Matheus Carlos Lima e Silva

Envelhecimento de características da voz

Goiânia
19 de dezembro de 2024

UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE ENGENHARIA ELÉTRICA, MECÂNICA E DE COMPUTAÇÃO

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Matheus Carlos Lima e Silva

Título do trabalho: **Envelhecimento de características da voz**

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Carlos Galvao Pinheiro Junior**, **Professor do Magistério Superior**, em 19/12/2024, às 10:07, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Matheus Carlos Lima E Silva**, **Discente**, em 19/12/2024, às 12:37, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Matheus Carlos Lima e Silva

Envelhecimento de características da voz

Projeto Final de Curso apresentado como requisito parcial para a obtenção do título de Bacharel em Engenharia no curso de graduação em Engenharia de Computação da Escola de Engenharia Elétrica, Mecânica e de Computação da Universidade Federal de Goiás.

Universidade Federal de Goiás
Escola de Engenharia Elétrica, Mecânica e de Computação

Prof. Orientador: Dr. Carlos Galvão Pinheiro Júnior

Goiânia
19 de dezembro de 2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Silva, Matheus Carlos Lima e
Envelhecimento de características da voz [manuscrito] / Matheus Carlos Lima e Silva. - 2024.
XV, 15 f.: il.

Orientador: Prof. Dr. Carlos Galvão Pinheiro Júnior.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de Computação (EMC), Engenharia da Computação, Goiânia, 2024.
Bibliografia. Apêndice.
Inclui gráfico, tabelas.

1. Envelhecimento Vocal. 2. Classificação Etária. 3. Conversão de Voz. 4. Processamento de Fala. 5. Transferência de Estilo. I. Júnior, Carlos Galvão Pinheiro, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE ENGENHARIA ELÉTRICA, MECÂNICA E DE COMPUTAÇÃO

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Aos 19 dias do mês de Dezembro do ano de 2024 iniciou-se a sessão pública de defesa do Projeto de Final de Curso (PFC2) intitulado "**Envelhecimento de características da voz"**, de autoria de **Matheus Carlos Lima e Silva**, do curso de Engenharia de Computação, da Escola de Engenharia Elétrica Mecânica e omputação da UFG. Os trabalhos foram instalados pelo Prof. Dr. Carlos Galvão Pinheiro Júnior (EMC/UFG) com a participação dos demais membros da Banca Examinadora: Me. Lucas Rafael Stefanel Gris e Prof. Dr. Alisson Assis Cardoso. Após a apresentação, a banca examinadora realizou a arguição do(a) estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de **9,3** , tendo sido o PFC considerado APROVADO**.**

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Carlos Galvao Pinheiro Junior**, **Professor do Magistério Superior**, em 19/12/2024, às 11:29, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Alisson Assis Cardoso**, **Professor do Magistério Superior**, em 19/12/2024, às 11:44, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Lucas Rafael Stefanel Gris**, **Usuário Externo**, em 19/12/2024, às 13:09, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5053694** e o código CRC **D741E2CA**.

Referência: Processo nº 23070.046367/2024-27 SEI nº 5053694

Envelhecimento de características da voz

Matheus Carlos Lima e Silva, *EMC*

Resumo—This work presents an approach based on linear models to recover the vocal identity of individuals with dysphonia, using recordings made before the onset of the disorder. The method applies transformations aimed at approximating the representation of vocal age to what the individual would have had if the disorder had not compromised voice quality. The classification models achieved a mean absolute error of 3.49 years in age prediction, although the analysis was limited by an imbalanced dataset with low representativeness across different age groups. Furthermore, a simple linear model applied to speaker embeddings showed promising performance within the scope of the available data. Despite the demonstrated potential, further validation with larger and more diverse datasets is needed to ensure its practical applicability and adaptation to individual particularities.

Index Terms— *Voice Aging*, *Speaker Identity*, *Style Transfer*, *Age Classification*, *Voice Conversion*, *Speech Processing*.

Resumo— Este trabalho apresenta uma abordagem baseada em modelos lineares para recuperar a identidade vocal de indivíduos com disfonia, utilizando gravações realizadas antes do surgimento do distúrbio. O método aplica transformações destinadas a aproximar a representação da idade vocal àquela que o indivíduo teria caso o distúrbio não tivesse comprometido a qualidade da voz. Os modelos de classificação alcançaram um erro absoluto médio de 3,49 anos na predição de idade, embora a análise tenha sido limitada por um conjunto de dados desbalanceado e com baixa representatividade entre diferentes faixas etárias. Além disso, um modelo linear simples aplicado a *speaker embeddings* apresentou desempenho promissor no escopo dos dados disponíveis. Apesar do potencial demonstrado, é necessária uma validação adicional com datasets mais amplos e diversificados para assegurar sua aplicabilidade prática e adaptada às particularidades individuais.

Palavras-chave— Envelhecimento Vocal, Identidade Vocal, Transferência de Estilo, Classificação Etária, Conversão de Voz, Processamento de Fala.

1 Introdução

A voz humana desempenha um papel fundamental na comunicação, influenciando diretamente as interações sociais e o desempenho em atividades profissionais. A disfonia, definida como uma alteração na qualidade da voz causada por fatores funcionais ou estruturais, afeta significativamente a capacidade comunicativa dos indivíduos (Neighbors & Song, 2020). Estudos, como o de Smith et al. (1995), demonstram que pessoas com distúrbios vocais frequentemente relatam impactos negativos na qualidade de vida, incluindo diminuição da auto-estima e dificuldades em carreiras que exigem comunicação eficaz. Além disso, conforme apresentado por Ana Lúcia Spina et al. (2009), a gravidade desses impactos independe do uso profissional da voz.

A disfonia pode ocorrer em qualquer idade e é experienciada por 1 a cada 13 adultos anualmente (Smith et al., 1995). Embora seja uma ocorrência natural em idades avançadas, conhecida como presbifonia, também pode ser resultante de patologias ou condições pós-operatórias que causam alta severidade no comprometimento vocal (Neighbors & Song, 2020). Problemas como dificuldades de conversação em ambientes ruidosos e redução na qualidade das interações sociais são comumente relatados entre pessoas afetadas.

O envelhecimento da voz é um processo multifacetado que ocorre devido a mudanças anatômicas naturais nas estruturas do trato respiratório (Tarafder et al., 2012). Fatores externos, como condições médicas, tabagismo e desgaste vocal, também podem afetar a qualidade vocal em qualquer idade. Pesquisas indicam que, em idades avançadas, a voz masculina tende a se tornar mais aguda devido à rigidez da cobertura das cordas vocais, enquanto a voz feminina pode se tornar mais grave em função do afinamento da mucosa laríngea (Tarafder et al., 2012).

Este trabalho propõe uma abordagem para recuperar a identidade vocal de indivíduos que tiveram sua sonoridade prejudicada em determinado momento da vida. Ao aplicar técnicas de envelhecimento sintético da voz em gravações anteriores ao surgimento da disfonia, busca-se gerar uma representação vocal que corresponda à voz que o indivíduo teria atualmente, na ausência do distúrbio. Esta metodologia considera a pré-existência de um áudio não afetado, permitindo uma solução personalizada para a recuperação vocal.

A motivação para este estudo é pessoal e deriva da experiência própria com problemas vocais. Acredita-se que a aplicação de ferramentas e técnicas bem estabelecidas na literatura possa contribuir para a melhoria da qualidade de vida de muitas pessoas afetadas pela disfonia, restaurando não apenas a voz, mas também a confiança e eficácia na comunicação diária.

2 Referencial Teórico

Para o desenvolvimento de um estudo sobre envelhecimento vocal e preservação de identidade sonora, é essencial compreender alguns conceitos e técnicas fundamentais que embasam a análise e o processamento de sinais de áudio. Esta seção apresentará os principais termos e métodos utilizados, desde conceitos acústicos básicos, como timbre e altura, até técnicas de aprendizado profundo aplicadas à representação e transformação de dados de áudio.

No contexto do envelhecimento da voz, técnicas de processamento de áudio e aprendizado de máquina são especialmente relevantes. O envelhecimento vocal impacta tanto características subjetivas, perceptíveis como o timbre e a altura, quanto as características objetivas, identificáveis em representações visuais do áudio, como os espectrogramas. Entender esses conceitos possibilita interpretar mudanças acústicas e visuais na voz, elementos que são cruciais na criação de modelos que mantêm a identidade vocal ao simular variações etárias.

Com essa base teórica, abordar-se-á conceitos fundamentais como vocoder, espectrograma e a escala Mel, além de técnicas de extração de características como MFCC e modelos de misturas gaussianas (GMM). Avançaremos para tecnologias de aprendizado profundo, incluindo redes neurais convolucionais (CNNs), *autoencoders* e redes adversárias generativas (GANs), que são cruciais para modelar, classificar e transformar dados de áudio. Esses conceitos, além de formarem a base para a construção de modelos de envelhecimento vocal, fornecem um entendimento essencial para o desenvolvimento de sistemas que visam preservar a identidade vocal e a autenticidade sonora ao longo de diferentes faixas etárias.

2.1 Conceitos de Áudio e Fala

Três aspectos básicos para categorizar uma fonte sonora são o tibre, a intensidade e a altura.

O Timbre é um conjunto de características únicas de um som que permite distinguir sua fonte das demais. Ele é normalmente determinado pela composição espectral do som, relacionando-se ao formato e padrões contidos nela (Lazzarini 1998).

Já a intensidade refere-se à energia do som e está relacionada à amplitude da onda sonora, determinando o volume para o ouvinte (Lazzarini 1998).

Por fim, a altura é uma qualidade do som relacionada à frequência fundamental do sinal acústico. Ele é o responsável pela escala de agudo à grave percebida aos ouvidos. Sua determinação está relacionada à análise de segmentos curtos de fala, aplicando-se autocorrelação ou cepstrum. A autocorrelação utiliza de picos na periodicidade da fala para encontrar a frequência fundamental. Já o cepstrum analisa picos, mas de uma representação que separa informação do trato vocal da fonte de excitação, a voz (Rabiner & Schafer, 2007).

2.2 Técnicas de Representação e Extração de Features em Áudios

2.2.1 Vocoder

Vocoder (*voice coder*) é um tipo de codificador de voz que utiliza modelos de produção e percepção da fala para representar de forma eficiente sinais da voz (Rabiner & Schafer, 2007). Ele é amplamente utilizado em tarefas de sintetização de fala, podendo ser por manipulação de parâmetros como a frequência fundamental (F0) e o envelope espectral (Morise et al., 2016) (Kawahara, 2006), ou por uma rede neural sem necessidade de parametrização, como na WaveNet (Van Den Oord et al., 2016).

2.2.2 Espectrograma

Espectrograma é uma representação visual da Transformada de Fourier de um sinal ao longo do tempo, exibindo a variação da amplitude do sinal em função da frequência e do tempo, proporcionando uma análise simplificada para componentes de frequência de um sinal sonoro (Rabiner & Schafer, 2007).

