CENTRO UNIVERSITÁRIO DA FEI

ARIEL GRAÇA FERREIRA

ESTIMATIVA DE FAIXA ETÁRIA PELA VOZ COM REDES NEURAIS CONVOLUCIONAIS

SÃO BERNARDO DO CAMPO

2021

Ariel Graça Ferreira

Estimativa de Faixa Etária Pela Voz com Redes Neurais Convolucionais

Qualificação apresentada como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica, pelo Programa de Pós-Graduação em Engenharia Elétrica do Centro Universitário da FEI.

Orientador: Prof. Dr. Ivandro Sanches

São Bernardo do Campo

2021

RESUMO

No decorrer dos últimos anos tem sido possível acompanhar a expansão e popularização do conceito de Internet das Coisas, onde os mais diversos objetos do cotidiano estão sendo interconectados. De forma similar, o progresso na área de Inteligência Artificial tem proporcionado uma interação cada vez maior entre as pessoas e esses objetos, formando assim uma grande rede que tende a crescer, e se desenvolver, cada vez mais. Visando o avanço tecnológico dentro desse contexto, é imprescindível que se busque uma interação homem-máquina mais refinada, e um dos principais meios para estabelecer tal interação é a voz humana. O presente trabalho propõem estudar os sistemas de estimação automática de faixa etária pela voz com Redes Neurais Convolucionais (RNC's), porém não é proposta apenas a busca pela compreensão e melhorias na arquitetura desse tipo de modelo, mas de se fazer também um contraponto com modelos mais tradicionais, que utilizem outras técnicas, como por exemplo Gaussian Mixture Models (GMM). Nos ensaios iniciais realizados com uma RNC, o modelo apresentou precisão de 60%, aproximadamente, e a partir desses resultados deve-se desenvolver o estudo proposto.

Palavras-chaves: Fala, Voz, Faixa Etária, Redes Neurais Convolucionais.

ABSTRACT

Over the last years, it has been possible to see the expansion and popularization of the concept called Internet of Things, where the most diverse ordinary objects are being interconnected. Similarly, studies and developments on Artificial Intelligence field are increasing the interaction between people and such objects, thus creating a large network that tends to grow, more and more. Aiming at technological progress within this context, it is essential to seek a more refined human-machine interaction, and the human voice, as source of personal information, is very relevant so it should be used on this purpose. The present work proposes to study the automatic estimation of age from speech signals with Convolutional Neural Networks (CNN's), but it is not only proposed to search for understanding and improvements in the architecture of this type of model, but also to make a comparison with more traditional models that use other techniques, such as Gaussian Mixture Models (GMM). The initial tests performed with an CNN, presented an accuracy of approximately 60%, then based on these results the proposed study should be developed.

Key-words: Speech, Voice, Age, Convolutional Neural Networks.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

MFCC	Mel-Frequency Cepstral Coefficients
LPC	Linear Predictive Coefficients
ASR	Automatic Speech Recognition
AL	Autenticação de Locutor
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
RNC	Rede Neural Convolucional
RNP	Rede Neural Profunda
RNR	Rede Neural Recorrente
URA	Unidade de Resposta Audível
IVR	Interactive Voice Response
API	Application Programming Interface
IoT	Internet of Things
IA	Inteligência Artificial

SUMÁRIO

1 INTRODUÇÃO

A estimação de faixa etária pela voz, possui diversas aplicações que visam criar uma interface homem-máquina para utilização em áreas como transporte, segurança, medicina e automação residencial.

Um exemplo bastante presente no cotidiano das pessoas atualmente, seriam as Inteligências Artificiais (IA) como a *Alexa*, desenvolvida pela Amazon, e o *Google Assistant*, criado pelo Google. São dispositivos espalhados em milhares de casas ao redor do mundo, e que recebem os mais diversos comandos de voz para execução de várias tarefas, que podem ser desde tocar uma música, até realizar uma complexa rotina de acionamento em conjunto com outros elementos e/ou equipamentos inteligentes, interconectados e espalhados pela casa de um indivíduo, ou fora dela. Para citar apenas uma aplicação, a estimação de faixa etária poderia funcionar como uma validação de segurança, evitando assim acesso ou acionamento indevido por alguma criança que tenha acesso ao dispositivo mas não possa ter acesso a todas as funcionalidades disponíveis.

Tomando como base esse cenário de crescimento tecnológico e expansão das aplicações de *Internet of Things* (IoT) e IA, este trabalho tem por finalidade estudar os sistemas de estimação de faixa etária por sinais de voz que utilizem técnicas de aprendizado profundo, como Redes Neurais Convolucionais (RNC). Busca-se entender e mostrar evidências do desempenho desse tipo de sistema, considerando a utilização de tais técnicas.

