –. Cote interne IRCAM: Herrera99a. Disponível em: https://hal.archives-ouvertes.fr/hal-01105710.

HOWARD, A. G. *et al.* **MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications**. 2017.

ICHIKAWA, A.; NAKANO, Y.; NAKATA, K. Evaluation of various parameter sets in spoken digits recognition. **IEEE Transactions on Audio and Electroacoustics**, v. 21, n. 3, p. 202–209, 1973.

JAIN, A.; ROSS, A.; PANKANTI, S. Biometrics: a tool for information security. **IEEE Transactions on Information Forensics and Security**, v. 1, n. 2, p. 125–143, 2006.

KANNALA, J.; RAHTU, E. Bsif: Binarized statistical image features. *In*: **2012 21st International Conference on Pattern Recognition (ICPR 2012)**. Los Alamitos, CA, USA: IEEE Computer Society, 2012. p. 1363–1366. ISSN 1051-4651. Disponível em: https://doi.ieeecomputersociety.org/.

LI, Y. *et al.* Research on a surface defect detection algorithm based on mobilenet-ssd. **Applied Sciences**, MDPI AG, v. 8, n. 9, p. 1678, Sep 2018. ISSN 2076-3417. Disponível em: http://dx.doi.org/10.3390/app8091678.

LOUKADAKIS, M.; CANO, J.; O'BOYLE, M. Accelerating deep neural networks on low power heterogeneous architectures. *In*: **Eleventh International Workshop on Programmability**

and Architectures for Heterogeneous Multicores (MULTIPROG-2018). Manchester, UK: Enlighten, 2018.

LU, J. *et al.* Deep convolutional neural network with transfer learning for environmental sound classification. *In*: **2021 International Conference on Computer, Control and Robotics (ICCCR)**. Shanghai, China: IEEE, 2021. p. 242–245.

LU, L.; ZHANG, H.-J.; LI, S. Z. Content-based audio classification and segmentation by using support vector machines. **Multimedia Systems**, v. 8, n. 6, p. 482–492, Apr 2003. ISSN 1432-1882. Disponível em: https://doi.org/10.1007/s00530-002-0065-0.

NAGRANI, A. *et al.* Voxceleb: Large-scale speaker verification in the wild. **Computer Science and Language**, Elsevier, 2019.

NAIK, J.; DODDINGTON, G. Evaluation of a high performance speaker verification system for access control. *In*: **ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing**. Dallas, TX, USA: IEEE, 1987. v. 12, p. 2392–2395.

NANNI, L.; GHIDONI, S.; BRAHNAM, S. Handcrafted vs. non-handcrafted features for computer vision classification. **Pattern Recognition**, v. 71, p. 158–172, 2017. ISSN 0031-3203. Disponível em: https://www.sciencedirect.com/science/article/pii/S0031320317302224.

PAULINO, M. A. D. *et al.* A brazilian speech database. *In*: **2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)**. Volos, Greece: IEEE, 2018. p. 234–241.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

REYNOLDS, D. A. Speaker identification and verification using gaussian mixture speaker models. **Speech Communication**, v. 17, n. 1, p. 91–108, 1995. ISSN 0167-6393. Disponível em: https://www.sciencedirect.com/science/article/pii/016763939500009D.

REYNOLDS, D. A. Automatic speaker recognition: Current approaches and future trends. *In*: . Pennsylvania: Citeseer, 2001. v. 5, p. 14–15.

SHAFIK, A. *et al.* Speaker identification based on radon transform and cnns in the presence of different types of interference for robotic applications. **Applied Acoustics**, v. 177, p. 107665, 2021. ISSN 0003-682X. Disponível em: https://www.sciencedirect.com/science/article/pii/S0003682X20307696.

SHUKLA, A.; TIWARI, R.; KALA, R. Speech signal analysis. Springer Berlin Heidelberg, Berlin, Heidelberg, p. 111–128, 2010. Disponível em: https://doi.org/10.1007/978-3-642-14344-1_5.

SIMONYAN, K.; ZISSERMAN, A. **Very Deep Convolutional Networks for Large-Scale Image Recognition**. 2015.

TAHLIRAMANI, N. V.; BHATT, N. Performance analysis of speaker identification system with and without spoofing attack of voice conversion. *In*: **2018 2nd International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE)**. Ghaziabad, India: IEEE, 2018. p. 130–135.

TING, C. *et al.* Text independent speaker identification using gaussian mixture model. *In*: **2007 International Conference on Intelligent and Advanced Systems**. Kuala Lumpur, Malaysia: IEEE, 2007. p. 194–198.

TURSUNOV, A. *et al.* Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. **Sensors**, v. 21, n. 17, 2021. ISSN 1424-8220. Disponível em: https://www.mdpi.com/1424-8220/21/17/5892.

VACHER, M. *et al.* Speech and speaker recognition for home automation: Preliminary results. *In*: **2015 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)**. Bucharest, Romania: IEEE, 2015. p. 1–10.

VELICHKO, V.; ZAGORUYKO, N. Automatic recognition of 200 words. **International Journal of Man-Machine Studies**, v. 2, n. 3, p. 223–234, 1970. ISSN 0020-7373. Disponível em: https://www.sciencedirect.com/science/article/pii/S0020737370800086.

ZHUANG, F. *et al.* A comprehensive survey on transfer learning. **Proceedings of the IEEE**, v. 109, n. 1, p. 43–76, 2021.