Já o mel espectrograma é uma variante do espectrograma que utiliza a escala mel para destacar as frequências mais próximas à percepção humana. Ele é bastante usado para tarefas envolvendo identificação de locutores e análise musical (Friedrich, 2024).

2.2.3 GMM (Gaussian Mixture Model)

O GMM é um modelo probabilístico que representa um conjunto de dados como uma combinação de vários subconjuntos, cada uma delas seguindo uma distribuição normal (Binu & Rajakumar, 2021). Devido à representação em distribuição normal, o GMM pode ser usado para comparar fontes sonoras diferentes em um mesmo contexto mesmo em uma complexa variação de características acústicas (Doi et al., 2010).

2.2.4 MFCC (Mel Frequency Cepstral Coefficient)

MFCC é um método de extração de *features* capaz de extrair informações do espectro de frequências de um sinal de áudio (Abdul & Al-Talabani, 2024). As *features* são bastante úteis em casos de reconhecimento automático de fala e análise musical, pois tem em sua modelagem uma separação entre fonte sonora e a estrutura que modela o som em sua passagem, chamada de filtro. Ao analisar essas estruturas separadamente, a tarefa de identificar características específicas do sinal, como o conteúdo linguístico ou timbre, torna-se mais eficiente e robusta. (Friedrich, 2024).

2.2.5 Speaker embeddings

Speaker embeddings são representações vetoriais fixas que capturam as características distintivas da voz de um locutor. Elas são amplamente empregadas em tarefas como reconhecimento e diarização de locutores, sendo utilizadas para calcular similaridades entre amostras por meio de técnicas como a similaridade cosseno ou análise discriminante linear probabilística (PLDA). Além disso, os embeddings de locutor têm aplicações em outras áreas do processamento de fala, incluindo adaptação de locutores em reconhecimento de fala, modelagem de locutores em síntese de texto para fala e conversão de voz (Wang et al., 2022).

2.3 Aprendizado de Máquina

2.3.1 Redes neurais

Redes neurais são um tipo de inteligência artificial inspiradas no funcionamento do cérebro humano. Elas consistem em unidades conectadas, chamadas neurônios, e os valores numéricos atribuídos às conexões, conhecidos como pesos, determinam a influência de um neurônio sobre outro no processamento dos sinais (Islam et al., 2019).

O uso das redes neurais nas tarefas de fala trouxe benefícios ao captar melhor as relações não lineares de dados da voz (Sisman et al.,2021).

2.3.2 Aprendizado Profundo

Aprendizado Profundo é uma subárea do aprendizado de máquina que se utiliza de várias camadas de redes neurais ou processamento de dados para ter uma capacidade de generalização maior que modelos tradicionais. Ele é adequado para grandes volumes de dados, apresentando melhor desempenho nesses casos (Sarker, 2021). A introdução do Aprendizado Profundo nos *pipelines* de conversão de voz levou a uma melhora na qualidade de áudio gerado, assim como visto na utilização de vocoders baseados em Aprendizado Profundo, como a WaveNet (Van Den Oord et al., 2016).

2.3.3 Autoencoders

Autoencoders são redes neurais projetadas para aprender uma representação latente e de interesse dos dados, com o objetivo de reconstruí-los de forma próxima ao original. São compostos por duas partes: o codificador, que gera o espaço latente representativo do dado, e o decodificador, que tenta reconstruí-lo a partir da representação codificada (Bank, Koenigstein & Giryes, 2020). Como demonstrado pelo AutoVC (Qian et al., 2019), os *autoencoders* levam vantagens em relações à outras arquiteturas em tarefas de conversão *zero-shot* pela sua capacidade de generalização. Além disso, possuem uma arquitetura bem mais simplificada, permitindo inferências mais rápidas que modelos tradicionais.

2.3.4 GANs (Generative Adversarial Networks)

GANs são uma classe de modelos de aprendizado de máquina compostos por duas redes neurais: uma que cria amostras falsas a partir de ruído aleatório, tentando assemelhar a um conjunto de dados real, e outra que determina se a amostra gerada pela primeira pode pertencer ao conjunto de dados real. Durante o treinamento, a rede de geração (geradora) tende a criar amostras cada vez mais próximas ao conjunto real e a de detecção (discriminadora) tende a refinar seu processo de distinção entre o real e o gerado (Goodfellow et al., 2014).

A introdução de GANs no contexto de síntese de fala trouxe uma melhora no tempo de inferência, com a HIFI-GAN (Kong et al., 2020) e suas posteriores variantes, como a SIFI-GAN (YONEYAMA et al., 2023), apresentando tempo de sintetização aprimorado em algumas dezenas de vezes se comparadas a outros vocoders de seu tempo, como a WaveNet (Van Den Oord et al., 2016) e WaveGlow (PRENGER et al., 2018).

2.3.5 CNNs (Convolutional Neural Networks)

CNNs são um tipo de arquitetura de rede neural projetada para processar dados estruturados espacialmente. Elas extraem e aprendem características hierárquicas por meio de técnicas de *pooling* e convolução (Alzubaidi et al., 2021). São bastante utilizadas para captar informações de espectrogramas ou mel espectrogramas, sendo muitas vezes combinadas com outras arquiteturas, como no AutoVC (Qian et al., 2019).

2.3.6 RNNs (Recurrent Neural Networks)

As RNNs são uma classe de arquiteturas neurais projetadas para processar sequências de dados ao longo do tempo, permitindo que informações de estados anteriores influenciem as decisões atuais. Essa característica é obtida por meio de conexões recorrentes em sua estrutura, que permitem o armazenamento de informações contextuais de entradas passadas (Schmidt, 2019).

As RNNs têm sido amplamente utilizadas em tarefas que envolvem dados sequenciais, como modelagem de linguagem, reconhecimento de fala e predição de séries temporais. No entanto, devido à propagação recorrente de informações, essas redes enfrentam desafios como o gradiente desaparecente ou explosivo, que limitam sua capacidade de aprendizado em dependências de longo prazo (Schmidt, 2019).

2.3.7 LSTM (Long Short-Term Memory)

As redes LSTM são uma variante das redes neurais recorrentes (RNN) projetadas para resolver problemas relacionados ao gradiente e à incapacidade de capturar dependências de longo prazo. Cada unidade LSTM é dividida em três elementos principais: o que controla quais informações da entrada são relevantes para o estado da célula; o que regula quais informações do estado anterior devem ser descartadas; e o que decide quais informações do estado da célula serão propagadas para o próximo passo temporal. Essa arquitetura permite que as LSTM aprendam padrões em dados sequenciais, tornando-as adequadas para tarefas que envolvem dependências temporais complexas, como processamento de fala e séries temporais (Hochreiter e Schmidhuber, 1997).

3 Revisão Bibliográfica

Um trabalho relacionado encontrado foi o de Wilson et al. (2021), que aborda o envelhecimento conjunto da imagem facial e do áudio do falante. Nele, uma rede de classificação baseada em camadas convolucionais e lineares processa o mel-espectrograma do áudio e a imagem do rosto para predizer a idade, enquanto uma rede GAN aplica envelhecimento tanto ao áudio quanto à imagem. O modelo de classificação alcançou acurácias de 24,7% para faixas etárias de 10 anos e 46% para faixas de 25 anos usando apenas áudio, enquanto a abordagem multimodal (áudio e visual) elevou esses valores para 26,1% e 52,7%, respectivamente.

O dataset criado por eles, assim como o desenvolvido por Hechimi et al. (2021), foi utilizado no presente trabalho por permitir a obtenção de pares de áudio de um mesmo falante em diferentes idades.

Indo na direção de abordagens para o problema, destacam-se também trabalhos focados na transferência de estilo de fala e na classificação de idade em áudios. Uma técnica notável é o AutoVC (Qian et al., 2019), uma rede *autoencoder* composta por dois codificadores para processar espectrogramas de áudio: um gerando o *speaker embedding*, que representa a identidade acústica, como o timbre, e outro gerando o *content embedding*, que captura informações fonéticas e de prosódia. Este modelo torna-se relevante ao possibilitar uma reconstrução fiel da voz, assegurando a preservação de características identitárias enquanto modifica o estilo vocal, como a idade.

Nesse sentido, o uso de *autoencoders* para separar identidade do locutor do conteúdo permite que o envelhecimento vocal seja tratado como uma transformação de estilo,

aplicando mudanças na idade sem perder a autenticidade do timbre original do falante. Em classificação de idade, Hechmi et al. (2021) também exploram técnicas que testam CNN 1-D com MFCC extraído do áudio, contribuindo para modelos mais robustos e específicos que utilizam a identidade de falante.

No objetivo de preservação e restauração da identidade de voz, pesquisas como a de Zhao et al. (2020) abordam a reconstrução da identidade vocal para pacientes com Esclerose Lateral Amiotrófica (ELA), controlando distorções espectrais para transformar uma voz disfônica em uma voz normal. De forma semelhante, Doi et al. (2010) utilizaram modelos de Mistura de Gaussianas (GMM) junto a parâmetros acústicos para modificar uma voz assistida por dispositivo, fazendo-a se aproximar da voz natural pré-disfonia.

Tais abordagens são essenciais no contexto do envelhecimento vocal e disfonia, pois buscam não apenas alterar o áudio, mas restaurar uma voz que o indivíduo teria naturalmente.

4 METODOLOGIA

Neste estudo, buscou-se explorar a aplicação de técnicas de envelhecimento sintético em gravações de voz, com o objetivo de aproximar características acústicas relacionadas à idade em gravações de falantes. Mais especificamente, foram realizadas transformações em *speaker embeddings* representativos ao timbre. A metodologia adotada envolve várias etapas principais: construção e preparação do dataset, processamento dos dados, definição da arquitetura dos modelos de classificação e envelhecimento e procedimentos de treinamento e avaliação. A avaliação dos modelos foi conduzida utilizando um classificador de idade para analisar o desempenho do envelhecimento vocal gerado, considerando como métrica a proximidade das características vocais com a idade-alvo definida.

4.1 Construção e Preparação do Dataset

4.1.1 Fontes de Dados

Para a realização dos experimentos, compilou-se um dataset composto por áudios de falantes em diferentes faixas etárias, provenientes de duas fontes principais:

- **VoxCeleb Dataset**: Utilizamos o VoxCeleb (Nagrani et al., 2019), um corpus público e amplamente utilizado que contém milhares de horas de gravações de fala de celebridades extraídas de vídeos do YouTube. A partir dos trabalhos de Wilson et al.(2021) e Hechmi et al.(2021), extraímos informações adicionais de idade e identidade dos falantes, resultando em 3.283 áudios que se encaixavam na nossa abordagem.
- **Coleta Manual de Áudios**: Complementamos o dataset com 168 áudios coletados manualmente de filmes e entrevistas disponíveis no YouTube. Esses áudios foram selecionados criteriosamente para incluir falantes com registros vocais em diferentes idades, permitindo a criação de pares correspondentes.