Através do sinal da voz de um indivíduo, é possível obter muitas informações sobre quem produz o som, como gênero, emoções e idade, porém essa tarefa também traz desafios. A incidência de distorções na fonte do sinal, geradas pelo próprio indivíduo (estado de saúde, emoção), ou geradas pelo ambiente (ruído externo), deformam o sinal de tal forma que a etapa de extração das características acústicas (*features*) fica prejudicada e, consequentemente, a etapa seguinte, a classificação do dado, também vai sofrer com a degradação.

Nesse sentido, as técnicas de aprendizado de máquina, mais especificamente aprendizado profundo, com base em pesquisas realizadas na literatura atual, mostram-se mais robustas para implementação desse tipo de sistemas que trabalham com sinais acústicos. Por tal razão, este trabalho busca estudar e avaliar a utilização de RNC's no problema de estimação automática de idade pela voz.

Para efeitos de comparação e posterior análise dos resultados, será utilizado como referência a pesquisa FAPESP desenvolvida pelo Prof. Dr. Ivandro Sanches (2019), que utiliza técnicas mais tradicionais de aprendizado de máquina, como o Gaussian Mixture Models (GMM) e i-Vector. Visto que essa pesquisa faz uso do mesmo banco de dados que será utilizado no desenvolvimento do presente trabalho com redes neurais, haverá oportunidade de comparar e discutir os resultados obtidos com ambos os sistemas.

Em relação ao banco de dados, será adotado um corpus com sinais de chamadas telefônicas que pertence a *Deutsche Telekom Laboratories*, de Berlim na Alemanha.

1.1 Objetivos

- Avaliar o desempenho de um sistema de estimação automática de faixa etária, o qual utiliza técnicas de aprendizado profundo para classificação de dados. Dentre tais técnicas, as Redes Neurais Convolucionais serão um dos principais objetos de estudo do trabalho visto que suas características são propícias para o processamento dos sinais de voz.

- Comparar o desempenho do sistema de estimação de faixa etária que utiliza RNC com o sistema que utiliza GMM e i-Vector, desenvolvido em Sanches (2019), visto que ambos os trabalhos executam simulações com a mesma base de dados.

- Analisar os impactos no desempenho do sistema em decorrência da alteração dos tipos de features utilizadas.

- Durante as pesquisas, ensaios e simulações, deve-se avaliar eventuais limitações dos elementos que compõem o sistema de estimação de faixa etária com RNC. Com base na identificação desses limitadores, deve-se buscar e pesquisar sobre formas de mitigar tais fatores, e assim propor otimizações.

2 REVISÃO BIBLIOGRÁFICA

Para auxiliar nas pesquisas do presente estudo, se faz necessário não somente ter uma visão sobre outros trabalhos que versem sobre a utilização de aprendizado profundo para estimação etária pela voz, mas também o entendimento da evolução das técnicas de aprendizado de máquina aplicadas ao processamento de sinais acústicos, de forma que seja possível compreender os desenvolvimentos que foram realizados no decorrer dos anos, e quais desafios ainda devem ser abordados.

Como pontuado por Deng (2016), apesar de as redes neurais artificiais já serem uma ferramenta conhecida por mais de meio século, e possuirem inclusive aplicações para processamento de fala, foi apenas com a conjunção de fatores como o surgimentos das redes "profundas", referenciadas aqui como Redes Neurais Profundas (RNP), disponibilização de grandes bases de dados, e popularização de computadores com grande capacidade de processamento (que utilizam GPU - *Graphics Processing Unit*), que os resultados mais expressivos começaram a aparecer e a se diferenciar em relação aos sistemas mais tradicionais, que até então utilizavam modelos com base em GMM e *Hidden Markov Models* (HMM). De acordo com o autor, esse momento de transição teve seu auge entre 2009 e 2010, e foi a grande revolução para o campo de Autenticação de Locutor (AL).

Ainda nessa linha, existem diversos trabalhos que trazem pontos importantes que devem ser considerados, e sempre fazendo um contraponto com os modelos tradicionais baseados principalmente em GMM. O próprio Deng (2016), comenta que um fator crucial para se obter bons resultados com sistemas baseados em aprendizado profundo para aplicações de AL é a quantidade de dados que se requer para a etapa de treinamento do modelo. Para o autor, a relação é simples, quanto mais dados para treinamento, melhor a precisão. Outra informação trazida no mesmo artigo diz que o aprendizado profundo reforça a idéia de que para aplicações de AL, as etapas de extração de *features* e classificação não devem ser tratadas de forma separada e independentes. Para ilustrar melhor esse conceito, deve-se considerar que os melhores resultados de AL com RNP são obtidos, não utilizando *Mel Frequency Cepstral Coeficients* (MFCC), por exemplo, mas sim *features* mais "primitivas"como banco de filtros na escala Mel. Entende-se que a utilização de dados em formato "raw", sem grande processamento, é benéfica nesse caso pois contém mais informação, visto que preserva dados que são expressos apenas no domínio da frequência. O autor finaliza esse raciocínio, argumentando ainda que o uso de RNC's só é efetivo para aplicações de AL, caso sejam utilizadas *features* espectrais, ao invés de features como MFCC. Esse mesmo conceito é abordado por Zhang et al. (2018).