4.1.2 Seleção de Falantes

Para assegurar a qualidade e relevância dos dados selecionamos apenas falantes com áudios disponíveis em diferentes idades, visando criar pares de áudio que representassem o mesmo indivíduo em momentos distintos da vida.

4.1.3 Formação dos Pares de Áudio-Idade

O dataset final consistiu em:

- **Total de Áudios**: 3.451 (3.283 do VoxCeleb e 168 da coleta manual). A Figura 2 ilustra a distribuição de idades no dataset, evidenciando um desbalanceamento pela concentração de amostras na faixa etária entre 20 e 40 anos, enquanto faixas em idades nos extremos apresentam uma representação consideravelmente menor.
- **Total de Pares**: 4.498 pares de áudio-idade, sendo 3.062 do VoxCeleb e 1.436 da coleta manual. A Figura 1 ilustra a dispersão de pares de idade, observe que há um bom volume com idade inicial baixa, o que é resultado do foco da coleta manual em atores com início de carreira na infância.

A dispersão dos pares de idade revelou uma média de diferença de 5,8 anos e uma variância de 48 anos. É crível que essa baixa dispersão impactou diretamente a capacidade de generalização dos modelos, especialmente em idades extremas, onde há poucas amostras disponíveis. Esse cenário reflete, em parte, a recente popularização do hábito de registrar áudios, limitando a diversidade temporal dos dados disponíveis. Para mitigar esse problema, exploramos diferentes arquiteturas, buscando isolar a análise para diferentes faixas etárias.

4.2 Processamento dos Dados

Para preparar os áudios presentes na coleta manual, seguimos um *pipeline* de processamento em várias etapas que foi ilustrado na Figura 3.

4.2.1 Separação de Fontes Sonoras

Utilizamos o Demucs, uma rede neural profunda para separação de fontes musicais, para isolar a voz de outros componentes sonoros nos áudios, como música de fundo e ruídos ambientais.

4.2.2 Diarização de Áudio

Utilizamos o Pyannote, uma biblioteca especializada em processamento de fala, para realizar a diarização dos áudios, segmentando e extraindo apenas as partes correspondentes ao falante de interesse.

4.2.3 Redução de Ruído

Aplicamos técnicas de redução de ruído com o Noise-Reduce, atenuando ruídos de fundo sem distorcer a voz do falante, no intuito de melhorar a qualidade do sinal de áudio.

4.2.4 Remoção de Silêncios

Utilizamos o PyDub para remover silêncios prolongadas, normalizando a duração dos segmentos de fala e facilitando o processamento subsequente.

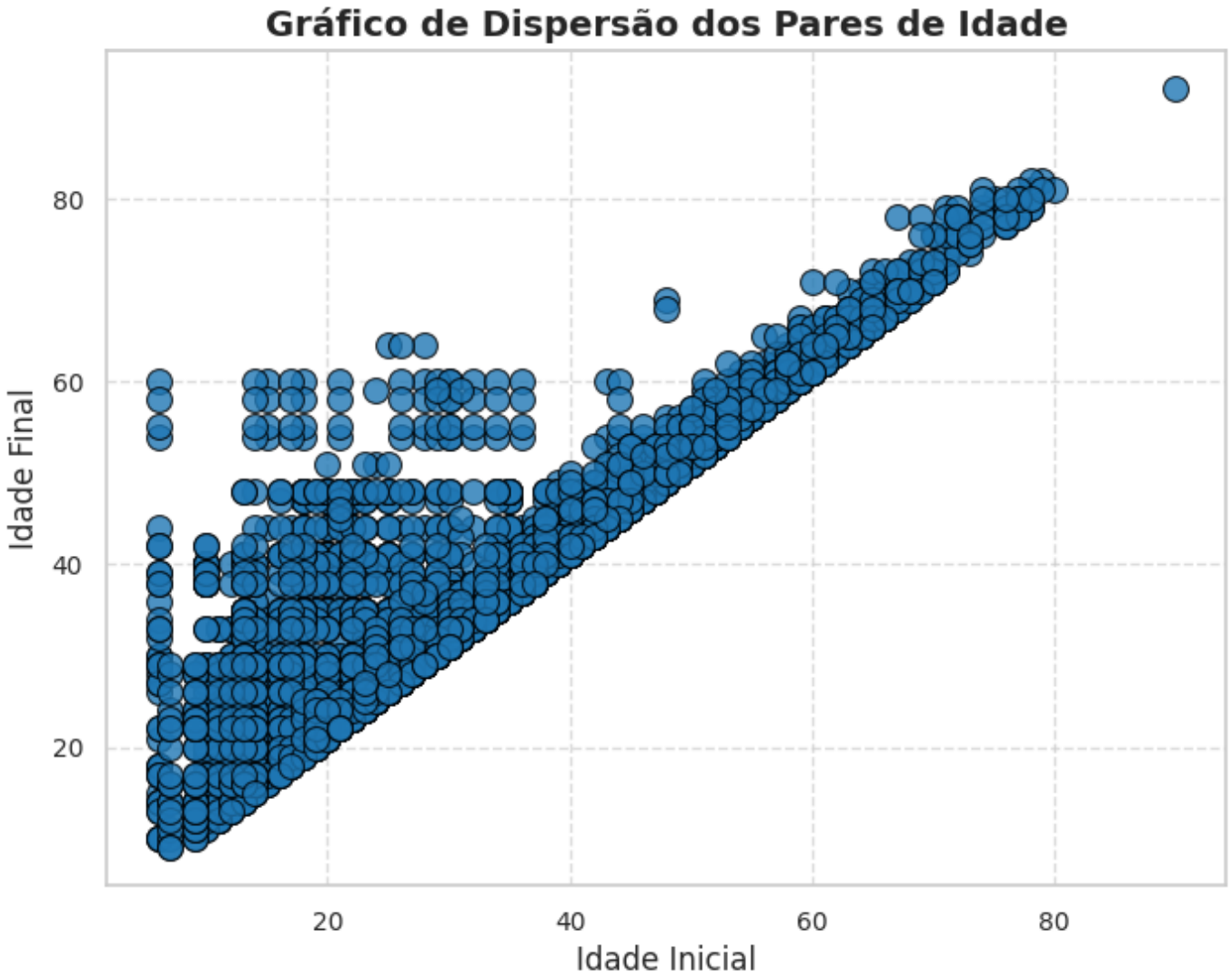

Figura 1. Dispersão de pares de idade no dataset

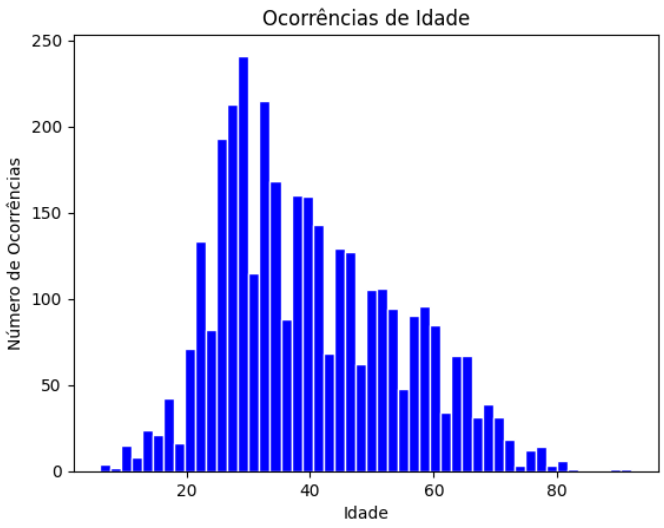

Figura 2. Distribuição de idade no dataset

4.2.5 *Supervisão Manual*

Revisamos manualmente os áudios processados com o Audacity, um editor de áudio digital, garantindo ajustes finos e a qualidade dos dados.

4.3 Extração de Características

4.3.1 *Padronização dos Áudios*

- **Reamostragem**: Todos os áudios foram reamostrados para uma taxa de 16.000 Hz, garantindo consistência nos dados.
- **Normalização**: Aplicamos normalização de amplitude para uniformizar os níveis de volume entre os diferentes áudios.

4.3.2 *Uso do Whisper-VITS*

Para este trabalho, utilizou-se o encoder de timbre disponibilizado no repositório do **Whisper-VITS**, um modelo de síntese de voz que combina as capacidades do Whisper (um modelo de reconhecimento de fala) com o VITS (*Variational Inference with adversarial learning for end-to-end Text-to-Speech*).

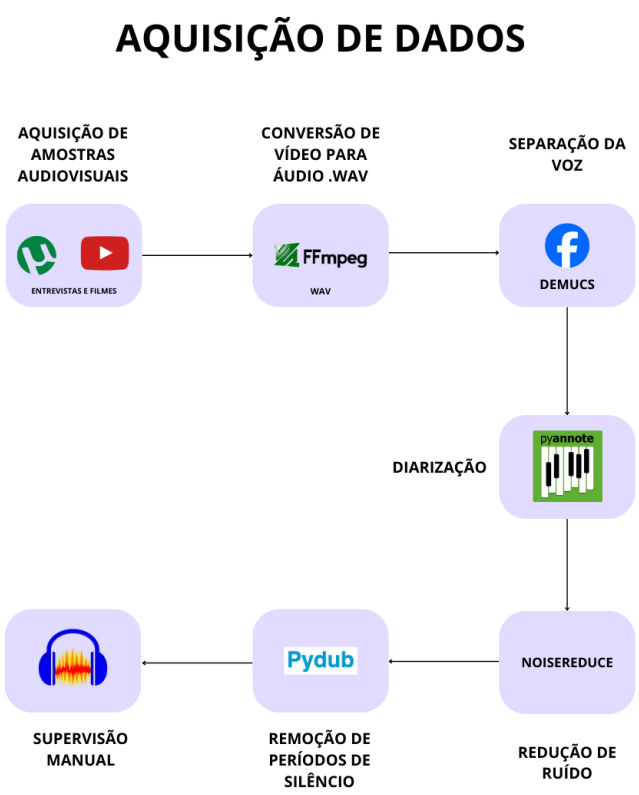

Figura 3. *Pipeline* do processamento de amostras coletadas

O encoder de timbre é baseado em uma arquitetura de LSTMs, composta por três camadas com projeções habilitadas, projetadas para extrair *embeddings* representativos das características acústicas dos locutores. A arquitetura utiliza 80 dimensões de entrada, 768 unidades ocultas em cada camada e gera *embeddings* em um espaço latente de 256 dimensões, permitindo a representação compacta e eficiente da identidade vocal.