Outra vantagem ao se utilizar RNP, citada em Deng (2016) e Zhang et al. (2018), seria a robustez desse tipo de modelo em relação a ruído, apesar de haver ressalvas de que esse tópico

ainda deve ser estudo para que haja uma compreensão mais abrangente sobre o tema. O artigo do Fayek, Lech e Cavedon (2017), apesar de não abordar o problema de estimação de faixa etária, mas sim o reconhecimento de emoções pela voz, mostra que RNC's alcançam melhor desempenho se comparadas com outras arquiteturas de RNP's. O autor obteve uma taxa global de precisão de 64.78% nos ensaios realizados com esse tipo de rede.

Outro fator de relevância, abordado em alguns trabalhos são as *short-term* e *long-term features*. Segundo Kalluri, Vijayasenan e Ganapathy (2020), *short-term* features seriam MFCC, LPC, frequências formantes, frequência fundamental, entre outras, que podem ser extraídas de sinais de voz de curta duração (< 5 segundos) e ainda possibilitar a obtenção de resultados satisfatórios. O autor demonstra que um sistema com arquitetura de RNP, destinado a estimar grupos etários, ao utilizar esse tipo de *features*, obtém, em termos de *Mean Absolute Error* (MAE), valores de 5,2 e 5,6 anos para locutores homens e mulheres, respectivamente. Outra parte interessante desse mesmo artigo, são as análises de estruturas corporais como altura, peso e comprimentos dos ombros e cintura, parâmetros esses também obtidos através do processamento da voz por RNP. Em Buyuk e Arslan (2018), os autores abordam com mais detalhes sobre a diferenciação das *short-terms* e *long-terms features*. Mais uma vez, *short-term features* são classificadas como as características extraídas de trechos curtos (com duração de poucos segundos) do sinal acústico. Novamente, MFCC e LPC são mencionadas como exemplos bastante difundidos desse tipo espectral de *features*. Além disso, argumenta-se ainda que o pitch (frequência de vibração das cordas vocais) também pode ser extraído de trechos curtos do sinal e, em geral, é combinado com as outras *features* já mencionadas para conferir maior robustez ao sistema. Ainda nesse trabalho, os autores fazem simulações com alguns modelos e concluem que um modelo composto pela combinação de GMM com RNP, ou seja, um sistema onde supervetores oriundos de um sistema GMM são utilizados para alimentar uma RNP, apresenta desempenho superior se comparado com um modelo que utiliza RNP alimentada diretamente com *features* MFCC. os resultados são 74,22% e 66,75% de taxa de acerto, para os sistemas GMM-RNP e MFCC-RNP, respectivamente.

Seguindo na linha de análise quanto a duração das amostras dos sinais de áudio submetidos aos sistemas que utilizam arquitetura com RNP para aplicações de AL, em Garain, Singh e Sarkar (2021), é feita uma relação direta de proporcionalidade entre a duração de tais amostras e a precisão obtida. Estabelece-se que quanto maior a duração, melhor serão os resultados, justificando simplesmente pelo fato de que amostras maiores fornecem uma quantidade maior de informação. Vale ressaltar que o sistema abordado nesse caso, possui a finalidade de identificar automaticamente a linguagem (idioma) utilizada pelo locutor, então devemos considerar que a natureza da aplicação neste caso requer também trechos mais longos do sinal da fala para que certos padrões possam ser detectados.

Avançando nas pesquisas e buscando trabalhos que tenham uma maior relação com o problema da estimação de faixa etária, fica clara a dificuldade que existe para que seja possível

determinar com precisão e de forma exata a idade do locutor. Até onde houve acesso na literatura, não foram encontrados registros de nenhum trabalho que mostrou a capacidade de obtenção da idade precisa do locutor, seja utilizando aprendizado profundo ou outros modelos. Em geral, o que se obtém são faixas etárias como crianças, jovens, adultos e idosos, dentro das quais se procura fazer a separação por gênero também.