O pré-treino do codificador foi realizado com a função de perda *Angular Prototypical Loss*, que otimiza diretamente a separação angular entre classes no espaço latente, garantindo maior discriminação entre locutores e coesão dentro das classes (Chung et al., 2020). Essa abordagem é uma evolução em relação às funções de perda tradicionais, como a *Generalized End-to-End* (GE2E) (Wan et al., 2018), que serviu como inspiração para o design inicial do encoder utilizado.

Este codificador foi essencial para criar um conjunto de dados que representasse as características sonoras de cada indivíduo em diferentes fases da vida. Os *embeddings* extraídos possibilita uma fácil manipulação e transformação da voz utilizando o *pipeline* de sintetização da voz do Whisper-VITS.

4.4 Definição da Arquitetura dos Modelos de Predição de Idade e Envelhecimento

Devido à natureza dos dados utilizados, que já têm informações latentes significativas, optou-se por abordagens simples, utilizando modelos lineares e convolucionais para as tarefas de predição de idade (classificação e regressão) e modelos lineares para a modelagem do envelhecimento, via regressão.

4.4.1 *Predição de Idade*

Conforme especificado anteriormente, o dataset utilizado baseia-se em um conjunto de pares de *speaker embeddings* de diferentes idades do mesmo falante. Para a tarefa de classificação de idade, foram testadas duas abordagens principais, cada uma empregando dois métodos distintos: classificação multiclasse e regressão linear.

- **Utilização de um único *speaker embedding***: Nesta abordagem, utilizamos apenas um *speaker embedding* para predizer a idade do falante. No entanto, esta estratégia foi rapidamente descartada, pois os resultados (Figura 4) demonstraram uma pobre generalização envolvendo os *speaker embeddings* e a idade dos falantes. Com o conjunto de dados utilizado, o erro médio na predição da idade foi de 11,07 anos.

  Para averiguar se a limitação era inerente ao método ou específica do dataset empregado, realizamos um teste adicional utilizando 408.043 áudios do *dataset* Common Voice (Ardila et al., 2020). Diferentemente do conjunto de dados inicial, o Common Voice apresenta as idades em faixas de 10 anos. Mesmo assim, a abordagem com um único *embedding* apresentou erros médios ainda maiores do que aqueles obtidos com o uso de pares de embeddings (método a ser apresentado posteriormente). Os resultados detalhados referentes ao teste com o Common Voice encontram-se no Apêndice.
- **Concatenação de *speaker embeddings* e idade inicial**: Nesta abordagem, utilizamos pares de *speaker embeddings* do mesmo falante juntamente com o valor da idade inicial, com o objetivo de predizer a idade do segundo *speaker embedding* (aquele cuja idade não foi concatenada). A ideia foi capturar as mudanças observadas nos embeddings devido ao envelhecimento. Treinamos modelos de regressão linear e classificação multiclasse para predizer idades entre 6 e 93 anos. Como resultado, um erro médio de 3,49 anos foi obtido (Figura 5).

Para ambos os métodos, realizamos um *Grid Search* envolvendo a combinação de camadas e parâmetros relevantes, treinando com a base de dados completa. Os testes foram conduzidos em falantes com idades variando de 6 a 80 anos, devido à baixa quantidade de dados nas idades extremas. A discrepância entre os resultados levou à escolha da classificação multiclasse para avaliar o modelo de envelhecimento.

O modelo escolhido para a predição da idade foi, portanto, o modelo de classificação multiclasse. Seus parâmetros foram definidos por meio de *Grid Search* e a arquitetura geral está representada na Figura 7.

Para simplificar a visualização dos dados e podermos comparar a outros trabalhos, subdividimos as idades em classes representando faixas etárias, obtendo a matriz de confusão da Figura 6.

É importante ressaltar que esses resultados são experimentais e visam apenas a definição da abordagem. Para uma avaliação adequada do envelhecimento, o conjunto de dados precisaria ser dividido de forma a não conter

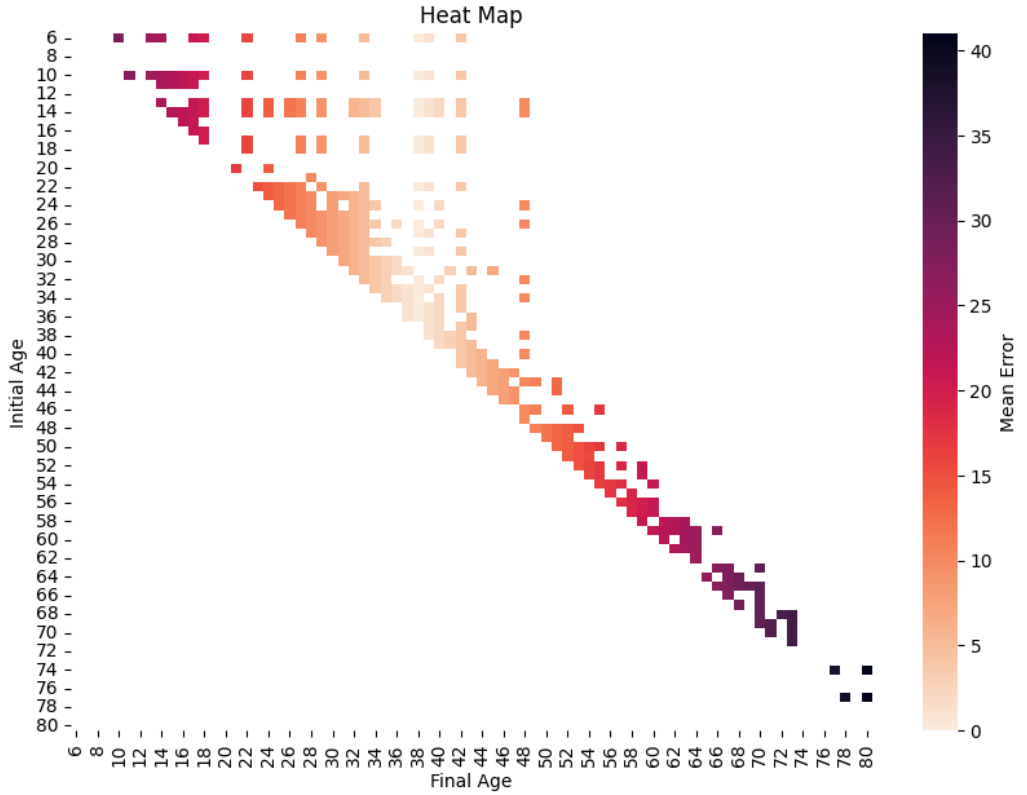

Figura 4. Mapa de calor da classificação com *speaker embeddings* sozinhos. Erro médio de idade: 11,07 anos.

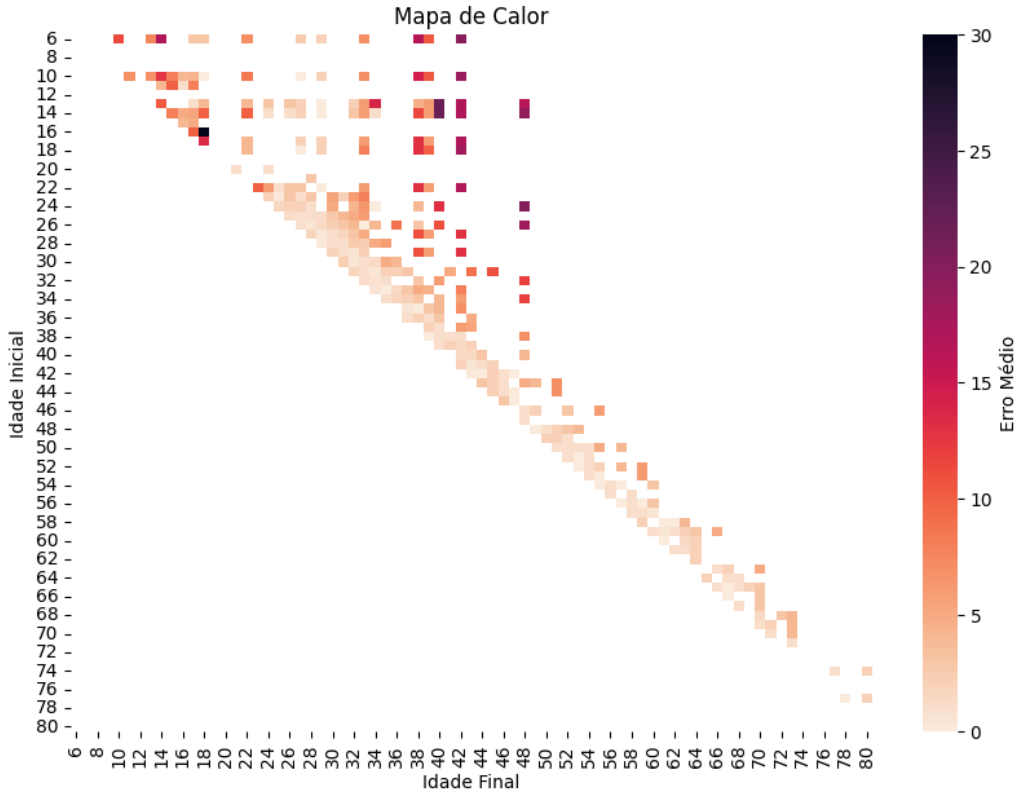

Figura 5. Mapa de calor da classificação usando par de *speaker embedding*. Erro médio de idade: 3,49 anos.