Existem diversos trabalhos de pesquisa que buscam determinar a idade de um indivíduo utilizando técnicas de aprendizado profundo, mas que não se restringem apenas a estudar os sinais de voz. Em Zaghbani, Boujneh e Bouhlel (2018), por exemplo, é apresentado um método de estimação de de idade através de imagens faciais de pessoas, porém utilizando um outro tipo de rede neural, os *autoencoders*. Esse método também é considerado uma técnica de aprendizado profundo, tendo em vista a complexidade de se extrair as informações necessárias de imagens para esse tipo de análise, porém como também discutido em Moyse (2014), os resultados obtidos ao se estimar a idade por imagens da face são superiores se comparados aos sistemas que utilizam apenas o sinal da voz do indivíduo, independente de qual modelo seja utilizado. A autora estima que, de modo geral, erros de estimativa de idade a partir de imagens de faces são na ordem de 5 anos, enquanto ao se analisar a voz, esse erro aumenta para 10 anos. Temos ainda uma abordagem diferente em Ilyas, Othmani e Nait-ali (2020), onde a análise para estimação etária é feita com base na percepção auditiva do indivíduo. Nesse caso particular, os autores demonstram a possibilidade de alcançar erros médios absolutos (MAE, do inglês *Mean Absolute Error*) entre 4,7 a 10 anos, o que se traduz em uma precisão de 50%, aproximadamente.

No trabalho apresentado por Graves, Mohamed e Hinton (2013) é apresentado um modelo para AL onde duas Redes Neurais Recorrentes (RNR's) são treinadas em conjunto. Durante os testes detalhados pelo autor, fica clara a vantagem da utilização de redes profundas para esse tipo de aplicação visto que, em uma das simulações, a taxa de erro decresce de 23,9% para 18,4% ao aumentar o número de *hidden layers* de 1 para 5.

Uma pesquisa bem detalhada sobre a aplicação de aprendizado profundo para estimação de faixa etária, foi elaborada por Qawaqneh, Mallouh e Barkana (2017). Nesse trabalho foram utilizadas duas redes neurais profundas, conectadas entre si, de forma que cada rede fosse treinada com diferentes conjuntos de características extraídos dos mesmos dados de entrada. Com essa configuração, o erro gerado pelo primeiro conjunto de características é diferente do erro gerado pelo segundo conjunto, os erros são então somados e utilizados para ajustar os pesos e vieses da rede. Os autores acreditam que dessa forma, o erro resultante é capaz de ajustar os parâmetros da rede de forma mais precisa. Com essa configuração, foi possível atingir uma precisão global de 56,06%, superior as demais variações ensaiadas para elaboração do artigo.

Por fim, deve-se ressaltar novamente a pesquisa do Prof. Ivandro Sanches (2019), que aborda o problema de estimação de faixa etária utilizando GMM e i-Vectors. Ainda para esse trabalho foi desenvolvida ainda uma técnica de estimação de pitch utilizando o algoritmo de Viterbi, e assim obtendo uma taxa de acerto, considerando 3 grupos etários (criança, homem,

mulher), de 87.8%. As simulações foram realizadas utilizando MFCC para caracterizar as amostras de áudio, visto que trata-se da técnica de extração de *features* mais consagrada para processamento desse tipo de sinal.

3 METODOLOGIA

3.1 Base de Dados

Corpus aGender da *Deutsche Telekom AG Laboratories*, criado por Burkhardt et al. (2010).

A base de dados escolhida para este trabalho foi utilizada tanto em Sanches (2019), como também em outros trabalhos pesquisado durante a revisão bibliográfica. Trata-se de um banco composto por 65364 arquivos de áudio, que correspondem a um total de 47 horas de gravações telefônicas (somatória de arquivos que possuem, em média, 2.58 segundos de duração), realizadas por 954 indivíduos, com uma distribuição igualitária de gênero dentro de quatro grupos etários: crianças, jovens, adultos e sêniores/idosos. Com base nessa composição, o banco possui 7 classes, divididas conforme tabela abaixo:

Classe	Idade	Gênero
1	7 - 14	f, m
2	15 - 24	f
3	15 - 24	m
4	25 - 54	f
5	25 - 54	m
6	55 - 80	f
7	55 - 80	m

Tabela 1 – Classes

f: feminino / m: masculino

Para elaboração desse banco, cada participante selecionado teve que realizar seis chamadas telefônicas utilizando dispositivos móveis, alternando entre locais internos (*indoor*) e externos (*outdoor*) para obtenção de diferentes níveis de ruídos de ambiente. A cada ligação, o participante interagia com uma Unidade de Resposta Audível (URA, em inglês conhecida como *Interactive Voice Response* - IVR) e tinha que ler as declarações/sentenças pré definidas e providenciadas antecipadamente pela equipe do laboratório. Havia um intervalo de um dia entre uma ligação e outra, para que fosse possível obter maior variação nas vozes dos indivíduos.

Todas as gravações são disponibilizadas em extensão .raw, armazenadas com 8 bits, frequência de amostragem de 8kHz e codificação PCMA.

3.1.1 Base de treinamento

Da base total, foram separados 53076 arquivos, que correspondem a 770 indivíduos e 38.16 horas de gravação. Esse conjunto foi definido como base para treinamentos, onde as

gravações são disponibilizadas já com a indicação de a qual classe cada uma pertence. A tabela abaixo, mostra a divisão dos locutores e gravações (o artigo oficial indicado nas referências desse trabalho, se refere as gravações como *utterances*, traduzido como *declarações*) por classe:

Classe	**1**	**2**	**3**	**4**	**5**	**6**	**7**
#locutores	106	99	88	113	107	123	134
#gravações	6804	7360	6189	7844	6911	8575	9575

Tabela 2 – Distribuições no banco de treinamento

3.2 Sistema de Estimação Automática de Faixa Etária com Aprendizado Profundo

Para a execução das simulações e ensaios com base em aprendizado profundo, será utilizado uma RNC derivada do programa criado por Badine (2020) e publicada na página Web do Keras, o qual é uma API, em linguagem de programação Python, desenvolvida para abstração do framework Tensorflow, criado pelo Google Brain Team, e que possui vasta aplicação na área de Ciência de Dados.

A RNC original do programa em questão é treinada para classificar arquivos de áudio em 5 classes, cada uma correspondendo a um locutor diferente. Trata-se, por tanto, de um sistema de *Automatic Speech Recognition* (ASR), ou seja, Autenticação de Locutor, com o objetivo de identificar o locutor a partir de um trecho de fala. A análise do sinal é realizada no domínio da frequência, e por isso utiliza-se FFT para obtenção dos espectros. Ponto importante que deve ser observado no artigo, seria o fato de ter-se acrescido ruído artificialmente ao sinal de voz para que assim a rede tivesse mais dados para processar. Esse é um ponto de relevância que a pesquisa deve se aprofundar, a fim de investigar a influência do tamanho do banco de dados no desempenho do sistema.

Visto que a finalidade da RNC original seria a de reconhecimento de locutor, o sistema apresenta precisão de, aproximadamente, 98%, porém deve-se levar em conta que as etapas de treinamento e validação foram realizadas com a mesma base de dados, e que a precisão foi obtida sobre os resultados da validação. Em outras simulações, com diferentes dados, essa taxa de precisão deve cair.

Tomando então a mesma RNC como base, bem como o princípio de realizar a análise do sinal de voz a partir do seu espectro em frequência, foram realizadas algumas alterações no código para criar um sistema de estimação de faixa etária capaz de trabalhar com a mesma base aGender da Deutsche Telekom Laboratories.

É importante lembrar que o corpus aGender é dividido em 7 classes: crianças (sem separação de gênero), jovem/masculino, jovem/feminino, adulto/masculino, adulto/feminino, sênior/masculino, sênior/feminino.

Para que fosse possível realizar ensaios e simulações iniciais de forma mais simples, o programa criado, e utilizado para o início da pesquisa, considera apenas 6 classes, portanto os resultados disponibilizados nessa dissertação, não consideram a classe "1"*crianças*. Vale ressaltar no entanto, que no decorrer dos trabalhos de pesquisa o código será melhorado para que as 7 classes sejam processadas pela RNC.

Na sequência, são apresentados alguns trechos do programa criado.

Figura 1 – Código em Python: declaração das bibliotecas e variáveis relevantes

```
import os
import pandas as pd
import numpy as np
import tensorflow as tf
from tensorflow import keras


DATASET_ROOT = os.path.join(os.path.expanduser("~"), '''Documents/Mestrado/tra
                            balho/redeNeural/agender_distribut
                            ion/''')
VALID_SPLIT = 0.1  # Percentage of samples to use for validation
SAMPLING_RATE = 16000
SHUFFLE_SEED = 43
BATCH_SIZE = 128
EPOCHS = 100
```

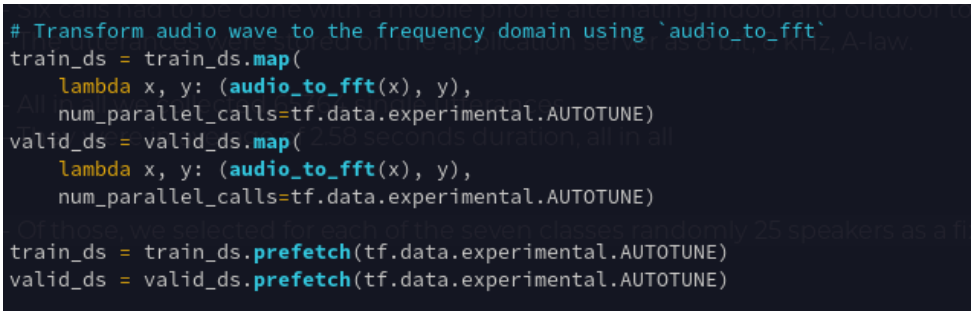
Trecho do código que mostra todas as bibliotecas utilizadas no código atual, e parâmetros importantes para a modelagem e treinamento da RNC.