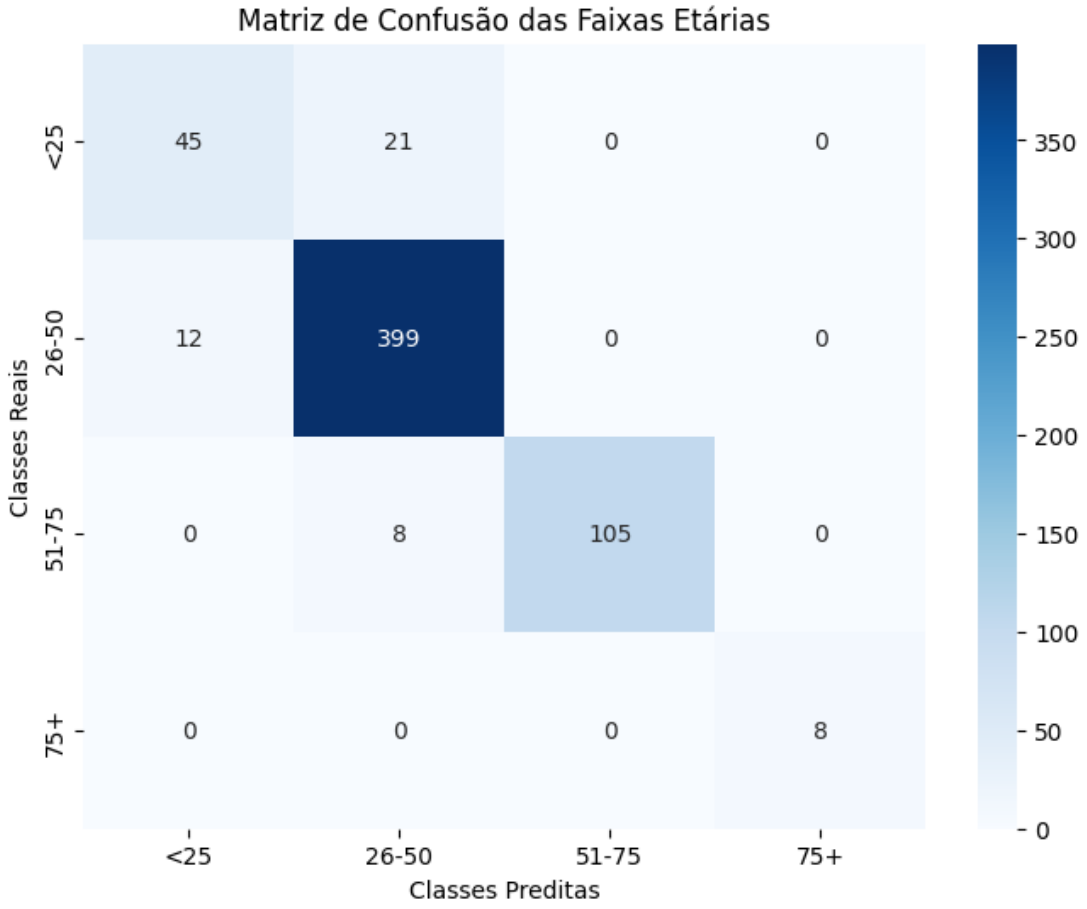

Figura 6. Avaliação do conjunto de teste subdividido em faixas etárias. 93% de precisão.

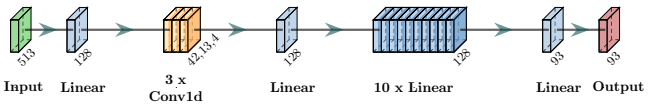

Figura 7. Arquitetura do modelo de classificação multiclasse da idade obtida por *Grid Search*.

amostras iguais entre os modelos de classificação e envelhecimento.

4.4.2 *Envelhecimento*

Assim como na tarefa de classificação, o desbalanceamento e a falta de dados representativos no dataset apresentaram desafios significativos no contexto do envelhecimento. Para mitigar essa distribuição desigual dos dados, foi proposta a criação de uma arquitetura dinâmica (Figura 8) que aloca camadas de rede neural com base nos intervalos de idade. Essa arquitetura, embora apresente alto custo computacional e baixa escalabilidade, realiza o treinamento de cada amostra apenas nos intervalos de idade desejados. O modelo foi implementado como um dicionário, onde cada elemento representa um conjunto de camadas dedicadas a uma transição unitária de idade. Durante o treinamento, uma rede separada é construída para cada par de amostras, utilizando os parâmetros correspondentes à diferença de idade. Os pesos iniciais são clonados das camadas associadas no dicionário. O ajuste dos pesos é realizado por meio de gradiente em uma única amostra (e não em *batch*), e os pesos atualizados são replicados de volta para o dicionário. Essa abordagem permite que o modelo trate separadamente as camadas associadas às classes minoritárias (idades extremas), evitando que suas características sejam ofuscadas pelas classes majoritárias presentes no conjunto de dados. Paralelamente, modelos lineares convencionais (Figura 9) foram utilizados como baseline para avaliar a arquitetura mais complexa proposta.

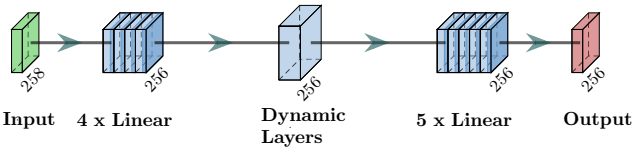

Figura 8. Rede de alocação dinâmica para envelhecimento.

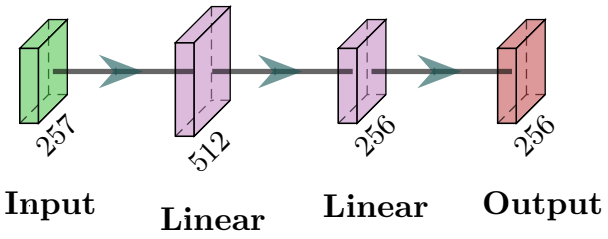

Figura 9. Exemplo de rede linear normal para envelhecimento.

4.5 Procedimentos de Treinamento e Avaliação

4.5.1 Divisão de Dados

Primeiramente, foi realizada a divisão dos 4.498 pares de *embedding*-idade entre o treinamento dos modelos de classificação e de envelhecimento. A divisão foi feita por falante, garantindo que não houvesse representatividade de dados de um modelo no outro. Os conjuntos de teste e validação foram mantidos iguais para as diferentes tarefas. O balanceamento de cada proveniência de dados e o cuidado com a divisão por falante foram mantidos. A seguir, apresentamos a divisão dos dados para cada modelo:

- **Modelo de Classificação:**
  - **Conjunto de Treino:** 1.490 pares
  - **Conjunto de Validação:** 327 pares
  - **Conjunto de Teste:** 516 pares
- **Modelo de Envelhecimento:**
  - **Conjunto de Treino:** 2165 pares
  - **Conjunto de Validação:** 327 pares
  - **Conjunto de Teste:** 516 pares

4.5.2 Treino

Durante o processo de treinamento, estendemos cada modelo por até 300 épocas. Ao longo dessas épocas, monitoramos continuamente a função de perda (*loss*) no conjunto de validação para avaliar o desempenho do modelo em dados não vistos durante o treinamento e determinar o número de épocas ideal para treinamento. Testamos também a introdução de regularização por *dropout*, procurando prevenir o *overfitting* e melhorar a capacidade de generalização.

Para as tarefas de classificação, utilizamos a função de perda *Cross Entropy Loss*, enquanto que, para as tarefas envolvendo regressão de dados, empregamos a *Mean Squared Error* (MSE).

4.5.3 Avaliação

A avaliação consistiu na aplicação da rede de classificação nos resultados obtidos pela de envelhecimento.

5 Resultados

Os resultados que serão mostrados refletem a divisão de um conjuto de dados já deficiente em idades nos extremos. Primeiro mostrar-se-á a primeira iteração e seus frutos, depois um conjunto de iterações para cobrir vieses nas análises de desempenho.

5.1 Classificação

Apesar da divisão dos dados, o primeiro resultado da classificação demonstrou-se promissor ao manter um erro médio de idade próximo ao conjunto de dados completo, com 3,30 +- 5,05 anos, assim como ilustra na Figura 10 com o mapa de calor. Ainda pelo mapa de calor, pode-se observar um problema na generalização para diferenças de idade muito grande, o que pode ser resultado da pouca diferença de idade observada nos pares. Na Figura 11 está a matriz de confusão obtida para faixas etárias de 25 anos, com 84% de precisão e 79% de F1-*score*.

5.2 Envelhecimento

Para a tarefa de envelhecimento, foram comparados a regressão da biblioteca Scikit-learn e dois tipos de arquitetura: redes lineares genéricas e a anteriormente citada rede dinâmica de alocação de camadas. Em um primeiro momento, já aplicando a rede de classificação, a regressão do Scikit-Learn apresentou um erro maior, sendo de 3,54 +- 5,28 anos, com a ilustração presente no mapa de calor da Figura 12. Sua matriz de confusão, presente na Figura 13 obteve 86% de precisão e 80% de F1-*score*. Com um resultado um pouco melhor, a melhor rede linear testada apresentou um erro de idade de 3,09 +- 4,35 anos, com ilustração do mapa de calor na Figura 14. Sua matriz de confusão, presente na Figura 15, obteve 87% de precisão e 83% de F1-*score*. Com um resultado próximo, a rede dinâmica apresentou erro de 3,04 +- 4,86 anos, com ilustração no mapa de calor da Figura 16. Já sua matriz de confusão, presente na Figura 17, obteve 74% de precisão e 78% de F1-*score*. Os testes dos modelos foram testados sobre a mesma distribuição de dados.

5.3 Mútipla Replicação dos Testes

O desbalanceamento potencializado pela divisão dos dados levou à hipótese dos modelos estarem adquirindo um grande bias decorrente de como a divisão aleatória dos dados ocorreu. Juntamente, com resultados bastante próximos da rede dinâmica e linear, optou-se por uma validação utilizando a replicação dos testes por meio de 6 iterações para obter métricas mais representativas. Nessa replicação, porém, os modelos lineares passaram por grid search para extrair o melhor resultado possível, e obtiveram desempenho ligeiramente superior em todas as métricas, demonstrando-se mais adequados ao conjunto de dados do estudo. Os resultados podem ser observados na Tabela 1.

O modelo linear demonstrou ser o mais adequado à tarefa e, portanto, foi utilizado para um último teste.

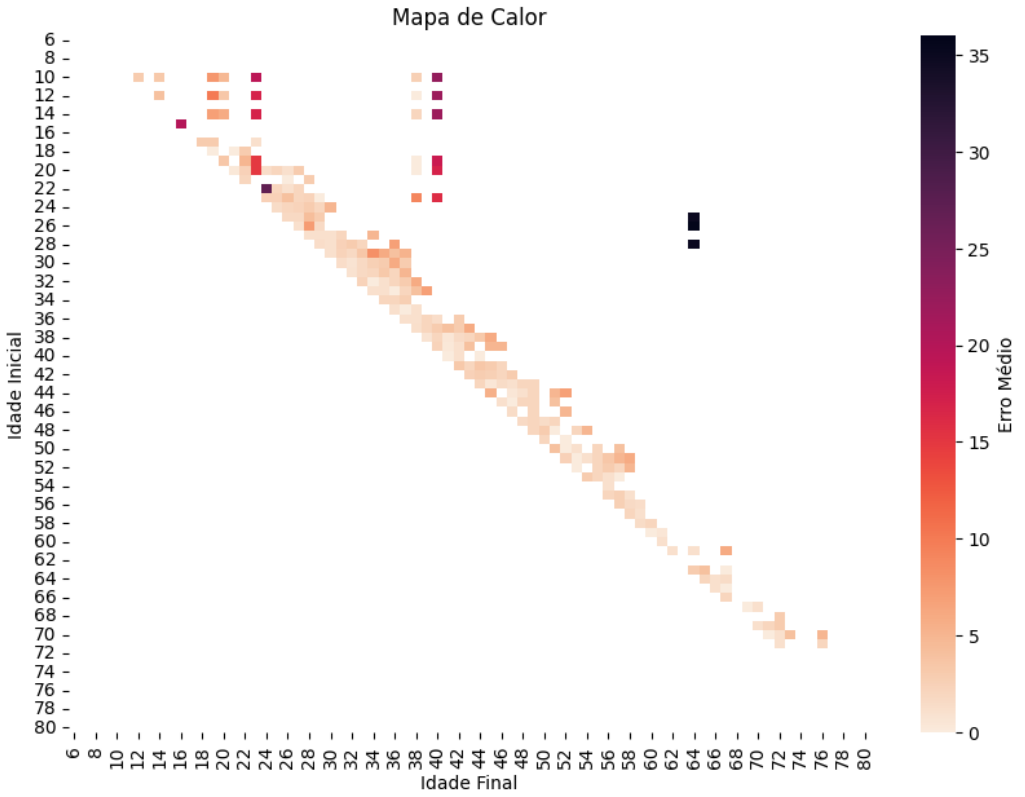

Figura 10. Mapa de calor da classificação. Erro médio de idade: 3,30 +- 5,05 anos.