Figura 2 – Código em Python: função audio_to_fft

```
def audio_to_fft(audio):
    # Since tf.signal.fft applies FFT on the innermost dimension,
    # we need to squeeze the dimensions and then expand them again
    # after FFT
    audio = tf.squeeze(audio, axis=-1)
    fft = tf.signal.fft(tf.cast(tf.complex(real=audio, imag=tf.zeros_like(audio
                                                                          )
                                           ), tf.complex64))
    fft = tf.expand_dims(fft, axis=-1)
    # Return the absolute value of the first half of the FFT
    # which represents the positive frequencies
    return tf.math.abs(fft[:, : (audio.shape[1] // 2), :])
```

Função para obtenção da FFT de cada arquivo de áudio a ser submetido à RNC.

Figura 3 – Código em Python: preparação dos dados para treinamento do modelo



```python
# Transform audio wave to the frequency domain using `audio_to_fft`
train_ds = train_ds.map(
    lambda x, y: (audio_to_fft(x), y),
    num_parallel_calls=tf.data.experimental.AUTOTUNE)
valid_ds = valid_ds.map(
    lambda x, y: (audio_to_fft(x), y),
    num_parallel_calls=tf.data.experimental.AUTOTUNE)

train_ds = train_ds.prefetch(tf.data.experimental.AUTOTUNE)
valid_ds = valid_ds.prefetch(tf.data.experimental.AUTOTUNE)
```

Trecho do código onde os dados de treinamento e validação são transformados de áudio para frequência.

Figura 4 – Código em Python: definição do modelo

```python
# MODEL DEFINITION


def residual_block(x, filters, conv_num=3, activation="relu"):
    # Shortcut
    s = keras.layers.Conv1D(filters, 1, padding="same")(x)
    for i in range(conv_num - 1):
        x = keras.layers.Conv1D(filters, 3, padding="same")(x)
        x = keras.layers.Activation(activation)(x)
    x = keras.layers.Conv1D(filters, 3, padding="same")(x)
    x = keras.layers.Add()([x, s])
    x = keras.layers.Activation(activation)(x)
    return keras.layers.MaxPool1D(pool_size=2, strides=2)(x)


def build_model(input_shape, num_classes):
    inputs = keras.layers.Input(shape=input_shape, name="input")

    x = residual_block(inputs, 16, 2)
    x = residual_block(x, 32, 2)
    x = residual_block(x, 64, 3)
    x = residual_block(x, 128, 3)
    x = residual_block(x, 128, 3)

    x = keras.layers.AveragePooling1D(pool_size=3, strides=3)(x)
    x = keras.layers.Flatten()(x)
    x = keras.layers.Dense(256, activation="relu")(x)
    x = keras.layers.Dense(128, activation="relu")(x)

    outputs = keras.layers.Dense(num_classes,
                                 activation="softmax", name="output")(x)

    return keras.models.Model(inputs=inputs, outputs=outputs)


model = build_model((SAMPLING_RATE // 2, 1), len(class_labels))

model.summary()

# Compile the model using Adam's default learning rate
model.compile(
    optimizer="Adam", loss="sparse_categorical_crossentropy",
    metrics=["accuracy"])
```

Funções e trechos do código que modelam a rede neural e seus parâmetros.

Figura 5 – Código em Python: treinamento e precisão do modelo

```
# TRAINING

history = model.fit(
    train_ds,
    epochs=EPOCHS,
    validation_data=valid_ds,
    callbacks=[earlystopping_cb, mdlcheckpoint_cb],
)


# Model 's precision
print(model.evaluate(valid_ds))
```

Trecho do código que executa o treinamento e mede a precisão do modelo.

Para treinamento do modelo foram utilizados 28120 arquivos de áudio da base de treino do corpus aGender, convertidos para o formato .wav. Além disso, conforme mencionado anteriormente, removeu-se ainda uma classe (crianças) dos dados para realizar as simulações iniciais. Dessa base, 90% dos dados foram utilizados para treinamento do modelo e 10% para validação. Utilizando um computador com as especificações indicadas na sequência, o modelo leva 4 horas para ser treinado. Considerando que está sendo utilizada a linguagem de programação Python, que não é compilada mas baseia-se em um interpretador, poderíamos esperar que esse tempo de treinamento diminuísse caso fosse utilizada outra linguagem como C, por exemplo. Mais pesquisas devem ser realizadas para avaliar se esse tempo obtido com o programa atual, pode ser considerado dentro dos padrões pela literatura (considerando a quantidade de dados e outros fatores).

- Processador: Intel(R) Core(TM) i7-8565U | CPU @ 1.80GHz
- Memória RAM: 8GB (DDR3)
- Armazenamento: 256GB (SSD)
- Placa gráfica: nVidia GM108M [GeForce MX110] | 0.97 - 0.99 GHz (velocidade CPU) / 1800 MHz (velocidade memória DDR3)

Finalizada a etapa de treinamento, é possível obter a precisão do modelo, que neste caso foi de 59.82%.