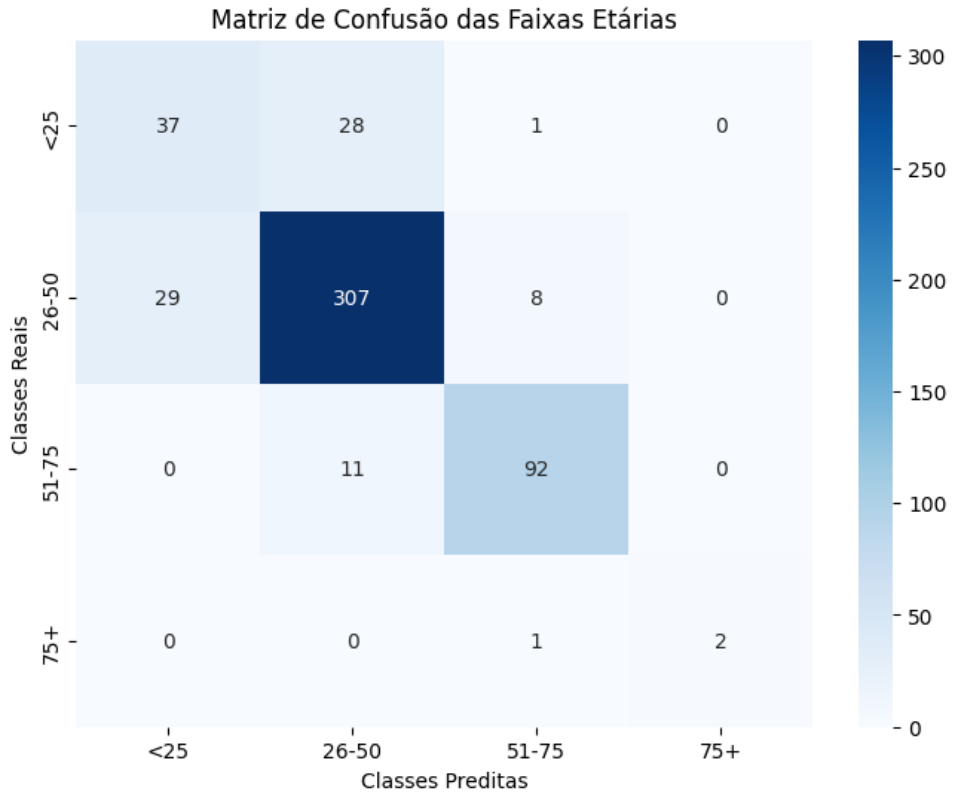

Figura 11. Matriz de confusão da classificação. 84% de precisão. 79% de F1-*score*.

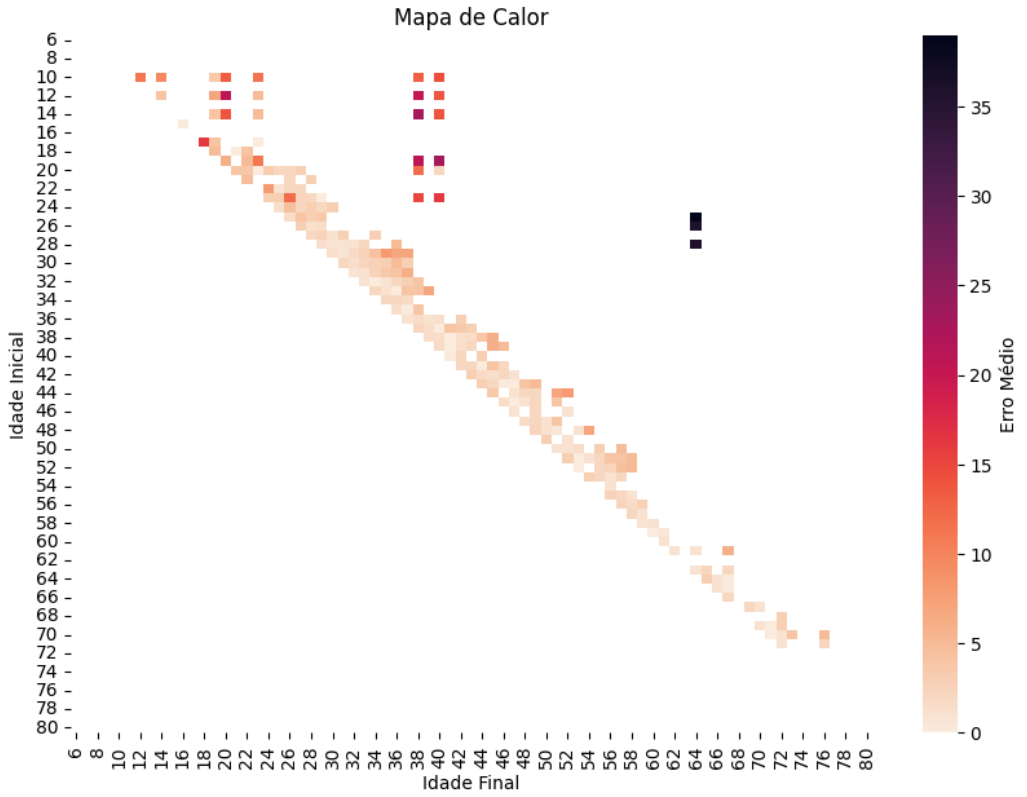

Figura 12. Mapa de calor do envelhecimento com LinearRegression Scikit Learn. Erro médio de idade pela avaliação do classificador: 3,54 +- 5,28 anos.

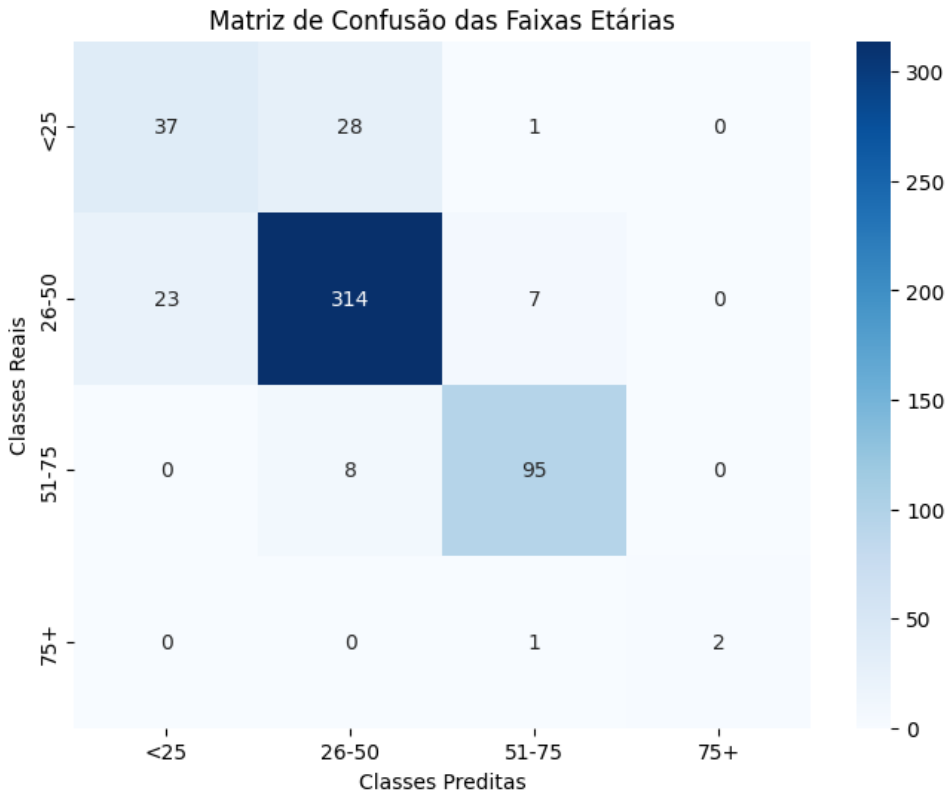

Figura 13. Matriz de confusão do envelhecimento com LinearRegression Scikit Learn. 86% de precisão e 80% de F1-*score* pela avaliação do classificador.

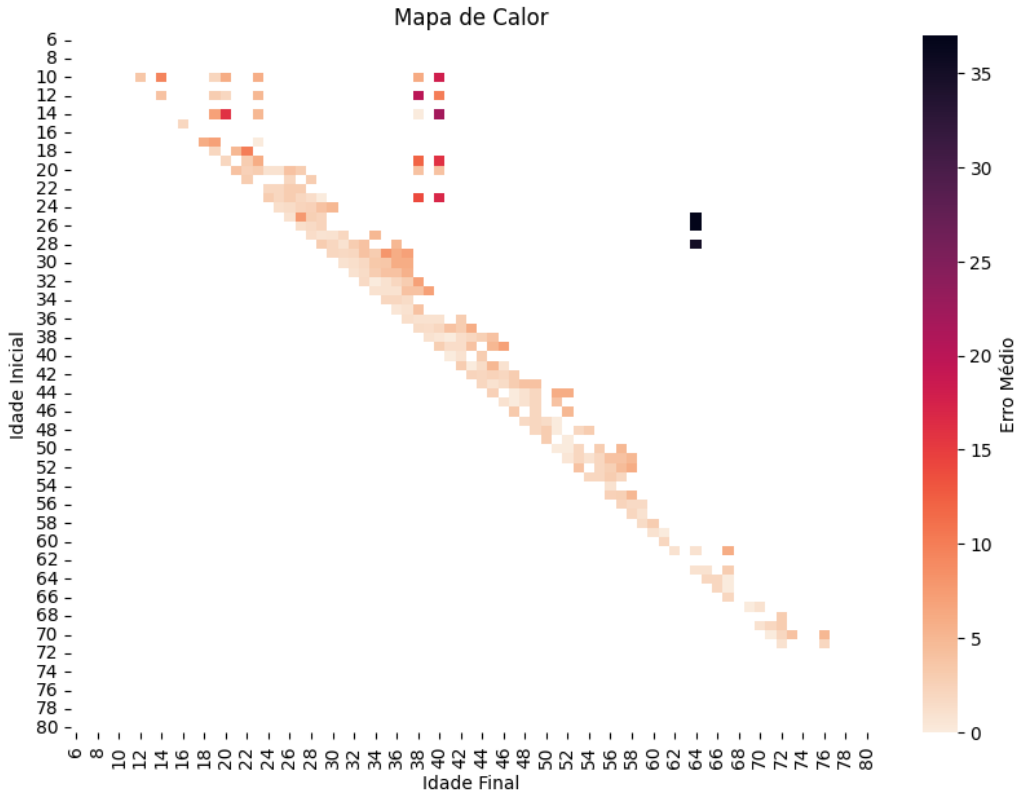

Figura 14. Mapa de calor do envelhecimento com o melhor modelo linear testado. Erro médio de idade pela avaliação do classificador: 3,09 +- 4,35 anos.