O planejamento para a sequência do trabalho, é estudar todos os aspectos e parâmetros desse sistema de forma a buscar formas de otimizar o desempenho do mesmo.

Figura 6 – Precisão obtida pelo modelo

```
print(model.evaluate(valid_ds))

88/88 [==============================] - 24s 258ms/step - loss: 1.0513 - accuracy: 0.5982
[1.0513083934783936, 0.5981507897377014]
```

Precisão obtida ao fim da etapa de treinamento da rede.

3.3 Sistema de Estimação Automática de Faixa Etária com GMM

O trabalho elaborado pelo Prof. Ivandro Sanches (2019) (FAPESP: 2016/18700-7), utiliza GMM e i-vector como técnicas de aprendizado de máquina. Outra diferença em relação ao trabalho proposto nessa dissertação, é que o sistema foi implementado em linguagem C, diferentemente do Python que está sendo utilizado na nova pesquisa.

Utiliza-se ainda a estimação de *pitch*, através do algoritmo de Viterbi, para conferir maior robustez ao sistema. O pitch é a frequência de vibração das cordas vocais durante a produção da voz, e funciona como parâmetro biométrico pois seu valor médio é função, por exemplo, do gênero e idade do indivíduo. Ao agregar o valor do pitch ao vetor de dados a ser processado, tem-se dados mais robustos para reforçar o padrão que precisa ser definido.

Nesse trabalho, foram realizados três tipos de experimentos conforme descrito a seguir:

- Exp 01: todo o sinal de voz de cada arquivo de áudio será processado e usado nos processos de treino, adaptação e teste
- Exp 02: apenas os quadros em que ocorre vibração das cordas vocais de cada arquivo de áudio serão processados.
- Exp 03: apenas os quadros em que ocorre vibração das cordas vocais de cada arquivo de áudio serão processados e, adicionalmente, o valor estimado de pitch será adicionado ao vetor de padrões do quadro correspondente.

A seguir, serão compartilhados alguns dos resultados obtidos. Os valores indicados, se referem a ensaios realizados com base de treinamento e teste distintas, que ao todo somam **20492** arquivos de áudio.

	GMM	i-vector
Exp 01	41.4	46.4
Exp 02	44.5	46.4
Exp 03	48.0	47.8

Tabela 3 – Resultados globais em porcentagem de acerto.

		estimado						
		1	2	3	4	5	6	7
	1	**1354**	467	155	192	29	122	69
	2	510	**1398**	24	544	10	303	14
	3	4	12	**949**	33	583	65	454
real	4	155	944	48	**1263**	19	907	13
	5	2	4	681	21	**901**	62	895
	6	221	559	32	947	17	**1663**	37
	7	7	4	400	42	920	120	**2317**

Tabela 4 – Matriz de confusão referente ao Exp 03, considerando GMM.

Considerando apenas os quatro grupos etários (crianças, jovens, adultos, sêniores), a taxa de acerto obtida foi de 49.2%.

		estimado			
		criança	jovem	adulto	sênior
real	criança	**1354**	622	221	191
	jovem	514	**2383**	1170	836
	adulto	157	1677	**2204**	1877
	sênior	228	995	1926	**4137**

Tabela 5 – Matriz de confusão referente ao Exp 03, considerando GMM e quatro grupos etários.

Somatória dos valores da diagonal principal: 10078.

$$\frac{10078}{20492} x 100\% = 49.2\%.$$

Considerando apenas três grupos (crianças, homens e mulheres), a taxa de acerto obtida foi de 87.8%.

		estimado		
		crianças	mulheres	homens
real	crianças	**1354**	78	253
	mulheres	886	**8528**	214
	homens	13	363	**8100**

Tabela 6 – Matriz de confusão referente ao Exp 03, considerando GMM e três grupos etários.

Somatória dos valores da diagonal principal: 17982.

$$\frac{17982}{20492} x 100\% = 87.8\%.$$

Pretende-se realizar ensaios similares e adicionais com o sistema utilizando a RNC e então comparar os resultados para posterior discussão.

3.4 Métricas

Para avaliação de desempenho do sistema, deverão ser utilizados os parâmetros *accuracy*, *precision*, *recall* e *F1-score*. Esses parâmetros podem ser calculados através da quantidade de positivos verdadeiros (TP, true positives), negativos verdadeiros (TN, true negatives), negativos falsos (FN, false negatives) e positivos falsos (FP, false positives). De qualquer forma, as API's e bibliotecas Python que serão utilizadas possuem ferramentas que possibilitam a obtenção dessas métricas de forma imediata.

3.5 Desenvolvimento do Trabalho

3.5.1 Desenvolvimento do código (A1)

O código atual da rede neural, deve ser desenvolvido para trabalhar com as sete classes disponíveis no corpus aGender. Feito isso, ensaios e simulações serão executados para que se possa comparar com os resultados obtidos através da técnica de GMM. Nessa etapa, as simulações com a rede devem considerar arquivos distintos nas bases de treino e teste, diferentemente dos ensaios que foram realizados até o momento.

3.5.2 Adição de ruído externo (A2)

A etapa seguinte, consistirá em adicionar artificialmente ruído aos sinais de áudio para que a rede tenha mais dados para processar. Nesse ponto, espera-se iniciar uma investigação quanto a influência do tamanho do volume de dados no desempenho desse tipo de técnica de aprendizado profundo com sinais acústicos.

3.5.3 Busca por otimizações (A3)

Conforme progresso das atividades e obtenção de dados mais concretos quanto a capacidade da RNC para esse tipo de problema (estimação de faixa etária), planeja-se definir as limitações que por ventura tenham sido identificadas no sistema, além de pesquisar possíveis otimizações que possam ser indicadas para a aplicação que está sendo abordada nesse trabalho. Essa etapa deve ocorrer em paralelo com as demais atividades , após a banca de qualificação, e durante todo o trabalho de pesquisa.

4 CRONOGRAMA

A Tabela 7 apresenta o cronograma de execução das atividades desta proposta.

Tabela 7 – Cronograma das atividades previstas

Etapa	Meses											
	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez
Revisão bibliográfica	■	■	■	■	■	■	■	■	■	■	■	■
A0		■	■	■	■							
Banca de qualificação						■						
A1							■					
A2								■				
A3							■	■	■	■	■	
A4												■

A0: Implementação do programa inicial e realização das primeiras simulações

A1: Desenvolvimento do código com melhorias.

A2: Adição de ruído externo.

A3: Busca por otimizações.

A4: Elaboração da dissertação, análise e discussão dos resultados.

REFERÊNCIAS

BADINE, F. *Speaker Recognition*. 2020. Acesso: 08/05/2021. Disponível em: <https://keras.io/examples/audio/speaker_recognition_using_cnn/>. Citado na página 15.

BURKHARDT, F. et al. A database of age and gender annotated telephone speech. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), 2010. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2010/pdf/262_Paper.pdf>. Citado na página 14.

BUYUK, O.; ARSLAN, M. L. Combination of long-term and short-term features for age identification from voice. *Advances in Electrical and Computer Engineering*, v. 18, n. 2, p. 101–108, 2018. Citado na página 11.

DENG, L. Deep learning: from speech recognition to language and multimodal processing. *APSIPA Transactions on Signal and Information Processing*, v. 5, p. e1, 2016. Citado na página 10.

FAYEK, H. M.; LECH, M.; CAVEDON, L. Evaluating deep learning architectures for speech emotion recognition. *Elsevier Neural Networks*, v. 92, p. 60–68, 2017. Citado na página 11.

GARAIN, A.; SINGH, P. K.; SARKAR, R. Fuzzygcp: A deep learning architecture for automatic spoken language identification from speech signals. *Elsevier Expert Systems with Applications*, v. 168, p. 114416, 2021. Citado na página 11.

GRAVES, A.; MOHAMED, A.; HINTON, G. *Speech Recognition with Deep Recurrent Neural Networks*. 2013. <https://arxiv.org/abs/1303.5778>. Accesso: 05/06/2021. Citado na página 12.

ILYAS, M.; OTHMANI, A.; NAIT-ALI, A. Auditory perception based system for age classification and estimation using dynamic frequency sound. *Multimedia Tools and Applications*, v. 79, p. 21603–21626, 2020. Citado na página 12.

KALLURI, S. B.; VIJAYASENAN, D.; GANAPATHY, S. Automatic speaker profiling from short duration speech data. *Elsevier Speech Communication*, v. 121, p. 16–28, 2020. Citado na página 11.

MOYSE, E. Age estimation from faces and voices: A review. *Psychologica Belgica*, v. 54, n. 3, p. 255–265, 2014. Citado na página 12.

QAWAQNEH, Z.; MALLOUH, A. A.; BARKANA, B. D. Age and gender classification from speech and face images by jointly fine-tune deep neural networks. *Elsevier Expert Systems with Applications*, v. 85, p. 76–86, 2017. Citado na página 12.

SANCHES, I. *Estimação automática de faixa etária pelo processamento do sinal de voz*. São Bernardo do Campo, SP, 2019. 14 p. Processo FAPESP 2016/18700-7. Citado 5 vezes nas páginas 8, 9, 12, 14 e 19.

ZAGHBANI, S.; BOUJNEH, N.; BOUHLEL, M. S. Age estimation using deep learning. *Elsevier Computers and Electrical Engineering*, v. 68, p. 337–347, 2018. Citado na página 12.

ZHANG, Z. et al. *Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments*. 2018. <https://arxiv.org/abs/1705.10874v3>. Accesso: 05/06/2021. Citado na página 10.