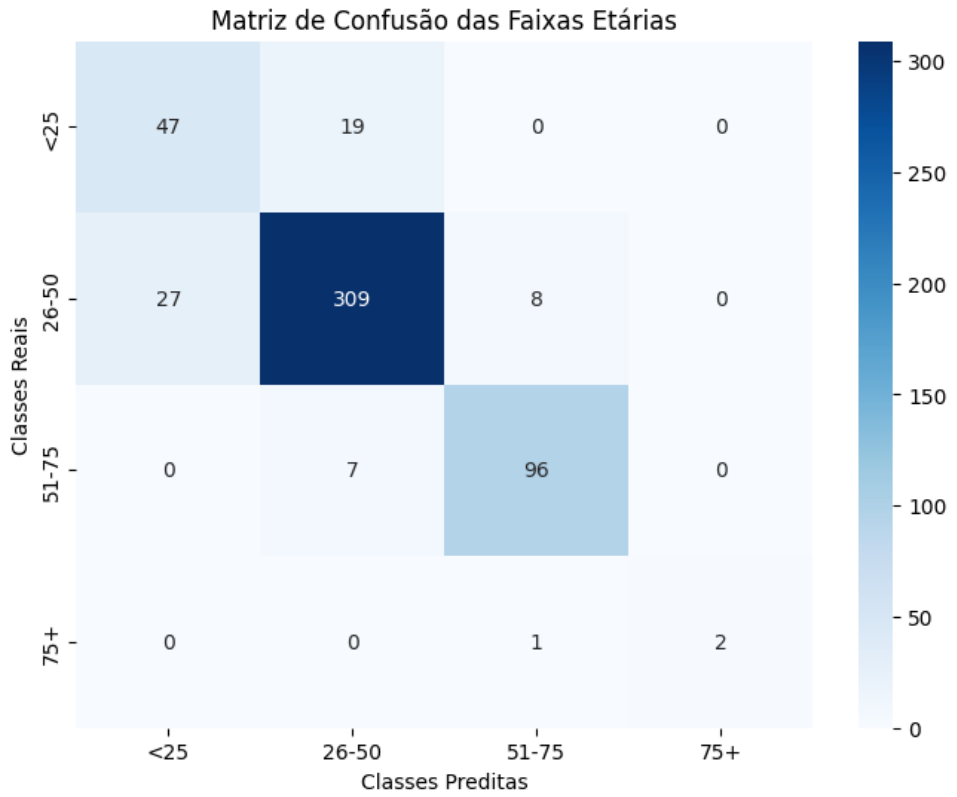

Figura 15. Matriz de confusão do envelhecimento com o melhor modelo linear testado. 87% de precisão e 83% de F1-*score* pela avaliação do classificador.

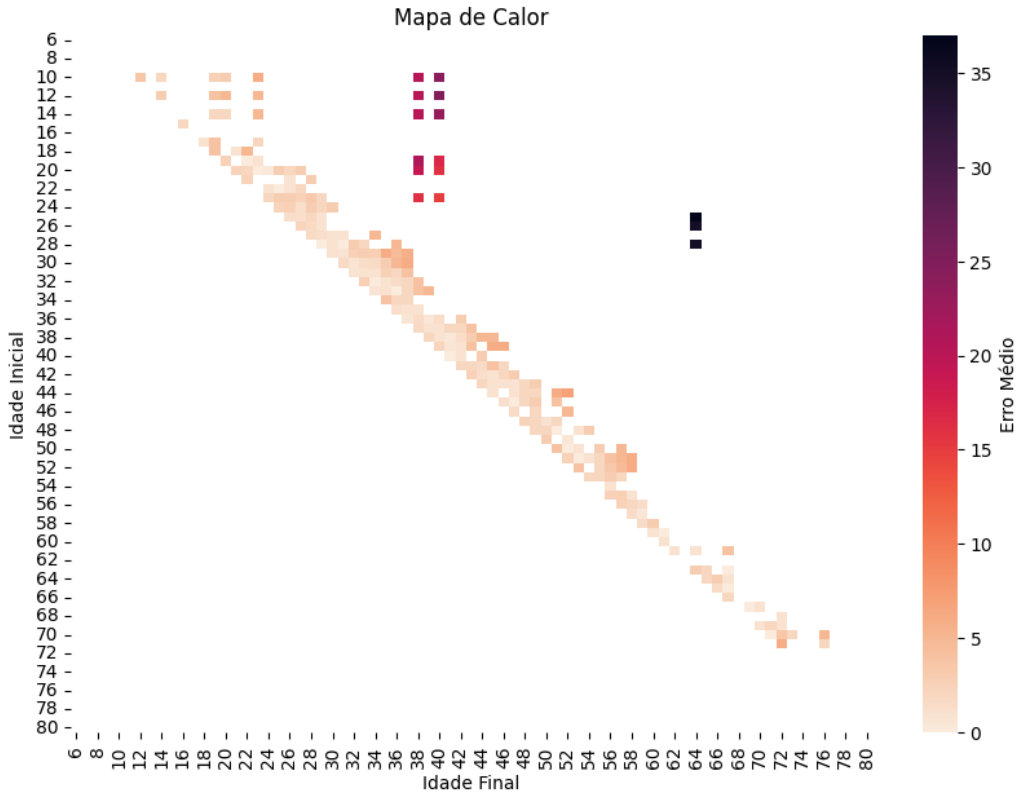

Figura 16. Mapa de calor do envelhecimento com modelo dinâmico. Erro médio de idade pela avaliação do classificador: 3,04 +- 4,86 anos.

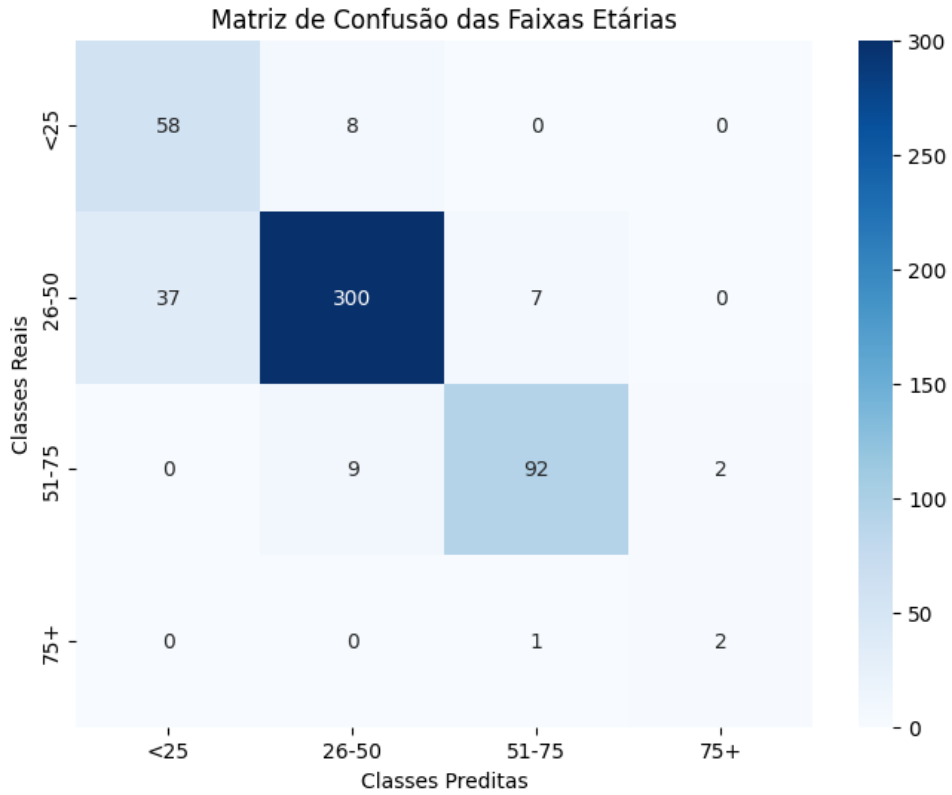

Figura 17. Matriz de confusão do envelhecimento com modelo dinâmico. 74% de precisão e 78% de F1-*score* pela avaliação do classificador.

Tabela 1
6 Diferentes Arranjos de Dados

	Erro de Idade da Classificação	Precisão da Classificação	Erro de Idade do Modelo Linear	Precisão do Modelo Linear	Erro de Idade do Modelo Dinâmico	Precisão do Modelo Dinâmico
1	3.30 ± 5.05	0.84	3.09 ± 4.35	0.87	3.04 ± 4.86	0.74
2	3.12 ± 4.45	0.64	2.67 ± 3.78	0.71	2.95 ± 4.23	0.67
3	3.56 ± 5.58	0.66	3.30 ± 5.30	0.67	3.42 ± 5.46	0.70
4	3.13 ± 4.57	0.70	3.13 ± 4.60	0.70	3.08 ± 4.31	0.63
5	3.58 ± 5.68	0.64	2.90 ± 3.97	0.68	3.59 ± 4.88	0.63
6	3.71 ± 5.58	0.68	3.26 ± 4.96	0.69	3.05 ± 4.39	0.69
Geral	3.40 ± 5.18	0.69 ± 0.07	3.06 ± 4.53	0.72 ± 0.07	3.19 ± 4.71	0.68 ± 0.04

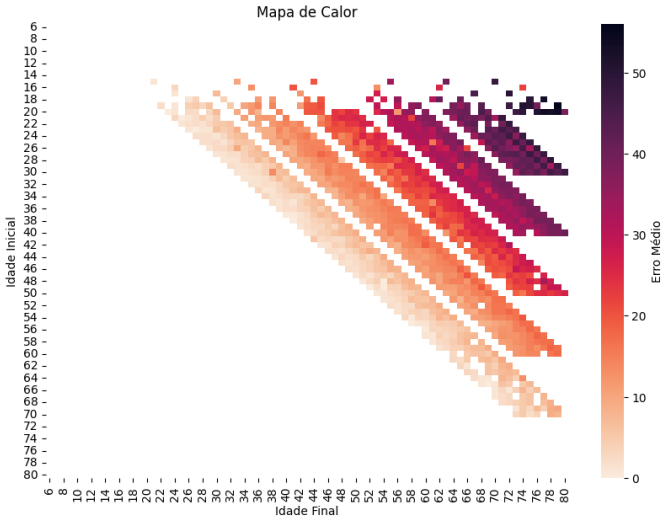

Figura 18. Mapa de calor do envelhecimento em múltiplos intervalos de idade. Erro médio: 17,80 +- 12,86 anos.

5.3 Múltiplos Intervalos de Idade para Teste

Em um último esforço para avaliar o desempenho do melhor modelo linear obtido em um conjunto de dados caracterizado por baixa variação etária, replicamos o envelhecimento para diversos intervalos de idade distribuindo-os igualmente. Para isso, usamos áudios de falantes do VoxCeleb que não possuíam par de idade e envelhecemos seu *speaker embedding* nesses intervalos, obtendo um mapa de calor (Figura 18) com diferença média de idade de 17,80 anos e com uma distribuição dessas diferenças que pode ser relacionada ao enviesamento carregado pela falta de grandes intervalos no conjunto de treino.

5.4 Comparação à Matriz de Confusão de Classificação de Idade de Outros Trabalhos

Por fim, comparamos na Tabela 2 as métricas de clasificação de idade obtidas pelo uso de par de *speaker embedding* no modelo linear com as do trabalho de Wilson et al. (2021).

Tabela 2
Comparação de Acurácia por Intervalos de 25 Anos

Método	Acurácia em 25 anos de intervalo
VANN-V	50.2%
VANN-AV Cat	46.0%
VANN-AV MFB	52.7%
Par de *speaker embeddings*	84.9%

6 Conclusão

Este trabalho propôs uma abordagem para a recuperação de identidade vocal em indivíduos com disfonia, aplicando técnicas de envelhecimento sintético em gravações anteriores ao surgimento do distúrbio. Os resultados indicaram que o uso de pares de *speaker embeddings* para a classificação trouxe métricas relevantes e que os modelos de envelhecimento linear, representatos por n *hidden layers*, apresentaram desempenho mais consistente e estável que a rede dinâmica proposta.

Embora os achados reforcem a viabilidade de utilizar redes lineares em tarefas de envelhecimento vocal, o desbalanceamento dos dados sugere a necessidade de ampliar o dataset para incluir uma representação mais homogênea das idades, o que pode beneficiar a precisão do modelo em futuras implementações, principalmente ao analisar as métricas associadas à classificação da idade treinada no conjunto completo e incompleto de dados.

Este trabalho apresentou um ponto de partida para a aplicação de técnicas de envelhecimento vocal em contextos de disfonia, destacando as limitações associadas ao desbalanceamento dos dados e à representatividade etária. Futuras investigações poderão explorar melhorias nos modelos, dados e até representações, avaliando sua viabilidade em contextos mais amplos e diversificados.

7 Apêndice

Este apêndice consta os dados obtidos para o teste de classificação de idade usando *speaker embeddings* dos áudios do *dataset* Common Voice. Os rótulos, assim como pode ser observado pela Figura 19, não são por idade, mas por faixa etária definida, no geral, por 10 anos. Na Figura 20 consta os resultados obtidos na tarefa de classificação para diferentes quantidades de dados por classe.

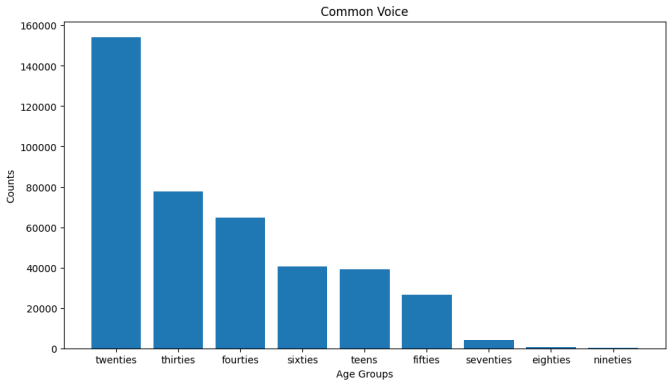

Figura 19. Distribuição de classes no dataset Common Voice.

Common Voice Classification Results

Max Elements	Age Error	Precision
80	17.22 +- 14.46	0.19
400	12.98 +- 11.44	0.33
1000	13.19 +- 11.16	0.28
2000	13.13 +- 10.98	0.28
5000	12.68 +- 10.30	0.26
10000	12.22 +- 9.99	0.32
20000	11.94 +- 10.02	0.39
-	9.57 +- 9.50	0.41

Figura 20. Resultados obtidos para classificação de idade por speaker embedding usando o Common Voice.

REFERÊNCIAS

[1] TARAFDER, Kamrul & DATTA, Pran & TARIQ, Ahmed. The Aging Voice. Bangabandhu Sheikh Mujib Medical University Journal, v. 5, 2012. DOI: 10.3329/bsmmuj.v5i1.11033.

[2] WILSON, J. et al. Voice Aging with Audio-Visual Style Transfer. Disponível em: https://arxiv.org/abs/2110.02411.Acesso em: 19 set. 2024.

[3] SPINA, Ana Lúcia et al. Correlation between voice and life quality and occupation. Brazilian Journal of Otorhinolaryngology, v. 75, n. 2, p. 275–279, 1 mar. 2009.

[4] SMITH, E. M. et al. Effects of Voice Disorders on Quality of Life. v. 113, n. 2, 1 ago. 1995.

[5] NEIGHBORS, C.; SONG, S. A. Dysphonia. Disponível em: https://www.ncbi.nlm.nih.gov/books/NBK565881/.Acesso em: 19 set. 2024.

[6] ZHAO, Y.; KURUVILLA-DUGDALE, M.; SONG, M. Voice Conversion for Persons with Amyotrophic Lateral Sclerosis. IEEE Journal of Biomedical and Health Informatics, v. 24, n. 10, p. 2942–2949, out. 2020.

[7] ISLAM, M.; CHEN, G.; JIN, S. An Overview of Neural Network. American Journal of Neural Networks and Applications, v. 5, n. 1, p. 7, 2019.

[8] RABINER, L. R.; SCHAFER, R. W. Introduction to Digital Speech Processing. Foundations and Trends in Signal Processing, v. 1, n. 1–2, p. 1–194, 2007.

[9] SARKER, I. H. Deep Learning: a Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Computer Science, v. 2, n. 6, 18 ago. 2021.

[10] HECHMI, K. et al. VOXCELEB ENRICHMENT FOR AGE AND GENDER RECOGNITION. [s.l.: s.n.]. Disponível em: https://arxiv.org/pdf/2109.13510. Acesso em: 19 set. 2024.

[11] DOI, H. et al. Speaking-Aid Systems Based on One-to-Many Eigenvoice Conversion for Total Laryngectomees. 1 dez. 2010.

[12] BINU, D.; RAJAKUMAR, B. R. Artificial Intelligence in Data Mining. Academic Press, 2021.

[13] QIAN, K. et al. AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. In: International Conference on Machine Learning, p. 5210–5219, 24 maio 2019.

[14] ABDUL, Zrar Kh.; AL-TALABANI, Abdulbasit K. Mel Frequency Cepstral Coefficient and Its Applications: A Review. Disponível em: https://ieeexplore.ieee.org/document/3223444. Acesso em: 19 set. 2024.

[15] BANK, D.; KOENIGSTEIN, N.; GIRYES, R. Autoencoders. [s.l.: s.n.]. Disponível em: https://arxiv.org/pdf/2003.05991.Acesso em: 19 set. 2024.

[16] GOODFELLOW, I. et al. Generative Adversarial Nets. [s.l.: s.n.]. Disponível em: https://arxiv.org/pdf/1406.2661.Acesso em: 19 set. 2024.

[17] ALZUBAIDI, L. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, v. 8, n. 1, 31 mar. 2021.

[18] VAN DEN OORD, A. et al. WAVENET: A GENERATIVE MODEL FOR RAW AUDIO. [s.l: s.n.]. Disponível em: https://arxiv.org/pdf/1609.03499.

[19] MORISE, M.; YOKOMORI, F.; OZAWA, K. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. IEICE Transactions on Information and Systems, v. E99.D, n. 7, p. 1877–1884, 2016.

[20] KAWAHARA, H. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. Acoustical Science and Technology, v. 27, n. 6, p. 349–353, 2006.

[21] Friedrich, W. Spectral and Rhythm Features for Audio Classification with Deep Convolutional Neural Networks. Disponível em: https://arxiv.org/html/2410.06927v1. Acesso em: 7 dez. 2024.

[22] SISMAN, B. et al. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, v. 29, n. 1, p. 132–157, 2021.

[23] KONG, J.; KIM, J.; BAE, J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. Neural Information Processing Systems, v. 33, p. 17022–17033, 12 out. 2020.

[24] YONEYAMA, R.; WU, Y.-C.; TODA, T. Source-Filter HiFi-GAN: Fast and Pitch Controllable High-Fidelity Neural Vocoder. arXiv (Cornell University), 4 jun. 2023.

[25] PRENGER, R.; VALLE, R.; CATANZARO, B. Waveglow: A Flow-based Generative Network for Speech Synthesis. Disponível em: ¡https://ieeexplore.ieee.org/abstract/document/8683143¿.

[26] QIAN, K. et al. AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. International Conference on Machine Learning, p. 5210–5219, 24 maio 2019.

[27] ARDILA, R. et al. Common Voice: A Massively-Multilingual Speech Corpus. Disponível em: ¡https://arxiv.org/abs/1912.06670¿. Acesso em: 10 dez. 2023.

[28] CHUNG, J. S. et al. In defence of metric learning for speaker recognition. Disponível em: ¡https://arxiv.org/abs/2003.11982¿. Acesso em: 10 maio. 2024.

[29] WAN, L. et al. Generalized End-to-End Loss for Speaker Verification. International Conference on Acoustics, Speech, and Signal Processing, 15 abr. 2018.

[30] LAZZARINI, V. Elementos de Acústica. 1998. Disponível em: ¡https://hugoribeiro.com.br/biblioteca-digital/Lazzarini-Elementos_Acustica.pdf¿. Acesso em: 20 dez. 2024.

[31] HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. Neural Computation, v. 9, n. 8, p. 1735–1780, nov. 1997.

[32] SCHMIDT, R. Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. 2019. Disponível em: ¡https://arxiv.org/pdf/1912.05911¿.

[33] WANG, H. et al. Wespeaker: A Research and Production Oriented Speaker Embedding Learning Toolkit. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 1–5, 5 maio 2023.

[34] NAGRANI, A. et al. VoxCeleb: Large-scale Speaker Verification in the Wild. Computer Speech & Language, p. 101027, out. 2019.